McGRAW-HILL ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

1 A - ANO

www.MHEST.com

McGRAW-HILL

ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

McGRAW-HILL ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

10th Edition

An international reference work in twenty volumes including an index

McGraw-Hill

New York Chicago San Francisco Lisbon London Madrid Mexico City
Milan New Delhi San Juan Seoul Singapore Sydney Toronto

On the front cover
TraR protein structure of *Agrobacterium tumefaciens*, the causal agent of crown gall disease. (*Image: Argonne National Laboratory*)

Library of Congress Cataloging-in-Publication Data

McGraw-Hill encyclopedia of science & technology—10th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-07-144143-8 (alk. paper)
1. Science—Encyclopedias. 2. Technology—Encyclopedias.
I. Title: Encyclopedia of science & technology.
II. Title: McGraw-Hill encyclopedia of science and technology.
Q121.M3 2007
503—dc22 2007006137

13-digit ISBN: 978-0-07-144143-8 (set)

10-digit ISBN: 0-07-144143-3 (set)

McGraw-Hill
A Division of The McGraw-Hill Companies

1 2 3 4 5 6 7 8 9 0 DOW/DOW 0 1 2 1 0 9 8 7

This book set was printed on acid-free paper.

It was set in Garamond Book and Neue Helvetica Black Condensed by Aptara, Falls Church, Virginia. The art was prepared by Aptara. The book was printed and bound by R. R. Donnelley, Willard, Ohio.

Editorial and Technical Staff

Mark D. Licker, Publisher

Elizabeth Geller, Managing Editor

Jonathan Weil, Senior Staff Editor

David Blumel, Staff Editor

Alyssa Rappaport, Staff Editor

Stefan Malmoli, Staff Editor

Jessa Netting, Staff Editor

Renee Taylor, Editorial Assistant

Charles Wagner, Manager, Digital Content

Editing, Design, & Production Staff

Roger Kasunic, Vice President—Editing, Design, and Production

Joe Faulk, Editing Manager

Frank Kotowski, Jr., Senior Editing Supervisor

Ron Lane, Art Director

Thomas G. Kowalczyk, Production Manager

Suppliers

Art prepared by Aptara, Falls Church, Virginia.

Colorplates printed by Lehigh Press Lithographers, Pennsauken, New Jersey.

Color maps of continents prepared by Mapping Specialists, Ltd., Madison, Wisconsin.

Index prepared by Stephen R. Ingle, WordCo Indexing Services, Norwich, Connecticut.

Consulting Editors

Prof. Chao-Jun Li. *Canada Research Chair in Green Chemistry, Department of Chemistry, McGill University, Montreal, Quebec, Canada.* ORGANIC CHEMISTRY.

Prof. Donald W. Linzey. *Wytheville Community College, Wytheville, Virginia.* VERTEBRATE ZOOLOGY.

Dr. Philip V. Lopresti. *Retired; formerly, Engineering Research Center, AT&T Bell Laboratories, Princeton, New Jersey.* ELECTRONIC CIRCUITS.

Dr. Dan Luss. *Cullen Professor of Engineering, Department of Chemical Engineering, University of Houston, Texas.* CHEMICAL ENGINEERING.

Prof. Scott M. McLennan. *Department of Geosciences, State University of New York at Stony Brook.* GEOLOGY (PHYSICAL, HISTORICAL, AND SEDIMENTARY).

Prof. Philip L. Marston. *Department of Physics and Astronomy, Washington State University, Pullman.* ACOUSTICS.

Dr. Ramon A. Mata-Toledo. *Professor of Computer Science, James Madison University, Harrisonburg, Virginia.* COMPUTERS.

Prof. Krzysztof Matyjaszewski. *J. C. Warner Professor of Natural Sciences, Department of Chemistry, Carnegie-Mellon University, Pittsburgh, Pennsylvania.* POLYMER SCIENCE AND ENGINEERING.

Prof. Joel S. Miller. *Department of Chemistry, University of Utah, Salt Lake City.* INORGANIC CHEMISTRY.

Dr. Orlando J. Miller. *Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, Michigan.* GENETICS.

Prof. Jay M. Pasachoff. *Director, Hopkins Observatory, Williams College, Williamstown, Massachusetts.* ASTRONOMY.

Prof. David J. Pegg. *Department of Physics, University of Tennessee, Knoxville.* ATOMIC AND MOLECULAR PHYSICS.

Prof. J. Jeffrey Peirce. *Department of Civil and Environmental Engineering, Edmund T. Pratt Jr. School of Engineering, Duke University, Durham, North Carolina.* ENVIRONMENTAL ENGINEERING.

Dr. William C. Peters. *Professor Emeritus, Mining and Geological Engineering, University of Arizona, Tucson.* MINING ENGINEERING.

Prof. Arthur N. Popper. *Department of Biology, University of Maryland, College Park.* NEUROSCIENCE.

Dr. Kenneth Pritzker. *Director, Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, Canada.* MEDICINE AND PATHOLOGY.

Prof. Justin Revenaugh. *Department of Geology and Geophysics, University of Minnesota, Minneapolis.* GEOPHYSICS.

Dr. Roger M. Rowell. *USDA-Forest Service, Forest Products Laboratory, Madison, Wisconsin.* FORESTRY.

Dr. John L. Safko. *Distinguished Professor Emeritus, Physics and Astronomy, Associated Faculty, School of the Environment, University of South Carolina, Columbia.* CLASSICAL MECHANICS.

Dr. Andrew P. Sage. *Founding Dean Emeritus and First American Bank Professor, University Professor, School of Information Technology and Engineering, George Mason University, Fairfax, Virginia.* CONTROL AND INFORMATION SYSTEMS.

Dr. Alfred S. Schlachter. *Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, California.* ATOMIC AND MOLECULAR PHYSICS.

Prof. Ivan Schuller. *Department of Physics, University of California-San Diego, La Jolla, California.* CONDENSED-MATTER PHYSICS.

Dr. David Sherman. *Department of Earth Sciences, University of Bristol, United Kingdom.* MINERALOGY.

Dr. Steven A. Slack. *Department of Plant Pathology, Cornell University, Ithaca, New York.* PLANT PATHOLOGY.

Dr. Arthur A. Spector. *Professor of Biochemistry and Internal Medicine, University of Iowa, Iowa City.* BIOCHEMISTRY.

Dr. Bruce A. Stanley. *Director, Scientific Programs, Section of Technology Development and Research Resources H093, The Pennsylvania State University College of Medicine, Hershey.* PHYSIOLOGY.

Prof. Anthony P. Stanton. *Carnegie-Mellon University, Pittsburgh, Pennsylvania.* GRAPHIC ARTS AND PHOTOGRAPHY.

Dr. Trent Stephens. *Professor of Anatomy and Embryology, Idaho State University, Pocatello.* DEVELOPMENTAL BIOLOGY.

Prof. John F. Timoney. *Department of Veterinary Science, University of Kentucky, Lexington.* VETERINARY MEDICINE.

Prof. Antonio J. Torres. *Department of Food Science and Technology, Oregon State University, Corvallis.* FOOD SCIENCE.

Dr. Sally E. Walker. *Associate Professor of Geology and Marine Science, University of Georgia, Athens.* INVERTEBRATE PALEONTOLOGY.

Prof. Pao K. Wang. *Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison.* METEOROLOGY AND CLIMATOLOGY.

Prof. Frank M. White. *Professor Emeritus, Department of Mechanical Engineering, University of Rhode Island, Kingston.* FLUID MECHANICS.

Prof. Mary Anne White. *Killam Research Professor in Materials Science, Department of Chemistry, Dalhousie University, Halifax, Nova Scotia, Canada.* MATERIALS SCIENCE AND METALLURGIC ENGINEERING.

Prof. Thomas A. Wikle. *Department of Geography, and Associate Dean, College of Arts and Sciences, Oklahoma State University, Stillwater.* PHYSICAL GEOGRAPHY.

Preface

As we approach the 50th anniversary of the first publication of the *McGraw-Hill Encyclopedia of Science & Technology* in 1960, we are reminded that the objectives expressed in the preface to the first edition remain valid to this day: *"To provide the widest possible range of articles that will be understandable and useful to any person of modest technical training who wants to obtain information outside his [or her] particular field of specialization."* This philosophy has guided the authors and editors of the Encyclopedia in their striving to serve the reference and educational needs of students, professionals, and librarians worldwide by offering reliable, comprehensive information in all major fields of science and engineering.

Advances in science and technology over the five years since the publication of the last edition (2002) have been remarkable in many respects, with some advances even at the center of hotly debated societal issues. Our legislatures, for example, weigh the potential benefits of stem cell research against the ethical concerns raised by some. Ever more sophisticated models running on powerful computing systems point to human-induced changes in global climate over this century in a range that could cause significant societal disruptions. How were these models developed? What technologies are available to mitigate these changes, and at what cost? Advances in information and telecommunications technology are made at an accelerating pace; how are they changing how people communicate, interact, learn, and do business? Dealing with such issues requires authoritative sources of scientific and technical information that are understandable to the nonspecialist, and thus the need for a comprehensive scientific encyclopedia seems as great as ever.

Even though the goal has been a broad revision of the previous edition, much effort was concentrated in certain rapidly advancing areas, particularly in cell and molecular biology; information technology and telecommunications; nanotechnology; the environmental, earth, and climate sciences; materials science; and cosmology, among others. The result of these efforts are the 7100 articles in the tenth edition, many new or rewritten, and others updated as needed, following the recommendations of our distinguished Board of Consulting Editors. Each article was prepared by one or more experts nominated by the Board and coming from universities, industry, and public agencies worldwide.

The Encyclopedia has always been intended to be a work *of*, not *about*, science and technology, with each article reflecting the scope and sophistication of the topic. However, since the Encyclopedia is used primarily by nonspecialists, care has been given to structure, edit, and illustrate the articles such that the reader has ready access to the key concepts of the field. Articles start with a definition and a concise overview of the topic. The subject matter is developed according to a clear outline and concludes with a bibliography of generally available publications for further study (nearly 25,000 bibliographic entries in total). One of the hallmarks of the Encyclopedia is the numerous cross references to related articles—more than 60,000 in this edition—which not only allow the reader to acquire background information, or to study in greater depth, but also help illuminate the context of the material and broader connections among topics.

More than 12,000 illustrations enhance the text and allow the reader to visualize concepts. These include photographic images, two-color graphs, charts, and maps redrawn for uniformity of style; and 88 full-color plates. Over 1400 tables provide useful data, and, where appropriate, chemical structures (900) and reactions (more than 2500) are included as well as mathematical equations (8500). An appendix with tables in volume 20 explains scientific notation and the dual systems for units of measurement—the U.S. Customary System and the International System (SI)—used throughout the text. The reader will also find an extensive table of conversion factors for the measurement systems, symbols used in scientific writing, and chemical nomenclature.

Ease of access to this tremendous amount of information is essential if the Encyclopedia is to fulfill its mission. There are several ways of finding information. The articles are arranged alphabetically, and of course the reader can go directly to an entry in the appropriate volume. To find all of the information available on a particular subject, the reader may use the Analytical Index, which occupies more than 500 pages. The Topical Index lists all of the articles pertaining to a particular discipline, for example, theoretical physics or microbiology. The Study Guides outline the major fields of science and technology and their subdisciplines, and provide for each a list of overview articles. They offer the opportunity not only to find a relevant article but also to systematically learn or review a field. Finally, a comprehensive list of the contributors and their affiliations along with the titles of the contributions allows the reader to find the works of particular authors. All of these indexes may be found in volume 20.

New to the tenth edition is a companion Web site

http://MHEST.com

which contains materials complementary to the printed text, much of it drawn from AccessScience, the McGraw-Hill Encyclopedia of Science & Technology Online: updates of critical articles, "rich media" (such as animations, videos, and audio presentations), commentaries on articles, explorations of topical themes, and more. The content of MHEST.com will be updated periodically, and we invite the reader to visit it often.

The editors of the Encyclopedia are grateful to the international authorship, the consulting editors, reviewers, and many others whose efforts have made this revision possible. As in past editions, we dedicate the product of this massive collaboration to our readers and trust that it will continue to uniquely serve their need for high-quality information as well as the objectives of scientific education and technological literacy.

Mark D. Licker
Publisher

Organization of the Encyclopedia

The *McGraw-Hill Encyclopedia of Science & Technology* presents pertinent information in every field of modern science and technology. The 7100 articles are arranged alphabetically in the 19 text volumes. The range of article titles included in each volume is indicated on the spine and front cover (for example, volume 1 contains articles with titles starting with "Aar" up to "Ano"). Thus the reader may quickly locate an article by its title. The 20th volume contains the indexes and ancillary materials.

Broad survey articles are available for each of the disciplines covered; even readers with little prior knowledge of that discipline will find the basic concepts covered in these articles. From the survey article, the reader may proceed to more specialized articles using the cross-referencing system. These cross references are set in small capital letters for emphasis and are inserted at the relevant points in the text. For example, in a survey article such as **Digital computer**, the reader is directed to numerous other articles such as COMPUTER PERIPHERAL DEVICES, COMPUTER STORAGE TECHNOLOGY, MICROPROCESSOR, and PROGRAMMING LANGUAGES. The references may lead to subjects that have not occurred to the reader. The article **Solvent** has such diverse cross references as COORDINATION CHEMISTRY, HALOGENATED HYDROCARBON, INDUSTRIAL HEALTH AND SAFETY, and WATER POLLUTION. The cross references not only lead to articles of greater specialization but also help illuminate the context of the article and the broader connections among topics. This edition contains more than 60,000 cross references.

The pattern of proceeding from the general to the specific has been employed not only in the plan of the Encyclopedia but within the body of the articles. Each article begins with a definition of the subject, followed by sufficient background material to give a frame of reference and permit the reader to move into the detailed text of the article. Within the text are centered heads and two levels of sideheads that outline the article; they are intended to enhance understanding and can guide the user that prefers to read selectively the sections of a long article.

Alphabetization of article titles is by word, not by letter, with a comma providing a stop in occasional inverted article titles (so that subject matter can be grouped). Two examples of sequence are:

Air	**Earth, age of**
Air-cushion vehicle	**Earth, heat flow of**
Air mass	**Earth crust**
Air-traffic control	**Earth tides**
Aircraft fuel	**Earthquake**

Numerous illustrations, both line drawings and images, contribute to the utility, clarity, and interest of the text. Each illustration (as well as each table) is called out in boldface at its first mention in the text. This emphasis enables the reader to move from an illustration to the point in the text where the illustration is often discussed in detail.

To meet the needs of the Encyclopedia's broad readership, measurements are given in dual systems of units: The U.S. Customary System is used throughout the text along with equivalent measurements in the International System of Units. In particular cases, such as references to measurements in some illustrations or tables, conversion factors may be given for simplicity.

The contributor's full name appears at the end of an article section or an entire article. Each author is identified in an alphabetical Contributors list in volume 20, which cites the university, laboratory, business, or other organization with which the author is affiliated and the titles of the articles written by that contributor.

Most of the articles contain bibliographies citing useful sources. The bibliographies are placed at the ends of articles or occasionally at the ends of major sections in long articles. For additional bibliographies, the reader should refer to related articles as indicated by cross references.

Thus, the alphabetical arrangement of article titles, the text headings, the cross references, and the bibliographies permit the reader to research a particular topic by simply taking a volume from the shelf. However, the reader can also find information in the Encyclopedia by using the Analytical Index and the Topical Index in volume 20. The Analytical Index—over 500 pages in length—contains each important term, concept, and person mentioned throughout the 19 text volumes. It guides the reader to the volume numbers and page numbers concerned with a specific point. The reader wishing to consult everything in the Encyclopedia on a particular aspect of a subject will find that the Analytical Index is the best approach. A broader survey may be made through the Topical Index, which groups all article titles of the Encyclopedia under 90 general headings. For example, under "Atomic and molecular physics," 90 articles are listed, and under "Biochemistry," 147. The Topical Index thus enables the reader quickly to identify all articles in the Encyclopedia in a particular subject area.

The Study Guides in volume 20 provide highly structured outlines of major scientific disciplines and relate groups of Encyclopedia articles to each outline heading. By following a guide, the reader is led through pertinent Encyclopedia articles in a sequence that provides an overall grasp of the discipline.

A useful feature is the section "Scientific Notation" in volume 20. It clarifies usage of symbols, abbreviations, and nomenclature, and is especially valuable in making conversions between the International System, U.S. Customary, and metric measurements.

With the 10th edition, the editors are introducing a new feature to enhance the usefulness of the Encyclopedia: a companion Web site

http://MHEST.com

containing periodically updated collections of articles, graphics, and multimedia content pertaining to a timely theme, as well as selected updates of Encyclopedia articles as developments in science and technology dictate. We encourage readers to visit the site to benefit from this material.

Table of Contents for Colorplates

Colorplates face the page numbers cited. The plates are relevant to the subject matter in the respective articles, but may not be specifically mentioned in the text.

A15 phases

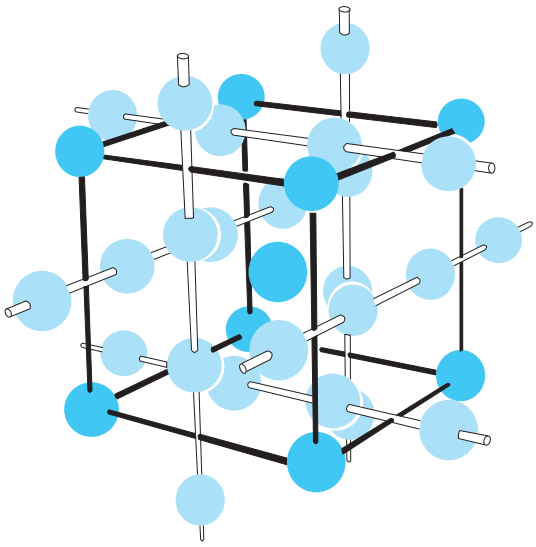
A series of intermetallic compounds that have a particular crystal structure and the chemical formula A_3B, where A represents a transition element and B can be either a transition element or a nontransition element. Many A15 compounds exhibit the phenomenon of superconductivity at relatively high temperatures in the neighborhood of 20 K (−424°F) and in high magnetic fields on the order of several tens of teslas (several hundred kilogauss). High-temperature–high-field superconductivity has a number of important technological applications and is a challenging fundamental research area in condensed-matter physics.

Crystal structure. The A15 compounds crystallize in a structure in which the unit cell, the repeating unit of the crystal structure, has the overall shape of a cube. The B atoms are located at the corners and in the center of the cube, while the A atoms are arranged in pairs on the cube faces (see **illus.**). A special characteristic of the A15 crystal structure is that the A atoms form mutually orthogonal linear chains that run throughout the crystal lattice, as shown in the illustration. The intrachain distance between A atoms is the shortest distance between atoms in the A15 crystal structure, and is about 22% less than the smallest interchain distance between A atoms. The extraordinary superconducting properties of the A15 compounds are believed to be primarily associated with these linear chains of transition-element A atoms.

At low temperature, some of the A15 phases undergo a diffusionless transformation from a cubic to a tetragonal structure, in which the length of one of the sides of the unit cell changes with respect to the length of the other two. This structural transformation was first observed in V_3Si at approximately 21 K (−422°F), and has also been found in Nb_3Sn at 43 K (−382°F), V_3Ga at above 50 K (−370°F), and Nb_3Al at 80 K (−316°F). *See* CRYSTAL STRUCTURE.

Superconductivity. Superconductivity is a phenomenon in which a metal, when cooled below its superconducting transition temperature T_c, loses all resistance to the flow of an electric current. The A15 compound Nb_3Ge has a T_c of 23.2 K (−417.9°F), the highest value of the A15 compounds, and, prior to 1986, the highest value ever observed for any known material. Other A15 compounds with high T_c include Nb_3Al (18.8 K or −425.8°F), Nb_3Ga (20.3 K or −423.1°F), Nb_3Si (approximately 19 K or −426°F), Nb_3Sn (18.0 K or −427.3°F), V_3Ga (15.9 K or −431°F), and V_3Si (17.1 K or −428.9°F).

In 1986, evidence of superconductivity with a T_c of approximately 30 K (−405°F) was reported for a material containing the elements lanthanum, barium, copper, and oxygen. Since then, several families of compounds containing copper, oxygen, and



A15 crystal structure of the A_3B intermetallic compounds. The light spheres represent the A atoms; the dark spheres represent the B atoms. The linear chains of A atoms are emphasized.

other elements have been discovered with yet higher superconducting transition temperatures, including a value of 135 K (−217°F) for a compound made of mercury, thallium, barium, calcium, copper, and oxygen. The origin of high-temperature superconductivity in the copper oxide compounds has not yet been established and may be different from that in the conventional superconductors such as the A15 compounds. In a conventional superconductor, the interaction of the negatively charged electrons with the positively charged ions of the crystal lattice produces an attractive interaction between electrons, causing them to form electron pairs, called Cooper pairs, that are responsible for the superconductivity.

The superconductivity of the A15 compounds can be destroyed with a sufficiently high magnetic field, called the upper critical magnetic field, $H_{c2}(T)$, which depends on the temperature T and varies from material to material. As the temperature is lowered, $H_{c2}(T)$ generally increases from zero at T_c to a maximum value, $H_{c2}(0)$, as the temperature approaches absolute zero (0 K or −459.67°F). The A15 compounds also have very high values of $H_{c2}(0)$, the highest of which is 44 teslas (T) for a pseudobinary A15 compound with the composition $Nb_{79}(Al_{73}Ge_{27})_{21}$. Other A15 compounds with high values of $H_{c2}(0)$ include Nb_3Al (32 T), Nb_3Ge (39 T), Nb_3Sn (23 T), V_3Ga (21 T), and V_3Si (25 T). For comparison, the highest value of $H_{c2}(0)$ ever observed for a conventional superconducting material is about 60 T for the Chevrel-phase compound $PbMo_6S_8$, while the values of $H_{c2}(0)$ for the highest-T_c copper oxide superconductors may reach several hundred teslas, values so large that they cannot readily be measured by using the magnetic fields and techniques that are presently available in the laboratory. *See* CHEVREL PHASES; SUPERCONDUCTIVITY.

Technological applications. Because of their high values of T_c and $H_{c2,}$ and critical current density J_c, the A15 compounds have a number of important technological applications. (The critical current density is the highest electric current per unit area that a material can conduct and still remain superconducting in a given magnetic field.) Processes have been developed for preparing multifilamentary superconducting wires that consist of numerous filaments of a superconducting A15 compound, such as Nb_3Sn, embedded in a nonsuperconducting copper matrix. Superconducting wires can be used in electric power transmission lines and to wind electrically lossless coils (solenoids) for superconducting electrical machinery (motors and generators) and magnets. Superconducting magnets are employed to produce intense magnetic fields for laboratory research, confinement of high-temperature plasmas in nuclear fusion research, bending beams of charged particles in accelerators, levitation of high-speed trains, mineral separation, nuclear magnetic resonance imaging, and energy storage. If conductors of the high-temperature, high-field oxide superconductors can eventually be fabricated with high values of J_c, they may revolutionize the technological applications of superconductivity. *See* SUPERCONDUCTING DEVICES.

M. Brian Maple

Bibliography. W. Buckle, *Superconductivity: Fundamentals and Applications*, 1991; P. Fulde (ed.), *Superconductivity of Transition Metals*, Springer Series in Solid State Sciences, vol. 27, 1982; T. H. Geballe and J. K. Hulm, Superconductivity: The state that came in from the cold, *Science*, 239:367–374, 1988; Y. Kao (ed.), *Superconductivity and Its Applications,* 1992; V. Z. Kresin, H. Morawitz, and S. A. Wolf, *Mechanics of Conventional and High T_c Superconductivity*, 1993.

Aardvark

A large piglike mammal in the order Tubulidentata. The order contains a single family, Orycteropodidae, and a single species, *Orycteropus afer*. Aardvarks occur throughout Africa south of the Sahara, wherever suitable habitat exists. In the past, tubulidentates were often considered closely related to ungulates. However, recent data from mitochondrial and nuclear genes support a relationship among aardvarks, elephant-shrews, paenungulates (hyraxes, sirenians, and proboscideans), and golden moles (Chrysochloridae). These ecologically divergent adaptive types probably originated in Africa; the molecular evidence implies that they may have arisen there from a common ancestor that existed in the Cretaceous Period, when Africa was isolated from other continents. This order shows the results of an extreme adaptation for burrowing and feeding on small food items (mainly termites). *See* MAMMALIA; TUBULIDENTATA.

Characteristics. Aardvarks, also known as antbears, resemble a medium-sized to large pig (see **illustration**). The body is massive with an elongate head and snout. The round, blunt, piglike muzzle ends in circular nostrils that possess fleshy tentacles on the nasal septum and dense tracts of white hairs that can serve as a filter and seal the nostrils to prevent soil from entering the lungs. The tough, thick skin is sparsely covered with bristly hair ranging from dull brownish gray to dull yellowish

Aardvark, *Orycteropus afer*. (*Photo by Lloyd Glenn Ingles; © 2001 California Academy of Sciences*)

gray. Numerous vibrissae (hairs with specialized erectile tissues) are present on the face. The waxy, smooth, tubular donkeylike ears are large and can be folded back to exclude dirt when the animal is burrowing, and they can be moved independently of each other. The strong, muscular tail is kangaroolike, tapering to a point. Incisors and canines are lacking in the dentition of adults but are present in the deciduous dentition. Teeth on the sides of the jaw are simple, peglike, open-rooted (not anchored in the jaw), and grow continuously during the animal's lifetime. The teeth are hollow; instead of a pulp cavity, they are transversed by a number of tubules. The dental formula is I 0/0 C 0/0 Pm 2/2 M 3/3 × 2 for a total of 20 teeth. Anal scent glands are present in both sexes. The senses of sight and hearing are acute, but their eyesight appears to be poor. Adult aardvarks have a head and body length of 100–158 cm (39–62 in.), a tail length of 44–71 cm (17–28 in.), and a shoulder height of 60–65 cm (23–25 in.). Most weigh 50–70 kg (110–153 lb), although some individuals may reach 100 kg (219 lb). Aardvarks exhibit a low basal metabolic rate and a low body temperature of about 34.5°C (94.1°F). *See* DENTITION; SCENT GLAND.

The aardvark is one of the world's great diggers. Its short, thick legs (with four toes on the front feet, five on the rear ones) are armed with long, powerful, sharp-edged, spoon-shaped claws that are intermediate between hoof and nail. The aardvark uses its claws to excavate the extensive burrows in which it dwells. Able to dig with amazing speed, it digs shallow holes when searching for food, larger burrows used for temporary shelter, and extensive tunnel systems in which the young are born. Tunnels may be 13 m (43 ft) long with numerous chambers and several entrances. Only mothers and their young share burrows. Burrows dug by aardvarks may be used by a variety of mammals, including warthogs, hyenas, porcupines, jackals, bat-eared foxes, hares, bats, ground squirrels, and civets, as well as monitor lizards and owls. One bird, the ant-eating chat (*Myrmecocicla formicivora*), nests in occupied burrows. In agricultural areas, burrows may cause damage to farming equipment and earthen dams.

Aardvarks are primarily nocturnal and solitary, and occupy a variety of habitats, including grassy plains, bush country, woodland, and savannah. They appear to prefer sandy soils. The main factor in their distribution is the presence of sufficient quantities of termites and ants. The aardvark is of considerable economic importance in keeping the great hordes of termites in check; if termites are not controlled, they do serious damage. The numbers of termites and ants consumed by the big-bodied aardvark in a single night are staggering. Studies have found that more than 50,000 insects may be consumed by an aardvark in one night. The long, round, thin, sticky, protrusible foot-long tongue is broader than that of the typical anteater or pangolin; together with the well-developed salivary glands, it can retrieve more termites in a shorter period of time than the two others. Since aardvarks do not need to chew their food, they have no need for incisors or canine teeth. Instead, their stomachs have a muscular pyloric area that serves a function similar to a gizzard in grinding up the food. They occasionally eat beetles, grasshoppers, and the fruit of the wild cucumber, known as the aardvark pumpkin, apparently as a source of water. *See* ANT; ANTEATER; ISOPTERA.

Breeding and development. A single altricial offspring (a young that is born immature and helpless, thus requiring extended parental care) is born after a gestation period of 7–9 months. Births generally occur from July to November. After remaining in the burrow for about 2 weeks, the young aardvark begins accompanying its mother on her nightly foraging excursions. The young can dig for itself when about 6 months old. Sexual maturity is attained at about 2 years of age. Captive aardvarks have lived up to 10 years.

Threats. Large flesheaters such as the lion, leopard, and ratel (a fierce, badgerlike animal) frequently prey on aardvarks. Pythons also take the young. African natives relish the aardvark's meat. They wear its teeth and claws as bracelets or necklaces as good-luck charms and for warding off illnesses. The tough hide may be made into straps and bracelets. Aardvarks have been greatly reduced in numbers and distribution and are now on Appendix 2 of the Convention on International Trade in Endangered Species (CITES). Termites have caused enormous damage to pasture and cereal crops in areas where aardvarks and other insectivorous animals have been exterminated. *See* ENDANGERED SPECIES.

Donald W. Linzey

Bibliography. D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Abaca

One of the strongest of the hard fibers, commercially known as Manila hemp. Abaca is obtained from the leafstalks of a member of the banana family, *Musa textilis*. The plant resembles the fruiting banana, but is a bit shorter in stature, bears small inedible fruits, and has leaves that stand more erect than those of the banana and are slightly narrower, more pointed, and 5–7 ft (1.5–2 m) long. Relatives of abaca grow wild throughout Southeast Asia, but the plant was domesticated long ago in the southern Philippines. Experiments have succeeded in growing plants yielding fiber in a few other parts of the world, chiefly Middle America, but commercial production has come almost entirely from the Philippines and Borneo. *See* ZINGIBERALES.

Environment and cultivation. Abaca prefers a warm climate with year-round rainfall, high humidity, and absence of strong winds. Soils must always be moist but the plant does not tolerate waterlogging. Abaca grows best on alluvial soils in the southern Philippines and northern Borneo below 1500 ft (457 m) elevation. The plant is best propagated by rootstalk

Appearance of diseased abaca. (*a*) Affected by mosaic; note ragged leaves and chlorotic areas on leaves. (*b*) Bunchy top; plant 1 on left is dead and lower leaves of plant 2 have been killed. (*USDA*)

suckers. There are about 75 varieties grown in the Philippines, grouped into seven categories, each of which varies slightly in height, length, and quality and yield of fiber. Newly planted stock reaches cutting size in 18 months, and a few stalks may be cut from each "hill" every 4 months thereafter for 12–15 years, after which new plantings are made. Plantings on good soils yield 90–150 lb (41–68 kg) of fiber per acre per year from 2 tons (1.8 metric tons) of stalks and the fiber yield ranges 2–4%. On large commercial holdings plants are grown in solid stands with the new stock spaced 15–20 ft (4.5–6 m) apart, but on small farms plants are scattered and mixed with bamboo and fruit trees.

Properties and uses. The fiber ranges 6–14 ft (1.8–4 m) in strand length, is lustrous, and varies from white to dull yellow. Pre-Spanish Filipinos made clothing with the fine fiber, which today has been replaced by softer, more pliable textiles. As one of the longest and strongest plant fibers, resistant to fresh and salt water, abaca is favored for marine hawsers and other high-strength ropes. Abaca is also used in sackings, mattings, strong papers, and handicraft art goods.

Processing and production. The fibers are carried in the outer sections of the leafstalks, and the fleshy material is stripped away mechanically. Traditionally, leafstalks were drawn by hand over a knife blade, under a wooden wedge, and a worker could clean about 15 lb (6.8 kg) of fiber per day. Small power machines strip about 175 lb (79 kg) per worker-day. Large mechanical decorticators save labor and recover more fiber, but the capital outlay is very large.

Diseases. Abaca is affected by several diseases, of which the chief are bunchy top, mosaic, and wilt (see **illus.**). Bunchy top is caused by a virus spread by the banana aphid (*Pentalonia nigronervosa*); leaves become smaller and clustered at the plant top, after which the plant dies. Mosaic is also caused by a virus spread by aphids (chiefly *Rhopalosiphum nymphaeae* and *Aphis gossypii*); the leaves yellow and dry out as the plant dies. Abaca wilt is caused by a soil or water-borne fungus, chiefly attacking plant roots. Filipino small farmers are not easily brought into disease-control cooperative measures, and postwar epidemics of bunchy top and mosaic have caused serious decline in acreage and production. *See* PLANT PATHOLOGY. Elton G. Nelson

Bibliography. R. E. Huke et al., *Shadows on the Land: An Economic Geography of the Philippines*, 1963; W. Manshard, *Tropical Agriculture*, 1975; B. B. Robinson and F. L. Johnson, *Abaca: A Cordage Fiber,* Agr. Monogr. 21, 1954; J. E. Spencer, The abaca plant and its fiber, Manila hemp, *Econ. Bot.*, 7:195–213, 1953; F. L. Wernstedt and J. E. Spencer, *The Philippine Island World: A Physical, Cultural and Regional Geography*, 1967.

Abacus

An early mechanical calculator whose design has evolved through the centuries, with two styles in use today. Both the Chinese and the Japanese styles consist of a frame with a crossbeam. They may be made from many different materials, such as wood

or brass. Rods or wires carrying sliding beads extend vertically through the crossbeam. The Chinese *suan pan* has two beads above the beam on each rod and five beads below. Each rod of the Japanese *soroban* carries one bead above and four below. Similar to the abacus in both construction and use, but much larger, are the counting frames used in elementary schools. Braille versions of the abacus are available for use by those without sight.

Operation. In working with whole numbers, the rightmost rod represents the ones position, with each rod to the left representing the tens, hundreds, thousands, and so forth, respectively. The beads below the crossbeam represent one of that rod's units (that is, a one, a ten, a hundred, and so forth), and those above represent five. Beads are moved from the outer position toward the crossbeam when used to represent a number (**Fig. 1**).

By a series of motions using the thumb and forefinger, numbers may be added and subtracted. Addition and subtraction both work from left to right, in contrast to the usual method, using substantial mental arithmetic. For example, to add 384 + 795, the user begins with the beads positioned to represent 384 (Fig. 1). Then to add 795, one lower bead on the fourth rod from the right is moved up, while simultaneously moving the three beads on the third rod down (700 = 1000 − 300). Next, moving to the tens position, the user adds the 90 by the mental method of "add 100 and subtract 10," thus moving one lower bead up on the hundreds rod and one lower bead down from the center on the tens rod. Finally, to add the 5, the user simply brings one of the upper beads down to the crosspiece. The result of the addition, 384 + 795 = 1179, then appears (**Fig. 2**).

Subtraction is done in a similar manner. In order to subtract 7 from any position, for example, the user subtracts 10 and adds 3. Both multiplication and division can also be performed with the help of the abacus, but these are much more complicated.

Use. The abacus, in contrast to the electronic calculator, is simply an aid to mental computation. A well-developed facility with numbers is required in order to use it effectively. For this reason, it is the calculator of choice for teachers in the Far East. The advent of the electronic calculator, while a boon to the scientific community, may have little impact on shop owners who are comfortable with the abacus. *See* CALCULATORS.

Marjorie K. Gross

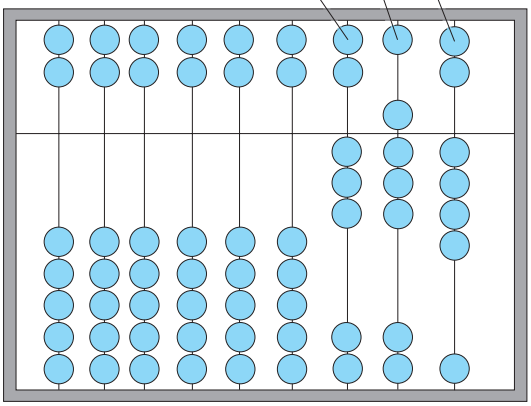


Fig. 1. Beads positioned to represent 384 on a Chinese abacus.

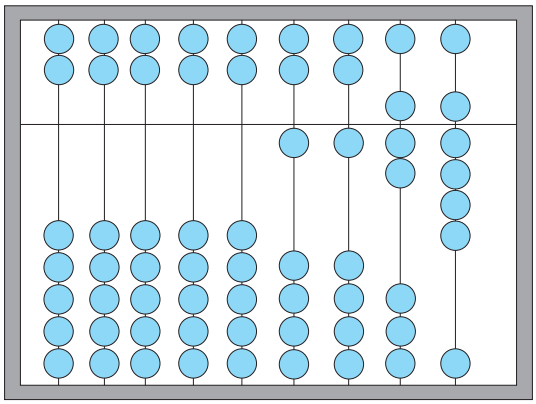


Fig. 2. Beads positioned to represent 384 + 795 = 1179 on a Chinese abacus.

Bibliography. American Printing House for the Blind, Inc., *Abacuses*, June 1999; C. B. Boyer (revised by U. C. Merzbach), *A History of Mathematics*, 3d ed., 1991; J. Dilson, *The Abacus*, St. Martin's Press, 1995.

Abalone

Common name in California for several species of marine gastropod mollusks in the family Haliotidae; also known as ear shells in many tropical and warm-temperate seas, as awabi in Japan, as paua in New Zealand, and as ormers in the Mediterranean region, Spain, and the English Channel.

The haliotid limpets constitute the most successful family of the primitive two-gilled gastropods included in the order Archaeogastropoda. They have a broad muscular sucker foot, an ear-shaped limpet shell (up to 30 cm or 12 in. long in some species) with a row of about seven shell apertures (see **illus.**), and a pair of featherlike gills (aspidobranch ctenidia). They live on rock surfaces in the low intertidal and sublittoral, and are grazing microherbivores, scraping and ingesting algal films with the relatively unspecialized dentition of a rhipidoglossan radula. Haliotids are among the relatively few living gastropods which are zygobranch (with two aspidobranch ctenidia) and diotocardiac (with two auricles in the heart). However, such paired symmetry of pallial, urinogenital, and cardiac structures is almost certainly characteristic of the earliest (ancestral) stocks of true gastropods. In the living forms, which include the rarer pleurotomariids and fissurellids as well as the haliotids, the symmetry of the organs requires a symmetrical respiratory flow through the branchial and pallial structures. The inhalant water current comes in to the mantle cavity from both sides ventrally, and the exhalant current passes out centrally and dorsally. In all of these living zygobranch gastropods

Shell of a typical haliotid limpet, or abalone, showing the characteristic curved series of exhalant apertures.

(and all the more numerous fossil forms) there are shell slits or complex apertures in the shell to ensure that the exhalant current, bearing with it both feces and urinogenital products, does not discharge directly over the head. In the haliotids (see illus.) these form a curved series of six or seven exhalant apertures, maintained during shell growth by the oldest (most apical) aperture being filled from below with shell material while a new aperture is formed initially as an indentation in the margin of the shell. Under the penultimate functioning aperture lies the anus, and thus feces are flushed out of the mantle cavity dorsally and well away from the mouth and sense organs. The sexes are separate, with broadcast spawning (again through the older apertures) and external fertilization. The planktonic larvae are typical veligers.

In all regions where they occur, flesh of the massive foot of these haliotids is prized as food, and the flattened ear-shaped shells, with characteristically green-blue iridescent inner surfaces (termed mother-of-pearl), have been used over many centuries in decorative arts by Maoris, Japanese, Amerindians (Wakashans), and Byzantine Greeks. Collected almost to extinction in some regions for their combined shell beauty and gourmet value, abalones and other haliotids, where protected, have recovered to become again very abundant. *See* GASTROPODA; LIMPET; PROSOBRANCHIA. W. D. Russell-Hunter

ABC lipid transporters

A family of adenosine triphosphate–binding cassette (ABC) transmembrane proteins that use energy to transport various molecules across extracellular and intracellular membranes (cytoplasmic membranes, endoplasmic reticulum, mitochondria, or peroxisomes). ABC lipid transporters consist of two hydrophobic transmembrane domains and two hydrophilic nucleotide-binding folds. These binding folds contain highly conserved sequence motifs (Walker A and B) and are separated by a linker sequence, also known as the signature (C) motif. ABC transporters are organized as full transporters or half transporters depending on the number of transmembrane domains and nucleotide-binding folds. Half transporters must form homodimers or heterodimers to be functional.

ABC transporters are grouped into seven major subfamilies named ABCA to ABCG. The transmembrane domains of the ABC proteins form a chamber within the membrane and act as a "flippase" that gives inner leaflet substrates access to their binding sites. The substrates transported by the ABC proteins can vary from small ions to large polypeptides and polysaccharides.

A great number of the ABC transporters are involved with cellular lipid transport and homeostasis, and the defect or absence of these transporters is often correlated with pathologic phenotypes. The **table** illustrates the major ABC transporters currently associated with lipid transport or homeostasis.

ABCA1. ABCA1 (ABC1; CERP, cholesterol efflux regulatory protein) is a member of the subfamily of ABC full-size transporters, containing two highly conserved ATP-binding cassettes (catalytic regions) and two transmembrane domains. ABCA1 is involved in the formation of high-density lipoprotein (HDL) and the cellular efflux of excess cholesterol and phospholipids. The 149-kb gene is located on the human chromosome 9q22-q31. It comprises 50 exons and codes for a calculated 220-kDa protein that is ubiquitously expressed, with highest expression in placenta, liver, lung, adrenal glands, and testes, and modest expression in small intestine, lung, and adipose. The expression of ABCA1 is highly regulated on a transcriptional as well as posttranslational level by several agents, including cholesterol, fatty acids, liver X receptor agonists, cAMP [cyclic adenosine monophosphate], and cytokines.

ABCA1 has been identified as the molecular defect in Tangier disease, a rare genetic disorder characterized by nearly complete HDL deficiency, high plasma triglyceride levels, orange tonsils, and increased susceptibility to atherosclerosis. Studies in ABCA1 knockout and transgenic mouse models as well as in-vitro studies have led to major advances in our understanding of the ABCA1-mediated formation of plasma HDL and its role in reverse cholesterol transport. Such transport is considered an antiatherogenic process, in which excess cholesterol is transported from peripheral tissues to the liver in the form of HDL, where it can be excreted from the body directly through the bile or indirectly after conversion into bile acids.

ABCA1 is responsible for the initial lipidation of apoA-I, the major apoprotein found on HDL. These nascent or pre-β HDL particles function as acceptors for further lipidation with cholesterol and phospholipids from peripheral tissues, such as lipid-laden foam cells found in atherosclerotic lesions. ABCA1 also interacts with serum amyloid A to form large, dense HDL particles associated with pathological conditions such as acute chronic inflammation.

Major ABC transporters currently associated with lipid transport or homeostasis

Transporter	Function	Disease
ABCA1	Cholesterol and phospholipid transport	Tangler disease
ABCA3	Surfactant metabolism	Neonatal surfactant deficiency
ABCA7	Cholesterol and phospholipid transport	Sjögren's syndrome
ABCG1	Cholesterol and phospholipid transport	Unknown
ABCG2	Drug resistance	Unknown
ABCG4	Cholesterol transport	Unknown
ABCG5	Sterol transport	Sitosterolemia
ABCG8	Sterol transport	Sitosterolemia

Plasma HDL concentrations are regulated primarily by liver-derived (~70%) and intestine-derived (~30%) ABCA1 activity and are inversely correlated with coronary artery disease. Complete ABCA1 deficiency in mice did not alter atherosclerosis, but repopulation of ABCA1-deficient mice with wild-type or ABCA1-overexpressing macrophages led to significantly decreased atherosclerosis. Accordingly, selective inactivation of macrophage ABCA1 in mice results in markedly increased atherosclerosis. However, overexpression of ABCA1 in other tissues, including liver, increased plasma cholesterol concentrations of both the antiatherogenic HDL and the proatherogenic non-HDL, and resulted in both pro- and antiatherogenic effects in different mouse models. Thus, changes in ABCA1 expression alter atherogenesis in a tissue-specific way. Based on these results, upregulation of ABCA1 for therapeutic use may be less desirable, or it may have to target specific tissues to achieve an overall antiatherogenic effect.

ABCA3. The ABCA3 gene maps to chromosome 16p13.3, consists of a 1704-amino-acid polypeptide, and is expressed predominantly in lung. ABCA3 gene mutations cause a fatal deficiency of surfactant in newborns. Pulmonary surfactant is mostly composed of lipids that are densely organized into multilamellar structures. These structures are critical to the transition of the lung from an aqueous to an air environment at birth. In adult lung, surfactants are also critical in preserving its homeostasis. Cellular electron micrographs of ABCA3-deficient human lung cells showed lipid accumulation in the multilamellar vesicles, suggesting that ABCA3 plays an important role in the formation of pulmonary surfactant, probably by transporting phospholipids and cholesterol.

ABCA7. ABCA7 is a highly expressed transporter in spleen, thymus, lymphoid cells, bone marrow, brain, and trachea. It is mapped to chromosome 19 and has 46 introns, and it is a full-size transporter. ABCA7 is highly homologous to ABCA1 and induces an apolipoprotein-mediated assembly of cholesterol similar to that induced by ABCA1, releasing cholesterol and phospholipids to HDL in an apoA-I–dependent mechanism. In mice, it was reported to promote the apoA-I–dependent release of phospholipids, but not cholesterol. Another observation in ABCA7–deficient females, but not male mice, reported less visceral fat and lower total serum and HDL cholesterol levels than wild-type, gender-matched littermates. Also, in these same mice, the lack of ABCA7 did not alter cholesterol and phospholipid efflux.

ABCG1. ABCG1 (ABC8, or human *white* gene) is a member of the subfamily of ABC half-size transporters involved in intracellular lipid transport. In order to create a functional full-size transporter, ABCG1 can homodimerize and most probably can also heterodimerize with other half-size transporters. The human ABCG1 spans ~98 kb and comprises 23 exons, located on chromosome 21q22.3. ABCG1 is expressed in most tissues in mice and humans, with high baseline expression in mouse macrophages, lung, brain, spleen, and other tissues containing macrophages, while in humans the expression is highest in lung, adrenal glands, spleen, heart, placenta, and liver. Several isoforms and splice variants have been found on the RNA level, yet the different protein levels and their functionality remain to be established. Expression of ABCG1 is strongly induced in many tissues by cholesterol loading, by ligands for the nuclear receptors LXRα and β, and by PPARγ-ligands.

ABCG1 is responsible for the control of tissue lipid levels and mediates cholesterol and phospholipid efflux to mature HDL or other phospholipid-containing acceptors, particularly in macrophages under conditions of cholesterol excess. The targeted disruption of ABCG1 in mice on a high-fat/high-cholesterol diet resulted in massive accumulation of neutral lipids and phospholipids in hepatocytes and tissue macrophages; however, no changes in plasma lipids accompanied these findings. In accordance with the regulation of cellular lipid homeostasis, the tissues in mice overexpressing human ABCG1 were protected from diet-induced lipid accumulation.

ABCG1 was found to be localized perinuclear as well as at the plasma membrane of cells. One potential mechanism for the ABCG1-facilitated lipid transfer is by redistributing membrane cholesterol to cell-surface domains that are accessible to removal by HDL. The role of ABCG1 in cholesterol export of lipid-laden cells seems to complement the action of ABCA1, a full-size transporter with similar tissue expression pattern, which mediates the initial lipidation in the formation of HDL. The nascent HDL particles are substrates for ABCG1-mediated lipid efflux, making ABCG1 part of the reverse cholesterol transport, a protective mechanism against the critical step of macrophage lipid overloading in atherogenesis.

ABCG2. Human ABCG2 is also referred to as placenta-specific ABC transporter, breast cancer

resistance protein, or mitoxantrone resistance–associated protein. (Mitoxantrone is a chemotherapy drug used to treat certain cancers.) It was initially identified as a messenger RNA expressed in placenta. The human ABCG2 gene is located on chromosome 4q22 and encodes a half transporter. In normal human tissues, ABCG2 has been found to be expressed on the apical membrane of trophoblast cells in placenta, the apical membrane of enterocytes, the bile canalicular membrane of hepatocytes, and the apical membrane of lactiferous ducts in the mammary gland. In addition, ABCG2 has been demonstrated to be expressed in a wide variety of hematopoietic cells. It is expected that ABCG2 homodimerizes to acquire transporter activity.

ABCG2 transporter functions as a xenobiotic transporter which may play a major role in multidrug resistance. It serves as a cellular defense mechanism in response to mitoxantrone and anthracycline (a chemotherapeutic agent) exposure. Although the function of ABCG2 has been studied extensively in terms of multidrug resistance, the physiological and/or pharmacological functions of this transporter have not been clarified yet. ABCG2 can also transport several dyes, such as rhodamine 123 and Hoechst 33462. Recently, it has been shown to transport steroids such as cholesterol, estradiol, progesterone, and testosterone. It also transports sulfated conjugates of steroids (esterone 3-sulfate, dehydroepiandrosterone sulfate) and certain chlorophyll metabolites.

ABCG4. The ABCG4 gene consists of 14 exons that span 12.6 kb on chromosome 11. ABCG4 gene sequence, protein, regulation, and function are very similar to ABCG1. The ABCG4 protein contains 646 amino acids and shares 72% amino acid sequence identity with ABCG1. However, the tissue distribution of ABCG1 and ABCG4 is not identical: ABCG1 is present in many tissues, including macrophages, lung, brain, liver, and spleen, while ABCG4 is mainly expressed in the brain and spleen. ABCG4 expression in the brain has been suggested to promote cholesterol efflux to HDL-like particles present in cerebrospinal fluids. Low level of ABCG4 has also been found in mouse macrophages, and its expression, similarly to ABCG1, was also upregulated by oxysterols (22-OH cholesterol), retinoids (9-*cis*-retinoic acid), and a (liver X receptor)–specific agonist (T0901713). Because of the different pattern of expression, it seems very unlikely that ABCG1 and ABCG4 act as heterodimers. Instead, evidence that ABCG4 acts as homodimer comes from studies that demonstrate that overexpression of ABCG4 stimulated cholesterol efflux to HDL.

Both ABCG4 and ABCG1 play a major role in the HDL-mediated cholesterol efflux in liver X receptor–induced macrophages. In-vitro studies showed that ABCG4 expression alone increased cholesterol efflux to both HDL-2 and HDL-3, and when both ABCG4 and ABCG1 were overexpressed, the increase in cholesterol efflux was even higher. ABCG4-mediated cholesterol efflux to LDL and cyclodextrin was also observed, but in lesser degree compared with the HDL efflux. ABCG4 suppression reduced the cholesterol efflux to HDL, but also decreased ABCG1 messenger RNA expression, probably due to homology in their sequences. These data indicate that ABCG4 alone or in association with ABCG1 may modulate the intracellular concentration of cholesterol by promoting cholesterol efflux to HDL.

ABCG5 and ABCG8. ABCG5 and ABCG8 are two half transporters that belong to the ABCG subfamily. They were identified as proteins mutated in sitosterolemia, a rare genetic disorder characterized by increased plasma and tissue levels of plant sterols. Patients with sitosterolemia have an increased absorption of dietary sterols and an impaired ability to secrete cholesterol and plant-derived sterols into bile. Many sitosterolemic individuals deposit cholesterol and plant sterols in the skin, tendons, and coronary arteries and suffer premature coronary artery disease.

ABCG5 and ABCG8 genes contain 13 exons and are organized in a head-to-head orientation on chromosome 2 with only 140 base pairs separating their two respective start-transcription sites. In-vivo and in-vitro data indicate that ABCG5 and ABCG8 function as obligate heterodimers to promote sterol excretion into bile. The major regulation of their expression appears to be at the transcriptional level: cholesterol feeding activates the liver X receptor, which increases both ABCG5 and ABCG8 messenger RNA expression. ABCG5/G8 are expressed mainly at the apical surface of hepatocytes and enterocytes, where they limit the accumulation of sterols by promoting sterol excretion into bile and reducing dietary sterol uptake. Overexpression of human ABCG5 and ABCG8 in wild-type mice led to increased cholesterol secretion into bile, and the knockouts of either ABCG5/G8 or ABCG8 alone prevented sterol secretion into bile. However, no changes in liver or plasma cholesterol levels were found. No protection from atherosclerosis was observed in these studies. However, in LDL receptor–deficient mice, which are prone to develop atherosclerosis on a high-fat/high-cholesterol diet, ABCG5/G8 overexpression decreased the absorption of dietary cholesterol, increased the excretion of cholesterol into bile and stool, and decreased cholesterol levels in plasma, leading to a significant reduction in the development of atherosclerotic lesions. *See* ADENOSINE TRIPHOSPHATE (ATP); ATHEROSCLEROSIS; CHOLESTEROL; GENE; LIPID METABOLISM; LIPOPROTEIN; LIVER; PROTEIN.

E. M. Wagner; F. Basso; C. S. Kim; M. J. A. Amar

Bibliography. H. B. Brewer, Jr., et al., Regulation of plasma high-density lipoprotein levels by the ABCA1 transporter and the emerging role of high-density lipoprotein in the treatment of cardiovascular disease, *Arterioscler. Thromb. Vasc. Biol.*, 24:1755–1760, 2004; M. Dean, The Human ATP-Binding Cassette (ABC) Transporter Superfamily, http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=&rid=mono_001.TOC&depth=2; T. Janvilisri et al., Sterol transport by the human breast cancer resistance protein (ABCG2) expressed in *Lactococcus*

lactis, *J. Biol. Chem.*, 278:20645–20651, 2003; W. Jessup et al., Roles of ATP binding cassette transporters A1 and G1, scavenger receptor BI and membrane lipid domains in cholesterol export from macrophages. *Curr. Opin. Lipidol.*, 17:247–257, 2006; J. Santamarina-Fojo et al., Complete genomic sequence of the human ABCA1 gene: Analysis of the human and mouse ATP-binding cassette A promoter, *Proc. Nat. Acad. Sci. USA*, 97:7987–7992, 2000; K. Takahashi et al., ABC proteins: Key molecules for lipid homeostasis, *Med. Mol. Morphol.*, 38:2–12, 2005; N. Wang et al., ATP-binding cassette transporters G1 and G4 mediate cellular cholesterol efflux to high-density lipoproteins, *Proc. Nat. Acad. Sci. USA*, 101:9774–9779, 2004; K. R. Wilund et al., High-level expression of ABCG5 and ABCG8 attenuates diet-induced hypercholesterolemia and atherosclerosis in Ldlr −/− mice, *J. Lipid Res.*, 45:1429–1436, 2004.

Abdomen

A major body division of the vertebrate trunk lying posterior to the thorax; and in mammals, bounded anteriorly by the diaphragm and extending to the pelvis. The diaphragm, found only in mammals, separates the abdominal or peritoneal cavity from the pleural and pericardial cavities of the thorax. In all pulmonate vertebrates (possessing lungs or lunglike organs) other than mammals, the lungs lie in the same cavity with the abdominal viscera, and this cavity is known as the pleuroperitoneal cavity.

The large coelomic cavity that occupies the abdomen contains the viscera, such as the stomach, liver, gallbladder, spleen, pancreas, and intestinal tract with their associated nerves and blood vessels. The viscera are contained in the peritoneal sac. This structure is adherent to the abdominal wall as the parietal peritoneum and encompasses all parts of the viscera as an outermost coat, the serosa or visceral peritoneum. Connecting sheets of peritoneum from the body wall to the various organs form the mesenteries, which are always double-walled; blood vessels, lymphatics, nerves, and lymph nodes are located between the two walls of the mesenteries. Other folds of the peritoneum form the omenta. At the lesser curvature of the stomach the serosa continues upward to the liver as the lesser omentum, and from the greater curvature it hangs down as a sheet in front of the intestine, forming the greater omentum.

The term abdomen is also applied to a similar major body division of arthropods and other animals. *See* ANATOMY, REGIONAL.

Walter Bock

Aberration (optics)

A departure of an optical image-forming system from ideal behavior. Ideally, such a system will produce a unique image point corresponding to each object point. In addition, every straight line in the object space will have as its corresponding image a unique straight line. A similar one-to-one correspondence will exist between planes in the two spaces.

This type of mapping of object space into image space is called a collinear transformation. A paraxial ray trace is used to determine the parameters of the transformation, as well as to locate the ideal image points in the ideal image plane for the system. *See* GEOMETRICAL OPTICS; OPTICAL IMAGE.

When the conditions for a collinear transformation are not met, the departures from that ideal behavior are termed aberrations. They are classified into two general types, monochromatic aberrations and chromatic aberrations. The monochromatic aberrations apply to a single color, or wavelength, of light. The chromatic aberrations are simply the chromatic variation, or variation with wavelength, of the monochromatic aberrations. *See* CHROMATIC ABERRATION.

Aberration measures. The monochromatic aberrations can be described in several ways. Wave aberrations are departures of the geometrical wavefront from a reference sphere with its vertex at the center of the exit pupil and its center of curvature located at the ideal image point. The wave aberration is measured along the ray and is a function of the field height and the pupil coordinates of the reference sphere (**Fig. 1**).

Transverse ray aberrations are measured by the transverse displacement from the ideal image point to the ray intersection with the ideal image plane.

Some aberrations are also measured by the longitudinal aberration, which is the displacement along the chief ray of the ray intersection with it. The use

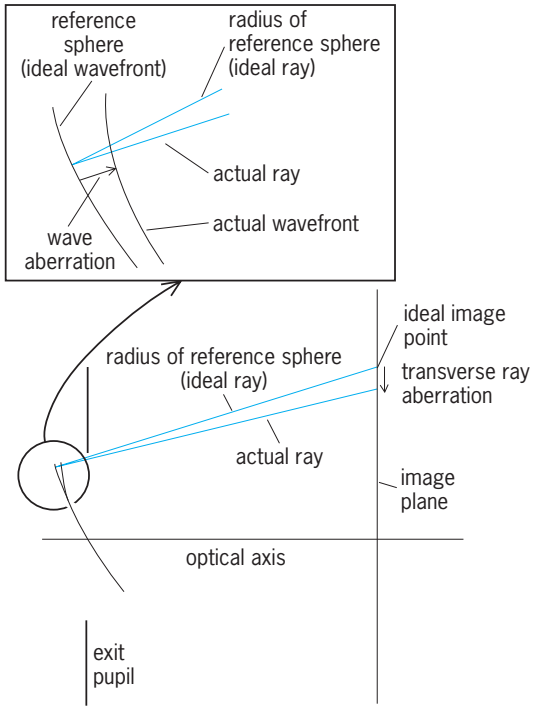


Fig. 1. Diagram of the image space of an optical system, showing aberration measures: the wave aberration and the transverse ray aberration.

of this measure will become clear in the discussion of aberration types.

Caustics. Another aberrational feature which occurs when the wavefront in the exit pupil is not spherical is the development of a caustic. For a spherical wavefront, the curvature of the wavefront is constant everywhere, and the center of curvature along each ray is at the center of curvature of the wavefront. If the wavefront is not spherical, the curvature is not constant everywhere, and at each point on the wavefront the curvature will be a function of orientation on the surface as well as position of the point. As a function of orientation, the curvature will fluctuate between two extreme values. These two extreme values are called the principal curvatures of the wavefront at that point. The principal centers of curvature lie on the ray, which is normal to the wavefront, and will be at different locations on the ray.

The caustic refers to the surfaces which contain the principal centers of curvature for the entire wavefront. It consists of two sheets, one for each of the two principal centers of curvature. For a given type of aberration, one or both sheets may degenerate to a line segment. Otherwise they will be surfaces.

The feature which is of greatest interest to the user of the optical system is usually the appearance of the image. If a relatively large number of rays from the same object point and with uniform distribution over the pupil are traced through an optical system, a plot of their intersections with the image plane represents the geometrical image. Such a plot is called a spot diagram, and a set of these is often plotted in planes through focus as well as across the field.

Orders of Aberrations

The monochromatic aberrations can be decomposed into a series of aberration terms which are ordered according to the degree of dependence they have on the variables, namely, the field height and the pupil coordinates. Each order is determined by the sum of the powers to which the variables are raised in describing the aberration terms. Because of axial symmetry, alternate orders of aberrations are missing from the set. For wave aberrations, odd orders are missing, whereas for transverse ray aberrations, even orders are missing. It is customary in the United States to specify the orders according to the transverse ray designation.

Monochromatically, first-order aberrations are identically zero, because they would represent errors in image plane location and magnification, and if the first-order calculation has been properly carried out, these errors do not exist. In any case, they do not destroy the collinear relationship between the object and the image. The lowest order of significance in describing monochromatic aberrations is the third order.

Chromatic variations of the first-order properties of the system, however, are not identically zero. They can be determined by first-order ray tracing for the different colors (wavelengths), where the refractive

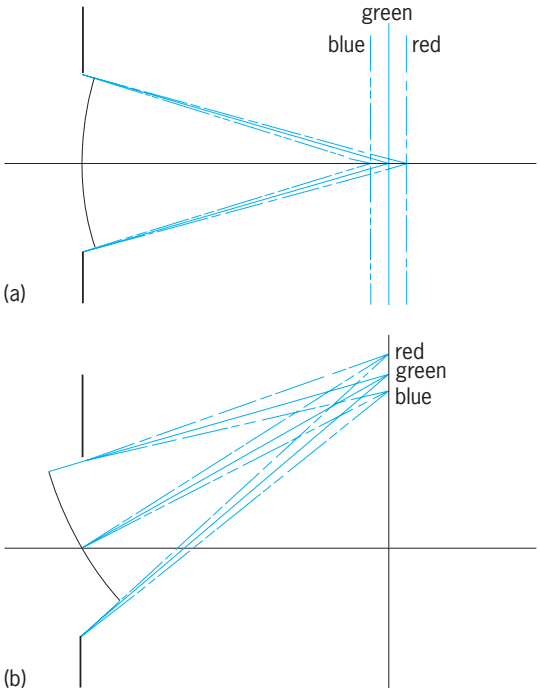


Fig. 2. Chromatic aberration. (*a*) Longitudinal chromatic aberration. (*b*) Transverse chromatic aberration.

indices of the media change with color. For a given object plane, the different colors may have conjugate image planes which are separated axially. This is called longitudinal chromatic aberration (**Fig. 2***a*). Moreover, for a given point off axis, the conjugate image points may be separated transversely for the different colors. This is called transverse chromatic aberration (Fig. 2*b*), or sometimes chromatic difference of magnification.

These first-order chromatic aberrations are usually associated with the third-order monochromatic aberrations because they are each the lowest order of aberration of their type requiring correction.

The third-order monochromatic aberrations can be divided into two types, those in which the image of a point source remains a point image but the location of the image is in error, and those in which the point image itself is aberrated. Both can coexist, of course.

Aberrations of geometry. The first type, the aberrations of geometry, consist of field curvature and distortion.

Field curvature. Field curvature is an aberration in which there is a focal shift which varies as a quadratic function of field height, resulting in the in-focus images lying on a curved surface. If this aberration alone were present, the images would be of good quality on this curved surface, but the collinear condition of plane-to-plane correspondence would not be satisfied.

Distortion. Distortion, on the other hand, is an aberration in which the images lie in a plane, but they are displaced radially from their ideal positions in the image plane, and this displacement is a cubic function of the field height. This means that any straight

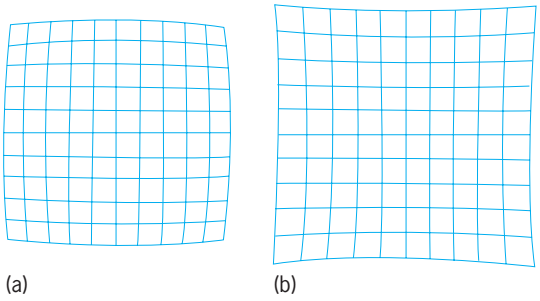


Fig. 3. Distortion. (*a*) Barrel distortion. (*b*) Pincushion distortion.

line in the object plane not passing through the center of the field will have an image which is a curved line, thereby violating the condition of straight-line to straight-line correspondence. For example, if the object is a square centered in the field, the points at the corners of the image are disproportionally displaced from their ideal positions in comparison with the midpoints of the sides. If the displacements are toward the center of the field, the sides of the figure are convex; this is called barrel distortion (**Fig. 3***a*). If the displacements are away from the center, the sides of the figure are concave; this is called pincushion distortion (Fig. 3*b*).

Aberrations of point images. There are three third-order aberrations in which the point images themselves are aberrated: spherical aberration, coma, and astigmatism.

Spherical aberration. Spherical aberration is constant over the field. It is the only monochromatic aberration which does not vanish on axis, and it is the axial case which is easiest to understand.

The wave aberration function (**Fig. 4***a*) is a figure of revolution which varies as the fourth power of the radius in the pupil. The wavefront itself has this wave aberration function added to the reference sphere centered on the ideal image. The rays (Fig. 4*b*) from any circular zone in the wavefront come to a common zonal focus on the axis, but the position of this focus shifts axially as the zone radius increases. This zonal focal shift increases quadratically with the zone radius to a maximum from the rays from the marginal zone. This axial shift is longitudinal spherical aberration. The magnitude of the spherical aberration can be measured by the distance from the paraxial focus to the marginal focus.

The principal curvatures of any point on the wavefront are oriented tangential and perpendicular to the zone containing the point. The curvatures which are oriented tangential to the zone have their centers on the axis, so the caustic sheet for these consists of the straight line segment extending from the paraxial focus, and is degenerate. The other set of principal centers of curvature lie on a trumpet-shaped surface concentric with the axis (**Fig. 5***a*). This second sheet is the envelope of the rays. They are all tangent to it, and the point of tangency for each ray is at the second principal center of curvature for the ray (Fig. 5*b*).

It is clear that the image formed in the presence of spherical aberration (Fig. 4*c*) does not have a well-defined focus, although the concentration of light outside the caustic region is everywhere worse than it is inside. Moreover, the light distribution in the image is asymmetric with respect to focal position, so the precise selection of the best focus depends on the criterion chosen. The smallest circle which can contain all the rays occurs one-quarter of the distance from the marginal focus to the paraxial focus, the image which has the smallest second moment lies one-third of the way from the marginal focus to the paraxial focus, and the image for which the variance of the wave aberration function is a minimum lies halfway between the marginal and paraxial foci.

Coma. Coma is an aberration which varies as a linear function of field height. It can exist only for off-axis field heights, and as is true for all the aberrations, it is symmetrical with respect to the meridional plane containing the ideal image point. Each zone in its wave aberration function (**Fig. 6***a*) is a circle, but each circle is tilted about an axis perpendicular to the meridional plane, the magnitude of the tilt increasing with the cube of the radius of the zone.

The chief ray (Fig. 6*b*) passes through the ideal

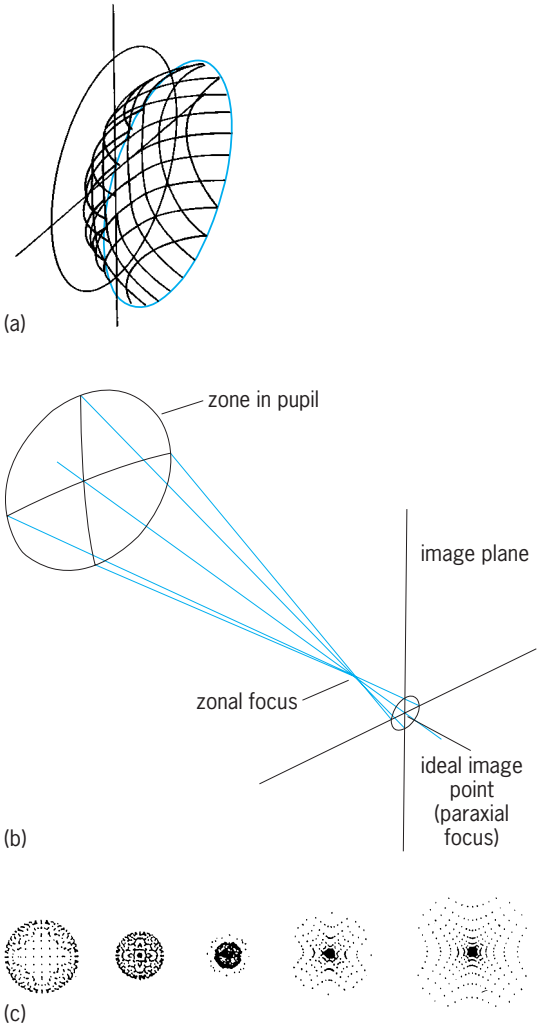


Fig. 4. System with spherical aberration. (*a*) Wave aberration function. (*b*) Rays. (*c*) Spot diagrams through foci showing transverse ray aberration patterns for a square grid of rays in the exit pupil.

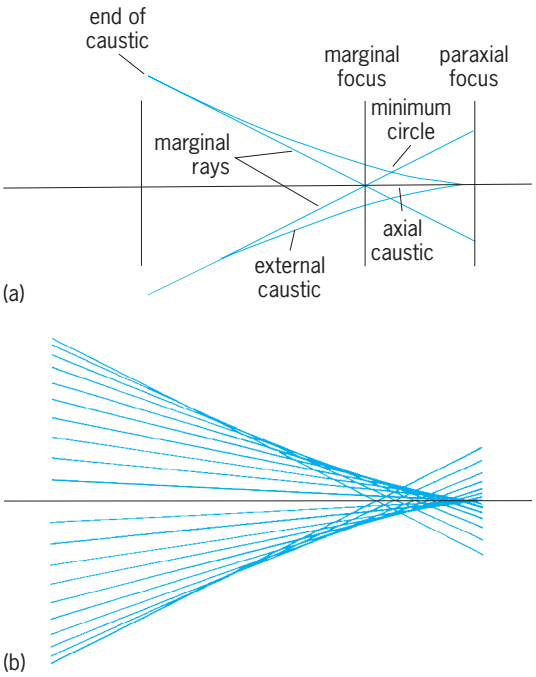


Fig. 5. Caustics in a system with spherical aberration. (*a*) Diagram of caustics and other major features. (*b*) Rays whose envelope forms an external caustic.

image point, but the rays from any zone intersect the image plane in a circle, the center of which is displaced from the ideal image point by an amount equal to the diameter of the circle. The diameter increases as the cube of the radius of the corresponding zone in the pupil. The circles for the various zones are all tangent to two straight lines intersecting at the ideal image point and making a 60° angle with each other. The resulting figure (Fig. 6*c*) resembles an arrowhead which points toward or away from the center of the field, depending on the sign of the aberration.

The upper and lower marginal rays in the meridional plane intersect each other at one point in the circle for the marginal zone. This point is the one most distant from the chief ray intersection with the image plane. The transverse distance between these points is a measure of the magnitude of the coma.

Astigmatism. Astigmatism is an aberration which varies as the square of the field height. The wave aberration function (**Fig. 7***a*) is a quadratic cylinder which varies only in the direction of the meridional plane and is constant in the direction perpendicular to the meridional plane. When the wave aberration function is added to the reference sphere, it is clear that the principal curvatures for any point in the wavefront are oriented perpendicular and parallel to the meridional plane, and moreover, although they are different from each other, each type is constant over the wavefront. Therefore, the caustic sheets both degenerate to lines perpendicular to and in the meridional plane in the image region, but the two lines are separated along the chief ray.

All of the rays must pass through both of these lines, so they are identified as the astigmatic foci. The astigmatic focus which is in the meridional plane is called the sagittal focus, and the one perpendicular to the meridional plane is called the tangential focus.

For a given zone in the pupil, all of the rays (Fig. 7*b*) will of course pass through the two astigmatic foci, but in between they will intersect an image plane in an ellipse, and halfway between the foci they will describe a circle (Fig. 7*c*). Thus, only halfway between the two astigmatic foci will the image be isotropic. It is also here that the second moment is a minimum, and the wave aberration variance is a minimum as well. This image is called the medial image.

Since astigmatism varies as the square of the field height, the separation of the foci varies as the square of the field height as well. Thus, even if one set of foci, say the sagittal, lies in a plane, the medial and tangential foci will lie on curved surfaces. If the field curvature is also present, all three lie on curved surfaces. The longitudinal distance along the chief ray from the sagittal focus to the tangential focus is a measure of the astigmatism.

The above description of the third-order aberrations applies to each in the absence of the other aberrations. In general, more than one aberration will be present, so that the situation is more complicated. The types of symmetry appropriate to each aberration will disclose its presence in the image.

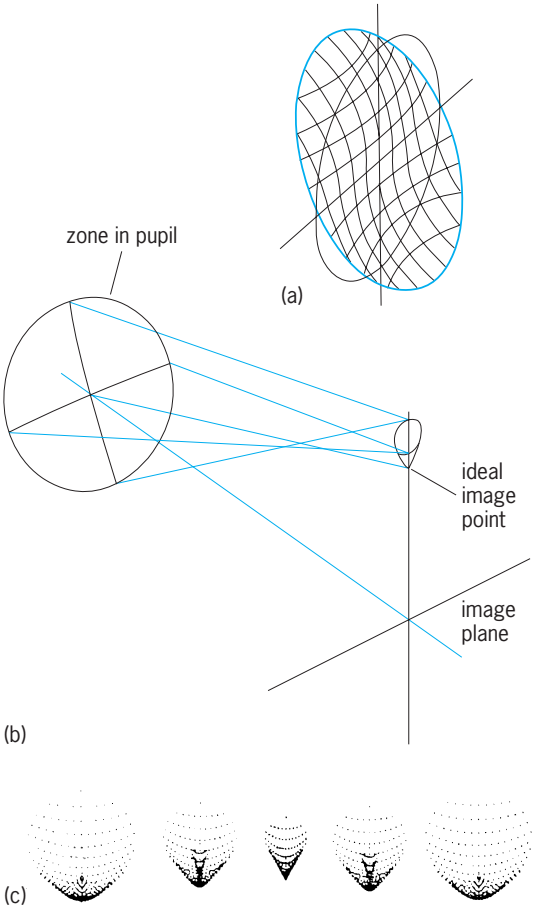


Fig. 6. System with coma. (*a*) Wave aberration function. (*b*) Rays. (*c*) Spot diagrams through foci, showing transverse ray aberration patterns for a square grid of rays in the exit pupil.

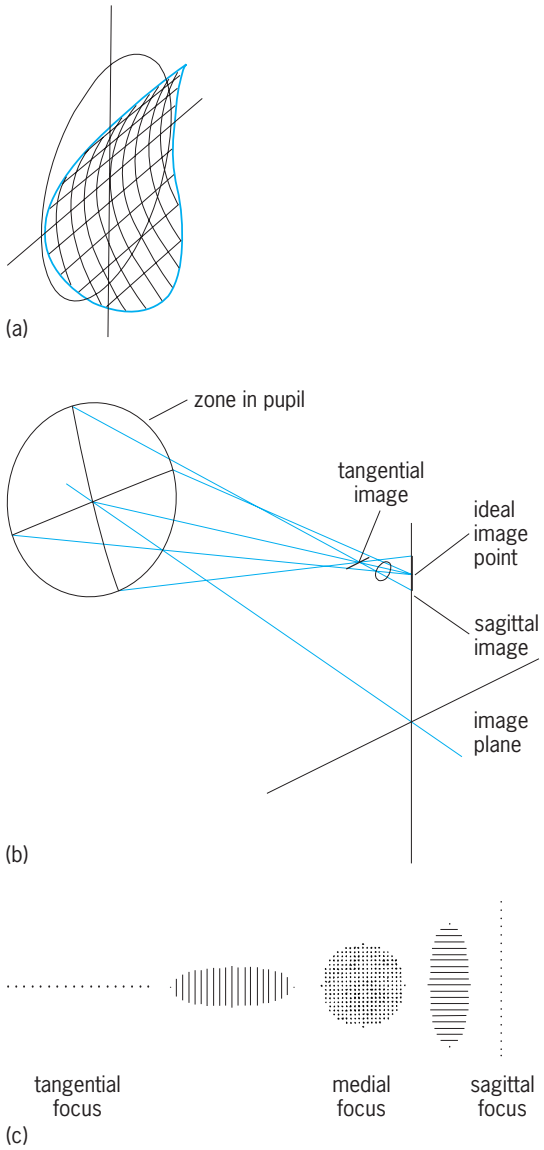


Fig. 7. System with astigmatism. (*a*) Wave aberration function. (*b*) Rays. (*c*) Spot diagrams through foci, showing transverse aberration patterns for a square grid of rays in the exit pupil.

Higher-order aberrations. The next order of aberration for the chromatic aberrations consists of the chromatic variation of the third-order aberrations. Some of these have been given their own names; for example, the chromatic variation of spherical aberration is called spherochromatism.

Monochromatic aberrations of the next order are called fifth-order aberrations. Most of the terms are similar to the third-order aberrations, but with a higher power dependence on the field or on the aperture. Field curvature, distortion, and astigmatism have a higher power dependence on the field, whereas spherical aberration and coma have a higher power dependence on the aperture. In addition, there are two new aberration types, called oblique spherical aberration and elliptical coma. These are not directly related to the third-order terms.

Expansions beyond the fifth order are seldom used, although in principle they are available. In fact, many optical designers use the third order as a guide in developing the early stages of a design, and then go directly to real ray tracing, using the transverse ray aberrations of real rays without decomposition to orders. However, the insights gained by using the fifth-order aberrations can be very useful.

Origin of Aberrations

Each surface in an optical system introduces aberrations as the beam passes through the system. The aberrations of the entire system consist of the sum of the surface contributions, some of which may be positive and others negative. The challenge of optical design is to balance these contributions so that the total aberrations of the system are tolerably small. In a well-corrected system the individual surface contributions are many times larger than the tolerance value, so that the balance is rather delicate, and the optical system must be made with a high degree of precision.

Insight as to where the aberrations come from can be gained by considering how the aberrations are generated at a single refracting surface. Although the center of curvature of a spherical surface lies on the optical axis of the system, it does not in itself have an axis. If, for the moment, the spherical surface is assumed to be complete, and the fact that the entrance pupil for the surface will limit the beam incident on it is ignored, then for every object point there is a local axis which is the line connecting the object point with the center of curvature. All possible rays from the object point which can be refracted by the surface will be symmetrically disposed about this local axis, and the image will in general suffer from spherical aberration referred to this local axis.

A small pencil of rays about this axis (locally paraxial) will form a first-order image according to the rules of paraxial ray tracing. If the first-order imagery of all the points lying in the object plane is treated in this manner, it is found that the surface containing the images is a curved surface. The ideal image surface is a plane passing through the image on the optical axis of the system, and thus the refracting surface introduces field curvature. The curvature of this field is called the Petzval curvature.

In addition to this monochromatic behavior of the refracting surface, the variation in the index of refraction with color will introduce changes in both the first-order imagery and the spherical aberration. This is where the chromatic aberrations come from.

Thus there are fundamentally only three processes operating in the creation of aberrations by a spherical refracting surface. These result in spherical aberration, field curvature, and longitudinal chromatic aberration referred to the local axis of each image. In fact, if the entrance pupil for the surface is at the center of curvature, these will be the only aberrations that are contributed to the system by this refracting surface.

In general, the pupil will not be located at the center of curvature of the surface. For any off-axis object point, the ray which passes through the center of the

pupil will be the chief ray, and it will not coincide with the local axis. The size of the pupil will limit the beam which can actually pass through the surface, and the beam will therefore be an eccentric portion of the otherwise spherically aberrated beam.

Since an aberration expansion decomposes the wave aberration function about an origin located by the chief ray, the eccentric and asymmetric portion of an otherwise purely spherically aberrated wave gives rise to the field-dependent aberrations, because the eccentricity is proportional to the field height of the object point.

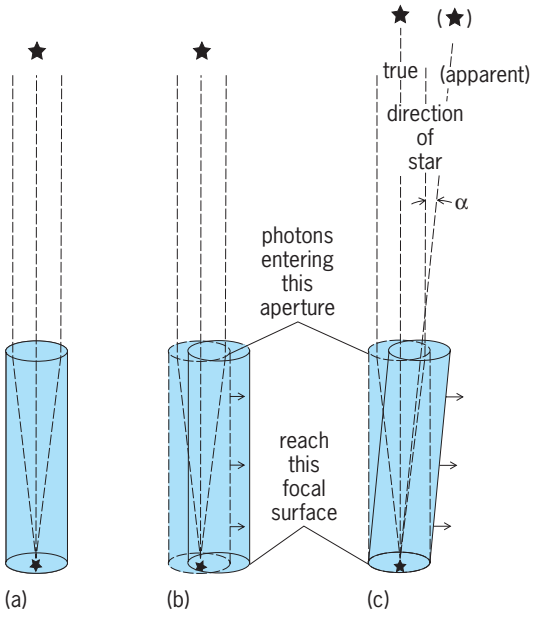
In this manner the aberrations which arise at each surface in the optical system, and therefore the total aberrations of the system, can be accounted for. *See* LENS (OPTICS); OPTICAL SURFACES. Roland V. Shack

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; B. D. Guenther, *Modern Optics*, 1990; R. Kingslake, *Optical System Design*, 1983; K. Narigi and Y. Matsui, *Fundamentals of Practical Aberration Theory*, 1993; W. Smith, *Modern Optical Engineering*, 2d ed., 1990; W. T. Welford, *Aberrations of Optical Systems*, 1986.

Aberration of light

The apparent change in direction of a source of light caused by an observer's component of motion perpendicular to the impinging rays.

To visualize the effect, first imagine a stationary telescope (**illus.** *a*) aimed at a luminous source such as a star, with photons traveling concentrically down the tube to an image at the center of the focal plane. Next give the telescope a component of motion perpendicular to the incoming rays (illus. *b*). Photons passing the objective require a finite time to travel the length of the tube. During this time the telescope has moved a short distance, causing the photons to reach a spot on the focal plane displaced from the former image position. To return the image to the center, the telescope must be tilted in the direction of motion by an amount sufficient to ensure that the photons once again come concentrically down the tube in its frame of reference (illus. *c*). The necessary tilt angle α is given by $\tan \alpha = v/c$, where v is the component of velocity perpendicular to the incoming light and c is the velocity of light. (An analogy illustrating aberration is the experience that, in order for the feet to remain dry while walking through vertically falling rain, it is necessary to tilt an umbrella substantially forward.)



Demonstration of aberration. (*a*) Fixed telescope; photons form image at center of focal plane. (*b*) Moving telescope; image is displaced from center. (*c*) Tilted moving telescope is required to restore image to center.

Aberration of light was discovered by the English astronomer James Bradley in 1725 while searching unsuccessfully for parallax of nearby stars, using the Earth's orbital diameter as baseline. He found instead the necessity to compensate continuously for the Earth's velocity in its elliptical orbit, an effect about a hundred times greater than the parallax of typical nearby stars. This discovery provided the first direct physical confirmation of the Copernican theory.

Annual aberration causes stars to appear to describe ellipses whose major axes are the same length, the semimajor axis (usually expressed in arc-seconds) being the constant of aberration $\alpha = 20.49552''$. The minor axis of the ellipse depends on the star's ecliptic latitude β, and is given by $2\alpha \sin \beta$. The small diurnal aberration arising from the Earth's axial rotation depends on geographic latitude ϕ, and is given by $0.31'' \cos \phi$.

A second important application of aberration has been its clear-cut demonstration that, as is axiomatic to special relativity, light reaching the Earth has a velocity unaffected by the relative motion of the source toward or away from the Earth. This is shown by the fact that the aberration effect is the same for all celestial objects, including some quasars with apparent recessional velocities approaching the speed of light. *See* EARTH ROTATION AND ORBITAL MOTION; LIGHT; PARALLAX (ASTRONOMY); QUASAR; RELATIVITY. Harlan J. Smith

Abrasive

A material of extreme hardness that is used to shape other materials by a grinding or abrading action. Abrasive materials may be used either as loose grains, as grinding wheels, or as coatings on cloth or paper. They may be formed into ceramic cutting tools that are used for machining metal in the same way that ordinary machine tools are used. Because of their superior hardness and refractory properties, they have advantages in speed of operation, depth of cut, and smoothness of finish.

Abrasive products are used for cleaning and machining all types of metal, for grinding and polishing ceramics and glass, for grinding logs to paper pulp, for cutting metals, glass, and cement, and for

manufacturing many miscellaneous products such as brake linings and nonslip floor tile.

Abrasive materials. These may be classified in two groups, the natural and the synthetic (manufactured). The latter are by far the more extensively used, but in some specific applications natural materials still dominate.

The important natural abrasives are diamond (the hardest known material), corundum (a relatively pure, natural aluminum oxide, Al_2O_3), and emery (a less pure Al_2O_3 with considerable amounts of iron). The last of these has been replaced to a great extent by synthetic materials. Other natural abrasives are garnet, an aluminosilicate mineral; feldspar, used in household cleansers; calcined clay; lime; chalk; and silica, SiO_2, in its many forms—sandstone, sand (for grinding plate glass), flint, and diatomite.

The synthetic abrasive materials are silicon carbide, SiC; aluminum oxide, Al_2O_3; titanium carbide, TiC; and boron carbide, B_4C. The synthesis of diamond puts this material in the category of manufactured abrasives. There are other carbides, as well as nitrides and cermets, which can be classified as abrasives but their use is special and limited. The properties of hardness and high melting point are related, and many abrasive materials are also good refractories. *See* DIAMOND; REFRACTORY.

Silicon carbide is still made in much the same way that E. G. Acheson first made it in 1891. The principal ingredients are pure sand, SiO_2; coke (carbon); sawdust (to burn and provide vent holes for the escape of the gaseous products); and salt (to react with the impurities and form volatile compounds). The batch is placed in a troughlike furnace, up to 60 ft (18 m) long, with a graphite electrode at each end. The charge is heated by an electric current, a core of carbon or graphite being used to carry the current. Temperatures up to 4400°F (2400°C) are reached. The net reaction is $SiO_2 + 3C \rightarrow SiC + 2CO$, but the details of the reaction are complicated and not thoroughly understood. After the furnace has cooled, the sides are removed and the usable silicon carbide picked out, crushed, and sized.

Abrasive aluminum oxide is made from calcined bauxite (a mixture of aluminum hydrates) or purified alumina by fusion in an electric-arc furnace with a water-cooled metal shell; the solid alumina that is around the edge of the furnace acts as the refractory.

Various grades of each type of synthetic abrasive are distinguishable by properties such as color, toughness, and friability. These differences are caused by variation in purity of materials and methods of processing.

The sized abrasive may be used as loose grains, as coatings on paper or cloth to make sandpaper and emery cloth, or as grains for bonding into wheels.

Abrasive wheels. A variety of bonds are used in making abrasive wheels: vitrified or ceramic, essentially a glass or glass plus crystals; sodium silicate; rubber; resinoid; shellac; and oxychloride. Each type of bond has its advantages. The more rigid ceramic bond is better for precision-grinding operations, and the tougher, resilient bonds, such as resinoid or rubber, are better for snagging and cutting operations.

Ceramic-bonded wheels are made by mixing the graded abrasive and binder, pressing to general size and shape, firing, and truing or finishing by grinding to exact dimensions. A high-speed rotation test is given to check the soundness of the wheel.

Grinding wheels are specified by abrasive type, grain size (grit), grade or hardness, and bond type; in addition, the structure or porosity may be indicated. The term hardness as applied to a wheel refers to its behavior in use and not to the hardness of the abrasive material itself. The wheel is a three-component system of abrasive, bond, and air; the hardness is a complex function of the type and amount of bond and of the density of the wheel.

Literally thousands of types of wheels are made with different combinations of characteristics, not to mention the multitude of sizes and shapes available; therefore, selecting the best grinding wheel for a given job is not simple. When the abrasive grains of a wheel become slightly dulled by use, the stresses in the grinding operation should increase enough to tear the grain from the wheel to expose a new cutting grain. Thus, too soft a wheel wears too fast, losing grains before they are dulled, whereas too hard a wheel develops a smooth, glazed surface that will not cut. In either case, efficiency drops.

Hardness tests. The hardness of materials is roughly indicated by the Mohs scale, in which 10 minerals were selected, arranged in order of hardness, and numbered from 1 to 10 (softest to hardest); diamond has a hardness number of 10 and talc a hardness number of 1. However, since this scale is not quantitative (for example, aluminum oxide, 9 on the Mohs scale, is not 10% softer than diamond), and since most abrasive materials fall in the region at the top of the scale, other hardness scales have been developed. One is the Knoop indentation test, in which a diamond pyramid of specified shape is pressed under a definite load into the material to be tested; the size of the indentation is taken as a measure of the material's hardness. *See* HARDNESS SCALES.

Ceramic cutting tools are generally made of aluminum oxide by hot-pressing and grinding to final size. *See* CERAMICS.

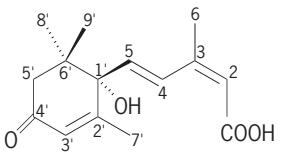
John F. McMahon

Bibliography. S. Krar and E. Ratterman, *Superabrasives: Grinding and Machining with Carbon and Diamond*,1990; J. Wang, *Abrasive Technology: Current Development and Applications I*,1999.

Abscisic acid

One of the five major plant hormones that play an important role in plant growth and development. Abscisic acid (ABA) is ubiquitous in higher plants; the highest levels are found in young leaves and developing fruits and seeds. Abscisic acid is also produced by certain algae and by several phytopathogenic fungi. It has been found in the brains of several mammals; however, animals probably do not produce abscisic acid but acquire it from plants in the diet.

Chemistry and biosynthesis. The chemical structure of (+)-abscisic acid is shown here; naturally



occurring abscisic acid is dextrorotary, hence the (+). It contains 15 carbon atoms and is a weak organic acid, as are two other plant hormones, auxin and gibberellin. *See* AUXIN; GIBBERELLIN.

Abscisic acid is a terpenoid. All chemicals belonging to this class of compounds have mevalonic acid as a common precursor and are built up of five carbon units (isoprenoid units). Other examples of terpenoids in plants are essential oils, gibberellins, steroids, carotenoids, and natural rubber. *See* CAROTENOID; ESSENTIAL OILS; RUBBER; STEROID.

In fungi, abscisic acid is synthesized by the so-called direct pathway, which means that abscisic acid molecules are built up by combining three isoprenoid units. The biosynthetic pathway in higher plants is not clear, but presumably it is an indirect pathway that involves the breakdown of a larger precursor molecule. Several lines of evidence suggest that this precursor may be a carotenoid: mutants lacking carotenoids also fail to produce abscisic acid, and inhibitors that block carotenoid biosynthesis at the same time inhibit abscisic acid formation. Furthermore, labeling studies with $^{18}O_2$ indicate that one ^{18}O atom is rapidly incorporated into the carboxyl group of abscisic acid, whereas the ring oxygens become labeled much more slowly. These observations are in accordance with the hypothesis that abscisic acid is derived from a large precursor molecule in which the ring oxygens are already present and an oxygen atom is introduced into the carboxyl group by oxidative cleavage.

Physiological roles. Abscisic acid was originally discovered by F. T. Addicott as a factor that accelerates abscission of young cotton fruits (hence the original name abscisin), and by P. F. Wareing as a hormone (called dormin) that causes bud dormancy in woody species. However, subsequent work established that abscisic acid has little or no effect on abscission of leaves and fruits and that it is not an endogenous regulator of bud dormancy in woody species. *See* ABSCISSION; DORMANCY.

Several approaches have been used to determine the roles of abscisic acid in plants. Application of abscisic acid results in a multitude of responses, but most of these may be of a pharmacological nature rather than reflect the roles of endogenous abscisic acid. The study of mutants that are deficient in abscisic acid, such as the flacca mutant of tomato and droopy in potato, has been more informative; both mutants synthesize much less abscisic acid than wild-type plants. These mutants wilt readily because they do not close their stomata. When they are treated with abscisic acid, the normal phenotype is restored, thus demonstrating that abscisic acid has a key role in preventing excessive water loss. Abscisic acid also functions in growth of roots and shoots, in heterophylly in aquatic plants, and in preventing premature germination of developing embryos. On the other hand, there is now increasing evidence against a role for abscisic acid in gravitropic curvature of roots. Seedlings in which abscisic acid synthesis is inhibited or mutants in which levels of abscisic acid are reduced, show the same gravitropic response of their roots as control seedlings.

Stress. Biosynthesis of abscisic acid is greatly accelerated under stress conditions. For example, in a wilted leaf the abscisic acid content can increase as much as 40 times the original level over a 4- to 5-h period. Following rehydration of such a leaf, the abscisic acid level returns to that of a turgid leaf after 5 to 6 h. Thus, the water status of the leaf, in particular the loss of turgor, regulates the rate at which abscisic acid is synthesized and degraded. The excess abscisic acid is either oxidized to phaseic acid, which in turn may be converted to dihydrophaseic acid, or it may be conjugated to glucose to give a complex that is stored in the vacuole.

A very rapid response to an increase in abscisic acid in a leaf is the closure of stomata: abscisic acid affects the permeability of the guard cells, which results in loss of solutes and turgor, so that the stomata close and the plant is protected from desiccation.

Abscisic acid is called a stress hormone, because it can ameliorate the effects of various stresses (drought, cold, salinity) to which plants are exposed. Freezing tolerance can be induced in certain plants by application of abscisic acid. *See* COLD HARDINESS (PLANT); PLANT-WATER RELATIONS; PLANTS, SALINE ENVIRONMENTS OF.

Growth. Abscisic acid inhibits the growth of stems and expansion of young leaves by decreasing cell-wall loosening, an effect opposite to that caused by auxin. On the other hand, growth of roots may be promoted by abscisic acid under certain conditions. These opposite effects of abscisic acid on the growth of roots and shoots are advantageous for the survival of plants under water stress conditions. By inhibiting shoot growth, turgor pressure is maintained, whereas a stimulation of root growth by abscisic acid enlarges the root system, thus increasing water uptake. *See* PLANT GROWTH.

Heterophylly. In aquatic plants two distinct types of leaves are produced on the individual. Submerged leaves are highly dissected or linear without stomata, whereas aerial leaves are entire and have stomata. When abscisic acid is added to the medium, aerial leaves with stomata are produced on submerged shoots. It is conceivable that under natural conditions shoot tips emerging from the water experience water stress, which causes production of abscisic acid, which in turn results in a change in leaf morphology.

Embryogenesis, seed maturation, and germination. Following fertilization in higher plants, the ovule develops into the seed with the embryo inside. During

late embryogenesis the seed desiccates and the embryo becomes dormant. However, if young embryos are excised from the seed prior to the onset of dormancy and cultured in the laboratory, they will germinate. This precocious germination of excised, immature embryos can be prevented by application of abscisic acid. There is convincing evidence that abscisic acid produced in the seed plays an important role in preventing precocious germination early in embryogenesis. Mutants deficient in abscisic acid all show the phenomenon of vivipary, which means that the embryo germinates precociously while the seed is still attached to the mother plant. Elegant genetic experiments with seeds of the small crucifer *Arabidopsis thaliana* have demonstrated that it is abscisic acid produced by the embryo, not by the maternal tissues of the seed, that controls dormancy of the embryo. Abscisic acid also induces the accumulation of specific proteins during late embryogenesis, but it is not clear at present that such proteins play a role in the onset of embryo dormancy. *See* PLANT MORPHOGENESIS; SEED.

During germination the reserve foods (carbohydrate, protein, and lipid) in the seed are degraded and mobilized to the young seedling. The hormonal regulation of enzymes involved in the mobilization of these reserves has been studied extensively in cereal grains, particularly in barley. Various hydrolytic enzymes, for example, α-amylase, are synthesized by aleurone cells in response to gibberellin produced by the germinating embryo. It has long been known that applied abscisic acid can inhibit the synthesis of these hydrolytic enzymes. Also, it has been found that abscisic acid produced in germinating seeds during water stress can have the same effect. So, the germinating seed not only has the means to turn on the metabolic machinery to mobilize food reserves but can also turn it off during conditions unfavorable for growth. *See* PLANT PHYSIOLOGY.

Mode of action. Two types of responses to abscisic acid can be distinguished. The rapid response, such as stomatal closure, occurs within a few minutes and can be induced only by naturally occurring (+)-abscisic acid. Since abscisic acid is active on guard cells over a wide pH range, while it is taken up by the cells only at low pH, the site of action is likely on the outer surface of the plasma membrane. Stomatal closure is caused by the rapid release of solutes, in particular the potassium ion (K^+) and malate, from the guard cells. Thus, in the rapid response abscisic acid affects the permeability of membranes.

The slow responses, which take hours or even days to become apparent, involve gene expression and can be induced equally well by (+)-abscisic acid and synthetic (−)-abscisic acid. This implies that the fast and slow responses have different receptor sites. The slow responses involve inhibition of the synthesis of certain enzymes, as well as induction of synthesis of a new group of proteins, the so-called abscisic acid–inducible proteins. These proteins appear during late embryogenesis when the seed begins to desiccate, and during water stress in other organs of the plant. The genes for these proteins are expressed only in the presence of abscisic acid or after water stress. When examined at the level of messenger ribonucleic acid, the effects of abscisic acid and water stress are not cumulative, indicating that both act by way of the same mechanism. Changes in protein pattern induced by water stress persist but are rapidly reversed by rehydration. The function of the abscisic acid–inducible proteins is not known, but it is thought that they have a function in drought tolerance. *See* PLANT HORMONES. Jan A. D. Zeevaart

Bibliography. F. T. Addicott (ed.), *Abscisic Acid*, 1983; J. A. D. Zeevaart and R. A. Creelman, Metabolism and physiology of abscisic acid, *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 39:439–473, 1988.

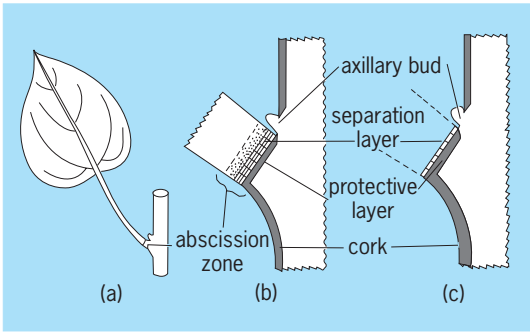
Abscission

The process whereby a plant sheds one of its parts. Leaves, flowers, seeds, and fruits are parts commonly abscised. Almost any plant part, such as very small buds and bracts to branches several inches in diameter, and scales or sheets of bark, may be abscised by some species. However, other species, including many annual plants, may show little abscission, especially of leaves.

Abscission may be of value to the plant in several ways. It can be a process of self-pruning, removing injured, diseased, or senescent parts. It permits the dispersal of seeds and other reproductive structures. It facilitates the recycling of mineral nutrients to the soil. It functions to maintain homeostasis in the plant, keeping in balance leaves and roots, and vegetative and reproductive parts. For example, trees living in climates that have a pronounced dry season are often adapted to abscising all or part of their leaves during the dry period, thus keeping the trees in balance with the water of the environment. Flowers and young fruit are usually produced in much larger numbers than the plant is capable of developing into mature fruit. Abscission of flowers and young fruits is the means whereby the plant keeps the number of fruits in balance with the ability of the leaves to nourish them. The selective abscission of branches and branchlets is an important factor in determining the architecture of a number of tree species.

Mechanisms. In most plants the process of abscission is restricted to an abscission zone at the base of an organ (see **illus**.); here separation is brought about by the disintegration of the walls of a special layer of cells, the separation layer. Less than an hour may be required to complete the entire process for some flower petals, but for most plant parts several days are required for completion of abscission. The portion of the abscission zone which remains on the plant commonly develops into a corky protective layer that becomes continuous with the cork of the stem.

In general, a healthy organ inhibits its own abscission. Only after it is injured, diseased, or senescent



Diagrams of the abscission zone of a leaf. (*a*) A leaf with the abscission zone indicated at the base of the petiole. (*b*) The abscission zone layers shortly before abscission and (c) the layers after abscission.

is an organ abscised. In many instances abscission is also a correlation phenomenon, resulting from events elsewhere in the plant. For example, the abscission of flower petals usually follows closely after pollination and fertilization; and in some broad-leaved evergreen trees, such as live oaks, the camphor tree, and *Magnolia grandiflora*, the major flush of leaf abscission follows closely after the appearance of new buds and leaves in the spring.

Abscission is a typical physiological process requiring favorable temperatures, oxygen, and energy-yielding respiration. The energy is required for the synthesis of hydrolytic enzymes. These enzymes function to soften the pectins and other constituents of the cell wall. When the cell walls of the separation layer are sufficiently weakened, the leaf (or other organ) falls from the plant by its own weight.

Abscission takes place in the lower plants (mosses, ferns, algae, and fungi) in the same way as in the higher plants. Enzymes soften and weaken the pectinlike materials that cement cells together. The cementing substances of higher and lower plants are quite similar in that all are composed of branching polysaccharides. Abscission in the lower plants is less conspicuous because of their simpler structure. However, propagules, such as buds and plantlets, are regularly abscised. In the algae, fragmentation by abscission of segments of filaments is a common method of vegetative propagation. Yeast buds, most mold spores, and bacterial colonies also are separated by the dissolution of polysaccharide cementing substances.

Evolution. The fossil record shows that abscission has been taking place for a very long time. The record contains many instances of the abscission of fruits, seeds, and seedlike structures going back millions of years. Leaf abscission became evident soon after leaves evolved in the Devonian Period, some 400 million years ago. Fossils of the oldest known green plants, blue-green algae of the Precambrian Period, 850 million years ago, show that their filaments were abscising (fragmenting), as do present-day blue-green algae. From the evidence summarized above, it is reasonable to conclude that the evolution of abscission began when two cells first separated. *See* FOSSIL SEEDS AND FRUITS; PALEOBOTANY.

Factors affecting abscission. Observations of higher plants show that abscission is affected by a number of environmental and internal factors. It can be initiated or accelerated by extremes of temperature, such as frost; by extremes of moisture, such as drought or flooding; by deficiency of mineral elements in the soil, particularly nitrogen, calcium, magnesium, potassium, and zinc; by the shortening photoperiod in the fall; by oxygen concentrations above 20%; by volatile chemicals, such as ethylene and air pollutants; by moderately toxic chemicals; and by the feeding of some injurious insects. Abscission can be retarded or inhibited by excessive nitrogen from the soil; by high carbohydrate levels in the plant; by oxygen concentrations below 20%; and by application of growth-regulator chemicals, such as naphthaleneacetic acid.

Hormones. Although a number of internal chemical changes regularly precede the abscission of an organ, at most only a few of the changes appear to be directly related to the process of abscission. Of the changes, a decrease in the level of the growth hormone auxin appears to be the most important. Auxin applied experimentally to the distal (organ) side of an abscission zone retards abscission, while auxin applied to the proximal (stem) side accelerates abscission. From this and related evidence, it is considered that the gradient of auxin across the abscission zone is the major internal factor controlling abscission. Further, many other factors affecting abscission appear to act through effects on the auxin gradient. *See* AUXIN.

The gibberellins are growth hormones which influence a number of plant processes, including abscission. When applied to young fruits or to leaves, they tend to promote growth, delay maturation, and thereby indirectly prevent or delay abscission. *See* GIBBERELLIN.

Abscisic acid is a different type of hormone and has the ability to promote abscission and senescence and to retard growth. Like auxin, it is synthesized in fruits and leaves; however, it tends to counteract the effects of the growth-promoting hormones, auxin, gibberellin, and cytokinin. *See* ABSCISIC ACID; CYTOKININS.

Another hormonal chemical is the gas ethylene. Small amounts of ethylene have profound effects on the growth of plants and can distort and reduce growth and promote senescence and abscission. *See* ETHYLENE.

Thus the process of abscission is strongly influenced by interaction of several plant hormones. Frequently the hormones that tend to retard or delay abscission (auxin, gibberellin) are counteracted by the hormones that tend to promote and accelerate abscission (abscisic acid, ethylene). These interactions are further influenced by many environmental and external factors, as mentioned above. Depending on circumstances, any one of several factors may at times be the key factor controlling the entire process of abscission. *See* PLANT HORMONES.

Agricultural applications. In agricultural practice, abscission is delayed or prevented by keeping soil

moisture and soil nitrogen at high levels. High soil moisture helps to keep abscisic acid at relatively low levels, and high soil nitrogen leads to high levels of auxin in the plant. Auxin-type chemical regulators are frequently used to retard abscission. Acceleration of abscission can be obtained by low soil nitrogen (leading to low auxin), by moisture stress (leading to high abscisic acid), by ethylene-releasing chemicals, or by mildly toxic chemicals.

To delay abscission of leaves, fruits, and other plant parts, the auxin-type regulator, naphthaleneacetic acid, or related compounds are commonly used. Some notable applications are the spraying of apple orchards and pear orchards to prevent abscission of nearly mature fruit and the dipping of cut holly to prevent abscission of berries and leaves during shipment.

To accelerate abscission a variety of chemicals find use in one or another practice. Cotton is chemically defoliated to facilitate mechanical harvest. To promote leaf abscission, chemicals such as magnesium chlorate, tributylphosphorotrithioate, and merphos have been widely used. In some regions, young nursery plants are defoliated to facilitate earlier digging and shipping. Certain varieties of apples, peaches, and similar fruits must be thinned to ensure that the remaining fruit can reach marketable size. Naphthaleneacetamide and related compounds are the most frequently used to thin fruit. Although these chemicals are of the auxin type, the dosages applied to the flowers and young fruit are somewhat toxic and lead to abortion and abscission of a portion of the young fruit. The slightly toxic insecticide carbaryl is also used in fruit thinning. Efficient harvesting of mature fruit often depends on some chemical promotion of abscission. Ethaphon and other compounds that can release ethylene are often used, although they sometimes show undesirable side effects of subsequent growth. For the harvesting of oranges in Florida, various combinations of cycloheximide, 5-chloro-3-methyl-4-nitro-1*H*-pyrazole, and chlorothalonil have been very effective in promoting abscission. *See* PLANT PHYSIOLOGY.

Fredrick T. Addicott

Bibliography. F. T. Addicott, *Abscission,* 1982; L. G. Nickell, *Plant Growth Regulators,* 1982.

Absolute zero

The temperature at which an ideal gas would exert no pressure. The Kelvin scale of temperatures is defined in terms of the triple point of water, $T_3 = 273.16^\circ$ (where the solid, liquid, and vapor phases coexist) and absolute zero. Thus, the Kelvin degree is 1/273.16 of the thermodynamic temperature of the triple point. Temperature is measured most simply via the constant-volume ideal-gas thermometer, in which a small amount of gas is introduced (in order to limit the effect of interactions between molecules) and then sealed off. The gas pressure P referenced to its value at the triple point $P(T_3)$ is measured, and temperature T may be inferred from Eq. (1). The

$$T(K) = 273.16 \lim_{P \to 0} \frac{P(T)}{P(T_3)} \tag{1}$$

ideal-gas law applies if the molecules in a gas exert no forces on one another and if they are not attracted to the walls. With helium gas, the constant-volume gas thermometer can be used down to a temperature on the order of 1 K. Absolute zero is the temperature at which the pressure of a truly ideal gas would vanish. *See* GAS THERMOMETRY; TEMPERATURE MEASUREMENT.

In order that a temperature be assigned to a particular specimen, there must exist mechanisms for the free transfer of energy within the system. Measurement of the distribution of speeds of particles can serve to define the temperature through the Maxwell distribution. The most probable speed gets smaller as the temperature decreases. According to classical physics, all motion would cease at absolute zero; however, the quantum-mechanical uncertainty principle requires that there be a small amount of residual motion (zero-point motion) even at absolute zero. *See* KINETIC THEORY OF MATTER; QUANTUM MECHANICS; UNCERTAINTY PRINCIPLE.

The Kelvin scale can be obtained by measurements of the efficiency of an ideal heat engine operating between reservoirs at two fixed temperatures. According to the second law of thermodynamics, the maximum possible efficiency ϵ is given by Eq. (2),

$$\epsilon = 1 - \frac{T_1}{T_2} \tag{2}$$

where T_1 and T_2 are the temperatures of the low- and the high-temperature reservoirs respectively. All comparisons between the gas-thermometer scales and the thermodynamic scales have shown the equivalence of the two approaches. *See* CARNOT CYCLE; THERMODYNAMIC PRINCIPLES.

Temperature can also be defined from the Boltzmann distribution. If a collection of spin 1/2 magnetic ions is placed in a magnetic field, the ratio of the occupancy of the lower to the higher energy state is given by Eq. (3). Here k is Boltzmann's constant, $|\Delta E|$ is the magnitude of the difference in energy between the states, and T is the Kelvin temperature. Thus, at high temperatures the two states have nearly equal occupation probability, while the lower energy state is progressively favored at lower temperatures. At absolute zero, only the lower energy level is occupied. By measuring the populations of the two states, the temperature of the spins can be inferred. This relation allows for the possibility of negative temperatures when the population of the higher energy state exceeds that of the lower state. From the point of view of energy content, negative temperatures correspond to an energy of the spin system that exceeds that of an infinite positive temperature, and thus they are hotter than ordinary

$$\frac{N_L}{N_H} = \exp \frac{|\Delta E|}{kT} \tag{3}$$

temperatures. *See* BOLTZMANN CONSTANT; BOLTZMANN STATISTICS; NEGATIVE TEMPERATURE.

Negative temperatures notwithstanding, the third law of thermodynamics states that the absolute zero of temperature cannot be attained by any finite number of steps. The lowest (and hottest) temperatures that have been achieved are on the order of a picokelvin (10^{-12}K). These are spin temperatures of nuclei which are out of equilibrium with the lattice vibrations and electrons of a solid. The lowest temperatures to which the electrons have been cooled are on the order of 10 microkelvins in metallic systems. *See* LOW-TEMPERATURE PHYSICS; TEMPERATURE.

Jeevak M. Parpia; David M. Lee

Absorption

Either the taking up of matter in bulk by other matter, as in the dissolving of a gas by a liquid; or the taking up of energy from radiation by the medium through which the radiation is passing. In the first case, an absorption coefficient is defined as the amount of gas dissolved at standard conditions by a unit volume of the solvent. Absorption in this sense is a volume effect: The absorbed substance permeates the whole of the absorber. In absorption of the second type, attenuation is produced which in many cases follows Lambert's law and adds to the effects of scattering if the latter is present. *See* ATTENUATION; SCATTERING OF ELECTROMAGNETIC RADIATION.

Energy. Absorption of electromagnetic radiation can occur in several ways. For example, microwaves in a waveguide lose energy to the walls of the guide. For nonperfect conductors, the wave penetrates the guide surface, and energy in the wave is transferred to the atoms of the guide. Light is absorbed by atoms of the medium through which it passes, and in some cases this absorption is quite distinctive. Selected frequencies from a heterochromatic source are strongly absorbed, as in the absorption spectrum of the Sun. Electromagnetic radiation can be absorbed by the photoelectric effect, where the light quantum is absorbed and an electron of the absorbing atom is ejected, and also by Compton scattering. Electron-positron pairs may be created by the absorption of a photon of sufficiently high energy. Photons can be absorbed by photoproduction of nuclear and subnuclear particles, analogous to the photoelectric effect. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; COMPTON EFFECT; ELECTRON-POSITRON PAIR PRODUCTION; PHOTOEMISSION.

Sound waves are absorbed at suitable frequencies by particles suspended in the air (wavelength of the order of the particle size), where the sound energy is transformed into vibrational energy of the absorbing particles. *See* SOUND ABSORPTION.

Absorption of energy from a beam of particles can occur by the ionization process, where an electron in the medium through which the beam passes is removed by the beam particles. The finite range of protons and alpha particles in matter is a result of this process. In the case of low-energy electrons, scattering is as important as ionization, so that range is a less well-defined concept. Particles themselves may be absorbed from a beam. For example, in a nuclear reaction an incident particle X is absorbed into nucleus Y, and the result may be that another particle Z, or a photon, or particle X with changed energy comes out. Low-energy positrons are quickly absorbed by annihilating with electrons in matter to yield two gamma rays. *See* NUCLEAR REACTION.

McAllister H. Hull, Jr.

Matter. The absorption of matter is a chemical engineering unit operation. In the chemical process industries and in related areas such as petroleum refining and fuels purification, absorption usually means gas absorption. This is a unit operation in which a gas (or vapor) mixture is contacted with a liquid solvent selected to preferentially absorb one, or in some cases more than one, component from the mixture. The purpose is either to recover a desired component from a gas mixture or to rid the mixture of an impurity. In the latter case, the operation is often referred to as scrubbing. The gas, once absorbed in the liquid, is in a completely dissolved state and therefore has become a full-fledged member of the liquid phase. In the chemical industry, gas absorption is the second-most common separation operation involving gas-liquid contacting. Only fractional distillation is employed more frequently.

When the operation is employed in reverse, that is, when a gas is utilized to extract a component from a liquid mixture, it is referred to as gas desorption, stripping, or sparging.

In gas absorption, either no further changes occur to the gaseous component once it is absorbed in the liquid solvent, or the absorbed component (solute) will become involved in a chemical reaction with the solvent in the liquid phase. In the former case, the operation is referred to as physical gas absorption, and in the latter case as gas absorption with chemical reaction. Examples of simple, or physical, gas absorption include the absorption of light oil (benzene, toluene, xylenes) from coke oven by-product gases by scrubbing with a petroleum oil, and recovery of valuable solvent vapors, such as acetone or those used in drycleaning processes, from gas streams by washing the gas with an appropriate solvent for the vapors. Examples of gas absorption with chemical reaction include scrubbing the flue gases from coal-fired generating stations with aqueous sodium carbonate solution to remove sulfur dioxide (a potential cause of acid rain), recovery of ammonia from coke oven by-product gases by scrubbing with dilute sulfuric acid, and the purification of natural gas by absorption of hydrogen sulfide in aqueous ethanolamine.

Gas absorption is normally carried out in the same types of columns as those used in fractional distillation, namely tray-type or packed towers, but with the latter used much more frequently than the former in distillation. The operation can be carried out by using cocurrent flow, that is, with both the gas and liquid phases flowing upward in the column, but it is normally countercurrent with the liquid phase flowing

downward and the gas phase rising. Gas absorption, unlike distillation, is not carried out at the boiling point. To the contrary, and because the solubility of gases in liquids increases with decreasing temperature, gas absorption is usually carried out cold, that is, at a temperature not far above the freezing point of the solvent. An evolution of heat occurs during gas absorption, as a result of the absorbed component giving up its latent heat of vaporization upon absorption, just as it would have if it had been condensed. The resulting temperature rise is minimized by a normally large solvent-to-gas ratio used in order to keep low the concentration of absorbed gas in the liquid phase, and therefore to keep high the concentration driving force causing the absorption to occur. Unlike fractional distillation, gas absorption does not require a reboiler, condenser, or reflux. However, it does require a subsequent solvent-recovery operation to separate solvent and solute, so that the solvent can be recycled. The energy requirements for gas absorption and solvent recovery combined are usually substantially less than for fractional distillation. The main basis for solvent selection is that the solvent have a high solubility for the gaseous component to be absorbed, and the lowest possible solubility for the other components of the gas stream—in other words, that the solvent have both a large absorptive capacity and a high degree of selectivity. *See* GAS ABSORPTION OPERATIONS; DISTILLATION; DISTILLATION COLUMN; UNIT OPERATIONS.

William F. Furter

Absorption (biology)

The net movement (transport) of water and solutes from outside an organism to its interior. The unidirectional flow of materials into an animal from the environment generally takes place across the alimentary tract, the lungs, or the skin, and in each location a specific cell layer called an epithelium regulates the passage of materials. *See* EPITHELIUM; RESPIRATION.

Structural aspects. The mammalian intestine is a good example of an epithelium which absorbs ions and water, along with a wide range of nutrient materials. The cells lining the gastrointestinal epithelium possess morphological features similar to those of many other epithelia.

Most intestinal epithelia have cells which are columnar in shape, have an extensive microvillar membrane facing the interior (or lumen) of the gut, and are functionally specialized to transport molecules from the gut contents to the blood (**Fig. 1**). The brush border or microvillar apical membrane has an expanded surface to maximize uptake of materials into the cytoplasm. Adjacent cells are attached to one another at the apical surface (facing the lumen) by "tight junctions," which serve as partial barriers to the passage of materials between the lumen and the blood by intercellular paths. Several types of enzymes, including peptidases, carbohydrases, and lipases, are located in the brush border membrane to break down complex peptides, carbohydrates, and fats before they are transported into the epithelial cytoplasm as molecular monomers (single units). *See* DIGESTIVE SYSTEM.

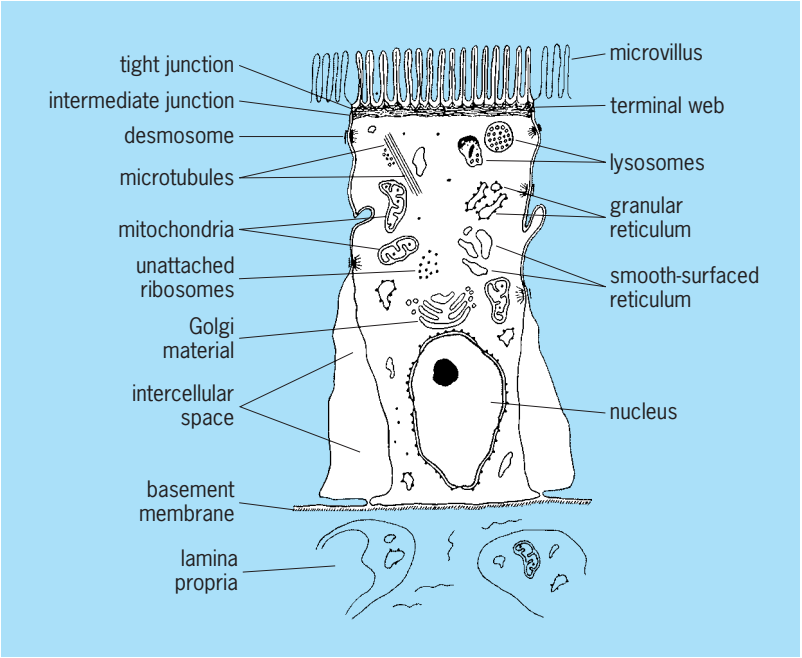


Fig. 1. Diagram of an absorptive epithelial cell from the mammalian small intestine. (*After L. R. Johnson, ed., Physiology of the Gastrointestinal Tract, Raven Press, 1981*)

The sides and bottoms of the cells with respect to the lumen are called the lateral and basal surfaces, or sometimes collectively the basolateral surfaces, and these may vary considerably among different epithelia and among species. While some epithelial cells have long, straight basolateral surfaces, others may be highly interdigitated with neighboring cells to increase their total surface area. In some epithelia these basolateral interdigitations contain abundant mitochondria, which provide energy for transport of materials across membranes. *See* MITOCHONDRIA.

During epithelial absorption, molecules must first pass across the apical membrane (from the mucosal side) into the cytoplasm. This entry process is then followed by ejection of the same molecule across the basolateral border to the blood side (called the serosal side) for eventual distribution to the body via the circulatory system. *See* CELL MEMBRANES.

Kinds of transport. Absorption across epithelia may occur by several different passive and active processes. Simple diffusion is the net movement of molecules from the apical to basolateral surfaces of an epithelium down chemical and electrical gradients without the requirement of cellular energy sources. Facilitated diffusion across the epithelium is similar to simple diffusion in that energy is not required, but in this process, molecular interaction with protein binding sites (carriers) in one or both membranes must occur to facilitate the transfer. Active molecular transport involves the use of membrane protein carriers as well as cellular energy supplies to move a transported molecule up an electrochemical gradient across the epithelium. Endocytosis and phagocytosis are also examples of active

transport because metabolic energy is required, but in these processes whole regions of the cell membrane are used to engulf fluid or particles, rather than to bring about molecular transfer using single-membrane proteins. *See* ENDOCYTOSIS; PHAGOCYTOSIS.

Mechanisms of ion absorption. Although a wide variety of ions are absorbed by different types of epithelial cells, the mechanisms of Na^+ and Cl^- transport in mammalian small intestine are perhaps best known in detail. Transepithelial transport of these two ions occurs in this tissue by three independent processes: active Na^+ absorption, not coupled directly to the flow of other solutes but accompanied indirectly by the diffusional absorption of Cl^-; coupled NaCl absorption; and cotransport of Na^+ with a wide variety of nutrient molecules.

During uncoupled active Na^+ transport, the cation crosses the apical membrane of the small-intestinal epithelial cell down an electrochemical gradient. Transfer across the brush border may be by simple or facilitated diffusion. Efflux of Na^+ up a concentration gradient out of the cell across the basolateral surface occurs by active transport, using a carrier protein called the Na^+/K^+–adenosine triphosphatase (Na^+ pump). Absorption of Na^+ by this two-step process results in the generation of a transepithelial electrical potential difference between the lumen and the blood of about 3–5 millivolts (blood electrically positive to apical surface). This electrical potential difference passively attracts Cl^- from the luminal side, most of this diffusion occurring paracellularly between the epithelial cells.

Two mechanisms have been proposed for coupled NaCl absorption in the mammalian small intestine. In the first model (**Fig. 2***a*), Na^+ and Cl^- bind to common apical membrane carriers and enter the epithelium together, with Na^+ diffusing down an electrochemical gradient and Cl^- moving up an electrochemical gradient. Efflux of Na^+ to the blood occurs by way of the energy-dependent Na^+ pump, while Cl^- moves passively across the basolateral border by facilitated diffusion. In the second model (Fig. 2*b*), Na^+ and Cl^- enter the cell by separate brush border carriers in exchange for intracellular H^+ and HCO_3^-, respectively. Their efflux to the blood occurs as described in Fig. 2*a*.

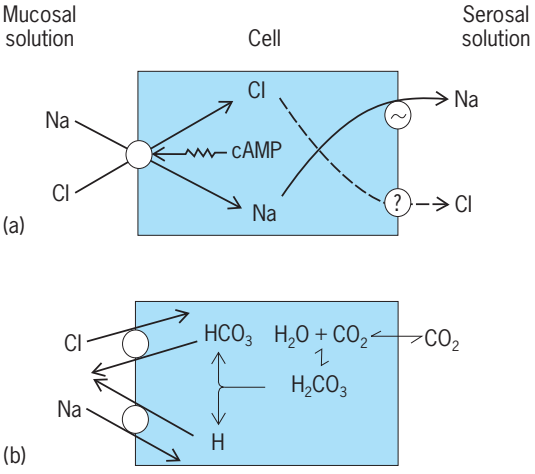


Fig. 2. Two models for electroneutral NaCl absorption by mammalian small-intestinal epithelial cells from the mucosal solution (food side) to the serosal solution (blood side). (*a*) Ions enter together by binding to common apical membrane carriers. (*b*) Ions enter by separate brush border carriers. (*After L. R. Johnson, ed., Physiology of the Gastrointestinal Tract, Raven Press, 1981*)

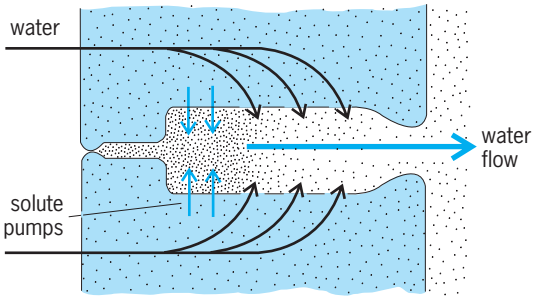


Fig. 3. Mechanism of solute-linked water absorption by epithelial cells. Density of dots refers to relative osmotic pressure in different compartments. (*After R. Eckert and D. Randall, Animal Physiology, Freeman, 1983*)

A final mechanism leading to the net flow of Na^+ across an epithelium is by coupling the entry of this ion into the cell with that of a wide variety of organic solutes. This pathway of Na^+ transport will be discussed below in conjunction with the mechanisms of nutrient absorption.

Mechanisms of water absorption. Net water transport across an epithelium is coupled to net ion transport in the same direction. As shown in **Fig. 3**, Na^+-pump sites are believed to be located along the lateral borders of epithelial cells. Energy-dependent Na^+ efflux from the cells to the intercellular spaces creates a local increase in osmotic pressure within these small compartments. An osmotic pressure gradient becomes established here, with greatest solute concentrations located nearest the tight junctions. Water flows into the cell across the brush border membrane and out the lateral membranes in response to the increased osmotic pressure in the paracellular spaces. Once water is in the intercellular compartment, a buildup of hydrostatic pressure forces the transported fluid to the capillary network. *See* OSMOREGULATORY MECHANISMS.

Mechanisms of nutrient absorption. It is believed that most nutrient molecules enter intestinal epithelial cells in conjunction with Na^+ on a common brush border membrane carrier protein (**Fig. 4**). Several types of proteins may exist on this membrane for each major category of organic solute (for example, a protein for sugars is functionally distinct from one for amino acids), but each carrier type requires Na^+ as a cotransported substrate. In all cases, Na^+ enters the cell down an electrochemical gradient, while the nutrient molecule is accumulated against a concentration gradient within the cytoplasm. The diffusional permeability of the brush border membrane toward organic solutes is very low, so that even though a considerable concentration gradient for the nutrient builds up across the apical cell pole, little passive loss

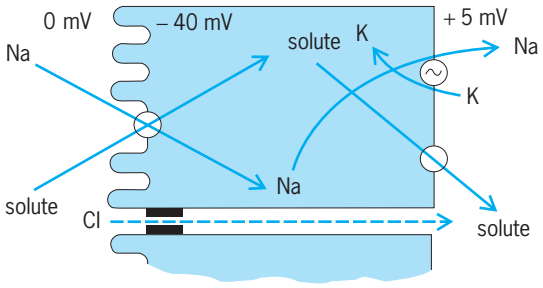


Fig. 4. Model for sodium-coupled absorption of organic solutes (D-hexoses, L-amino acids, dipeptides, tripeptides, water-soluble vitamins, bile salts) by mammalian small-intestinal epithelial cells. Cl^- ions diffuse between epithelial cells, down the electrochemical gradient. (*After L. R. Johnson, ed., Physiology of the Gastrointestinal Tract, Raven Press, 1981*)

to the lumen takes place. Intracellular Na^+ concentration is kept very low due to energy-dependent basolateral Na^+-pump activities transferring this cation up a concentration gradient out of the cell. Accumulated intracellular nutrients are transferred across the basolateral border by facilitated diffusion on a carrier protein which does not have a Na^+ requirement. No cellular metabolic energy is used directly for the absorption of nutrients. Instead, energy resources are used to maintain a steep Na^+ gradient across the brush border membrane to facilitate the accumulation of nutrient across this cell pole as a result of its cotransport with the cation. *See* ION TRANSPORT.

Gregory A. Ahearn

Bibliography. G. Giebisch, D. C. Tosteson, and H. H. Ussing (eds.), *Membrane Transport in Biology*, vols. 3 and 4, 1978; L. R. Johnson (ed.), *Physiology of the Gastrointestinal Tract*, vols. 1 and 2, 1981; H. Murer, U. Hopfer, and R. Kinne, Sodium-proton antiport in brush border membrane vesicles isolated from rat small intestine, *Biochem. J.*, 154:597–604, 1976; H. N. Nellans, R. A. Frizzell, and S. G. Schultz, Coupled sodium-chloride influx across the brush border of rabbit ileum, *Amer. J. Physiol.*, 225:467–475, 1973; S. G. Schultz and P. F. Curran, Coupled transport of sodium and organic solutes, *Physiol. Rev.*, 50:637–718, 1970.

Absorption of electromagnetic radiation

The process whereby the intensity of a beam of electromagnetic radiation is attenuated in passing through a material medium by conversion of the energy of the radiation to an equivalent amount of energy which appears within the medium; the radiant energy is converted into heat or some other form of molecular energy. A perfectly transparent medium permits the passage of a beam of radiation without any change in intensity other than that caused by the spread or convergence of the beam, and the total radiant energy emergent from such a medium equals that which entered it, whereas the emergent energy from an absorbing medium is less than that which enters, and, in the case of highly opaque media, is reduced practically to zero.

No known medium is opaque to all wavelengths of the electromagnetic spectrum, which extends from radio waves, whose wavelengths are measured in kilometers, through the infrared, visible, and ultraviolet spectral regions, to x-rays and gamma rays, of wavelengths down to 10^{-13} m. Similarly, no material medium is transparent to the whole electromagnetic spectrum. A medium which absorbs a relatively wide range of wavelengths is said to exhibit general absorption, while a medium which absorbs only restricted wavelength regions of no great range exhibits selective absorption for those particular spectral regions. For example, the substance pitch shows general absorption for the visible region of the spectrum, but is relatively transparent to infrared radiation of long wavelength. Ordinary window glass is transparent to visible light, but shows general absorption for ultraviolet radiation of wavelengths below about 310 nanometers, while colored glasses show selective absorption for specific regions of the visible spectrum. The color of objects which are not self-luminous and which are seen by light reflected or transmitted by the object is usually the result of selective absorption of portions of the visible spectrum. Many colorless substances, such as benzene and similar hydrocarbons, selectively absorb within the ultraviolet region of the spectrum, as well as in the infrared. *See* COLOR; ELECTROMAGNETIC RADIATION.

Laws of absorption. The capacity of a medium to absorb radiation depends on a number of factors, mainly the electronic and nuclear constitution of the atoms and molecules of the medium, the wavelength of the radiation, the thickness of the absorbing layer, and the variables which determine the state of the medium, of which the most important are the temperature and the concentration of the absorbing agent. In special cases, absorption may be influenced by electric or magnetic fields. The state of polarization of the radiation influences the absorption of media containing certain oriented structures, such as crystals of other than cubic symmetry. *See* STARK EFFECT; ZEEMAN EFFECT.

Lambert's law. Lambert's law, also called Bouguer's law or the Lambert-Bouguer law, expresses the effect of the thickness of the absorbing medium on the absorption. If a homogeneous medium is thought of as being constituted of layers of uniform thickness set normally to the beam, each layer absorbs the same fraction of radiation incident on it. If I is the intensity to which a monochromatic parallel beam is attenuated after traversing a thickness d of the medium, and I_0 is the intensity of the beam at the surface of incidence (corrected for loss by reflection from this surface), the variation of intensity throughout the medium is expressed by Eq. (1), in which α is a

$$I = I_0 e^{-\alpha d} \tag{1}$$

constant for the medium called the absorption coefficient. This exponential relation can be expressed in

an equivalent logarithmic form as in Eq. (2), where

$$\log_{10}(I_0/I) = (\alpha/2.303)d = kd \qquad (2)$$

$k = \alpha/2.303$ is called the extinction coefficient for radiation of the wavelength considered. The quantity $\log_{10}(I_0/I)$ is often called the optical density, or the absorbance of the medium.

Equation (2) shows that as monochromatic radiation penetrates the medium, the logarithm of the intensity decreases in direct proportion to the thickness of the layer traversed. If experimental values for the intensity of the light emerging from layers of the medium of different thicknesses are available (corrected for reflection losses at all reflecting surfaces), the value of the extinction coefficient can be readily computed from the slope of the straight line representing the logarithms of the emergent intensities as functions of the thickness of the layer.

Equations (1) and (2) show that the absorption and extinction coefficients have the dimensions of reciprocal length. The extinction coefficient is equal to the reciprocal of the thickness of the absorbing layer required to reduce the intensity to one-tenth of its incident value. Similarly, the absorption coefficient is the reciprocal of the thickness required to reduce the intensity to $1/e$ of the incident value, where e is the base of the natural logarithms, 2.718.

Beer's law. This law refers to the effect of the concentration of the absorbing medium, that is, the mass of absorbing material per unit of volume, on the absorption. This relation is of prime importance in describing the absorption of solutions of an absorbing solute, since the solute's concentration may be varied over wide limits, or the absorption of gases, the concentration of which depends on the pressure. According to Beer's law, each individual molecule of the absorbing material absorbs the same fraction of the radiation incident upon it, no matter whether the molecules are closely packed in a concentrated solution or highly dispersed in a dilute solution. The relation between the intensity of a parallel monochromatic beam which emerges from a plane parallel layer of absorbing solution of constant thickness and the concentration of the solution is an exponential one, of the same form as the relation between intensity and thickness expressed by Lambert's law. The effects of thickness d and concentration c on absorption of monochromatic radiation can therefore be combined in a single mathematical expression, given in Eq. (3), in which k' is a constant for a

$$I = I_0 e^{-k'cd} \qquad (3)$$

given absorbing substance (at constant wavelength and temperature), independent of the actual concentration of solute in the solution. In logarithms, the relation becomes Eq. (4). The values of the constants k'

$$\log_{10}(I_0/I) = (k'/2.303)cd = \epsilon cd \qquad (4)$$

and ϵ in Eqs. (3) and (4) depend on the units of concentration. If the concentration of the solute is expressed in moles per liter, the constant ϵ is called the molar extinction coefficient. Some authors employ the symbol a_M, which is called the molar absorbance index, instead of ϵ.

If Beer's law is adhered to, the molar extinction coefficient does not depend on the concentration of the absorbing solute, but usually changes with the wavelength of the radiation, with the temperature of the solution, and with the solvent.

The dimensions of the molar extinction coefficient are reciprocal concentration multiplied by reciprocal length, the usual units being liters/(mole)(cm). If Beer's law is true for a particular solution, the plot of log (I_0/I) against the concentrations for solutions of different concentrations, measured in cells of constant thickness, will yield a straight line, the slope of which is equal to the molar extinction coefficient.

While no true exceptions to Lambert's law are known, exceptions to Beer's law are not uncommon. Such exceptions arise whenever the molecular state of the absorbing solute depends on the concentration. For example, in solutions of weak electrolytes, whose ions and undissociated molecules absorb radiation differently, the changing ratio between ions and undissociated molecules brought about by changes in the total concentration prevents solutions of the electrolyte from obeying Beer's law. Aqueous solutions of dyes frequently deviate from the law because of dimerization and more complicated aggregate formation as the concentration of dye is increased.

Absorption measurement. The measurement of the absorption of homogeneous media is usually accomplished by absolute or comparative measurements of the intensities of the incident and transmitted beams, with corrections for any loss of radiant energy caused by processes other than absorption. The most important of these losses is by reflection at the various surfaces of the absorbing layer and of vessels which may contain the medium, if the medium is liquid or gaseous. Such losses are usually automatically compensated for by the method of measurement employed. Losses by reflection not compensated for in this manner may be computed from Fresnel's laws of reflection. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

Scattering. Absorption of electromagnetic radiation should be distinguished from the phenomenon of scattering, which occurs during the passage of radiation through inhomogeneous media. Radiant energy which traverses media constituted of small regions of refractive index different from that of the rest of the medium is diverted laterally from the direction of the incident beam. The diverted radiation gives rise to the hazy or opalescent appearance characteristic of such media, exemplified by smoke, mist, and opal. If the centers of inhomogeneity are sufficiently dilute, the intensity of a parallel beam is diminished in its passage through the medium because of the sidewise scattering, according to a law of the same form as the Lambert-Bouguer law for absorption, given in Eq. (5), where I is the intensity of

$$I = I_0 e^{-\tau d} \qquad (5)$$

the primary beam of initial intensity I_0, after it has traversed a distance d through the scattering medium. The coefficient τ, called the turbidity of the medium, plays the same part in weakening the primary beam by scattering as does the absorption coefficient in true absorption. However, in true scattering, no loss of total radiant energy takes place, energy lost in the direction of the primary beam appearing in the radiation scattered in other directions. In some inhomogeneous media, both absorption and scattering occur together. *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

Physical nature. Absorption of radiation by matter always involves the loss of energy by the radiation and a corresponding gain in energy by the atoms or molecules of the medium.

The energy of an assembly of gaseous atoms consists partly of kinetic energy of the translational motion which determines the temperature of the gas (thermal energy), and partly of internal energy, associated with the binding of the extranuclear electrons to the nucleus, and with the binding of the particles within the nucleus itself. Molecules, composed of more than one atom, have, in addition, energy associated with periodic rotations of the molecule as a whole and with oscillations of the atoms within the molecule with respect to one another.

The energy absorbed from radiation appears as increased internal energy, or in increased vibrational and rotational energy of the atoms and molecules of the absorbing medium. As a general rule, translational energy is not directly increased by absorption of radiation, although it may be indirectly increased by degradation of electronic energy or by conversion of rotational or vibrational energy to that of translation by intermolecular collisions.

Quantum theory. In order to construct an adequate theoretical description of the energy relations between matter and radiation, it has been necessary to amplify the wave theory of radiation by the quantum theory, according to which the energy in radiation occurs in natural units called quanta. The value of the energy in these units, expressed in ergs or calories, for example, is the same for all radiation of the same wavelength, but differs for radiation of different wavelengths. The energy E in a quantum of radiation of frequency ν (where the frequency is equal to the velocity of the radiation in a given medium divided by its wavelength in the same medium) is directly proportional to the frequency, or inversely proportional to the wavelength, according to the relation given in Eq. (6), where h is a universal con-

$$E = h\nu \tag{6}$$

stant known as Planck's constant. The value of h is 6.63×10^{-34} joule-second, and if ν is expressed in s^{-1}, E is given in joules per quantum. *See* QUANTUM MECHANICS.

The most energetic type of change that can occur in an atom involves the nucleus, and increase of nuclear energy by absorption therefore requires quanta of very high energy, that is, of high frequency or low wavelength. Such rays are the γ-rays, whose wavelength varies downward from 0.01 nm. Next in energy are the electrons nearest to the nucleus and therefore the most tightly bound. These electrons can be excited to states of higher energy by absorption of x-rays, whose range in wavelength is from about 0.01 to 1 nm. Less energy is required to excite the more loosely bound valence electrons. Such excitation can be accomplished by the absorption of quanta of visible radiation (wavelength 700 nm for red light to 400 nm for blue) or of ultraviolet radiation, of wavelength down to about 100 nm. Absorption of ultraviolet radiation of shorter wavelengths, down to those on the border of the x-ray region, excites electrons bound to the nucleus with intermediate strength.

The absorption of relatively low-energy quanta of wavelength from about 1 to 10 micrometers suffices to excite vibrating atoms in molecules to higher vibrational states, while changes in rotational energy, which are of still smaller magnitude, may be excited by absorption of radiation of still longer wavelength, from the short-wavelength radio region of about 1 cm to long-wavelength infrared radiation, some hundredths of a centimeter in wavelength.

Gases. The absorption of gases composed of atoms is usually very selective. For example, monatomic sodium vapor absorbs very strongly over two narrow wavelength regions in the yellow part of the visible spectrum (the so-called D lines), and no further absorption by monatomic sodium vapor occurs until similar narrow lines appear in the near-ultraviolet. The valence electron of the sodium atom can exist only in one of a series of energy states separated by relatively large energy intervals between the permitted values, and the sharp-line absorption spectrum results from transitions of the valence electron from the lowest energy which it may possess in the atom to various excited levels. Line absorption spectra are characteristic of monatomic gases in general. *See* ATOMIC STRUCTURE AND SPECTRA.

The visible and ultraviolet absorption of vapors composed of diatomic or polyatomic molecules is much more complicated than that of atoms. As for atoms, the absorbed energy is utilized mainly in raising one of the more loosely bound electrons to a state of higher energy, but the electronic excitation of a molecule is almost always accompanied by simultaneous excitation of many modes of vibration of the atoms within the molecule and of rotation of the molecule as a whole. As a result, the absorption, which for an atom is concentrated in a very sharp absorption line, becomes spread over a considerable spectral region, often in the form of bands. Each band corresponds to excitation of a specific mode of vibration accompanying the electronic change, and each band may be composed of a number of very fine lines close together in wavelength, each of which corresponds to a specific rotational change of the molecule accompanying the electronic and vibrational changes. Band spectra are as characteristic of the absorption of molecules in the gaseous state, and frequently in the liquid state, as line spectra

are of gaseous atoms. *See* BAND SPECTRUM; LINE SPECTRUM; MOLECULAR STRUCTURE AND SPECTRA.

Liquids. Liquids usually absorb radiation in the same general spectral region as the corresponding vapors. For example, liquid water, like water vapor, absorbs infrared radiation strongly (vibrational transitions), is largely transparent to visible and near-ultraviolet radiation, and begins to absorb strongly in the far-ultraviolet. A universal difference between liquids and gases is the disturbance in the energy states of the molecules in a liquid caused by the great number of intermolecular collisions; this has the effect of broadening the very fine lines observed in the absorption spectra of vapors, so that sharp-line structure disappears in the absorption bands of liquids.

Solids. Substances which can exist in solid, liquid, and vapor states without undergoing a temperature rise to very high values usually absorb in the same general spectral regions for all three states of aggregation, with differences in detail because of the intermolecular forces present in the liquid and solid. Crystalline solids, such as rock salt or silver chloride, absorb infrared radiation of long wavelength, which excites vibrations of the electrically charged ions of which these salts are composed; such solids are transparent to infrared radiations of shorter wavelengths. In colorless solids, the valence electrons are too tightly bound to the nuclei to be excited by visible radiation, but all solids absorb in the near- or far-ultraviolet region. *See* CRYSTAL OPTICS; INTERMOLECULAR FORCES.

The use of solids as components of optical instruments is restricted by the spectral regions to which they are transparent. Crown glass, while showing excellent transparency for visible light and for ultraviolet radiation immediately adjoining the visible region, becomes opaque to radiation of wavelength about 300 nm and shorter, and is also opaque to infrared radiation longer than about 2000 nm in wavelength. Quartz is transparent down to wavelengths about 180 nm in the ultraviolet, and to about 4 μm in the infrared. The most generally useful material for prisms and windows for the near-infrared region is rock salt, which is highly transparent out to about 15 μm. For a detailed discussion of the properties of optical glass *see* OPTICAL MATERIALS.

Fluorescence. The energy acquired by matter by absorption of visible or ultraviolet radiation, although primarily used to excite electrons to higher energy states, usually ultimately appears as increased kinetic energy of the molecules, that is, as heat. It may, however, under special circumstances, be reemitted as electromagnetic radiation. Fluorescence is the reemission, as radiant energy, of absorbed radiant energy, normally at wavelengths the same as or longer than those absorbed. The reemission, as ordinarily observed, ceases immediately when the exciting radiation is shut off. Refined measurements show that the fluorescent reemission persists, in different cases, for periods of the order of 10^{-9} to 10^{-5} s. The simplest case of fluorescence is the resonance fluorescence of monatomic gases at low pressure, such as sodium or mercury vapors, in which the reemitted radiation is of the same wavelength as that absorbed. In this case, fluorescence is the converse of absorption: Absorption involves the excitation of an electron from its lowest energy state to a higher energy state by radiation, while fluorescence is produced by the return of the excited electron to the lower state, with the emission of the energy difference between the two states as radiation. The fluorescent radiation of molecular gases and of nearly all liquids, solids, and solutions contains a large component of wavelengths longer than those of the absorbed radiation, a relationship known as Stokes' law of fluorescence. In these cases, not all of the absorbed energy is reradiated, a portion remaining as heat in the absorbing material. The fluorescence of iodine vapor is easily seen on projecting an intense beam of visible light through an evacuated bulb containing a few crystals of iodine, but the most familiar examples are provided by certain organic compounds in solution—for instance, quinine sulfate, which absorbs ultraviolet radiation and reemits blue, or fluorescein, which absorbs blue-green light and fluoresces with an intense, bright-green color. *See* FLUORESCENCE.

Phosphorescence. The radiant reemission of absorbed radiant energy at wavelengths longer than those absorbed, for a readily observable interval after withdrawal of the exciting radiation, is called phosphorescence. The interval of persistence, determined by means of a phosphoroscope, usually varies from about 0.001 s to several seconds, but some phosphors may be induced to phosphorescence by heat days or months after the exciting absorption. An important and useful class of phosphors is the impurity phosphors, solids such as the sulfides of zinc or calcium which are activated to the phosphorescent state by incorporating minute amounts of foreign material (called activators), such as salts of manganese or silver. So-called fluorescent lamps contain a coating of impurity phosphor on their inner wall which, after absorbing ultraviolet radiation produced by passage of an electrical discharge through mercury vapor in the lamp, reemits visible light. The receiving screen of a television tube contains a similar coating, excited not by radiant energy but by the impact of a stream of electrons on the surface. *See* FLUORESCENT LAMP; PHOSPHORESCENCE.

Luminescence. Phosphorescence and fluorescence are special cases of luminescence, which is defined as light emission that cannot be attributed merely to the temperature of the emitting body. Luminescence may be excited by heat (thermoluminescence), by electricity (electroluminescence), by chemical reaction (chemiluminescence), or by friction (triboluminescence), as well as by radiation. *See* LUMINESCENCE.

Absorption and emission coefficients. The absorption and emission processes of atoms were examined from the quantum point of view by Albert Einstein in 1916, with some important results that have been realized practically in the invention of the maser and the laser. Consider an assembly of atoms undergoing absorption transitions of frequency ν s^{-1} from the

ground state 1 to an excited state 2 and emission transitions in the reverse direction, the atoms and radiation being at equilibrium at temperature T. The equilibrium between the excited and unexcited atoms is determined by the Boltzmann relation $N_2/N_1 = \exp(-h\nu/kT)$, where N_1 and N_2 are the equilibrium numbers of atoms in states 1 and 2, respectively, and the radiational equilibrium is determined by equality in the rate of absorption and emission of quanta. The number of quanta absorbed per second is $B_{12}N_1\rho(\nu)$, where $\rho(\nu)$ is the density of radiation of frequency ν (proportional to the intensity), and B_{12} is a proportionality constant called the Einstein coefficient for absorption. Atoms in state 2 will emit radiation spontaneously (fluorescence), after a certain mean life, at a rate of $A_{21}N_2$ per second, where A_{21} is the Einstein coefficient for spontaneous emission from state 2 to state 1. To achieve consistency between the density of radiation of frequency ν at equilibrium calculated from these considerations and the value calculated from Planck's radiation law, which is experimentally true, it is necessary to introduce, in addition to the spontaneous emission, an emission of intensity proportional to the radiation density of frequency ν in which the atoms are immersed. The radiational equilibrium is then determined by Eq. (7), where B_{21} is

$$B_{12}N_1\rho(\nu) = A_{21}N_2 + B_{21}N_2\rho(\nu) \tag{7}$$

the Einstein coefficient of stimulated emission. The Einstein radiation coefficients are found to be related by Eqs. (8*a*) and (8*b*).

$$B_{12} = B_{21} \tag{8a}$$

$$A_{21} = (8\pi h\nu^3/c^3) \cdot B_{21} \tag{8b}$$

In the past when one considered radiation intensities available from terrestrial sources, stimulated emission was very feeble compared with the spontaneous process. Stimulated emission is, however, the fundamental emission process in the laser, a device in which a high concentration of excited molecules is produced by intense illumination from a "pumping" source, in an optical system in which excitation and emission are augmented by back-and-forth reflection until stimulated emission swamps the spontaneous process. *See* LASER; OPTICAL PUMPING.

There are also important relations between the absorption characteristics of atoms and their mean lifetime τ in the excited state. Since A_{21} is the number of times per second that a given atom will emit a quantum spontaneously, the mean lifetime before emission in the excited state is $\tau = 1/A_{21}$. It can also be shown that A_{21} and τ are related, as shown in Eq. (9),

$$A_{21} = 1/\tau = \frac{(8\pi^2\nu^2e^2)}{mc^3} \cdot f = 7.42 \times 10^{-22} f\nu^2 \qquad (\nu \text{ in } s^{-1}) \tag{9}$$

to the f number or oscillator strength for the transition that occurs in the dispersion equations shown as Eqs. (13) to (17). The value of f can be calculated from the absorption integrated over the band according to Eq. (18).

Dispersion. A transparent material does not abstract energy from radiation which it transmits, but it always decreases the velocity of propagation of such radiation. In a vacuum, the velocity of radiation is the same for all wavelengths, but in a material medium, the velocity of propagation varies considerably with wavelength. The refractive index n of a medium is the ratio of the velocity of light in vacuum to that in the medium, and the effect of the medium on the velocity of radiation which it transmits is expressed by the variation of refractive index with the wavelength λ of the radiation, $dn/d\lambda$. This variation is called the dispersion of the medium. For radiation of wavelengths far removed from those of absorption bands of the medium, the refractive index increases regularly with decreasing wavelength or increasing frequency; the dispersion is then said to be normal.

In regions of normal dispersion, the variation of refractive index with wavelength can be expressed with considerable accuracy by Eq. (10), known as

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \tag{10}$$

Cauchy's equation, in which A, B, and C are constants with positive values. As an approximation, C may be neglected in comparison with A and B, and the dispersion, $dn/d\lambda$, is then given by Eq. (11). Thus, in

$$\frac{dn}{d\lambda} = \frac{-2B}{\lambda^3} \tag{11}$$

regions of normal dispersion, the dispersion is approximately inversely proportional to the cube of the wavelength.

Dispersion by a prism. The refraction, or bending, of a ray of light which enters a material medium obliquely from vacuum or air (the refractive index of which for visible light is nearly unity) is the result of the diminished rate of advance of the wavefronts in the medium. Since, if the dispersion is normal, the refractive index of the medium is greater for violet than for red light, the wavefront of the violet light is retarded more than that of the red light. Hence, white light entering obliquely into the medium is converted within the medium to a continuously colored band, of which the red is least deviated from the direction of the incident beam, the violet most, with orange, yellow, green, and blue occupying intermediate positions. On emergence of the beam into air again, the colors remain separated. The action of the prism in resolving white light into its constituent colors is called color dispersion. *See* OPTICAL PRISM; REFRACTION OF WAVES.

The angular dispersion of a prism is the ratio, $d\theta/d\lambda$, of the difference in angular deviation $d\theta$ of two rays of slightly different wavelength which pass through the prism to the difference in wavelength $d\lambda$ when the prism is set for minimum deviation.

The angular dispersion of the prism given in Eq. (12) is the product of two factors, the variation,

$$\frac{d\theta}{d\lambda} = \frac{d\theta}{dn} \cdot \frac{dn}{d\lambda} \tag{12}$$

$d\theta/dn$, the deviation θ with refractive index n, and

the variation of refractive index with wavelength, the dispersion of the material of which the prism is made. The latter depends solely on this material, while $d\theta/dn$ depends on the angle of incidence and the refracting angle of the prism. The greater the dispersion of the material of the prism, the greater is the angular separation between rays of two given wavelengths as they leave the prism. For example, the dispersion of quartz for visible light is lower than that of glass; hence the length of the spectrum from red to violet formed by a quartz prism is less than that formed by a glass prism of equal size and shape. Also, since the dispersion of colorless materials such as glass or quartz is greater for blue and violet light than for red, the red end of the spectrum formed by prisms is much more contracted than the blue.

The colors of the rainbow result from dispersion of sunlight which enters raindrops and is refracted and dispersed in passing through them to the rear surface, at which the dispersed rays are reflected and reenter the air on the side of the drop on which the light was incident. *See* METEOROLOGICAL OPTICS; RAINBOW.

Anomalous dispersion. The regular increase of refractive index with decreasing wavelength expressed by Cauchy's equation breaks down as the wavelengths approach those of strong absorption bands. As the absorption band is approached from the long-wavelength side, the refractive index becomes very large, then decreases within the band to assume abnormally small values on the short-wavelength side, values below those for radiation on the long-wavelength side. A hollow prism containing an alcoholic solution of the dye fuchsin, which absorbs green light strongly, forms a spectrum in which the violet rays are less deviated than the red, on account of the abnormally low refractive index of the medium for violet light. The dispersion of media for radiation of wavelengths near those of strong absorption bands is said to be anomalous, in the sense that the refractive index decreases with decreasing wavelength instead of showing the normal increase. The theory of dispersion shows, however, that both the normal and anomalous variation of refractive index with wavelength can be satisfactorily described as aspects of a unified phenomenon, so that there is nothing fundamentally anomalous about dispersion in the vicinity of an absorption band. *See* DISPERSION (RADIATION).

Normal and anomalous dispersion of quartz are illustrated in **Fig. 1**. Throughout the near-infrared, visible, and near-ultraviolet spectral regions (between P and R on the curve), the dispersion is normal and adheres closely to Cauchy's equation, but it becomes anomalous to the right of R. From S to T, Cauchy's equation is again valid.

Relation to absorption. Figure 1 shows there is an intimate connection between dispersion and absorption; the refractive index rises to high values as the absorption band is approached from the long-wavelength side and falls to low values on the short-wavelength side of the band. In fact, the theory of dispersion shows that the complete dispersion curve

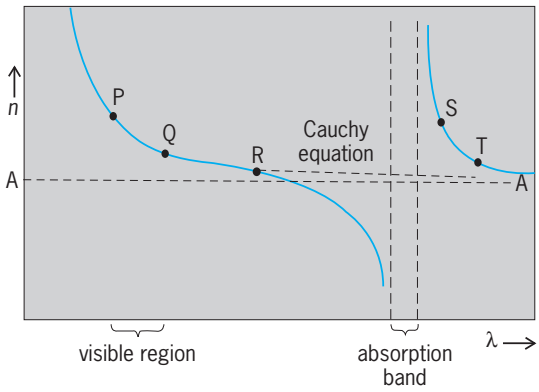


Fig. 1. Curve showing anomalous dispersion of quartz. A is limiting value of *n* as λ approaches infinity. (*After F. A. Jenkins and H. E. White, Fundamentals of Optics, 4th ed., McGraw-Hill, 1976*)

as a function of wavelength is governed by the absorption bands of the medium. In classical electromagnetic theory, electric charges are regarded as oscillating, each with its appropriate natural frequency ν_0, about positions of equilibrium within atoms or molecules. Placed in a radiation field of frequency ν per second, the oscillator in the atom is set into forced vibration, with the same frequency as that of the radiation. When ν is much lower or higher than ν_0, the amplitude of the forced vibration is small, but the amplitude becomes large when the frequency of the radiation equals the natural frequency of the oscillator. In much the same way, a tuning fork is set into vibration by sound waves corresponding to the same note emitted by another fork vibrating at the same frequency. To account for the absorption of energy by the medium from the radiation, it is necessary to postulate that in the motion of the atomic oscillator some frictional force, proportional to the velocity of the oscillator, must be overcome. For small amplitudes of forced oscillation, when the frequency of the radiation is very different from the natural period of the oscillator, the frictional force and the absorption of energy are negligible. Near resonance between the radiation and the oscillator, the amplitude becomes large, with a correspondingly large absorption of energy to overcome the frictional resistance. Radiation of frequencies near the natural frequency therefore corresponds to an absorption band. *See* SYMPATHETIC VIBRATION.

To show that the velocity of the radiation within the medium is changed, it is necessary to consider the phase of the forced vibration, which the theory shows to depend on the frequency of the radiation. The oscillator itself becomes a source of secondary radiation waves within the medium which combine to form sets of waves moving parallel to the original waves. Interference between the secondary and primary waves takes place, and because the phase of the secondary waves, which is the same as that of the atomic oscillators, is not the same as that of the primary waves, the wave motion resulting from the interference between the two sets of waves is different in phase from that of the primary waves incident

on the medium. But the velocity of propagation of the waves is the rate of advance of equal phase; hence the phase change effected by the medium, which is different for each frequency of radiation, is equivalent to a change in the velocity of the radiation within the medium. When the frequency of the radiation slightly exceeds the natural frequency of the oscillator, the radiation and the oscillator becomes 180° out of the phase, which corresponds to an increase in the velocity of the radiation and accounts for the fall in refractive index on the short-wavelength side of the absorption band.

The theory leads to Eqs. (13) through (17) for the refractive index of a material medium as a function of the frequency of the radiation. In the equations the frequency is expressed as angular frequency, $\omega = 2\pi\nu\ \mathrm{s}^{-1} = 2\pi c/\lambda$ where c is the velocity of light. When the angular frequency ω of the radiation is not very near the characteristic frequency of the electronic oscillator, the refractive index of the homogeneous medium containing N molecules per cubic centimeter is given by Eq. (13*a*), where e and m are

$$n^2 = 1 + \frac{4\pi Ne^2}{m} \cdot \frac{f}{\omega_0{}^2 - \omega^2} \qquad (13a)$$

$$n^2 = 1 + 4\pi Ne^2 \sum_i \frac{f_i/m_i}{\omega_i{}^2 - \omega^2} \qquad (13b)$$

the charge and mass of the electron, and f is the number of the oscillators per molecule of characteristics frequency ω_0. The f value is sometimes called the oscillator strength. If the molecule contains oscillators of different frequencies and mass (for example, electronic oscillators of frequency corresponding to ultraviolet radiation and ionic oscillators corresponding to infrared radiation), the frequency term becomes a summation, as in Eq. (13*b*), where ω_i is the characteristic frequency of the ith type of oscillator, and f_i and m_i are the corresponding f value and mass. In terms of wavelengths, this relation can be written as Eq. (14), where A_i is a constant for the medium,

$$n^2 = 1 + \sum_i \frac{A_i\lambda^2}{\lambda^2 - \lambda_i{}^2} \qquad (14)$$

λ is the wavelength of the radiation, and $\lambda_i = c/\nu_i$ is the wavelength corresponding to the characteristic frequency ν_i per second (Sellmeier's equation).

If the medium is a gas, for which the refractive index is only slightly greater than unity, the dispersion formula can be written as Eq. (15).

$$n = 1 + 2\pi Ne^2 \sum_i \frac{f_i/m_i}{\omega_i{}^2 - \omega^2} \qquad (15)$$

So long as the absorption remains negligible, these equations correctly describe the increase in refractive index as the frequency of the radiation begins to approach the absorption band determined by ω_i or λ_i. They fail when absorption becomes appreciable, since they predict infinitely large values of the refractive index when ω equals ω_i, whereas the refractive index remains finite throughout an absorption band.

The absorption of radiant energy of frequency very close to the characteristic frequency of the medium is formally regarded as the overcoming of a frictional force when the molecular oscillators are set into vibration, related by a proportionality constant g to the velocity of the oscillating particle; g is a damping coefficient for the oscillation. If the refractive index is determined by a single electronic oscillator, the dispersion equation for a gas at radiational frequencies within the absorption band becomes Eq. (16).

$$n = 1 + \frac{2\pi Ne^2}{m} \frac{f(\omega_0{}^2 - \omega^2)}{(\omega_0{}^2 - \omega^2)^2 + \omega^2 g^2} \qquad (16)$$

At the same time an absorption constant κ enters the equations, related to the absorption coefficient α of Eq. (1) by the expression $\kappa = \alpha c/2\omega\mu$. Equation (17) shows the relationship. For a monatomic

$$\kappa = \frac{2\pi Ne^2}{m} \frac{f\omega g}{(\omega_0{}^2 - \omega^2)^2 + \omega^2 g^2} \qquad (17)$$

vapor at low pressure, Nf is about 10^{17} per cubic centimeter, ω_0 is about 3×10^{15} per second, and g is about 10^{11} per second. These data show that, when the frequency of the radiation is not very near ω_0, ωg is very small in comparison with the denominator and the absorption is practically zero. As ω approaches ω_0, κ increases rapidly to a maximum at a radiational frequency very near ω_0 and then falls at frequencies greater than ω_0. When the absorption is relatively weak, the absorption maximum is directly proportional to the oscillator strength f. In terms of the molar extinction coefficient ϵ of Eq. (4), it can be shown that this direct relation holds, as seen in Eq. (18). The integration in Eq. (18) is carried out

$$f = 4.319 \times 10^{-9} \int \epsilon d\,\overline{\nu} \qquad (18)$$

over the whole absorption spectrum. The integral can be evaluated from the area under the curve of ϵ plotted as a function of wave number $\overline{\nu}\ \mathrm{cm}^{-1} = \nu\ (\mathrm{s}^{-1})/c = 1/\lambda$.

The width of the absorption band for an atom is determined by the value of the damping coefficient g; the greater the damping, the greater is the spectral region over which absorption extends.

The general behavior of the refractive index through the absorption band is illustrated by the dotted portions of **Fig. 2**. The presence of the damping term $\omega^2 g^2$ in the denominator of Eq. (17) prevents

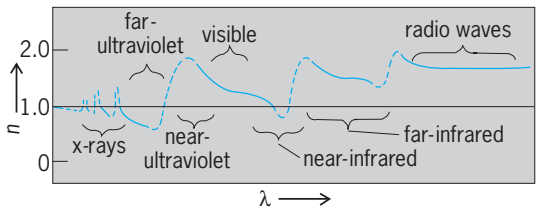


Fig. 2. Complete dispersion curve through the electromagnetic spectrum for a substance. (*After F. A. Jenkins and H. E. White, Fundamentals of Optics, 4th ed., McGraw-Hill, 1976*)

the refractive index from becoming infinite when $\omega = \omega_0$. Its value increases to a maximum for a radiation frequency less than ω_0, then falls with increasing frequency in the center of the band (anomalous dispersion) and increases from a relatively low value on the high-frequency side of the band.

Figure 2 shows schematically how the dispersion curve is determined by the absorption bands throughout the whole electromagnetic spectrum. The dotted portions of the curve correspond to absorption bands, each associated with a distinct type of electrical oscillator. The oscillators excited by x-rays are tightly bound inner elections; those excited by ultraviolet radiation are more loosely bound outer electrons which control the dispersion in the near-ultraviolet and visible regions, whereas those excited by the longer wavelengths are atoms or groups of atoms.

It will be observed in Fig. 2 that in regions of anomalous dispersion the refractive index of a substance may assume a value less than unity; the velocity of light in the medium is then greater than in vacuum. The velocity involved here is that with which the phase of the electromagnetic wave of a single frequency ω advances, for example, the velocity with which the crest of the wave advances through the medium. The theory of wave motion, however, shows that a signal propagated by electromagnetic radiation is carried by a group of waves of slightly different frequency, moving with a group velocity which, in a material medium, is always less than the velocity of light in vacuum. The existence of a refractive index less than unity in a material medium is therefore not in contradiction with the theory of relativity.

In quantum theory, absorption is associated not with the steady oscillation of a charge in an orbit but with transitions from one quantized state to another. The treatment of dispersion according to quantum theory is essentially similar to that outlined, with the difference that the natural frequencies ν_0 are now identified with the frequencies of radiation which the atom can absorb in undergoing quantum transitions. These transition frequencies are regarded as virtual classical oscillators, which react to radiation precisely as do the oscillators of classical electromagnetic theory.

Selective reflection. Nonmetallic substances which show very strong selective absorption also strongly reflect radiation of wavelengths near the absorption bands, although the maximum of reflection is not, in general, at the same wavelength as the maximum absorption. The infrared rays selectively reflected by ionic crystals are frequently referred to as reststrahlen, or residual rays. For additional information on selective reflection. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

William West

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; R. W. Ditchburn, *Light*, 3d ed., 1977, reprint 1991; I. S. Grant and W. R. Phillips, *Electromagnetism*, 2d ed., 1991; B. D. Guenther, *Modern Optics*, 1990; E. Hecht, *Optics*, 4th ed., 2001; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; E. E. Kriezis, D. P. Chrissoulidis, and A. C. Papagiannakis (eds.), *Electromagnetics and Optics*, 1992; Optical Society of America, *Handbook of Optics*, 2d ed., 4 vols., 1995, 2001; A. Sommerfeld, *Lectures of Theoretical Physics*, vol. 4, 1954.

Abstract algebra

The study of systems consisting of arbitrary sets of elements of unspecified type, together with certain operations satisfying prescribed lists of axioms. Abstract algebra has been developed since the mid-1920s and has become a basic idiom of contemporary mathematics. In contrast to the earlier algebra, which was highly computational and was confined to the study of specific systems generally based on real and complex numbers, abstract algebra is conceptual and axiomatic. A combination of the theories of abstract algebra with the computational speed and complexity of computers has led to significant applications in such areas as information theory and algebraic geometry.

Insight into the difference between the older and the abstract approaches can be obtained by comparing the older matrix theory with the more abstract linear algebra. Both deal with roughly the same area of mathematics, the former from a direct approach which stresses calculations with matrices, and the latter from an axiomatic and geometric viewpoint which treats vector spaces and linear transformations as the basic notions, and matrices as secondary to these. *See* LINEAR ALGEBRA.

Algebraic structures. Abstract algebra deals with a number of important algebraic structures, such as groups, rings, and fields.

An algebraic structure consists of a set S of elements of unspecified nature endowed with a number of finitary operations on S. If r is a positive integer, an r-ary operation associates with every r-tuple $(a_1, a_2, \ldots, a_r)$ of elements a_i in S a unique element $\omega(a_1, a_2, \ldots, a_r)$ in S. It is convenient to consider also nullary operations, that is, the selection of constants from S.

Binary operations. Many of the operations considered in abstract algebra are binary (2-ary) operations, in which case the symbol for the operation is usually written between the elements, or juxtaposition is used to indicate the operation, so that $\omega(a, b)$ is typically represented by $a * b$, $a + b$, or ab, depending on the context. Algebraic structures are classified by properties that the operations of the structure may or may not satisfy. Among the most common properties are:

1. Associativity: A binary operation $*$ of an algebraic structure S is associative if $a * (b * c) = (a * b) * c$ for all elements a, b, c of S.

2. Commutativity: A binary operation $*$ of an algebraic structure S is commutative if $a * b = b * a$ for all elements a, b of S.

3. Existence of an identity: An element e of an algebraic structure S is an identity for the binary

operation $*$ if $a * e = e * a = a$ for all elements a of S.

4. Existence of inverses: Assuming there is an element e of an algebraic structure S that is an identity for the binary operation $*$, an element a of S has an inverse if there is an element a^{-1} of S such that $a * a^{-1} = a^{-1} * a = e$.

5. Distributivity: A binary operation $*$ distributes over a binary operation $+$ of an algebraic structure S if $a * (b + c) = (a * b) + (a * c)$ for all elements a, b, c of S. The property of distributivity thus provides a link between two operations.

Groups. If the structure S is a group, a single binary operation is assumed to satisfy several simple conditions called the group axioms: the associativity of the operation, the existence of an identity element for the operation, and the existence of inverses for the operation. In this case, ab is usually written for $\omega(a,b)$, and the operation is called multiplication; or $a + b$ is written, and the operation is called addition, if the operation ω is commutative. Alternatively, a group may be described as a set with three operations rather than only one. In addition to the associative binary operation mentioned above, it is possible to consider a unary operation providing the inverse of an element, and a nullary operation providing the identity element of the group. Examples of groups include the rigid motions (symmetries) of a square with operation given by function composition, the integers under addition, and the nonzero real numbers under multiplication.

Rings and fields. In the case of a ring R, there are two binary operations, ab and $a + b$, which are assumed to satisfy a set of conditions known as the ring axioms, ensuring the associativity of multiplication and addition, the commutativity of addition, the distributivity of multiplication over addition, and the existence of an additive identity and additive inverses. Examples of rings include the integers, the real numbers, and the continuous real-valued functions of a real variable. *See* RING THEORY.

If axioms asserting the commutativity of multiplication and the existence of a multiplicative identity and multiplicative inverses are adjoined to the ring axioms, the ring is also a field. Examples of fields include the real numbers, the complex numbers, the finite set of integers modulo a prime number, and the rational polynomials with complex coefficients. *See* FIELD THEORY (MATHEMATICS).

Modules. In addition to the intrinsic study of algebraic structures, algebraists are interested in the study of the action of one algebraic structure on another. An important instance is the theory of modules and its special case of vector spaces. A left module over a ring R is defined to be a commutative group M on which R acts on the left in the sense that, given a pair of elements (a,x) where a is in R and x is in M, then the action determines a unique element ax in M. Moreover, the module product ax is assumed to satisfy the module axioms $a(x + y) = ax + ay$, $(a + b)x = ax + bx$, and $(ab)x = a(bx)$, where a and b are arbitrary elements of R, and x and y are arbitrary elements of M. Consideration of the modules admitted by a ring can provide significant information about the structure of the ring. Specializing from an arbitrary ring to a field yields the special type of module called a vector space.

Comparison of similar structures. Along with the study of particular algebraic structures, algebraists are interested in comparing structures satisfying similar axioms or sharing certain properties. Those structures most intimately related to a particular structure are its substructures and quotient structures.

Substructures. A substructure of an algebraic structure S is a subset T that is closed with respect to the operations of S; that is, if ω is any r-ary operation of S and $a_1, a_2, \ldots, a_r$ are elements of the subset T of S, then $\omega(a_1, a_2, \ldots, a_r)$ is again an element of T. In this case, T with operations defined by the operations of S restricted to T becomes an algebraic structure itself. Equivalently S may be referred to as an extension of T. As an example, if S is taken as the commutative group of integers under addition, the subset T of multiples of any particular integer forms a subgroup of S.

Quotient structures. A quotient structure is defined on an algebraic structure S by partitioning the elements of S into a family $\mathbf{Q}$ of (possibly infinitely many) disjoint congruence classes Q_α. As a result, (1) each element of S is a member of one and only one of the subsets Q_α of S, and (2) for any r-ary operation ω of S, no matter what elements $a_1, a_2, \ldots, a_r$ of $Q_{\alpha 1}$, $Q_{\alpha 2}, \ldots, Q_{\alpha r}$ respectively are chosen, the subset Q_β of S containing $\omega(a_1, a_2, \ldots, a_r)$ is the same. Then the collection of congruence classes $S/\mathbf{Q} = \{Q_\alpha\}$ can be made into an algebraic structure with operations determined by the action of the operations of S on the elements of the congruence classes, called a quotient structure of S. For example, an r-ary operation ω' would act on an r-tuple $(Q_{\alpha 1}, Q_{\alpha 2}, \ldots, Q_{\alpha r})$ by assigning to it the congruence class Q_β if it contains the element $\omega(a_1, a_2, \ldots, a_r)$ for elements $a_1, a_2, \ldots, a_r$ of $Q_{\alpha 1}, Q_{\alpha 2}, \ldots, Q_{\alpha r}$, respectively. An example of a quotient structure is one in which S is the ring of integers under addition and multiplication, and the family of congruence classes $\mathbf{Q}$ consists of the subset E of even integers and the subset O of odd integers. Then the quotient structure $S/\mathbf{Q}$ contains two elements E and O; the operations are given by $E + E = E = O + O$, $E + O = O = O + E$, $E * E = E = E * O = O * E$, and $O * O = O$. These operations symbolically represent the familiar facts that the sum of two even or two odd integers is even, the sum of an even integer and an odd integer is odd, the product of an even integer with either an even or an odd integer is even, and the product of two odd integers is odd.

Comparisons by means of functions. Other comparisons between similar algebraic structures are made by means of functions between the structures that preserve the action of the corresponding operations. Two principal ways of organizing this study are universal algebra and category theory.

1. *Universal algebra.* A considerable part of the study of algebraic structures can be developed in a unified way that does not specify the particular

structure, although deeper results do require specialization to the particular systems whose richness is to a large extent accounted for by their applicability to other areas of mathematics and to other disciplines. The general study of algebraic structures is called universal algebra. A basic notion here is that of a homomorphism of one algebraic structure S to a second one S'. Here there is a one-to-one correspondence $\omega \leftrightarrow \omega'$ between the sets of operations of S and of S' such that ω and ω' are both r-ary for the same $r = 0, 1, 2, 3, \ldots$. Then a homomorphism f from S to S' is a function mapping S to S' such that $\omega(a_1, \ldots, a_r) = \omega'(\mathrm{f}(a_1), \ldots, f(a_r))$ for all a_i in S and all the corresponding operations ω and ω'. A substantial part of the theory of homomorphisms of groups and rings is valid in this general setting. Another basic notion of universal algebra is that of a variety of algebraic structures, which is defined to be the class of structures satisfying a given set of identities defined by means of the given compositions (that is, the associative or the commutative laws). Examples are the variety of groups and the variety of rings.

2. *Category theory*. A second approach to the comparison of related structures is that of category theory. The notion of a category is applicable throughout mathematics, for example, in set theory and topology. The notion is formulated in order to place on an equal footing a class of mathematical objects and the basic mappings between these; in contrast is the more elemental focus of universal algebra. A category is defined to be a class of objects (generally mathematical structures) together with a class of morphisms (generally mappings of the structures) subject to certain axioms. An example of a category is the class of all groups as objects and the class of all homomorphisms of pairs of groups as the class of morphisms. Another example is the class of all sets as objects and of mappings of one set into another as the class of morphisms.

As an example of the ever-increasing levels of abstraction available and useful in algebra, once categories have been defined, they can themselves be compared by means of functors, which act as homomorphisms of categories. If C and C' are categories, a functor from C to C' is defined as a mapping F of the objects of C into the objects of C', together with a mapping ϕ of the morphisms of C into the morphisms of C', such that ϕ preserves the compositions and identity morphisms. At this point, functors can be compared via natural transformations, which thus act as homomorphisms of functors.

The concepts of category, functor, and natural transformation appear in many areas of mathematics. Many of the homomorphisms that arise in the study of algebraic structures prove to be instances of natural transformations. The frequency of the occurrences of the naturalness of the homomorphisms has implications for any adequate philosophy of mathematics. Like set theory, category theory can provide a foundation for much of mathematics. Joel K. Haack

Bibliography. M. Artin, *Algebra*, 1995; J. R. Durbin, *Modern Algebra: An Introduction*, 4th ed., 1999; J. Gallian, *Contemporary Abstract Algebra*, 2d ed., 1990; T. W. Hungerford, *Algebra*, rev. ed., 1980, reprint 1993; S. MacLane and G. Birkhoff, *Algebra*, 3d ed., 1987; C. C. Pinter, *A Book of Abstract Algebra*, 2d ed., 1990.

Abstract data type

A mathematical entity consisting of a set of values (the carrier set) and a collection of operations that manipulate them. For example, the Integer abstract data type consists of a carrier set containing the positive and negative whole numbers and 0, and a collection of operations manipulating these values, such as addition, subtraction, multiplication, equality comparison, and order comparison. *See* ABSTRACT ALGEBRA.

Abstraction. To abstract is to ignore some details of a thing in favor of others. Abstraction is important in problem solving because it allows problem solvers to focus on essential details while ignoring the inessential, thus simplifying the problem and bringing to attention those aspects of the problem involved in its solution. Abstract data types are important in computer science because they provide a clear and precise way to specify what data a program must manipulate, and how the program must manipulate its data, without regard to details about how data are represented or how operations are implemented. Once an abstract data type is understood and documented, it serves as a specification that programmers can use to guide their choice of data representation and operation implementation, and as a standard for ensuring program correctness.

A realization of an abstract data type that provides representations of the values of its carrier set and algorithms for its operations is called a data type. Programming languages typically provide several built-in data types, and usually also facilities for programmers to create others. Most programming languages provide a data type realizing the Integer abstract data type, for example. The carrier set of the Integer abstract data type is a collection of whole numbers, so these numbers must be represented in some way. Programs typically use a string of bits of fixed size (often 32 bits) to represent Integer values in base two, with one bit used to represent the sign of the number. Algorithms that manipulate these strings of bits implement the operations of the abstract data type. *See* ALGORITHM; DATA STRUCTURE; NUMERICAL REPRESENTATION (COMPUTERS); PROGRAMMING LANGUAGES.

Realizations of abstract data types are rarely perfect. Representations are always finite, while carrier sets of abstract data types are often infinite. Many individual values of some carrier sets (such as real numbers) cannot be precisely represented on digital computers. Nevertheless, abstract data types provide the standard against which the data types realized in programs are judged.

Example. Data processing problems are rarely so simple that data abstraction is not helpful in their solution. Suppose a program must process data in its

order of arrival but that data may arrive before its predecessors have been completely processed. Data values must wait their turn in storage in a way that preserves their arrival order. Before considering the structure needed to store the data and the algorithms required for creating and maintaining the structure, it is useful to specify an abstract data type for this problem.

What is needed is a well-known abstract data type called a Queue. The carrier set of the Queue abstract data type is the set of all queues, that is, the set of all lists of data values whose elements enter at the rear and are removed from the front. Although many operations might be included in this abstract data type, the following four suffice:

1. The enter operation takes a queue and a data value as parameters and returns the input queue with the new data value added to its rear.
2. The remove operation takes a queue as its parameter and returns the input queue with the data value at its front removed.
3. The front operation takes a queue as its parameter and returns the data value at the front of the queue (the queue is unaltered).
4. The size operation takes a queue as its parameter and returns a whole number greater than or equal to 0 reporting the number of data values in the queue.

These informal operation descriptions should be augmented by specifications that precisely constrain them:

1. For any queue q, if size (q) = 0 then remove (q) is invalid. That the operation is invalid is reported as an error.
2. For any queue q, if size (q) > 0 then size (q) = size (remove (q)) + 1.
3. For any queue q and data value v, size (q) = size (enter (q,v)) − 1.
4. For any queue q and data value v, let $q' =$ enter (q,v). Then after an arbitrary sequence of enter and remove operations applied to q' containing exactly size (q) remove operations, and yielding q'', $v =$ front (q'').

The first of these specifications states that nothing can be removed from an empty queue. The second and third state how queues grow and shrink. The final specification captures the crucial property that all data values eventually leave a queue in the order that they enter it. Some reasoning will show that these four specifications completely determine the behavior of queues.

Usefulness. Such specifications of abstract data types provide the basis for their realization in programs. Programmers know which data values need to be represented, which operations need to be implemented, and which constraints must be satisfied. Careful study of program code and the appropriate selection of tests help to ensure that the programs are correct. Finally, specifications of abstract data types can be used to investigate and demonstrate the properties of abstract data types themselves, leading to better understanding of programs and ultimately higher-quality software. *See* COMPUTER PROGRAMMING; SOFTWARE ENGINEERING; SHRINK FIT.

Relation to object-oriented paradigm. A major trend in computer science is the object-oriented paradigm, an approach to program design and implementation using collections of interacting entities called objects. Objects incorporate both data and operations. In this way they mimic things in the real world, which have properties (data) and behaviors (operations). Objects that hold the same kind of data and perform the same operations form a class.

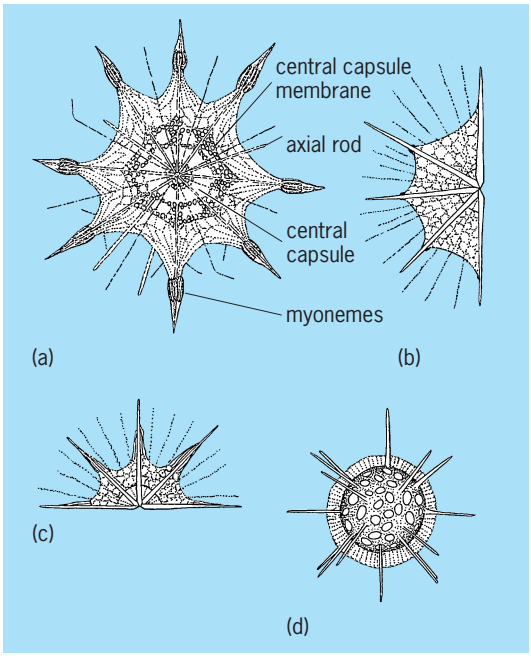
Abstract data values are separated from abstract data type operations. If the values in the carrier set of an abstract data type can be reconceptualized to include not only data values but also abstract data type operations, then the elements of the carrier set become entities that incorporate both data and operations, like objects, and the carrier set itself is very much like a class. The object-oriented paradigm can thus be seen as an outgrowth of the use of abstract data types. *See* OBJECT-ORIENTED PROGRAMMING.

Christopher Fox

Bibliography. J. Guttag, Abstract data types and the development of data structures, *Commun. ACM*, 20(6):396–404, June 1977; P. Thomas, H. Robinson, and J. Emms, *Abstract Data Types*, Oxford University Press, Oxford, 1988.

Acantharea

A class of the phylum Radiozoa in the Actinopoda. The kingdom Protozoa contains 18 phyla. One of the parvkingdoms (a hierarchical classification between kingdom and superphylum that is controversial and not officially recognized as such) is the Actinopoda (originally a class) containing two phyla: Heliozoa and Radiozoa. Within the Radiozoa is the class Acantharea. These marine protozoans possess a nonliving, organic capsular wall surrounding a central mass of cytoplasm. The intracapsular cytoplasm is connected to the extracellular cytoplasm by fine cytoplasmic strands passing through pores in the capsular wall. When viewed with the electron microscope, the capsular wall in some species appears to be made of many layers, each composed of a fine fibrillar network. The skeletons are made of celestite (strontium sulfate) instead of silica. The basic structural elements are 20 rods that pass through the capsule to the center in regular arrangements (polar and equatorial; see **illustration**). An equatorial rod forms a 90° angle with a polar rod, and other groups are arranged with similar exactness. This type of cytoskeleton may be modified by the addition of a latticework, composed of plates, each fused with a skeletal rod. Some genera show a double latticework, concentric with the central capsule. The skeletons do not contribute substantially to the paleontological record in marine sediments since celestite is dissolved by seawater.



Representatives of the Acantharea; *Acanthometra pellucida*, (*a*) showing axial rods extending through the central capsule, and myonemes (myophrisks) extending from the rods to the sheath; (*b*) contracted myophrisks with expanded sheath; and (*c*) relaxed myophrisks with contracted sheath. (*d*) *Dorotaspis heteropora*, with perforated shell and rods. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

Dissolution is pronounced below 617 ft (188 m) depth. While the protozoan is still alive, however, the cytoplasm appears to protect the skeleton from dissolution.

Pseudopodia are more or less permanent. A pseudopodium either is composed of an axial rod containing microtubules and a cytoplasmic envelope (forming an axopodium, or small foot; long slender pseudopodia are found in certain protozoa) or is fine-branched (forming a reticulopodium). Mitochondria are always present and contain flattened cristae. No cilia are present in the tropic phase. The central endoplasm contains numerous nuclei and other organelles.

When present, zooxanthellae (algae found within the cytoplasm) are of at least two kinds: Dinophyceae containing trichocysts in the cytoplasm; and a group of algae characterized by numerous discoid plastids, each of which contains an interlamellar pyrenoid. The systematics of the second group is still uncertain.

Myonemes (myophrisks) are significant components of the hydrostatic apparatus and regulate the buoyancy of the Acantharea by expanding or contracting portions of the extracapsular sheath. When observed with the electron microscope, they are composed of microfilaments arranged in a cone around the spine. Microfilaments contain actin, a contractile protein also found in metazoan muscle. Their contraction extends the sheath (illus. *b*); their relaxation flattens it against the rods (illus. *c*).

Although essentially pelagic, Acantharea may move vertically with the help of their hydrostatic apparatus. Symbiont-bearing species (species that contain zooxanthellae) have been reported from depths as great as 13,120 ft (4000 m). Little is known about the ecology or distribution of the Acantharea. The Gulf Stream is rich with these protozoans in the spring and summer but poor during other seasons. When compared to other plankton, Acantharea have a low average percentage of abundance even at peak periods (about 6–7%). Their approximate average abundance in the North Atlantic has been reported as 1–5%. *See* GULF STREAM.

Information about life cycles is very limited. For a few species, flagellate stages, syngamy, and zygotes have been reported. Certain stages of fission have been reported for some primitive types. The taxonomy of the Acantharea is still under review. *See* ACTINOPODEA; CELESTITE; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA; STRONTIUM.

O. Roger Anderson; John P. Harley

Bibliography. T. Cavalier-Smith, Kingdom Protozoa and its 18 phyla, *Microbiol. Rev.*, 57(4):953–994, 1993; M. Grell, in H.-E. Gruner (ed.), *Lehrbuch der Speziellen Zoologie*, 4th ed., Gustav Fischer, Stuttgart, 1980; K. Hausmann and N. Hulsmann, *Protozoology*, Georg Thieme Verlag, New York, 1996.

Acanthocephala

A distinct phylum of helminths, the adults of which are parasitic in the alimentary canal of vertebrates. They are commonly known as the spiny-headed worms. The phylum comprises the orders Archiacanthocephala, Palaeacanthocephala, and Eoacanthocephala. Over 500 species have been described from all classes of vertebrates, although more species occur in fish than in birds and mammals and only a relatively few species are found in amphibians and reptiles. The geographical distribution of acanthocephalans is worldwide, but genera and species do not have a uniform distribution because some species are confined to limited geographic areas. Host specificity is well established in some species, whereas others exhibit a wide range of host tolerance. The same species never occurs normally, as an adult, in cold-blooded and warm-blooded definitive hosts. More species occur in fish than any other vertebrate; however, Acanthocephala have not been reported from elasmobranch fish. The fact that larval development occurs in arthropods gives support to the postulation that the ancestors of Acanthocephala were parasites of primitive arthropods during or before the Cambrian Period and became parasites of vertebrates as this group arose and utilized arthropods for food. *See* ARCHIACANTHOCEPHALA; EOACANTHOCEPHALA; PALAEACANTHOCEPHALA.

Morphology

Adults of various species show great diversity in size, ranging in length from 0.04 in. (1 mm) in some species found in fish to over 16 in. (400 mm) in some mammalian species (**Fig. 1**). Most of the larger

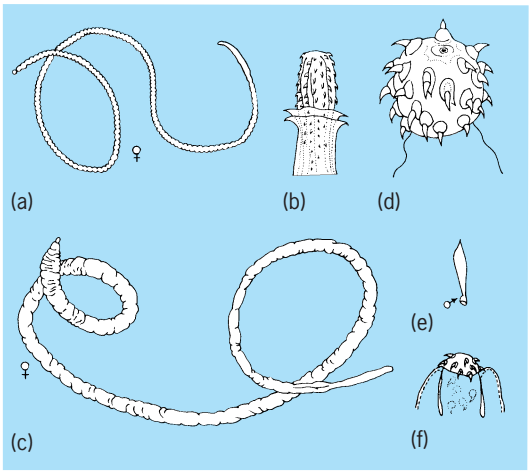


Fig. 1. Various acanthocephalan examples, adult forms drawn to same scale. (*a*) *Moniliformis dubius*, 4–12 in. (10–30 cm) long. (*b*) Proboscis of same. (*c*) *Macracanthorhynchus hirudinaceus*, 10–24 in. (25–60 cm) long. (*d*) Proboscis of same. (*e*) *Oncicola canis*, 0.5–0.8 in. (1–2 cm) long. (*f*) Proboscis of same. (*After A. C. Chandler and C. P. Read, Introduction to Parasitology, 10th ed., Wiley, 1961*)

species are from mammals; however, some mammalian forms are relatively small (*Corynosoma*) even though they parasitize seals and whales. When observed undisturbed in the intestine, the elongate body is distinctly flattened, but on removal to water or saline and after preservation the body becomes cylindrical. Living worms are translucent or milky white, although they may take on a distinctive color from the intestinal contents. The body of both sexes has three divisions: the proboscis armed with hooks, spines, or both; an unspined neck; and the posterior trunk.

Proboscis. The proboscis is the primary organ for attachment of Acanthocephala to the intestinal wall of the host. In some species the proboscis is parallel with the main axis of the body, but frequently it is inclined ventrally. The proboscis may be globular, or elongate and cylindrical in shape and is invariably armed with sharp pointed hooks, spines, or both (Fig. 1). These structures vary in shape and number and are usually arranged radially or in spiral rows. Electron microscopy studies have revealed that proboscis hooks consist of a central core of cytoplasm covered by hard nonliving material. In species with an elongate proboscis the hooks seem to be in longitudinal rows with a quincuncial arrangement.

In most species the proboscis is capable of introversion into a saclike structure, the proboscis receptacle. When in this position, the tips of the hooks are all directed anteriorly and are on the inside of the introverted proboscis. As the proboscis is everted and extended, these hooks engage the mucosa of the host's intestine, with the tips turning posteriorly as they reach their functional position on the exterior of the proboscis. The recurved points of the hooks anchor the worm firmly to the intestinal wall (**Fig. 2**). The proboscis receptacle and neck can be retracted into the body cavity but without inversion.

Within the anterior end of the trunk are found the proboscis receptacle; musculature for retracting the proboscis, receptacle, and neck; and the lemnisci (**Fig. 3**). The proboscis receptacle is attached to the inner surface of the proboscis. Muscles known as inverters of the proboscis are attached to the anterior tip of this structure and pass through the proboscis and receptacle, emerging from the posterior of the receptacle, and continue some distance posteriorly, where they attach to the trunk wall. Usually one inverter is attached dorsally and one ventrally. Contraction of the inverter fibers introverts the proboscis into the receptacle, while further contraction of these fibers pulls the receptacle deep into the body cavity.

Body wall. Electron microscopy studies show that the body wall is covered by a thin mucopolysaccharide known as the epicuticle secreted onto the surface of the worm. Beneath the epicuticle is the thin cuticle with an outer membrane and a homogeneous layer perforated by numerous pores of canals which pass through the striped layer beneath. Many of the canals are filled with an electron-dense material. Below the striped layer is the felt layer composed of many fibrous strands extending in various directions. Mitochondria and numerous vesicles are present in the felt layer. Some vesicles seem to be connected to the canals of the striped layer. The radial layer beneath the felt layer contains fewer fibrous strands, large thin-walled lacunar channels with mitochondria arranged around the channels. Lipids and glycogen are present in the radial layer. Circular and longitudinal muscle layers are found beneath the radial layer separated by plasma membranes. Some investigators consider all body-wall layers beneath the stripped layer as hypodermis.

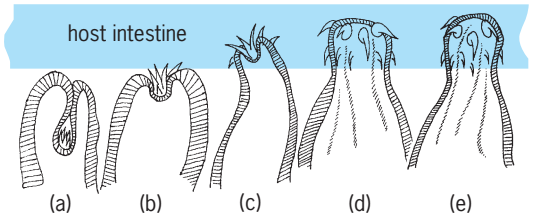


Fig. 2. Diagrams showing mechanics of proboscis attachment of *Neoechinorhynchus emydis* in intestinal wall of a turtle. (*a*) Proboscis fully introverted. (*b*) Most basal hooks in contact with host tissue. (*c–e*) Progressive stages in extroversion of the proboscis. (*After H. J. Van Cleave, Experimental Parasitology, vol. 1, 1952*)

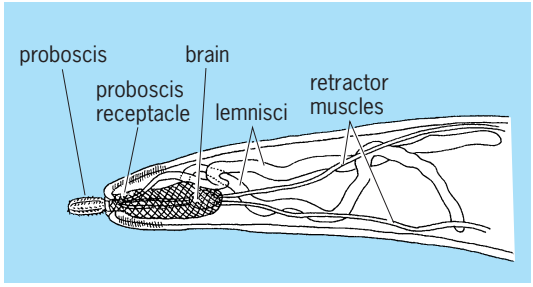


Fig. 3. Anterior end of *Moniliformis dubius*.

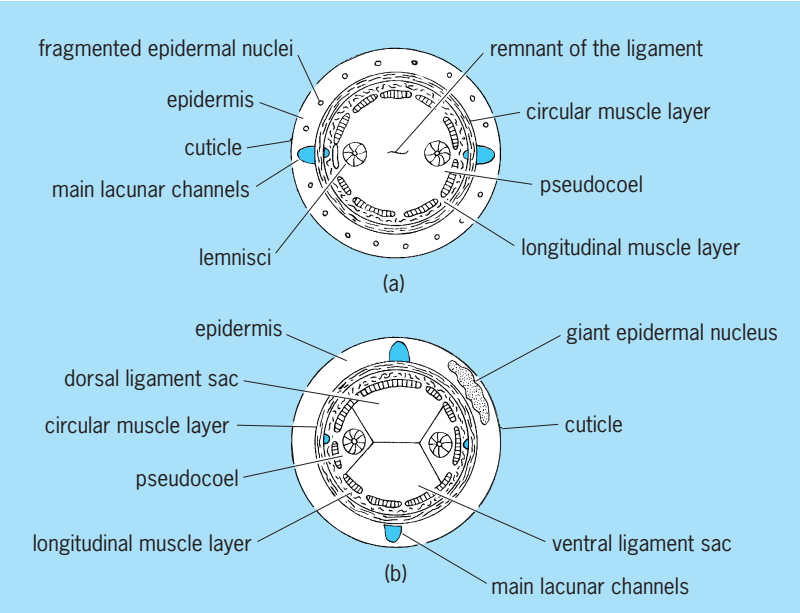


Fig. 4. Cross sections showing body plans of acanthocephalans. (*a*) Palaeacanthocephalan. (*b*) Archiacanthocephalan. (*After L. H. Hyman, The Invertebrates, vol. 3, McGraw-Hill, 1951*)

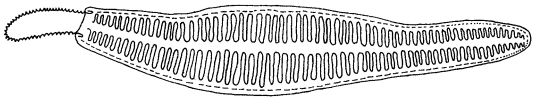


Fig. 5. Lacunar system of *Moniliformis*, dorsal view, showing regular circular branches. (*After L. H. Hyman, The Invertebrates, vol. 3, McGraw-Hill, 1951*)

Giant nuclei. The wall of the trunk has an external cuticula beneath which lies the thick syncytial hypodermis, or subcuticula. The hypodermis contains a relatively small number of giant nuclei which may be round, oval, ameboid, or elongate with a number of lateral branches (**Fig. 4**).

Lacunae. Within the hypodermis is a series of intercommunicating spaces, the lacunar system. Usually the lacunar system is composed of two longitudinal vessels either dorsal and ventral or ventral and lateral. In some species only the dorsal vessel is present. The longitudinal vessels are connected by a series of smaller vessels, the lacunae, which ramify throughout the body. In species exhibiting pseudosegmentation (*Moniliformis*), the regularly spaced lateral lacunae and the body musculature divide the wall into transverse folds (**Fig. 5**). These folds have no effect on the arrangement of internal organs.

Pseudocoele. The body cavity, or pseudocoele, contains all the internal organs, the most conspicuous of which are the reproductive organs enclosed in axial connective-tissue ligament sacs. These ligament sacs are hollow tubes which extend most of the length of the cavity of the trunk. They are single in both males and females of the Palaeacanthocephala, but are divided into dorsal and ventral sacs which communicate anteriorly in females of the other orders. There is no vestige of a digestive system in any stage of the life cycle.

Reproductive system. The reproductive organs of the male consist of a pair of testes and specialized cells, the cement glands. In most species there is a saclike structure behind the cement glands, Saefftigen's pouch, through which run the sperm ducts and the ducts from the cement glands (**Fig. 6***a*). The products of the testes and cement glands are discharged through a penis, which is surrounded by a posteriorly located bell-shaped copulatory bursa which is usually introverted into the posterior extremity of the trunk. At copulation the bursa is extended and applied to the posterior extremity of the female, where it is held firmly in place by the secretions of the cement glands. This material hardens, forming a copulatory cap on the female which is an internal cast of the bursa of the male, and remains attached to the posterior extremity of the female for some time following copulation.

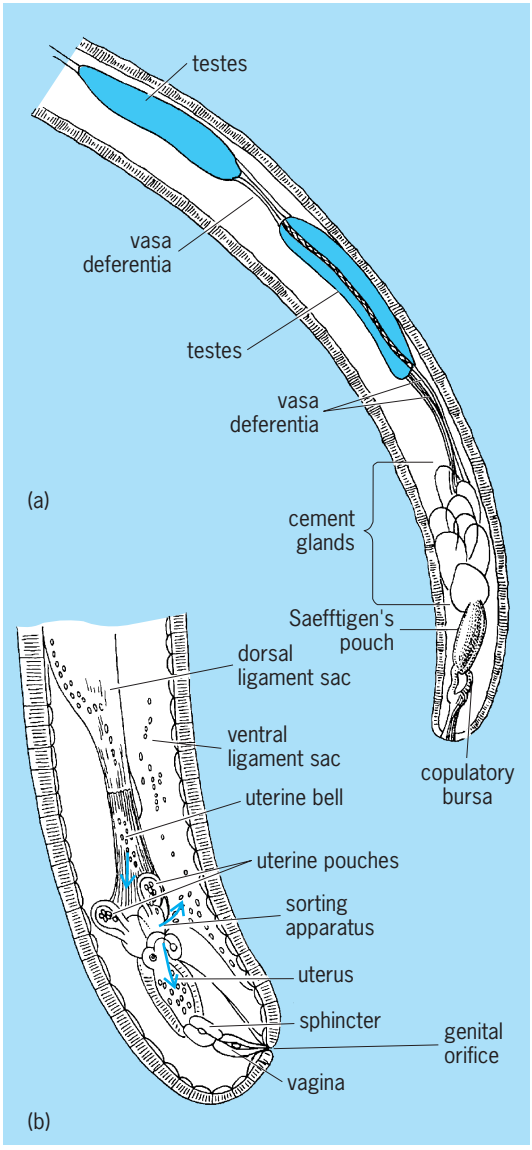


Fig. 6. Reproductive system of *Moniliformis dubius*. (*a*) Posterior end of male. (*b*) Posterior end of female. (*After A. C. Chandler and C. P. Read, Introduction to Parasitology, 10th ed., Wiley, 1961*)

Female Acanthocephala are unique in that the ovary exists as a distinct organ only in the very early stages of development and later breaks up to form free-floating egg balls. The eggs are fertilized as they are released from the egg balls and are retained within the ligament sacs until embryonation is complete. The genital orifice is posterior. A vagina provided with a strong sphincter extends anteriorly from the orifice and a saccular uterus is anterior to the vagina. The anterior end of the uterus is surrounded by a series of guard cells, the selective apparatus. From this extends a funnellike structure, the uterine bell, the broad anterior end of which opens into the body cavity or one of the ligament sacs holding the developing eggs (Fig. 6*b*). During embryonation in the body cavity the eggs acquire a series of membranes and a hard outer shell. Eggs which have not completed embryonation are passed back into the body cavity through special openings in the selective apparatus, whereas eggs which are fully mature are passed into the saccular uterus and eliminated through the vaginal sphincter and genital orifice into the intestinal contents of the host.

Lemnisci. The trunk and presoma are demarcated by an infolding of the cuticula and the presence of two elongate contractile structures, the lemnisci, which arise from the hypodermis and extend posteriorly into the body cavity. The lemnisci are provided with a definite number of nuclei which migrate from the hypodermis into the lemnisci as they are formed during larval development. The function of the lemnisci is unknown. One explanation offered for their function is that they act as reservoirs for the fluid of the lacunar system when the presoma and proboscis are invaginated.

Nervous system. The nervous system is composed of a chief ganglion or brain located within the proboscis receptacle. Most of the nerve trunks are inconspicuous but two of them, the retinacula, associated with muscle fibers, pass through the wall of the proboscis receptacle to innervate the trunk wall.

Excretory system. In members of the Archiacanthocephala modified protonephridial organs are found closely adherent to the reproductive system, but in most species specialized excretory organs are completely lacking. The protonephridial organs consist of a mass of flame bulbs attached to a common stem. The canals unite to form a single canal which joins the sperm duct in the male and the uterus in the female.

Embryology

The Acanthocephala are obligatory parasites throughout their entire life cycle; no known member of this phylum exists as a free-living organism.

Acanthor. The eggs passed in the feces of the host contain a mature embryo, the acanthor, which is surrounded by three membranes (**Fig. 7***a*). The life cycle always involves an intermediate host, which is usually an arthropod. Among these are small crustaceans for parasites of aquatic vertebrates, and grubs, roaches, and grasshoppers for those which parasitize terrestrial vertebrates. The eggs hatch after being ingested by the appropriate arthropod and release the spindle-shaped acanthor, which has a spiny body and is armed with retractile bladelike rostellar hooks used in the penetration of the intestinal wall of the intermediate host (Fig. 7*b*). The central portion of the body of the acanthor contains a column of dense undifferentiated nuclei, the entoblast, from which develop the structures of the ensuing stages. After penetration of the wall of the digestive tract, the acanthor comes to lie beneath the delimiting membrane of the gut wall, becomes quiescent, and begins to grow (Fig. 7*c*). It soon drops free into the host's hemocoele, where it undergoes a gradual transformation into the various larval stages (Fig. 7*d* and *e*).

Acanthella. The term acanthella is applied to the series of stages in which the rudiments of the reproductive organs, lemnisci, proboscis, and proboscis receptacle are formed (Fig. 7*f*). As development progresses, the acanthella becomes elongate, and is surrounded by a hyaline cyst produced by the larva.

Cystacanth. When the proboscis and the proboscis receptacle of the acanthella develop to the point of retractility, the larva becomes infective and is known as the cystacanth (Fig. 7*g* and *h*). The mature or infective cystacanth lies in the hemocoele of the intermediate host with the proboscis and receptacle retracted into the body cavity. The body cavity of the cystacanth contains all of the structures of the adult worm in an immature form.

In the least complicated life cycle the cystacanth remains in the hemocoele of the arthropod until it is ingested by a suitable vertebrate host, in which it excysts and attaches to the intestinal wall by the proboscis and grows to sexual maturity. In some forms the life cycle may be prolonged and complicated by the introduction of one or more transport hosts, usually a vertebrate, in which the cystacanth becomes established as a visceral cyst awaiting ingestion by a suitable vertebrate host in which sexual maturity can be attained.

Cell constancy. The somatic tissues of Acanthocephala are made up of a restricted number of nuclei which is more or less fixed for each tissue. This condition is termed cell constancy or more correctly eutely or nuclear constancy, because the majority of acanthocephalan tissues are syncytial in nature. The nuclei of the early embryonic cells which are destined to become syncytial in nature are large, regular, and spheroidal or elliptical in shape. During the transformation of the acanthella into the cystacanth the nuclei begin to assume the shape and form of the hypodermal nuclei seen in the adult by passing through an integrated series of ovoid, rosette, ameboid, elongate with lateral branches, and arboroid or dendritic forms. These diverse expressions of nuclear shape and modifications are instrumental in maintaining a favorable ratio between nuclear surface and surrounding cytoplasm. At the conclusion of the cleavage process the embryo, or acanthor, contains all of

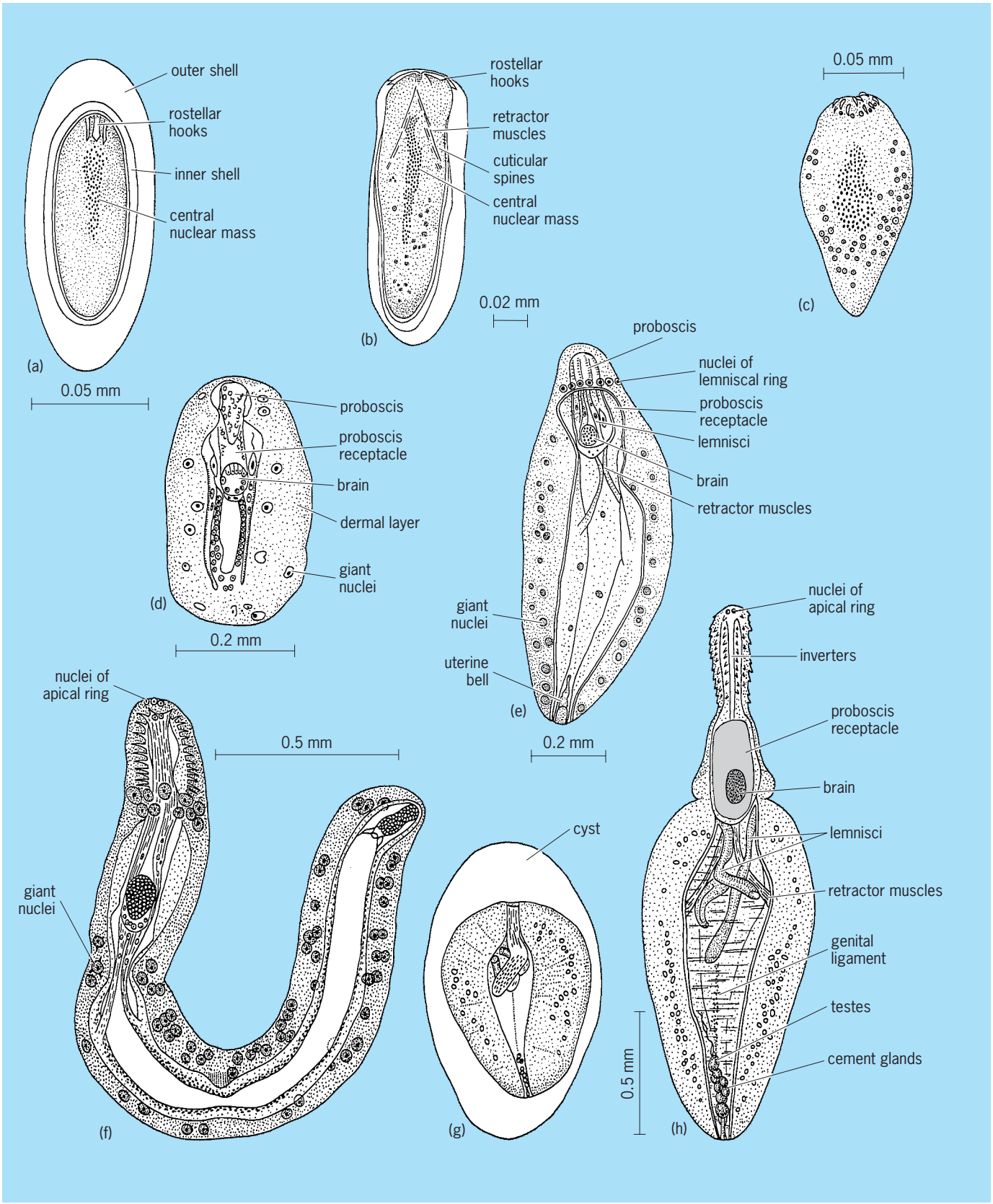


Fig. 7. Life history of *Moniliformis dubius.* (*a*) Mature egg containing acanthor. (*b*) Acanthor in process of escaping from egg shells and membranes. (*c*) Acanthor dissected from gut wall tissue of *Periplaneta americana* 5 days after infection. (*d*) Median sagittal section of larva removed from the body cavity of *P. americana* 29 days after infection. (*e*) Larva removed from body cavity of *P. americana* 42 days after infection. (*f*) Acanthella dissected from its enveloping sheath. (*g*) Encysted cystacanth from body cavity of *P. americana* with proboscis invaginated. (*h*) Cystacanth freed from its cyst and with proboscis evaginated.

the cellular elements of the adult. No mitotic divisions take place, although in the rearrangement of nuclei during larval development certain nuclei may undergo amitotic divisions. The extent of the metamorphoses of the embryonic cells and nuclei varies greatly in the different groups and genera. In the more highly specialized genera, the nuclei of the embryo lose all of their original appearance during larval development. In contrast, the nuclei of the adults of the less specialized genera retain the distinctive embryonic nuclei with little change. Varying changes, intermediate between these two extremes, are observed in other groups. *See* CELL CONSTANCY.

Physiology

Since acanthocephalans lack a gut, they undoubtedly obtain nutrients through the metasomal surface from material in the host's intestinal lumen. The role of the pores in this process is not really known, but it is likely that absorption of nutrients is facilitated by the increased surface area resulting from the plasma membrane lining the pores. It has been suggested that digestion may take place in the region of the presoma, which is in continuous contact with the host tissue. In this area, digestive enzymes are believed to be membrane-bound, allowing products of digestion to be absorbed at the host-parasite interface.

Acanthocephalans contain large quantities of glycogen and in general have a pronounced carbohydrate metabolism. Glycogen is localized in the proboscis, body wall, and muscles. They appear to utilize the Embden-Meyerhoff scheme of phosphorylating glycolysis. Chief waste products of carbohydrate metabolism in *Moniliformis dubius* are succinic, lactic, formic, and acetic acids, along with large quantities of ethanol. It has been suggested that the end products of carbohydrate metabolism excreted by acanthocephalans may be metabolized by the host or may be effective in preventing other helminths from establishing in that area of the host's intestine. *See* CARBOHYDRATE METABOLISM; GLYCOGEN.

A significant Krebs cycle has not been demonstrated in acanthocephalans, but evidence indicates that it may be partially operative in some species. Pyruvate or phosphenalpyruvate formed from glucose by glycolysis may be carboxylated to form malate, which is reduced via fumarate to form succinate. The reduction of fumarate could result in reoxidation of hydrogenated nicotinamide adenine dinucleotide (NADH) under anaerobic conditions, allowing glycolysis and synthesis of adenosine triphosphate (ATP) to be maintained. *See* CITRIC ACID CYCLE.

Donald V. Moore

Bibliography. H. G. Bronn (ed.), *Klassen und Ordnungen des Tierreichs,* vol. 4, 1932–1933; T. Dunagan and D. Miller, Acanthocephala, in *Microscopic Anatomy of Invertebrates*, vol. 4, pp. 299–332, Wiley, 1991; L. H. Hyman, *The Invertebrates*: *Acanthocephala, Aschelminthes, and Entoprocta,* vol. 3, 1951; H. J. Van Cleave, *Acanthocephala of North American Mammals,* 1953.

Acanthodii

A diverse group of early jawed fishes from the Paleozoic Era, usually classified as a distinct class of vertebrates. Most acanthodians have tiny body scales with a cross section like that of an onion, the layers reflecting incremental scale growth. Most acanthodians have strong spines at the leading edges of all fins except for the caudal (tail) fin. Acanthodians mostly were small fishes with large mouths and large eyes, suggesting an active, predatory lifestyle feeding on small prey, including other fishes. The oldest well-preserved examples are of Early Silurian age, and the last surviving examples are of Early Permian age. Primitive forms tended to be from marine environments and to have stouter bodies, larger stouter fin spines, and larger numbers of paired spines in front of the pelvic fins. Later forms are found also in freshwater and estuarine habitats and had more slender bodies with fewer and more slender spines.

Although experts do not currently consider acanthodians to be the most primitive of jawed vertebrates (the Placodermi have that distinction), acanthodians are the earliest jawed vertebrates to be represented in the fossil record by complete specimens. Superficially sharklike in appearance, they are instead currently grouped in the Teleostomi as the closest relatives of the bony fishes (Osteichthyes), with which they share characteristics such as small nasal capsules, large eyes, scales capable of continuous growth (in most species), and (in some specialized forms) three pairs of otoliths (ear stones). *See* OSTEICHTHYES; PLACODERMI; TELEOSTOMI.

Anatomy. Most acanthodians are small fishes, with size ranging from a few centimeters to half a meter. A few species reached lengths of more than 2 m (6.6 ft). The body usually was streamlined, with a blunt head, a terminal mouth, and a long, tapered, upturned tail. The large eyes were set close to the front of the head, behind a pair of small nasal capsules. The head usually was covered by large, flattened scales (tesserae), between which ran sensory canals similar to those in bony fishes. The braincase was ossified in advanced forms but not in primitive forms, where it remains cartilaginous. The ear region included three pairs of growing ear stones in some advanced acanthodians; primitive acanthodians had an open connection (endolymphatic duct) between the inner ear and the external environment, through which tiny sand grains entered and were used for the sense of balance, as in many sharks.

All acanthodians had jaws (sometimes calcified) consisting of a pair each of upper palatoquadrate cartilages and lower meckelian cartilages. In some acanthodians, jawbones, usually with teeth, were attached to these jaw cartilages (**Fig. 1**). In many species, teeth were absent; where present, they were of several types: multiple cusps in a whorl attached to a single base, usually near the front of the mouth; separate teeth in whorl-shaped arrangements along the margins of the mouth; or permanent conical teeth and denticles fixed to the jawbones. Gills were positioned behind the head in a compact gill chamber,

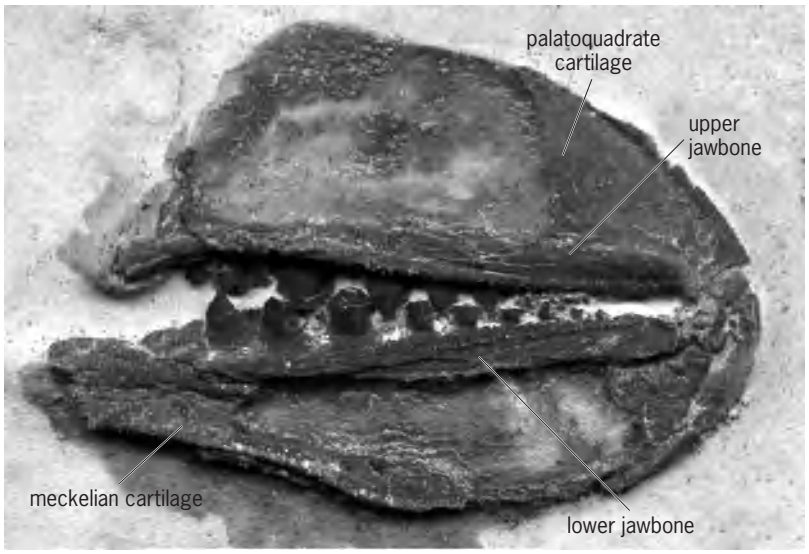


Fig. 1. Jaws of an Early Devonian ischnacanthid from northern Canada. The upper palatoquadrate and lower meckelian cartilages have tooth-bearing jawbones along their facing edges. (*From University of Alberta, Laboratory for Vertebrate Paleontology collection*)

as in bony fishes; a few acanthodians appear to have had multiple external gill slits on each side, but most had ornamented, platelike bony armor that enclosed the gills, leaving only a single pair of gill slits for water flow. *See* GILL.

Primitive acanthodians had two dorsal fins, an anal fin, and paired pectoral and pelvic fins, all with leading-edge spines (**Fig. 2***a–c*). Later forms retained a single dorsal fin (Fig. 2*d*). The tail usually was long and flexible. In primitive forms, a series of paired spines, the prepelvic series, was located along the belly in front of the pelvic fins (Fig. 2*a*). Paired prepectoral spines were present anterior and ventral to the pectoral fins, or were united by bony bases to form a ventral armor between the pectoral fins. These additional paired spines and associated bony plates were reduced or lost in later acanthodians.

Scales of acanthodians were small and continuously growing. Most had an onionskin-like structure internally as successive increments of growth were added to the outside of the scale. The crown of the scale was rhombic in outline, made of dentine, and either smooth or ornamented with ridges, while the bulbous base was composed of cellular or acellular bone. Scales of acanthodians often are abundant and useful for dating and correlating rocks.

Diversity. At one time there was controversy about whether the most primitive acanthodians were those with many fin spines or those with fewer, but more recently it has been recognized that the most primitive acanthodians were those with multiple pairs of stout, heavily ornamented spines, usually classified in the order Climatiiformes. More advanced acanthodians include those in the orders Ischnacanthiformes and Acanthodiformes. Many other acanthodian species are of uncertain relationships because they are known only from isolated scales or fin spines.

Climatiiformes. This order includes the most primitive acanthodians, typically with broad, heavily ridged spines, multiple pairs of prepelvic spines, and either prepectoral spines or well-developed pectoral armor. Teeth, where present, were numerous and arranged in whorl-like sets superficially similar to those of sharks. Examples are the well-known *Climatius, Ptomacanthus, Euthacanthus, Brochoadmones*, and *Lupopsyrus* (Fig. 2*a*). Diplacanthids, including *Diplacanthus*, *Gladiobranchus*, *Uraniacanthus*, and *Tetanopsyrus*, sometimes included as a suborder within Climatiiformes, had a pair of large dermal plates on the side of the head, and had their prepelvic spines reduced to one pair or none (Fig. 2*b*). Climatiiformes appeared in the Silurian; they flourished and then declined in the Devonian. The gyracanthid fishes, sometimes considered to be climatiiform acanthodians, have scales more like those of early chondrichthyans; gyracanthids survived into the Pennsylvanian.

Ischnacanthiformes. These acanthodians had tooth whorls at the front of their mouth as well as tooth-bearing jawbones borne on the upper and lower jaw cartilages (Fig. 1). They had slender fin spines and no prepelvic or prepectoral paired spines. Some ischnacanthiforms were rather large predators. Many species are known only from their toothed jawbones, which often are found separately. Examples include *Poracanthodes* and *Ischnacanthus* (Figs. 1, 2*c*). They are known from the Silurian to the end of the Devonian.

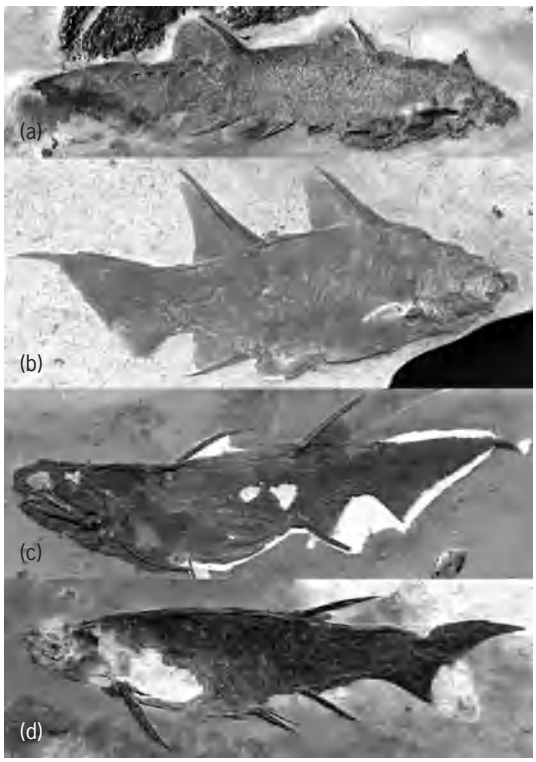


Fig. 2. Examples of acanthodians from the Early Devonian of northern Canada. (*a*) Primitive climatiiform *Lupopsyrus*. (*b*) Diplacanthid *Tetanopsyrus*. (c) Ischnacanthiform *Ischnacanthus*. (*d*) Unnamed mesacanthid acanthodiform. (*From the University of Alberta, Laboratory for Vertebrate Paleontology collection*)

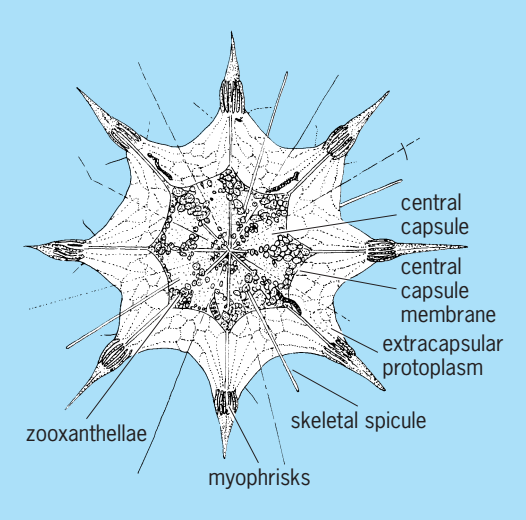
Acanthodiformes. These are advanced, streamlined, toothless forms with only one dorsal fin and at most one pair of prepelvic spines. One of the best-known acanthodians of any age is *Acanthodes*, thanks to its ossified internal cranial structures. Later acanthodiforms had long gill rakers and are thought to have been plankton feeders. The group appeared in the Early Devonian and survived until the Early Permian. Examples include *Cheiracanthus* in Cheiracanthidae, *Acanthodes* and *Homalacanthus* in Acanthodidae, and *Triazeugacanthus*, *Melanoacanthus*, and *Mesacanthus* in Mesacanthidae (Fig. 2*d*). Mark V. H. Wilson

Bibliography. R. H. Denison, *Acanthodii: Handbook of Paleoichthyology*, vol. 5, Gustav Fischer Verlag, Stuttgart, 1979; P. Janvier, *Early Vertebrates*, Clarendon Press, Oxford, 1996; J. A. Long, *The Rise of Fishes: 500 Million Years of Evolution*, Johns Hopkins University Press, Baltimore, 1995; J. A. Moy-Thomas and R. S. Miles, *Palaeozoic Fishes*, 2d ed., W. B. Saunders, Philadelphia, 1971.

Acanthometrida

A genus of marine protozoans in the class Acantharea. The kingdom Protozoa contains 18 phyla. One of the parvkingdoms (a hierarchical classification between kingdom and superphylum that is controversial and not officially recognized as such) is the Actinopoda (originally a class) containing two phyla, Heliozoa and Radiozoa. Within the Radiozoa is the class Acantharea, in which the genus *Acanthometrida* is placed. All members are unicellular planktonic marine protozoa with axopodia. Long slender pseudopodia are found in certain protozoans. Mitochondria are always present and contain flattened cristae. No cilia are present in the trophic phase. Members have a skeleton, made up of strontium sulfate (celestite) limited to 20 radially arranged rods that extend from the center, forming a characteristic pattern in which angles are quite exact, as in *Acanthometra* (see **illustration**). Since strontium sulfate is soluble in seawater, no acantharean fossils exist. The central endoplasm contains numerous nuclei and other organelles. The cytoplasm surrounding the spines contains a conical array of contractile microfilaments (myophrisks or myonemes that are Ca^{2+}-activated) containing the contractile protein actin. The microfilaments expand or contract the gelatinous sheath surrounding the cell, a phenomenon apparently responsible for changes in level of flotation. Electron microscopic research with related genera shows that the microfilaments are arranged in two sets of bands: one set runs parallel to the axis of the spine, and the other runs crosswise, thus producing a cross-fibrillar network. Cysts and flagellated swarmers represent two stages of the life cycle, which are not completely known at present. The endoplasm often contains zooxanthellae (small dinoflagellates living in the protozoan's endoplasm). *See* ACANTHAREA; ACTINOPODEA; CELESTITE; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA; STRONTIUM.

O. Roger Anderson; John P. Harley



***Acanthometra*. (*After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940*)**

Bibliography. T. Cavalier-Smith, Kingdom Protozoa and its 18 phyla, *Microbiol. Rev.*, 57(4):953–994, 1993; M. Grell, in H.-E. Gruner (ed.), *Lehrbuch der Speziellen Zoologie*, 4th ed., Gustav Fischer, Stuttgart, 1980; K. Hausmann and N. Hulsmann, *Protozoology*, Georg Thieme Verlag, New York, 1996.

Acanthophractida

An order of Acantharea. In this group of protozoa, skeletons typically include a latticework shell, although the characteristic skeletal rods are recognizable. The latticework may be spherical or ovoid, is fused with the skeletal rods, and is typically concentric with the central capsule. The body is usually covered with a single or double gelatinous sheath through which the skeletal rods emerge. Myonemes (usually a specific number) extend from the gelatinous sheath to each skeletal rod. These marine forms live mostly below depths of 150–200 ft (45–60 m). The order includes *Coleaspis, Diploconus, Dorotaspis*, and many other genera. *See* ACANTHAREA; ACTINOPODEA; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA. Richard P. Hall

Acari

A subclass of Arachnida, the mites and ticks. All are small (0.004–1.2 in. or 0.1–30 mm, most less than 0.08 in. or 2 mm in length), have lost most traces of external body segmentation, and have the mouthparts borne on a discrete body region, the gnathosoma (**Fig. 1**). They are apparently most closely related to Opiliones and Ricinulei.

General characteristics. The chelicerae may be chelate or needlelike. Pedipalps are generally smaller than the walking legs, and are simple or sometimes

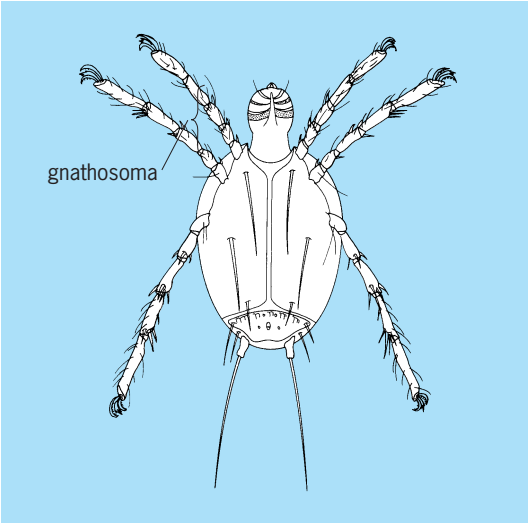


Fig. 1. Ventral view of *Laelaps echidnina*, the spiny rat mite (Parasitiformes), showing major body divisions.

weakly chelate. Legs typically move in the slow walking gait of arachnids; they may, however, be modified by projections, enlarged claws, heavy spines, or bladelike rows of long setae for crawling, clinging to hosts, transferring spermatophores, or swimming. Mites have a variety of integumentary mechanoreceptors and chemoreceptors (often modified setae), and one or two simple eyes (ocelli) frequently occur anterolaterally on the idiosoma (occasionally one anteromedially as well). The ganglia of the central nervous system are coalesced into a single "brain" lying around the esophagus. Excretion is by guanine crystals secreted into the hindgut and voided through a ventral anus. Respiration is by a tracheal network fed by one to four pairs of spiracles that vary in position. A simple dorsal heart is present in a few larger forms.

Life cycle. The typical acarine life cycle consists of six instars: egg; six-legged larva (**Fig. 2**); three nymphal instars (protonymph, deutonymph, and tritonymph); and adult, usually similar to active nymphs. Variations include suppression of the larva, one or more inactive nymphs for internal structural reorganization or phoresy, and change in number of nymphal stages. Male spermatophores may be deposited on the substrate, transferred to the female gonopore by modified legs or chelicerae, or inserted directly by a male copulatory organ. Females are normally oviparous, depositing 4–10,000 eggs singly or in clusters. The generation time varies from 4 days for some terrestrial forms to as much as 2 years in some fresh-water species.

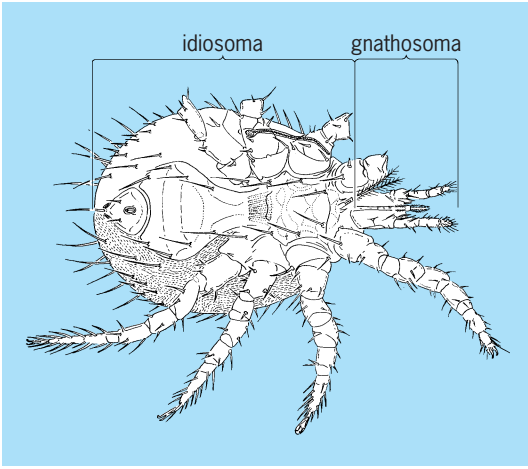


Fig. 2. Ventral view of a *Hygrobates* larva (Acariformes) modified for aquatic life.

Ecology and feeding. Acarine habitats and feeding habits are much more diverse than in other arachnids. Mites are ubiquitous, occurring from oceanic trenches below 13,000 ft (4000 m) to over 21,000 ft (6300 m) in the Himalayas and suspended above 3300 ft (1000 m) in the atmosphere; there are species in Antarctica. Soil mites are taxonomically varied, show the least specialized adaptations to habitat, and are frequently well-sclerotized predators. Inhabitants of stored grains, cheese, and house dust (many causing allergic reactions in humans) are also relatively unspecialized. The dominant forms inhabiting mosses are heavily sclerotized beetle mites (Oribatei), some of which can curl into a ball to reduce moisture loss from the ventral surface. Flowering plant associates include spider mites (Tetranychidae), which suck the contents of leaf epidermal cells, and gall mites (Eriophyidae), which cause a neoplastic growth of host tissue that provides both shelter and food. Fungivores are usually weakly sclerotized inhabitants of moist or semiaquatic habitats, while the characteristic fresh-water mites include sluggish crawlers, rapid swimmers, and planktonic drifters, all with reduced idiosomal setation. Mites associated with invertebrate animals include internal, external, and social parasites as well as inactive phoretic stages of Insecta, Crustacea, Myriapoda, Chelicerata, Mollusca, and Parazoa. A similar diversity of species are parasites or commensals of vertebrates; microhabitats include the nasal mucosa and lungs of reptiles and mammals, burrows in the skin of humans and most other vertebrates (scabies and mange mites), hair follicles (Demodicidae), feathers (both externally and within the feather shaft), fur (bat mites), and host nests and burrows. Some parasitic mites are disease vectors.

Origin and classification. Mites appear in the Devonian among the first true arachnids and thus apparently diverged early from the basic chelicerate stock. Differences in optical activity of the setae and degree of idiosomal segmentation suggest the possibility of a diphyletic origin. Over 30,000 species have been described, and it is estimated that as many as 500,000 may exist. *See* ARACHNIDA.

David Barr

Bibliography. D. R. Cook, *Water Mite Genera and Subgenera*, Mem. Amer. Entomol. Inst. 21, 1974; A. Kaestner, *Invertebrate Zoology*, vol. 2: *Arthropod Relatives, Chelicerata, Myriapoda*, 1968; G. W. Krantz, *A Manual of Acarology*, 2d ed., 1978; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; A. E. Treat, *Mites of Moths and Butterflies*, 1975.

Accelerating universe

Cosmic expansion is speeding up. In 1998, astronomers showed evidence from exploding stars that the expansion of the universe is not slowing down because of the gravity of matter in the universe. By observing the light from distant supernovae, astronomers can measure the distance to each explosion. By studying the redshift in the supernova's spectrum, they can measure the expansion of the universe since the light was emitted and, from a set of these measurements, can trace the history of cosmic expansion. Instead of seeing a slowing universe, as expected, observations indicate there is a component of the universe, called the dark energy, that makes cosmic expansion speed up. Dark energy comprises about two-thirds of the energy density of the universe. While it may be related to Albert Einstein's controversial cosmological constant, its true nature remains to be discovered by observation and thought. *See* COSMOLOGICAL CONSTANT; DARK ENERGY; SUPERNOVA.

Astronomers have known since the 1920s that galaxies are separated by millions of light-years and that the space between them is stretching, moving the galaxies apart: the nearest galaxies are moving away from us slowly and distant galaxies are receding more rapidly. This relation was discovered empirically by Edwin Hubble using the 100-in. (2.5-m) telescope at the Mount Wilson Observatory. If our patch of the universe is not unique, the simplest explanation for this observation is that the universe as a whole is expanding in all directions. What is new about the observations made in the 1990s is that the distances and times probed by the supernovae are large enough for the effects of cosmic acceleration to be measurable.

Since light travels at a finite speed, the history of cosmic expansion can be traced by looking at distant objects, whose light left their source hundreds of millions or even billions of years in the past. This procedure makes a telescope into a time machine capable of showing how the expansion of the universe has changed over time. Einstein's general theory of relativity, applied in the cosmic setting, shows that the presence of matter in the universe should lead to the gradual slowing of cosmic expansion. Since the 1950s, astronomers had sought to measure this deceleration of the universe, hoping to use it as a guide to the density of matter in the universe and as a predictor of the future behavior of cosmic expansion—whether the universe would coast outward indefinitely or would be drawn back to a dense state like the big bang from which it emerged 14×10^9 years ago. *See* BIG BANG THEORY; RELATIVITY.

In the 1990s, the perfection of accurate distance-measuring techniques based on supernova explosions and the development of powerful observing techniques made it possible to carry through the astronomers' quest to measure changes in the cosmic expansion rate. Instead of finding that the expansion has slowed over time because of the decelerating effects of gravity, as most expected, the new results from two teams showed the opposite: we live in an accelerating universe. Contrary to all expectation, the universe's expansion is speeding up. Subsequent work on supernovae, on the clustering of galaxies, and on the faint cosmic background light left over from the big bang itself converge on the same solution: the universe today is dominated by a strange dark energy whose role in fundamental physics is certainly important, but of which our present understanding is very limited. *See* COSMIC BACKGROUND RADIATION.

Expanding universe. Before 1920, astronomers thought that the Milky Way Galaxy, of which the Sun is an inconspicuous member, constituted the entire universe. When Einstein applied his recently developed theory of gravitation, the general theory of relativity, to the universe in 1916, he was guided by Willem DeSitter, a leading astronomer, to think of a static system. Einstein found that a static solution to his equations could be constructed if he added in a term that we now call the cosmological constant. This term had the effect of balancing out the attractive force of gravity with a kind of cosmic repulsion to produce a static universe.

Hubble, working at the Mount Wilson Observatory, showed that the distances to what he called the spiral nebulae were much larger than the extent of the Milky Way, and that it was more correct to think of these as independent systems, galaxies, each equivalent to the Milky Way. He did this by measuring the apparent brightness of stars whose intrinsic brightness he knew—Cepheid variables, whose intrinsic brightness can be found by measuring their periods of variation. The fainter a star appears, for a given intrinsic brightness, the more distant it must be. By observing this, a particular type of star, Hubble was able to show that the distances to even the nearest nebulae were a few million light-years, far larger than the size of the Milky Way, whose own size also had to be correctly measured. *See* CEPHEID; GALAXY, EXTERNAL; MILKY WAY GALAXY.

Hubble was also able to use measurements of the spectra of galaxies to show that their spectra have a redshift, the signature that they are receding from us. By combining the measurements of distance and of recession, Hubble discovered that the more distant galaxies are moving away from us more rapidly. The modern interpretation of these observations is cosmic expansion, with the universe stretching out in all directions at a rate of about 1 part in 14,000,000,000 of its size each year. *See* HUBBLE CONSTANT; REDSHIFT.

Naturally, Einstein had to adjust to this new picture of the cosmos. He had inserted his cosmological constant to make a static world, but later observations showed that the world was not static. Einstein was then inclined to banish the cosmological constant, which he knew was not mathematically justified and which he never did like much anyway. But DeSitter was not so sure this was a good idea. He thought it was possible that the universe had started out slowly but had accelerated as a result of the repulsive properties of the cosmological constant, denoted by Λ

Fig. 1. Supernova, 1994D. This type Ia supernova is in a galaxy at a distance of about 5×10^7 light-years in the Virgo cluster of galaxies. For a month, the light from a single exploding white dwarf is as bright as 4×10^9 stars like the Sun. (*P. Challis, Center for Astrophysics/STScI/NASA*)

(lambda). By 1932, Einstein and DeSitter were ready to stop discussing Λ, leaving that for future generations. In a joint paper they said of the cosmological constant, "An increase in the precision of data. . . will enable us in the future to fix its sign and determine its value."

Type Ia supernovae. That future arrived in 1998, 66 years later. The measurements on which the evidence for cosmic acceleration rests are similar to those carried out by Hubble: except that the stars being used are a million times brighter and the distances to the galaxies in which they are located are a thousand times larger. The claim that the universe is accelerating is so extraordinary that it is worth examining the evidence on which it is based: the brightness of exploding stars astronomers call type Ia supernovae (**Fig. 1**).

Type Ia supernovae are a particular type of exploding star that can be recognized by its spectrum. They are found in all types of galaxies and are thought to be the result of a sudden thermonuclear burning wave that propagates through a white dwarf star. This flame burns the carbon and oxygen of the star's interior into radioactive nickel. The subsequent decay into cobalt and iron releases energy gradually as the star is blown apart. For about a month, a type Ia supernova shines as brightly as 4×10^9 stars like the Sun shining simultaneously. This property makes them visible over very large distances, large enough to detect the change in cosmic expansion rate that has taken place while the light was on its way from a type Ia supernova halfway across the universe.

But being bright is only half the story. In order to measure cosmic distances from apparent brightness with good precision, the ideal tool would be a "standard candle" for which all the objects had the same intrinsic brightness. Then a measurement of the apparent brightness would give the distance. Type Ia supernovae are pretty good standard candles: the intrinsic light output at their peak luminosity varies by only a factor of 3 or so. This uniform behavior probably is connected to the theoretical prediction that there is a well-defined upper mass limit for white dwarfs of about 1.4 solar masses that is set by the quantum-mechanical properties of electrons. *See* WHITE DWARF STAR.

Even better precision comes from empirical approaches to type Ia supernovae. It turns out that the brightest supernovae have distinctly slower declines in brightness after they reach their peak, while the dimmer type Ia supernovae decline more rapidly. By measuring the rise and fall in the brightness of a supernova, which does not depend on its distance, one can tell whether an extrabright supernova or an extradim one is being observed, and take that information into account in estimating the distance. When the shape of the light curve is taken into account, the uncertainty in the distance to a single supernova shrinks to under 10%, making this type of supernova the very best distance-measuring tool for surveying the universe (**Fig. 2**). *See* LIGHT CURVES.

Technological advances. The disadvantage of type Ia supernovae is that they are rare. In a typical large galaxy, the rate of type Ia supernova explosions is about one in a century. In our own Milky Way Galaxy, the last event that was probably a type Ia supernova was observed by Tycho Brahe in 1572. So if you want to find and measure many supernovae to survey the history of cosmic expansion, you need either to be very patient or to be able to search many galaxies for the desired events.

This point is where technology has made great strides. The electronic light detectors used in modern telescopes are charge-coupled devices (CCDs), similar to those found in digital cameras. These are

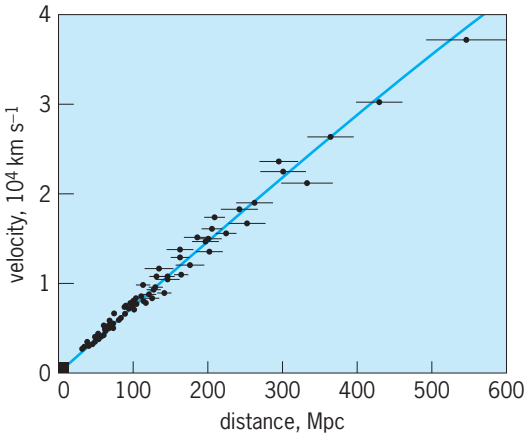


Fig. 2. Modern Hubble diagram for 95 type Ia supernovae. Recession velocity is plotted against distance in megaparsecs (1 Mpc = 3.26×10^6 light-years = 3.09×10^{19} km). The points represent supernovae up to 2×10^9 light-years away, with recession velocities up to one-tenth of the speed of light. The small scatter about the diagonal line shows that type Ia supernovae make excellent standard candles for measuring distances in the universe. This nearby patch does not show the effects of acceleration. (*Based on data from Saurabh Jha*)

more than 100 times more sensitive to light than the photographic plates used in Hubble's time. As the technology for fabricating computer chips has improved, the CCDs available to astronomers have grown larger as well as more sensitive, so that now arrays of 1000 megapixels are being applied to supernova studies and still larger cameras are in devlopment. These large and sensitive devices allow a modern telescope to image many thousands of galaxies in a single exposure. By coming back night after night to the same part of the sky, supernova searches in a single night gather the data needed to find many of these rare events (**Fig. 3**). *See* ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICE.

But gathering the light is only part of the story of finding supernovae to do cosmology. You also need to detect the new starlike events among the many fuzzy galaxies that do not change from night to night. This is the place where advances in computer hardware and software have turned the task of supernova searching from handicraft into something on an industrial scale. By using a fast computer to subtract digital images obtained last month (or last year) from the ones obtained tonight, it is relatively straightforward to sift out overnight a handful of new objects from the hundreds of thousands that remain unchanged.

Measuring the apparent brightness of these supernovae and the shape of their rise and fall over a month gives enough information to measure the distance to a supernova. Measuring the spectrum of the supernovae can show whether it is of type Ia and measure the redshift. By measuring the redshift at various distances, one can trace the history of cosmic expansion to look for the effects of slowing that results from gravity or acceleration caused by dark energy.

The first attempt at this type of searching was in 1988 by a Danish group. Because only small CCD chips were available, they published data on only one type Ia supernova. But the Supernova Cosmology Project based at Lawrence Berkeley Laboratory developed effective search techniques, and by 1997 they had a preliminary result. Their first indication, based on seven supernovae found in 1994 and 1995, was that the universe was slowing down because of the braking effects of dark matter. In 1998, the High-Z Supernova Search Team published a result indicating the opposite: that the universe was accelerating. Results from the Supernova Cosmology Project published in 1999 showed the same thing: we live in a universe where the rate of cosmic expansion is increasing as the universe grows older.

Subsequent observations have confirmed this latter result, and both teams agree. The samples have grown larger, the span of cosmic time probed by the supernova observations has increased, and methods for dealing with the pernicious effects of absorption by dust have been developed. One especially interesting result comes from a supernova search and follow-up carried out using the *Hubble Space Telescope* both to find and to get the light curves and spectra of very distant supernovae. This work showed that cosmic acceleration is a comparatively recent phenomenon, acting over only the last 5×10^9 years or so. Supernovae discovered at larger distances than 5×10^9 light-years show that the universe was probably slowing down when it was younger. So there is evidence that the universe was slowing down when it was young and dense, presumably resulting from the presence of dark matter, but then there came a shift, about 9×10^9 years after the big bang (and 5×10^9 years before the present), from deceleration to acceleration. In the lexicon of elementary physics, change in position is velocity, change in velocity is acceleration, and change in acceleration is called jerk. This program with the *Hubble Space Telescope* has produced evidence for the cosmic jerk. *See* DARK MATTER; HUBBLE SPACE TELESCOPE.

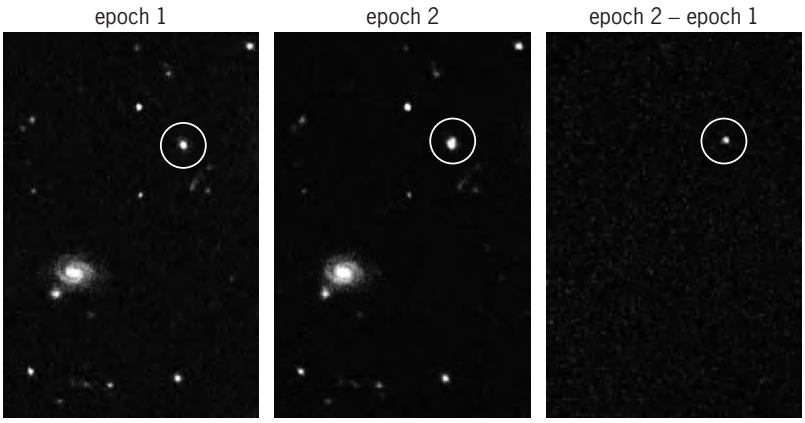


Fig. 3. Subtracting images to find supernovae. The image from a month ago (epoch 1) is subtracted from last night's image (epoch 2) to reveal a new supernova. The image area shown is about 1/1000 of the full area provided by the CCD camera. Rapid processing of dozens of image pairs demands nimble software and a dedicated team of searchers. (*Brian Schmidt, Australian National University*)

Dark energy? By combining measurements of supernova distances with results from galaxy clustering and from fluctuations in the cosmic microwave background, consensus has emerged that the universe at present has about 73% of its energy density in dark energy, about 23% in dark matter, and only 4% in the form of atoms made of neutrons, protons, and electrons. While it is satisfying that there is a single solution that matches a wide variety of current data, an honest appraisal is that we do not know what the dark energy is or what the dark matter is.

The dark energy could be a form of Einstein's cosmological constant. The cosmological constant has the right property: a tendency to make the universe expand faster over time. This tendency is why DeSitter thought the cosmological constant might be responsible for the cosmic expansion observed by Hubble. But the modern view of the cosmological constant is that it has its origin in the quantum properties of the vacuum. When the same principles that work so well for electromagnetism are applied to gravitation, the result is a quantitative disaster of unprecedented proportions. The observed dark energy is about 10^{120} times smaller than the predicted amount.

Another possibility is that the dark energy is not something that is constant in time, as the cosmological constant would be, but something that has been changing its value as the universe unfolds. This property could be detected by observations of distant supernovae.

The next round of observational programs for distant supernovae is aimed at observing the history of cosmic expansion with enough precision to tell the difference between a dark energy that does not change (like the cosmological constant) or something that varies with time. Large samples of supernovae at moderate distances and moderate samples of supernovae at large distances are being constructed. The goal is to pin down the properties of the dark energy well enough to tell if it is different in any way from a cosmological constant with a very small numerical value. So far, no significant deviation has been found, but the samples published to date are small. Dark energy is so new, so important, and so mysterious that it will be the focus of strenuous efforts in observational astronomy for decades to come. *See* COSMOLOGY; UNIVERSE.

Robert P. Kirshner

Bibliography. D. Goldsmith, *The Runaway Universe*, Perseus Books, Cambridge, MA, 2000; R. P. Kirshner, *The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos*, Princeton University Press, 2002; L. Krauss, *Quintessence*, Basic Books, New York, 2000.

Acceleration

The time rate of change of velocity. Since velocity is a directed or vector quantity involving both magnitude and direction, a velocity may change by a change of magnitude (speed) or by a change of direction or both. It follows that acceleration is also a directed, or vector, quantity. If the magnitude of the velocity of a body changes from v_1 ft/s to v_2 ft/s in t seconds, then the average acceleration a has a magnitude given by Eq. (1). To designate it fully the direction should be

$$a = \frac{\text{velocity change}}{\text{elapsed time}} = \frac{v_2 - v_1}{t_2 - t_1} = \frac{\Delta v}{\Delta t} \qquad (1)$$

given, as well as the magnitude. *See* VELOCITY.

Instantaneous acceleration is defined as the limit of the ratio of the velocity change to the elapsed time as the time interval approaches zero. When the acceleration is constant, the average acceleration and the instantaneous acceleration are equal.

If a body, moving along a straight line, is accelerated from a speed of 10 to 90 ft/s (3 to 27 m/s) in 4 s, then the average change in speed per second is $(90 - 10)/4 = 20$ ft/s or $(27 - 3)/4 = 6$ m/s in each second. This is written 20 ft per second per second or 20 ft/s^2, or 6 m per second per second or 6 m/s^2. Accelerations are commonly expressed in feet per second per second, meters per second per second, or in any similar units.

Whenever a body is acted upon by an unbalanced force, it will undergo acceleration. If it is moving in a constant direction, the acting force will produce a continuous change in speed. If it is moving with a constant speed, the acting force will produce an acceleration consisting of a continuous change of direction. In the general case, the acting force may produce both a change of speed and a change of direction.

R. D. Rusk

Angular acceleration. This is a vector quantity representing the rate of change of angular velocity of a body experiencing rotational motion. If, for example, at an instant t_1, a rigid body is rotating about an axis with an angular velocity ω_1, and at a later time t_2, it has an angular velocity ω_2, the average angular acceleration $\overline{\alpha}$ is given by Eq. (2), expressed

$$\overline{\alpha} = \frac{\omega_2 - \omega_1}{t_2 - t_1} = \frac{\Delta \omega}{\Delta t} \qquad (2)$$

in radians per second per second. The instantaneous angular acceleration is given by $\alpha = d\omega/dt$.

Consequently, if a rigid body is rotating about a fixed axis with an angular acceleration of magnitude α and an angular speed of ω_0 at a given time, then at a later time t the angular speed is given by Eq. (3).

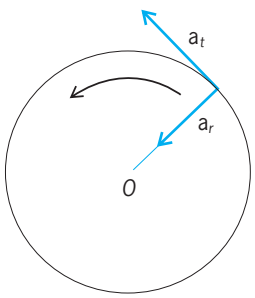
$$\omega = \omega_0 + \alpha t \qquad (3)$$

A simple calculation shows that the angular distance θ traversed in this time is expressed by Eq. (4).

$$\theta = \overline{\omega} t = \left[\frac{\omega_0 + (\omega_0 + \alpha t)}{2}\right] t = \omega_0 t + {}^1/_2 \alpha t^2 \qquad (4)$$

In the **illustration** a particle is shown moving in a circular path of radius R about a fixed axis through O with an angular velocity of ω radians/s and an angular acceleration of α radians/s^2. This particle is subject to a linear acceleration which, at any instant, may be considered to be composed of two components: a radial component $\mathbf{a}_r$ and a tangential component $\mathbf{a}_t$.

Radial acceleration. When a body moves in a circular path with constant linear speed at each point in its path, it is also being constantly accelerated toward the center of the circle under the action of the force required to constrain it to move in its circular path. This acceleration toward the center of path is called radial acceleration. In the illustration the radial acceleration, sometimes called centripetal acceleration, is shown by the vector $\mathbf{a}_r$. The magnitude of its value is v^2/R, or ω^2/R, where v is the instantaneous linear



Radial and tangential accelerations in circular motion.

velocity. This centrally directed acceleration is necessary to keep the particle moving in a circular path.

Tangential acceleration. The component of linear acceleration tangent to the path of a particle subject to an angular acceleration about the axis of rotation is called tangential acceleration. In the illustration, the tangential acceleration is shown by the vector $\mathbf{a}_t$. The magnitude of its value is αR. *See* ACCELERATION MEASUREMENT; ROTATIONAL MOTION.

C. E. Howe; R. J. Stephenson

Acceleration analysis

A mathematical technique, often done graphically, by which accelerations of parts of a mechanism are determined. In high-speed machines, particularly those that include cam mechanisms, inertial forces may be the controlling factor in the design of members. An acceleration analysis, based upon velocity analysis, must therefore precede a force analysis. Maximum accelerations are of particular interest to the designer. Although an analytical solution might be preferable if exact maxima were required, graphical solutions formerly tended to be simpler and to facilitate visualization of relationships. Today, for advanced problems certainly, a computer solution, possibly one based on graphical techniques, is often more effective and can also produce very accurate graphical output. *See* FORCE; KINEMATICS; MECHANISM.

Accelerations on a rigid link. On link OB (**Fig. 1***a*) acceleration of point B with respect to point O is the vector sum of two accelerations: (1) normal acceleration A^n_{BO} of B with respect to O because of displacement of B along a path whose instantaneous center of curvature is at O, and (2) tangential acceleration of A^t_{BO} of B with respect to O because of angular acceleration α.

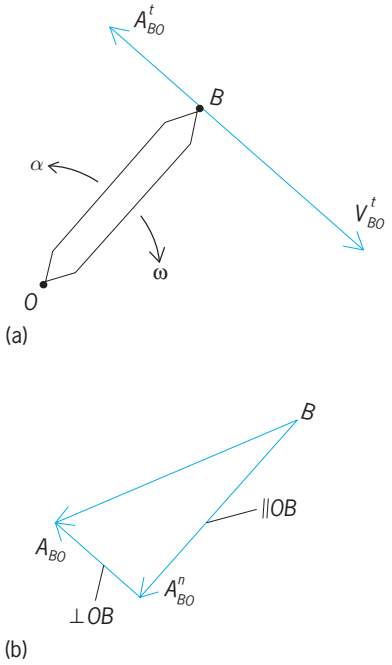


Fig. 1. Elementary condition. (*a*) Point *B* on rigid member rotates about center *O*. (*b*) Vector diagram shows normal, tangential, and resultant accelerations.

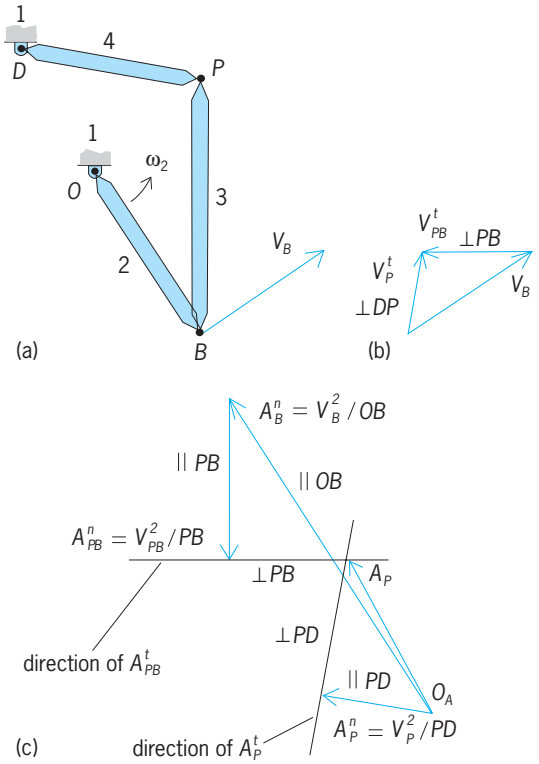


Fig. 2. Four-bar linkage. (*a*) Given are the linkages 2, 3, and 4 between fixed supports 1–1 and the velocity of point *B*, $\omega_2 = 0$. (*b*) A vector polygon is used to determine the velocity at point *P*. (*c*) A graphic solution then finds acceleration at *P*.

For conditions of Fig. 1*a* with link OB rotating about O at angular velocity ω and angular acceleration α, the accelerations can be written as Eqs. (1) and (2). The vector sum or resultant is A_{BO} (Fig. 1*b*).

$$A^n_{BO} = (OB)\omega^2 = V^2_B/(OB) \tag{1}$$

$$A^t_{BO} = (OB)\alpha \tag{2}$$

Accelerations in a linkage. Consider the acceleration of point P on a four-bar linkage (**Fig. 2***a*) with $\alpha_2 = 0$ and hence $\omega_2 = k$, the angular velocity of input link 2. First, the velocity problem is solved yielding V_B and, by Fig. 2*b*, V_{PB}. Two equations can be written for A_P; they are solved simultaneously by graphical means in Fig. 2*c* by using Fig. 2*b*; that is, normal accelerations of P with respect to B and D are computed first. Directions of tangential acceleration vectors A^n_{PB} and A^n_P are also known from the initial geometry. The tip of vector A_P must lie at their intersection, as shown by the construction of Fig. 2*c*. *See* FOUR-BAR LINKAGE.

Explicitly, acceleration A_P of point P is found on the one hand by beginning with acceleration A_B of point B, represented by Eq. (3). To this acceleration

$$A_p = A_B \mapsto A_{PB} \tag{3}$$

is added vectorially normal acceleration A^n_{PB} and tangential acceleration A^t_{PB}, which can be written as Eq. (4). Also for link 2, $\alpha_2 = 0$ and $A^t_B = 0$; and A_B need not be split up.

$$A_P = A_B \mapsto A^n_{PB} \mapsto A^t_{PB} \tag{4}$$

On the other hand, A_P can also be expressed as Eq. (5), or in Fig. 2*c*. The intersection of the two tangential components defines the tip of A_P. This problem illustrates the generalization that any basic acceleration analysis can be thoroughly performed only after completion of the underlying velocity analysis.

$$A_P = A^n_P \mapsto A^t_p \tag{5}$$

Acceleration field. Acceleration is a vector and hence has magnitude and direction but not a unique position. A plane rigid body, such as the cam of **Fig. 3***a*, in motion parallel to its plane will possess an acceleration field in that plane differing from, but resembling, a rigid-body velocity vector field. *See* CAM MECHANISM.

Every acceleration vector in this field, absolute or relative, makes with its instantaneous radius vector the same angle γ (tangent $\gamma = \alpha/\omega^2$) and is proportional in magnitude to the length of this radius vector (Fig. 3*b*). The acceleration field at any instant is thus defined by four parameters: magnitude and direction of accelerations at any two points on the body.

From an acceleration field defined by a_A and a_B (Fig. 3*c*), one can quickly determine the instantaneous center of zero acceleration. From point A', construct vector $A'A''$ parallel and equal to acceleration a_B, represented by vector $B'B$. The relative acceleration a_{AB} is given by the vector of Eq. (6).

$$a_{AB} = a_A - a_B \tag{6}$$

Angle γ between resultant a_{AB} and line AB is common to all radii vectors; therefore, construct radii vectors through A and through B both at γ (properly signed) to their respective acceleration vectors. These lines intersect at Γ, the center of zero acceleration. The geometry contains two similar triangles; $\Delta AB\Gamma$ is similar to $\Delta AA''A'$ because corresponding sides are equally inclined by the angle γ. From these similar triangles comes the generalization that acceleration magnitudes are proportional to radii vectors, just as for the velocity field.

Angle γ (tan $\gamma = \alpha/\omega^2$) is also used in orientation of acceleration image polygons, especially if the preceding velocity analysis used the image polygon method.

Points not on the same link. Acceleration of follower link 4 of the cam mechanism (Fig. 3*a*) is determined by considering relative accelerations of coincident points not on the same link. Thus point P on link 4, designated P_4, and coincident point P on link 2, designated P_2, have instantaneous velocities, as shown by vectors V_{P4} and V_{P2} (Fig. 3*a*). Relative velocity of P_4 with respect to P_2 is V_{P4P2} as indicated. Acceleration of P_4 can now be determined by Eq. (7), where

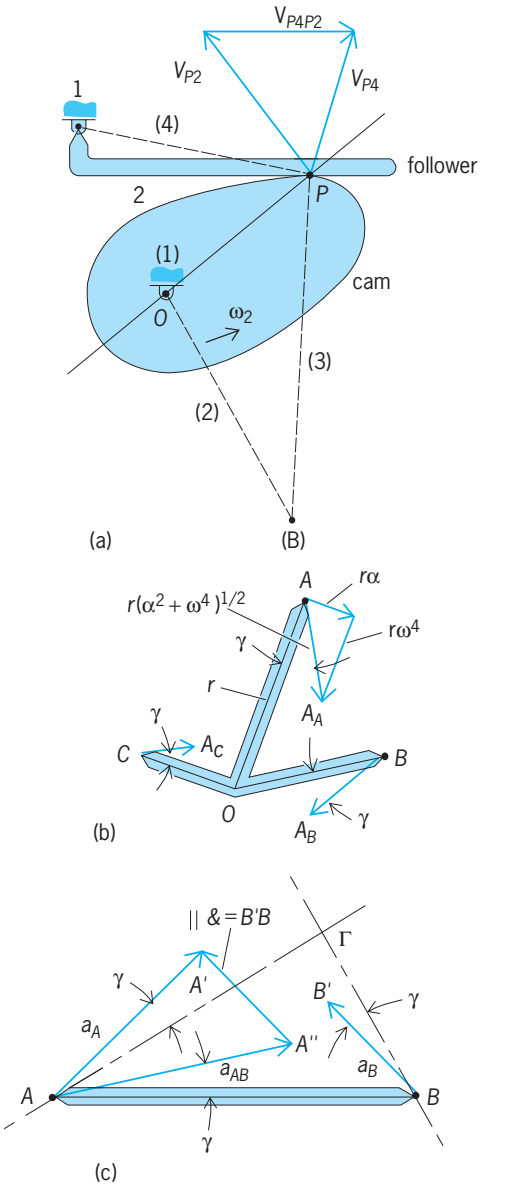


Fig. 3. Accelerations on a cam and follower. (*a*) The cam mechanism. (*b*) The acceleration field about a rigid rotating structure. (*c*) An acceleration field determining an instantaneous center of rotation.

$$A_{P_4} = A_{P_2} \mapsto A^n_{P_4P_2} \mapsto A^t_{P_4P_2} \mapsto 2\omega_2 V_{P_4P_2} \tag{7}$$

the last term is the Coriolis component. This component results from referring motion of P_4 on body 4 to P_2 on body 2. Serious errors of analysis can result from omission of the Coriolis component. *See* CORIOLIS ACCELERATION.

Occasionally, mechanisms that appear to require the analysis described in the preceding paragraph may also be analyzed by constructing an instantaneously equivalent linkage shown by broken lines (2), (3), and (4) in Fig. 3*a*. The instantaneous equivalent linkage is then analyzed by the method of Fig. 2.

Alternative procedures. Other methods may also be used. It is not always necessary to resolve vectors

into normal and tangential components; they can be solved by conventional orthogonal component techniques. For points A and B in part m, Eq. (8) holds.

$$a_A^{AB} = a_B^{AB} + a_{AB}^{AB} = a_B^{AB} + \omega^2 \cdot AB \qquad (8)$$

A convenient approximate solution is obtained by considering the difference between velocities due to a small angular displacement. This difference in velocities divided by the elapsed short time interval approximates the acceleration. Typically, the angular displacement is taken to be 0.1 radian (0.05 radian on either side of the position under study). *See* ACCELERATION; VECTOR (MATHEMATICS). Douglas P. Adams

Bibliography. J. Angeles, *Spatial Kinematics Chains: Analysis-Synthesis-Optimization,* 1982; J. S. Beggs, *Kinematics,* 1983; C. W. Ham et al., *Mechanics of Machinery,* 4th ed., 1958; D. Lent, *Analysis and Design of Mechanisms,* 2d ed., 1970; A. Sloane, *Engineering Kinematics,* 1966.

Acceleration measurement

The technique of measuring the magnitude and direction of acceleration. Measurement of acceleration includes translational acceleration with tangential and radial components and angular acceleration. *See* ACCELERATION.

Translation acceleration. The components of translational acceleration are the tangential one due to a change in magnitude and the normal one due to a change in direction of the velocity vector. Acceleration measurement by instruments is commonly expressed in terms of g, a dimensionless ratio of acceleration divided by the local value of gravity.

When all points of a rigid body are traveling along parallel straight line paths, the body is said to be in linear motion. For such motion the normal component of acceleration is absent, and if the velocity of the body is changing, the resulting translational acceleration is called linear acceleration. There are many familiar applications of linear acceleration measurement, for example, in land vehicles such as automobiles, in aerospace vehicles, and in moving machinery components.

One of the purposes of acceleration measurement is to determine the force causing the acceleration as manifested by Newton's second law. Also, from acceleration versus time data it is possible by either graphical or numerical integrations to determine the corresponding velocity-time and then the displacement-time histories. Since such techniques are comparatively accurate, it is evident that acceleration measurement is also important in the determination of velocity and displacement.

An acceleration-measuring instrument, or accelerometer, essentially consists of a reference mass mounted by means of a suitable configuration within a housing. A triaxial accelerometer will have three such reference masses so that they respond individually in three mutually perpendicular directions. Accelerometers with a single reference mass respond to accelerations along a predefined axis. To measure translational acceleration whose exact direction is unknown or whose direction varies, either a single triaxial accelerometer or three ordinary accelerometers are required. *See* ACCELEROMETER.

An indirect determination of acceleration is possible by measuring the displacement-time history of a system undergoing linear acceleration. A graphical or a numerical determination of the slope or derivative at selected values of time will produce the corresponding velocity-time relationship, and a repetition of the procedure yields the acceleration history. The accuracy of this means of obtaining acceleration values is poor because the second set of slope determinations by such approximate methods is inherently inaccurate.

The acceleration g due to gravity on the Earth can be determined by various means. A reasonably accurate means of determining g at a given location is by accurately measuring the period T and length L of a simple pendulum on a long cord and subject to a small oscillation in still air. The value of g is then readily determined from the formula $T = 2\pi\sqrt{L/g}$. Another means of determining g is offered by measuring the deflection of a reference mass supported by a highly accurate spring with known characteristics.

Angular acceleration. Technically, a body is said to have angular motion if some fixed reference line between two points rotates or sweeps out an angle. With this definition it is evident that a body can travel along a curved path and still not rotate. For example, the chairs on a moving ferris wheel are not in angular motion with respect to ground. *See* ANGULAR MOMENTUM.

The measurement of angular acceleration is useful in applications such as the angular motion of aerospace vehicles and the components of rotating machinery. Since angular and translational motion are kinematically related, it is possible, when the exact location of the axis of rotation is known, to determine angular acceleration by measuring the corresponding tangential component of translational acceleration.

For determining angular acceleration, a suitably suspended reference mass responds to an angular acceleration about an axis parallel to, or coincident with, its own predefined axis of rotation. As is the case for the translational acceleration measuring instruments, various designs of the angular versions are possible. A direct means of measuring angular acceleration is afforded by the angular accelerometer.

An indirect means of determining angular acceleration is to measure the angular velocity of a body. This technique is applicable in situations where a body is in pure rotation about a fixed axis. The direction of the rotating body can be determined from the output of an electric generator connected to it on the same shaft. A stroboscope can also be used for the same purpose. The rotational speed versus time history can then be graphically or numerically

differentiated once to yield the angular acceleration. *See* ROTATIONAL MOTION; STROBOSCOPE.

Teruo Ishihara; Roger C. Duffield

Bibliography. T. G. Beckwith, R. D. Marangoni, and J. H. Lienhard, *Mechanical Measurements,* 6th ed., 2006; R. S. Figliola and D. E. Beasley, *Theory and Design for Mechanical Measurements*, 4th ed., 2006; R. C. Hibbeler, *Engineering Mechanics: Statics and Dynamics*, 10th ed., 2003; W. F. Riley and L. D. Sturges, *Engineering Mechanics*: *Dynamics*, 2d ed., 1995.

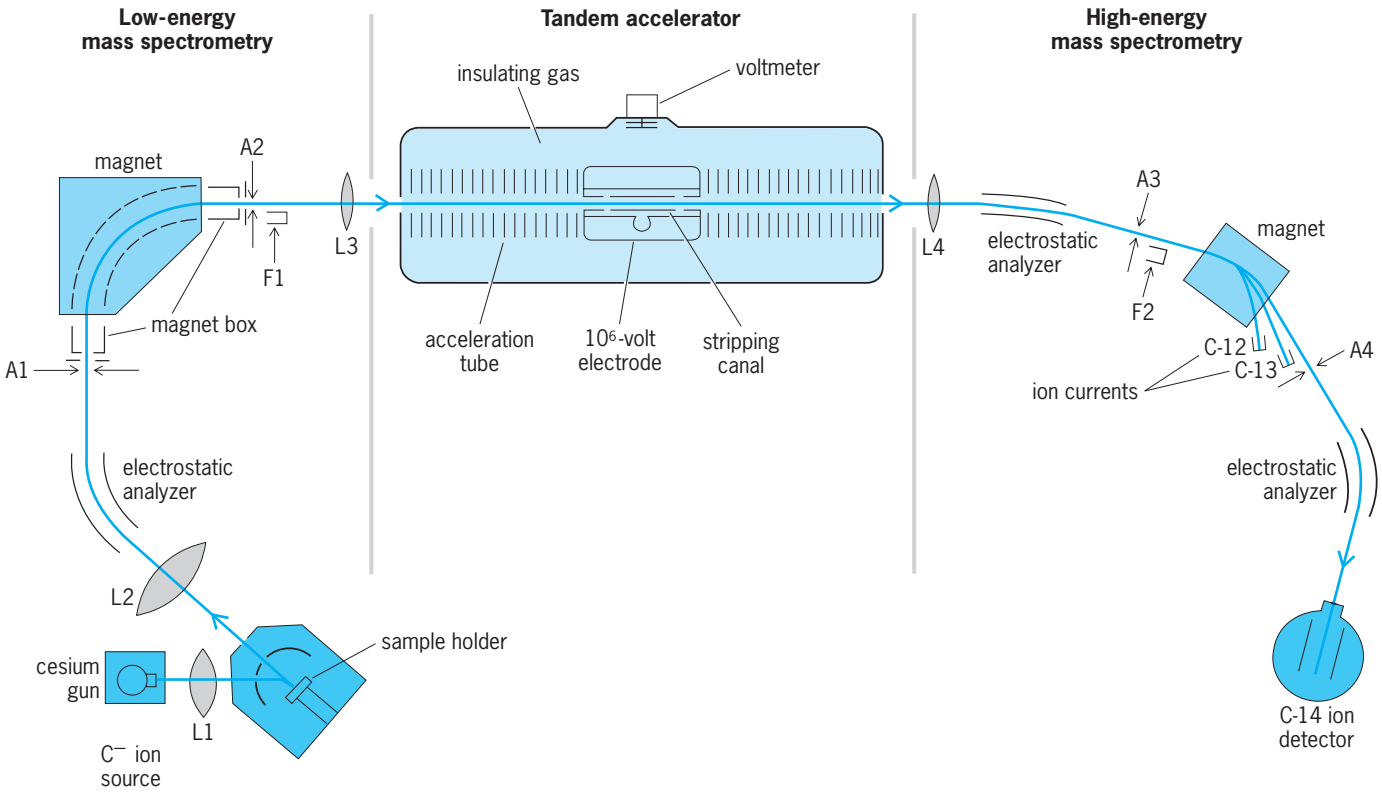
Accelerator mass spectrometry

The use of a combination of mass spectrometers and an accelerator to measure the natural abundances of very rare radioactive isotopes. These abundances are frequently lower than parts per trillion. The most important applications of accelerator mass spectrometry are in archeological, geophysical, environmental, and biological studies, such as in radiocarbon dating by the counting of the rare carbon-14 (radiocarbon; ^{14}C) isotope. *See* MASS SPECTROSCOPE; PARTICLE ACCELERATOR.

The advantage of counting the radioactive atoms themselves rather than their decay products is well illustrated by radiocarbon dating, which requires the measurement of the number of ^{14}C atoms in a sample. This number can be inferred from the beta-particle emission rate or by counting directly the number of radioactive atoms. The long half-life of 5730 years for ^{14}C implies that only 15 beta-particle emissions per minute are observed from 1 g (0.035 oz) of contemporary carbon which contains 6×10^{10} atoms of ^{14}C. Conventional mass spectrometry has not been able to accomplish the task of counting the ^{14}C atoms directly. However, an accelerator mass spectrometer can be used to count the ^{14}C atoms at over 30 per second from a milligram sample of carbon. Consequently, accelerator mass spectrometry can be used to date samples that are a thousand times smaller than those that are dated by using the beta-particle counting method, and the procedure is carried out about 120 times faster. The advantage of accelerator mass spectrometry over beta-particle counting is even greater for very long-lived radioactive isotopes such as aluminum-26 (7×10^5 year half-life) and iodine-129 (1.6×10^7 year half-life). *See* RADIOACTIVITY; RADIOCARBON DATING.

Apparatus. The success of accelerator mass spectrometry results from the use of more than one stage of mass spectrometry and at least two stages of ion acceleration. The layout of an ideal accelerator mass spectrometer for radiocarbon studies is generally divided for convenience into three stages (see **illus.**).

In the first, low-energy, mass spectrometry stage, there is the usual acceleration at the negative-ion source to produce a directed beam of ions with



Simplified diagram of an accelerator mass spectrometer used for radiocarbon dating. The equipment is divided into three sections. Electric lenses L1–L4 are used to focus the ion beams. Apertures A1–A4 and charge collection cups F1 and F2 are used for setting up the equipment. The cesium ions from the gun create the negative ions of carbon at the surface of the sample.

an energy of several kiloelectronvolts. This result is achieved by raising the sample holder to the required negative voltage. The first acceleration is followed by an electrostatic lens and an electrostatic analyzer for further ion focusing and energy selection, so that the subsequent mass analysis by a magnetic analyzer can select, uniquely, the ion mass of interest. The first part of the accelerator mass spectrometer is therefore very similar to a conventional mass spectrometer. In this case, the ^{12}C, ^{13}C, and ^{14}C isotopes are analyzed sequentially by varying the voltage on the magnet vacuum box for each isotope in order to analyze each isotope without changing the magnetic field. *See* ION SOURCES.

In the second stage, there is further acceleration by a tandem accelerator. This first accelerates negative ions to the central high-voltage electrode, converts them into positive ions by several successive collisions with gas molecules in a region of higher gas pressure, known as a stripping canal, and then further accelerates the multiply charged positive ions through the same voltage difference back to ground potential. If the acceleration voltage used here is 2×10^6 V, the ions emerging from the accelerator have 8×10^6 eV in energy when they are triply charged in the positive-ion stage.

In the third stage, the accelerated ions are analyzed further by the high-energy mass spectrometer. Following an electrostatic analyzer and a magnetic analyzer, the ion currents of ^{12}C and ^{13}C ions can be measured in the two Faraday cups; additional electrostatic or magnetic analysis then completes the isolation of the rare ionic species. In a typical arrangement (see illus.), the currents of the isotopes ^{12}C and ^{13}C are measured in sequence for a short time, and the much rarer ^{14}C ions then are counted for a longer time in a suitable ion detector. The $^{14}C/^{12}C$ and $^{13}C/^{12}C$ ratios can then be determined from these measurements. Finally, these ratios determine the age of the sample.

Distinguishing features. These three stages of analysis illustrate the features that clearly distinguish accelerator mass spectrometry from conventional mass spectrometry. These are the elimination of molecular ions and isobars from the mass spectrometry.

Elimination of molecular ions. The abundant molecular ions, such as $^{13}CH^-$ and $^{12}CH_2^-$, which have similar masses to the $^{14}C^-$ ions, either must be resolved by a very high-resolution mass spectrometer or must be separated by some other method. A tandem accelerator provides a convenient way of completely eliminating molecular ions from the mass spectrometry because ions of a few megaelectronvolts can lose several electrons on passing through the region of higher gas pressure in the stripping canal. Molecules with more than two electrons missing have not been observed in this case, so that the accelerator mass spectrometry of radiocarbon, utilizing charge-3 ions, is free of molecular interferences. The molecular fragments are removed by the high-energy mass spectrometry.

Elimination of isobars. The use of a negative-ion source, which is necessary for tandem acceleration, can also ensure the complete separation of atoms of nearly identical mass (isobars). In the case of radiocarbon analysis, the abundant stable ^{14}N ions and the very rare radioactive ^{14}C ions, which differ in mass by only 1 part in 86,000, are separated completely because the negative ion of nitrogen is unstable whereas the negative ion of carbon is stable. Other examples of rare radioisotopes that can be completely separated at the negative-ion source are ^{26}Al and ^{129}I. In addition, it is frequently possible to count the ions without any background in the ion detectors because of their high energy. In many cases, it is also possible to identify the ion. For example, ions of the isobars ^{10}Be and ^{10}B can easily be distinguished by their differing ranges in matter in the final ion detection system. Both beryllium and boron form negative ions so they cannot be distinguished at the ion source.

Advantages. As a result of the destruction of molecules in the accelerator and the elimination of the interfering isobars at the negative-ion source or their separation by the ion detection system, the modest mass-resolution requirements of accelerator mass spectrometry can be achieved with high efficiency. This is a great advantage of accelerator mass spectrometry in the measurement of rare radioactive species.

For the study of many rare radioactive atoms, accelerator mass spectrometry also has the important advantage that there can be no background except for contamination with the species being studied. For example, significant interference with the beta-particle counting of radiocarbon from cosmic rays and natural radioactivity occurs for carbon samples about 25,000 years old. In contrast, accelerator mass spectrometer measurements are affected only by the natural contamination of the sample which becomes serious for samples about 50,000 years old.

Applications. Accelerator mass spectrometry has been applied to an increasing variety of problems. In particular, radiocarbon dating is carried out at many laboratories, with thousands of dates being determined each year. These dates approach the accuracy of those produced by the beta-particle counting method. The principal advantage of accelerator mass spectrometry for radiocarbon dating is the ability to date smaller samples of carbon and to do it more quickly. The advantage of the lower background is very useful, and the background limit has been the subject of further studies.

An example of the use of accelerator mass spectrometry can be found in the dating of a small piece of wood, cut by a metal implement, from an archeological site in Newfoundland. This confirmed directly that the site had been occupied by people from Europe about the year 1000. It was presumably one of the early Viking settlements.

There has been an international study of an artifact known as the Shroud of Turin. Tradition holds that this was the burial cloth of Jesus; however, it is now widely believed to be an icon dating from about 1300. Only very small samples from this valuable artifact were made available for dating. *See* ART CONSERVATION CHEMISTRY.

Radiocarbon dating has had an increasing impact on geophysical measurements and is particularly suitable for studying the events associated with the most recent ice age. For example, it is possible to date raised beaches by measuring the radiocarbon from single shells instead of being forced to use heterogeneous mixtures of a large number of shells. Accelerator mass spectrometry is used for oceanographic studies, such as measuring the age of the dissolved carbon dioxide (CO_2) in ocean water, and the age of foraminifera from ocean sediments. Much smaller samples of water or foraminifera can be collected for analysis than would otherwise be possible, facilitating important studies—for example, of global ocean circulation—that are closely related to the understanding of climate.

Radiocarbon has been used extensively as a biomedical tracer in order to detect the effects of toxins, mutagens, carcinogens, or chemotherapeutics in living systems. Radiocarbon is used because it can readily be incorporated into molecules without causing a change in chemical behavior. Accelerator mass spectrometry is important for these studies because of the small sample sizes that can be used, which result in a much lower radiation dose being received by the organism.

Beryllium-10. The long-lived radioactive beryllium isotope ^{10}Be is created in the atmosphere by cosmic rays. Since its half-life is 1.6×10^6 years, it is extremely difficult to study by beta-particle counting, but it was first studied at natural abundances by accelerator mass spectrometry, using a cyclotron as an accelerator and range separation to eliminate the abundant natural isobar ^{10}B. It is now measured routinely by accelerator mass spectrometry in many laboratories by using tandem accelerators. The ^{10}Be atoms in sediments subducted beneath continents and later emitted from volcanoes have been used in studies of the time taken for such subterranean transport. Studies of ^{10}Be in ice cores and sediments are used in determining accumulation rates over millions of years.

Chlorine-36. The long-lived isotope of chlorine ^{36}Cl with a half-life of 350,000 years has also been extensively studied by accelerator mass spectrometry. Unlike ^{10}Be, the ^{36}Cl atoms readily dissolve in water and so are of importance to the science of hydrology. The ^{36}Cl is also created by cosmic rays in the atmosphere and so can be used to study the transport of water from the atmosphere and through underground aquifers. The ^{36}Cl, like ^{26}Al, is also created in surface rocks by cosmic rays, making possible the dating of the exposure of glacial moraines and certain types of rock. For example, the rocks ejected from a meteor crater have been used to estimate that the crater was formed about 30,000 years ago.

Iodine-129. The isotope ^{129}I has a half-life of 1.6×10^7 years and can be studied most easily by accelerator mass spectrometry. The ^{129}I is created in the atmosphere by cosmic rays but is generated mainly by the neutron-induced fission of the uranium isotope ^{235}U. This fission occurs naturally in uranium ores, during the production of nuclear power, and from the fallout following the use of nuclear weapons. This isotope has been used to study the dispersal in the oceans of radioactive wastes from nuclear fuel reprocessing sites and reactors in Europe and North America. The ^{129}I has been measured over a wide area of the North Atlantic, and it can be detected easily in liter quantities of water. It has also been studied in connection with the dispersal of radioactive wastes in Russia and the seas north of Russia.

Many applications of accelerator mass spectrometry will continue to be made in the future. The small sample size required and the high sensitivity for detecting certain long-lived radioactive atoms or very rare stable atoms such as iridium will facilitate their use as environmental and biological tracers. Methods for the detection of very rare atoms in high-purity materials are also being developed. *See* DATING METHODS; GEOCHRONOMETRY; MASS SPECTROMETRY.

Albert E. Litherland

Bibliography. L. Brown, Applications of accelerator mass spectrometry, *Annu. Rev. Earth Planet. Sci.,* 12:39–59, 1984; C. E. Chen et al. (eds.), *Accelerator Mass Spectrometry*, 1992; D. Elmore and F. M. Phillips, Accelerator mass spectrometry for measurement of long-lived radioisotopes, *Science*, 236:543–550, 1987; A. E. Litherland, Fundamentals of accelerator mass spectrometry, *Phil. Trans. Roy. Soc. London,* A323:5–21, 1987.

Accelerometer

A mechanical or electromechanical instrument that measures acceleration. The two general types of accelerometers measure either the components of translational acceleration or angular acceleration. *See* ACCELERATION MEASUREMENT.

Translational accelerometers. Most translational accelerometers fall into the category of seismic instruments, which means the accelerations are not measured with respect to a reference point. Of the two types of seismic instruments, one measures the attainment of a predefined acceleration level, and the other measures acceleration continuously. **Figure 1** shows the first type of instrument, in which a seismic mass is suspended from a bar made of brittle material which fails in tension at a predetermined

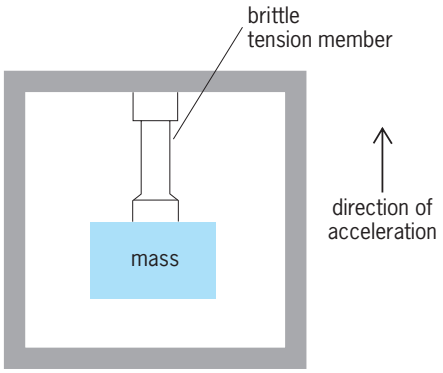


Fig. 1. Brittle-member acceleration-level indicator.

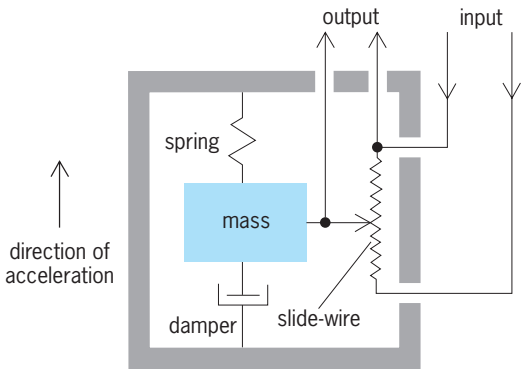


Fig. 2. Slide-wire potentiometer.

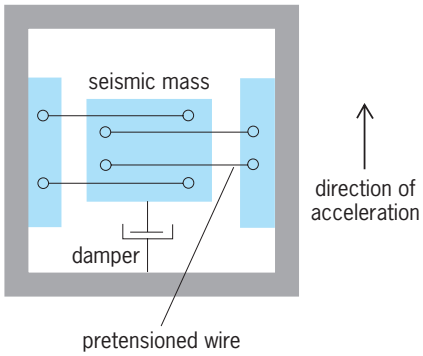


Fig. 3. Unbounded strain gage accelerometer.

acceleration level. *See* SEISMOGRAPHIC INSTRUMENTATION.

Continuously measuring seismic instruments are composed of a damped or an undamped spring-supported seismic mass which is mounted by means of the spring to a housing. The seismic mass is restrained to move along a predefined axis. Also provided is some type of sensing device to measure acceleration. To assure that the deflection of the seismic mass is proportional to the acceleration, such accelerometers are rated so that the excitation frequency associated with the acceleration is only 0.1 to 0.4 of the instrument's resonance frequency, depending on the amount of damping provided. For most accelerometers the damping coefficient varies between 0 and 0.7 of critical damping.

The type of sensing device used to measure the acceleration determines whether the accelerometer is a mechanical or an electromechanical instrument. One type of mechanical accelerometer consists of a liquid-damped cantilever spring-mass system, a shaft attached to the mass, and a small mirror mounted on the shaft. A light beam reflected by the mirror passes through a slit, and its motion is recorded on moving photographic paper. The type of electromechanical sensing device classifies the accelerometer as variable-resistance, variable-inductance, piezoelectric, piezotransistor, or servo type of instrument or transducer.

The sensing device for variable-resistance accelerometers operates on the principle that electrical resistance of an conductor is a function of its dimensions. When the dimensions of the conductor are varied mechanically, as constant current flows through it, the voltage developed across it varies as a function of this mechanical excitation. One type of variable-resistance transducer makes use of a slide-wire potentiometer (**Fig. 2**). This type of accelerometer has low acceleration-measuring ability and low resonance frequency. A more common type of variable-resistance transducer makes use of the fact that the resistance in a wire (called a strain gage) changes when it is strained. Most strain gage accelerometers are either of the unbounded or bounded type. In unbounded strain gage instruments fine pretensioned wires support a fluid-damped seismic mass, and the wires act as both the spring element and the strain gage (**Fig. 3**). The bounded strain gage transducers have the strain gages mounted directly on the spring element, and the gages measure the strain that takes place in the spring due to the relative deflection of the seismic mass. Wire strain gage accelerometers are characterized by limited high resonance frequency, acceleration measurement up to 1000 g, and ability to measure acceleration signals down to zero (dc) frequencies. *See* POTENTIOMETER.

Another type of variable-resistance accelerometer uses as the spring a piezoresistive element which is a crystal semiconductor whose resistivity changes with applied force (**Fig. 4**). This type of instrument has much greater sensitivity and resonance frequency than the wire strain gage transducer and is capable of measuring accelerations of 10,000 g or less. Variable-resistance instruments are usually arranged electrically to form a Wheatstone bridge. The bridge input excitation is from dry cells or commercial power supplies, and the output which is usually amplified is recorded by means of an oscilloscope or an oscillograph recorder.

A third type of variable-resistance accelerometer employs the Rolamite concept. This accelerometer is composed of two rolling cylinders supported by a thin flexible band, and this low-friction mechanical system is constrained by two parallel guide surfaces (**Fig. 5**). The flexible band has a cutout which provides a negative spring against the displacement of the rollers, which are also the seismic mass. The sensing device consists of a roller making electrical contact across two resistive elements located on one of

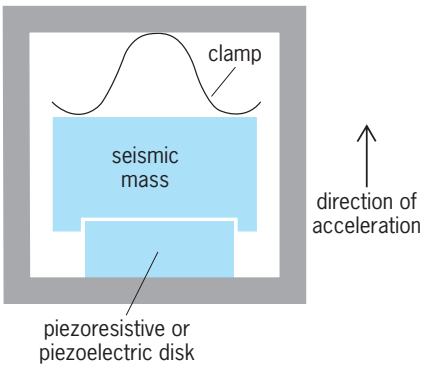


Fig. 4. Piezoresistive or piezoelectric accelerometer.

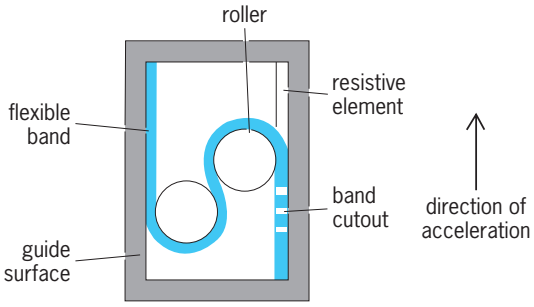


Fig. 5. Rolamite type of accelerometer.

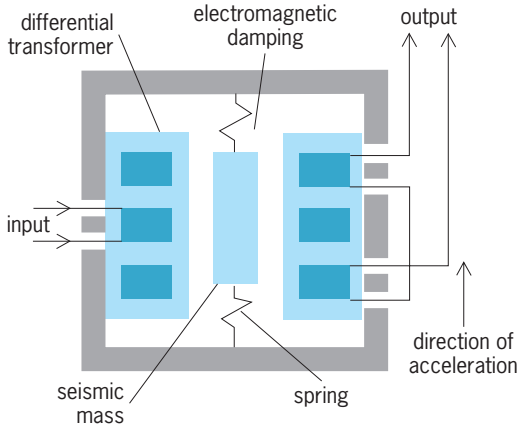


Fig. 6. Variable-inductance accelerometer.

the guide surfaces. The Rolamite accelerometer can be designed to be miniature in size and to measure accelerations in the range 1–500 *g*.

The variable-inductance accelerometer is a differential transformer which makes use of three coils arranged as shown in **Fig. 6**. The center coil is excited from an external ac power source, and the two end coils connected in series opposition are used to produce an ac output which is proportional to the displacement of a seismic mass passing through the coils. This instrument is characterized by fairly low resonance frequencies, acceleration measurements of 50 *g* or less, and high voltage outputs at low frequencies which eliminate the need for preamplifiers.

Piezoelectric accelerometers utilize a piezoelectric crystal which supports the seismic mass in such a way that the crystal acting as a spring is strained in either compression, flexure, or shear (Fig. 4). Many crystals, such as quartz, Rochelle salts, and barium titanate, have these properties. A number of ceramic materials can be used when they contain certain impurities. Piezoelectric instruments are noted for their small size, large acceleration measurement up to 5000 *g*, very high resonance frequency, inability to measure dc frequency acceleration signals, and the need for specialized circuitry. The acceleration signal is usually amplified by means of a charge or voltage amplifier and then recorded.

A piezotransistor accelerometer consists of a seismic mass supported by a stylus which transmits a concentrated force to the upper diode surface of the transistor (**Fig. 7**). Because of this force, localized stresses occur in the upper diode surface which, in turn, cause a very large reversible change in current across the underlying *pn* junction of the transistor. This change of current is a measure of the acceleration. These instruments are characterized by small size, very high resonance frequencies, acceleration measurements of 400 *g* or less, and voltage outputs that are high enough to eliminate need for amplifiers.

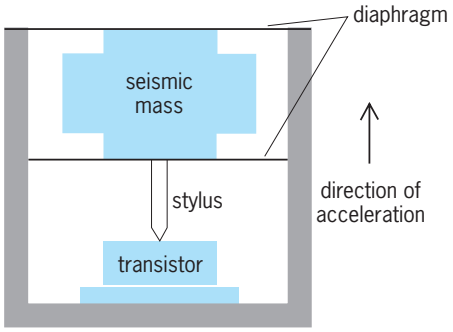


Fig. 7. Piezotransistor accelerometer.

A servo accelerometer is a high-accuracy transducer, containing an electromechanical servomechanism which holds a mass in a fixed position as the instrument is accelerated. The force required to restrain the mass is a measure of the acceleration. This instrument has limited high-frequency response, acceleration outputs of 100 *g* or less, and capability of measuring dc frequency responses. *See* SERVOMECHANISM.

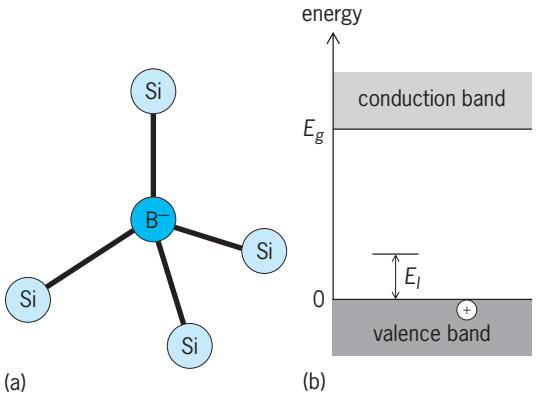
Angular accelerometers. There are several different types of angular accelerometers. In one type the damping fluid serves as the seismic mass. Under angular acceleration the fluid rotates relative to the housing and causes on two symmetrical vanes a pressure which is a measure of the angular acceleration. Another type of instrument has a fluid-damped symmetrical seismic mass in the form of a disk which is so mounted that it rotates about the normal axis through its center of gravity. The angular deflection of the disk, which is restrained by a spring, is proportional to the angular acceleration. The sensing devices used in these instruments to measure the angular acceleration are of the variable-resistance and servo types.

Roger C. Duffield; Teruo Ishihara

Bibliography. T. G. Beckwith, R. D. Marangoni, and J. H. Lienhard, *Mechanical Measurements,* 6th ed., 2006; W. Boyes, *Instrumentation Reference Book*, 3d ed., 2002; E. O. Doebelin, *Measurement Systems: Application and Design,* 5th ed., 2003; C. P. Wright, *Applied Measurement Engineering,* 1994.

Acceptor atom

An impurity atom in a semiconductor which can accept or take up one or more electrons from the crystal and become negatively charged. An atom which substitutes for a regular atom of the material but has one less valence electron may be expected to be an acceptor atom. For example, atoms of boron, aluminum, gallium, or indium are acceptors in germanium and silicon (**illus.** *a*), and atoms of antimony



Trivalent acceptor atom, boron (B), in the elemental semiconductor silicon (Si). (*a*) Boron atom in a substitutional position, that is, replacing silicon, a tetravalent host atom, by completing the four tetrahedral covalent bonds with its nearest neighbor silicon atoms. This requires an electron to be accepted from the valence band, thus making boron negatively charged. (*b*) Energy diagram showing that the absence of an electron in the valence band is equivalent to a positive charge carrier, a hole, which is bound to boron via Coulomb attraction with an ionization enerzy E_I. E_g = energy gap separating valence band from conduction band.

and bismuth are acceptors in tellurium crystals. Acceptor atoms tend to increase the number of holes (positive charge carriers) in the semiconductor (illus. *b*). The energy gained when an electron is taken up by an acceptor atom from the valence band of the crystal is the ionization energy of the atom. *See* DONOR ATOM; SEMICONDUCTOR.

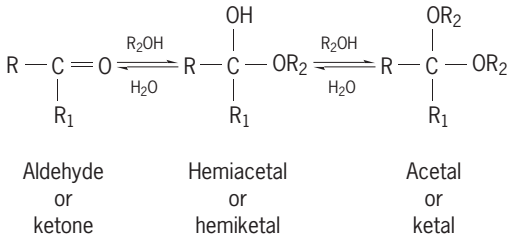
H. Y. Fan; Anant K. Ramdas

Bibliography. C. Kittel, *Introduction to Solid State Physics*, 7th ed., Wiley, 1996; A. K. Ramdas and S. Rodriguez, Spectroscopy of the solid-state analogues of the hydrogen atom: Donors and acceptors in semiconductors, *Rep. Prog. Phys.*, 44:1297–1387, 1981.

Acetal

A geminal diether (R_1 = H). Ketals, considered a subclass of acetals, are also geminal diethers (R_1 = C, aliphatic or aromatic). Acetals are (1) independent structural units or a part of certain biological and commercial polymers, (2) blocking or protecting groups for complex molecules undergoing selective synthetic transformations, and (3) entry compounds for independent organic chemical reactions.

Acetals are easily prepared by the reaction of aldehydes with excess alcohol, under acid-catalyzed conditions. This is usually a two-step process, as shown below, in which an aldehyde is treated with an



alcohol to yield a less stable hemiacetal, which then reacts with additional alcohol to give the acetal. Protonic or Lewis acids are effective catalysts for acetal formation; dehydrating agents, such as calcium chloride and molecular sieves, can also be used for molecules, such as sugars, where acids may cause problems. Less common acetal preparations are Grignard reagent condensation with orthoformates and mercuric-catalyzed additions of alcohols to acetylenes. *See* GRIGNARD REACTION.

Two well-known hemiacetals (monomers) are α- and β-glucose: the 1,4-polysaccharide from α-glucose is amylose (amylopectin has an acetal 1,6-linkage branch from the linear polymer backbone), and the 1,4-polysaccharide from β-glucose is cellulose. The acetal link connects the monomers incorporated into the polymer in each case, and it is formed by the enzymatic condensation of the hemiacetal functional group (C_1—OH) of one monomer with the (C_4—OH) hydroxyl functional group of another monomer. Several other biopolymers containing similar acetal links between repeating units are chitin, inulin, and pectin. Synthetic polymers with 1,3-diol units, such as polyvinyl alcohol, can be transformed into polymeric acetals, which are stable and useful plastics (for example, formal and butyral). Difunctional aldehydes and ketones can be condensed with certain tetrols (for example, pentaerythritol) for preparation of polyketals, and polyspiroketals, which are highly crystalline materials. Polyoxymethylene, which is the acetal resin homopolymer of formaldehyde, is prepared by cationic polymerization of trioxane, the cyclic acetal of formaldehyde. Homopolymerization of acetaldehyde or acetone, where the acetal or ketal oxygen is also part of the polymer backbone, is possible, but requires specialized conditions. *See* CELLULOSE; POLYACETAL; POLYMER; POLYMERIZATION; POLYSACCHARIDE; STARCH.

Their use as protective groups is attributed to the fact that acetals (or ketals) are readily cleaved by aqueous acids to the respective aldehyde (or ketone) and that they are stable to alkalies and nucleophiles. Cyclic acetals, such as *meta*-dioxolanes and *meta*-dioxanes, are very convenient to work with for blocking–unblocking types of syntheses. During these multistep processes, the carbonyl portion of the molecule is transformed to its acetal (or ketal), while synthetic reactions that do not utilize acid are effected on other functional groups of the molecule. The acetal (or ketal) part of the molecule is then readily hydrolyzed to free the carbonyl functional group. The cyclic acetals can also be used as aprotic solvents and plasticizers. *See* ALDEHYDE; CARBONYL; KETONE; ORGANIC SYNTHESIS.

Acetals undergo a variety of useful reactions such as α-acylation, α-bromination, condensations with enol acetates and silyl enolates, conversion to enol ethers, oxidation to esters, reduction to ethers, and reactions with active hydrogen compounds (for example, diethyl malonate). Acetal additions to vinyl ethers followed by hydrolysis result in a preparation of α,β-unsaturated aldehydes. Dioxolanes (from aldehydes) are also treated with alkyllithiums for the preparation of ketones.

Thioacetals, especially cyclic thioacetals, can be reduced to alkanes with Raney nickel, and 1,3-dithianes, prepared from aldehydes, can be metalated, condensed with a variety of electrophilic reagents, and hydrolyzed to free the functionalized aldehyde. Charles F. Beam

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; D. S. Kemp and F. Vellaccio, *Organic Chemistry*, 1980; R. W. Lenz, *Organic Chemistry of Synthetic High Polymers*, 1967, reprint 1988; J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structures*, 4th ed., 1992; R. B. Seymour and C. E. Carraher, Jr., *Polymer Chemistry*, 1982; M. P. Stevens, *Polymer Chemistry*, 3d ed., 1998.

Acetic acid

A colorless, pungent liquid, CH_3COOH, melting at 16.7°C (62.1°F) and boiling at 118.0°C (244.4°F). Acetic acid is the sour principle in vinegar. Concentrated acid is called glacial acetic acid because of its readiness to crystallize at cool temperatures. Acetic acid in vinegar arises through an aerobic fermentation of dilute ethanol solutions, such as wine, cider, and beer, with any of several varieties of *Acetobacter*.

Chemical properties. Pure acetic acid is completely miscible with water, ethanol, diethyl ether, and carbon tetrachloride, but is not soluble in carbon disulfide. Freezing of acetic acid is accompanied by a remarkable volume contraction: the molar volume of liquid acetic acid at the freezing point is 57.02 cm^3/mole, but at the same temperature the crystalline solid is 47.44 cm^3/mole. It is a strongly proton-donating solvent with a relatively small dipole moment and a low dielectric constant. In a water solution, acetic acid is a typical weakly ionized acid ($K_a = 1.8 \times 10^{-5}$). *See* CARBOXYLIC ACID.

The vapor density of acetic acid indicates a molecular weight considerably higher than would be expected for a compound with a formula weight of 60.05. The acid probably exists largely as the dimer in the vapor and liquid states.

Acetic acid neutralizes many oxides and hydroxides, and decomposes carbonates to furnish acetate salts, which are used in textile dyeing and finishing, as pigments, and as pesticides; examples are verdigris, white lead, and paris green. *See* ARSENIC; LEAD.

Over two-thirds of the acetic acid manufactured is used in production of either vinyl acetate or cellulose acetate. Acetic anhydride, the key intermediate in making cellulose acetate, is prepared commercially by pyrolysis of acetic acid in the presence of trialkyl phosphate catalyst. Considerable amounts of acetic acid are consumed in the manufacture of terephthalic acid by liquid-phase oxidation of xylene, and in the preparation of esters for lacquer solvents, paints and varnishes, pharmaceuticals, and herbicides.

Production. Acetic acid was formerly manufactured from pyroligneous acid obtained in destructive distillation of wood. These processes are of historical interest because many modern chemical engineering operations developed through the study of acetic acid production. Today acetic acid is manufactured by three main routes: butane liquid-phase catalytic oxidation in acetic acid solvent, palladium–copper salt–catalyzed oxidation of ethylene in aqueous solution, and methanol carbonylation in the presence of rhodium catalyst. Large quantities of acetic acid are recovered in manufacturing cellulose acetate and polyvinyl alcohol. Some acetic acid is produced in the oxidation of higher olefins, aromatic hydrocarbons, ketones, and alcohols. *See* ESTER; OXIDATION PROCESS; SOLVENT; WOOD CHEMICALS. Frank Wagner

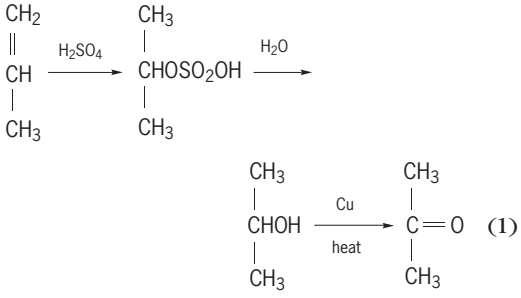
Bibliography. W. H. Brown, *Introduction to Organic and Biochemistry*, 4th ed., 1987; T. A. Geissman, *Principles of Organic Chemistry*, 4th ed., 1977; J. D. Roberts and M. C. Caserio, *Basic Principles of Organic Chemistry*, 2d ed., 1977.

Acetone

A chemical compound, CH_3COCH_3; the first member of the homologous series of aliphatic ketones. It is a colorless liquid with an ethereal odor. Its physical properties include boiling point 56.2°C (133°F), melting point −94.8°C (−138.6°F), and specific gravity 0.791. Acetone is an extremely important, low-cost raw material that is used for production of other chemicals.

Acetone is used as a solvent for cellulose ethers, cellulose acetate, cellulose nitrate, and other cellulose esters. Cellulose acetate is spun from acetone solution. Lacquers, based on cellulose esters, are used in solution in mixed solvents including acetone. Acetylene is safely stored in cylinders under pressure by dissolving it in acetone, which is absorbed on inert material such as asbestos. It has a low toxicity.

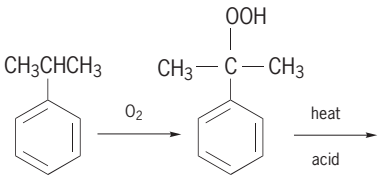
Production. The principal method of acetone production uses propylene, obtained from the cracking of petroleum. Addition of sulfuric acid to propylene yields isopropyl hydrogen sulfate, which upon hydrolysis yields isopropyl alcohol. Oxidation or dehydrogenation over metal catalysts, such as copper, converts the alcohol to acetone, as shown in reaction (1).

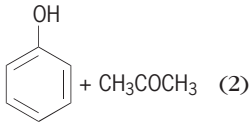


(1)

Acetone is also produced by passage of acetic acid vapor over metallic oxide catalysts at 400–450°C (750–840°F), by partial oxidation of the lower alkane hydrocarbons, and by the decomposition of cumene

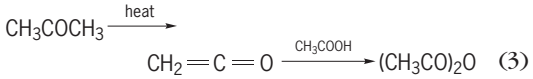
hydroperoxide, as shown in reaction (2). Phenol is the other product of this last process.





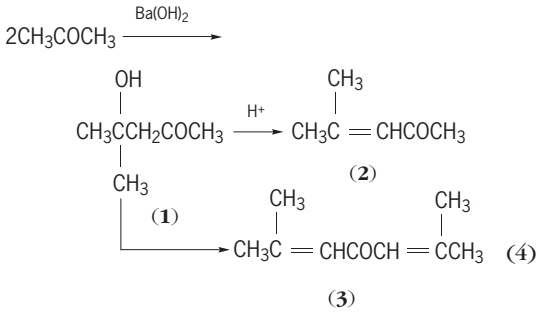
(2)

Chemical uses. Pyrolysis of acetone vapor at 700°C (1300°F) produces ketene, which reacts with acetic acid to produce acetic anhydride, as shown in reaction (3).



(3)

Aldol-type condensation of acetone with $Ba(OH)_2$ yields diacetone alcohol (**1**), mesityl oxide (**2**), and phorone (**3**), shown in reaction (4). Catalytic hydrogenation of mesityl oxide gives methyl isobutyl ketone. All these products are of commercial use as solvents or as chemical intermediates. *See* KETONE.



(4)

David A. Shirley

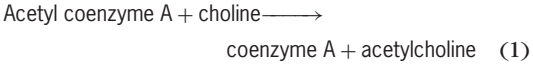
Acetylcholine

A naturally occurring quaternary ammonium cation ester, with the formula $CH_3(O)COC_2H_4N(CH_3)_3^+$, that plays a prominent role in nervous system function.

Neurotransmission. The great importance of acetylcholine derives from its role in physiology as a neurotransmitter for cholinergic neurons. Cholinergic nerves innervate many tissues, including smooth muscle and skeletal muscle, the heart, ganglia, and glands. The effect of stimulating a cholinergic nerve, for example, the contraction of skeletal muscle or the slowing of the heartbeat, results from the release of acetylcholine from the nerve endings. Other types of nerves release other transmitters, for example, norepinephrine.

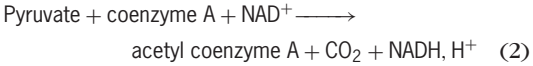
The release of a neurotransmitter was first clearly shown by Otto Loewi, who demonstrated in 1921 that the perfusate from a frog's heart that was slowed by stimulation of the vagus nerve slowed the heart of a second frog. The active substance in the perfusate was expected to be acetylcholine, which was known to slow the heart; this proved to be correct.

Biosynthesis. Acetylcholine is synthesized at axon endings from acetyl coenzyme A and choline by the enzyme choline acetyltransferase which transfers the acetyl group to choline, reaction (1). Acetylcholine



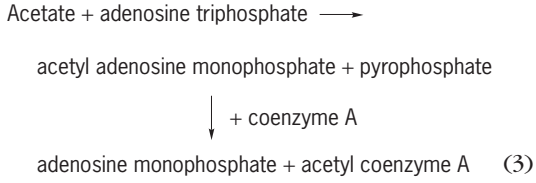
(1)

is stored at each ending in hundreds of thousands of membrane-enclosed synaptic vesicles. The axon ending is also rich in mitochondria that supply acetyl coenzyme A using pyruvate (from glucose) as the source of the acetyl group, in a reaction catalyzed by the pyruvate dehydrogenase complex, reaction (2).



(2)

Mitochondria also supply adenosine triphosphate for the acetylation of coenzyme A by acetate in a two-step reaction (3).



(3)

Molecular biology of axon. Neurons, like other cells, have a high internal concentration of potassium ions and a low internal concentration of sodium ions maintained by a sodium–potassium "pump." An axon at rest is permeable mainly to potassium, and so a very small amount of potassium leaves the cell and establishes a membrane potential of about −80 millivolts (the inside negative) that nearly blocks further leakage of potassium. This small potential produces a strong electric field in the membrane because the membrane is thin and has a low dielectric constant. The potential and high external concentration of sodium ions could drive a large inward sodium current, if the membrane were permeable to sodium. *See* ION TRANSPORT.

Axonal membranes contain voltage-gated sodium and potassium channels. These channels are composed of proteins that span the membrane and allow the transient flow of their specific ion, depending on the membrane potential. When an axon is stimulated at some point by slightly increasing the membrane potential to a less negative value, there is a transient opening of sodium channels followed by a transient opening of potassium channels. As a result, sodium ions enter the cell and drive the membrane potential positive. As the sodium current subsides, the potassium channels open and the outward flow of potassium restores the resting potential. This sequence of potential changes, the action potential, is self-propagating as a nerve impulse that moves along

the axon away from the cell body toward the numerous branched endings (toward the cell body in a dendrite). Thus, although the electric currents are perpendicular to the axon, the impulse moves parallel to the axon. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Molecular biology of synapse. When an impulse reaches an axon ending, voltage-gated calcium channels open and calcium, which is extremely low inside the cell, enters the nerve ending. The increase in calcium-ion concentration causes hundreds of synaptic vesicles to fuse with the cell membrane and expel acetylcholine into the synaptic cleft (exocytosis). The acetylcholine released at a neuromuscular junction binds reversibly to acetylcholine receptors in the muscle end-plate membrane, a postsynaptic membrane that is separated from the nerve ending by a very short distance. The receptor is a cation channel. In *Torpedo*, an electric fish, the receptor is composed of five protein subunits of four kinds, $\alpha_2\beta\gamma\delta$. The channel opens when two acetylcholine molecules are bound, one to each α subunit. A sodium current enters the cell and depolarizes the membrane, producing an end-plate potential which triggers an action potential. The action potential, as in the nerve, is self-propagating. The resulting impulse indirectly causes the muscle to contract.

Acetylcholine must be rapidly removed from a synapse in order to restore it to its resting state. This is accomplished in part by diffusion but mainly by the enzyme acetylcholinesterase which hydrolyzes acetylcholine, reaction (4). The enzyme mechanism

$$\text{Acetylcholine} + H_2O \longrightarrow \text{acetic acid} + \text{choline} \qquad (4)$$

involves the transfer of the acetyl group from choline to the hydroxyl group of a specific serine residue of the enzyme located at the active site. The reaction is completed when the acetyl enzyme, itself an ester, hydrolyzes to form acetic acid and free enzyme. As expected, acetylcholinesterase is a very fast enzyme: one enzyme molecule can hydrolyze 10,000 molecules of acetylcholine in 1 s.

Red cells contain acetylcholinesterase and blood serum contains a different cholinesterase, but the function of these enzymes is not clearly known. *See* SYNAPTIC TRANSMISSION.

Poisons, insecticides, and drugs. It is evident that any substance that efficiently inhibits acetylcholinesterase will be extremely toxic. Thus, synthetic organophosphonates such as Sarin, Tabun, and Soman are military nerve gases. These substances, which are volatile liquids, transfer a phosphonyl group to the specific serine side chain of the enzyme to produce a phosphonyl-enzyme derivative analogous to the acetyl enzyme. But unlike the acetyl enzyme, the phosphonyl enzyme hydrolyzes only extremely slowly and is trapped as an inactive derivative. Some organophosphates (similar to phosphonates) are useful: Parathion and Malathion are selectively toxic to insects and are important insecticides. Organophosphate poisoning is treated with atropine and 2-pyridine aldoxime methiodide (2-PAM): atropine antagonizes some of the effects of acetylcholine, and 2-PAM reactivates the inhibited enzyme by removing the phosphoryl group. *See* INSECTICIDE.

Similarly, certain carbamates form carbamyl derivatives of acetylcholinesterase. Carbaryl is an important carbamate insecticide. An increase in the persistence of acetylcholine at synapses is sometimes medically desirable, and carbamates are used to treat a number of medical disorders. Pyridostigmine is extensively used in the treatment of myasthenia gravis, a disease characterized by muscle weakness which is believed to be an autoimune disease, that is, antibodies are produced against the acetylcholine receptor. The proper dosage of the carbamate increases the persistence of acetylcholine at the neuromuscular junction and improves muscle strength. Carbamates, once extensively used in treating glaucoma, usually in conjunction with pilocarpine which activates the acetylcholine receptor now find limited use. Other types of acetylcholinesterase inhibitors are used in managing Alzheimer's disease. *See* MYASTHENIA GRAVIS.

Succinyldicholine, which reacts with the acetylcholine receptor of skeletal muscle, is used to produce paralysis and relaxation of muscle during surgery. This ester is not hydrolyzed by acetylcholinesterase but is hydrolyzed by serum cholinesterase which limits the duration of paralysis. Because a small fraction of the population has a variant cholinesterase, patients who will be given succinyldicholine are tested to be sure their serum will hydrolyze this drug.

There are a number of exotic, naturally occurring poisons that interfere with the nerve impulse or with synaptic transmission. Many of these substances have been very useful in studying the mechanism of these processes. Botulinus toxin prevents the release of acetylcholine, and black widow spider toxin causes the release of acetylcholine. The jungle-dart poison curare binds to the acetylcholine receptor of skeletal muscle and causes paralysis. Bungarotoxin and cobratoxin found in snake venoms have a similar effect. Muscarine from some poisonous mushrooms and arecoline from the betel nut react with the acetylcholine receptor of other tissues. Tetrodotoxin from the Japanese puffer fish, saxitoxin from the bacterium that causes the "red tide" of mussels and clams, and batrachotoxin from a South American frog bind to and block the function of sodium channels, thus preventing the nerve impulse. *See* POISON; TOXICOLOGY; TOXIN.

Electric organs. Some aquatic animals stun or kill their prey by producing a high-voltage electrical discharge. Their electric organs are composed of thousands of flat cells arranged in series. These cells are derived from skeletal muscle and have a conducting membrane on one face which also contains many nerve junctions. The nearly simultaneous excitation of these faces produces action potentials that are summed by the series arrangement to produce hundreds of volts. *See* ELECTRIC ORGAN (BIOLOGY); NERVOUS SYSTEM (VERTEBRATE).

Irwin B. Wilson

Bibliography. B. Alberts et al., *Molecular Biology of the Cell*, 1994; A. G. Gilman et al. (eds.), *The Pharmacological Basis of Therapeutics*, 9th ed., 1996; L. Stryer, *Biochemistry*, 4th ed., 1995.

Acetylene

An organic compound with the formula C_2H_2 or HC≡CH. The first member of the alkynes, acetylene is a gas with a narrow liquid range; the triple point is −81°C (−114°F). The heat of formation (ΔH°_f) is +227 kilojoules/mole, and acetylene is the most endothermic compound per carbon of any hydrocarbon. The compound is thus extremely energy-rich and can decompose with explosive force. At one time acetylene was a basic compound for much chemical manufacturing. It is highly reactive and is a versatile source of other reactive compounds.

Preparation. The availability of acetylene does not depend on petroleum liquids, since it can be prepared by hydrolysis of calcium carbide (CaC_2), obtained from lime (CaO), and charcoal or coke (C) [reaction (1)]. Other preparative methods are based

$$CaO + 3C \longrightarrow CO + CaC_2 \xrightarrow{H_2O} C_2H_2 + Ca(OH)_2 \quad (1)$$

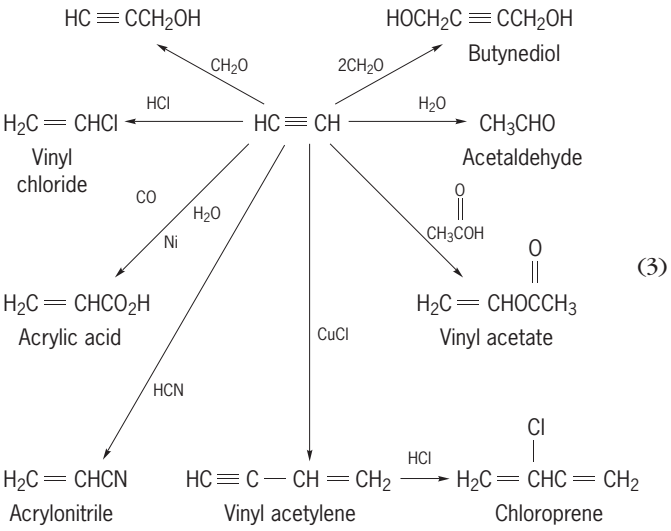
on pyrolysis of some carbon source. In modern practice, methane (CH_4) is passed through a zone heated to 1500°C (2732°F) for a few milliseconds, and acetylene is then separated from the hydrogen in the effluent gas [reaction (2)].

$$2CH_4 \longrightarrow C_2H_2 + 3H_2 \quad (2)$$

Uses. There are processes for the conversion of acetylene to a number of important compounds, including vinyl chloride, vinyl acetate, acrylonitrile, and chloroprene, all of which are monomer intermediates for polymer production [reaction (3)]. Much of this chemistry is applicable also to higher alkynes.

Although acetylene is a versatile and flexible raw material, nearly all of these processes became obsolete as the petrochemical industry developed and ethylene became a more economical feedstock for large-volume vinyl monomers and other C_2 chemicals. A major consideration is that acetylene is a very dangerous substance and handling costs are high. Production of acetylene dropped about 10% per year from a high of 1.1×10^9 lb in 1969 to 3.5×10^8 lb (1.6×10^8 kg) in 1990. *See* ETHYLENE.

The main use of acetylene is in the manufacture of compounds derived from butyne-1,4-diol. The latter is obtained by condensation of acetylene with two moles of formaldehyde and is converted to butyrolactone, tetrahydrofuran, and pyrrolidone. Two additional products, vinyl fluoride and vinyl ether, are also based on acetylene. Cyclooctatetraene, the tetramer of acetylene, is potentially available in quantity if a demand should develop. A superior grade of carbon black that is obtained by pyrolysis of acetylene above 1500°C (2732°F) has good electrical conductivity and is produced for use in dry-cell batteries. *See* PYROLYSIS.



(3)

Because of the very high heat of formation and combustion, an acetylene-oxygen mixture provides a very high temperature flame for welding and metal cutting. For this purpose acetylene is shipped as a solution in acetone, loaded under pressure in cylinders that contain a noncombustible spongy packing. *See* ALKYNE; TORCH.

James A. Moore

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 1, 4th ed., 1991; *Ullmann's Encyclopedia of Industrial Chemistry*, vol. A1, 5th ed., 1985.

Acid and base

Two interrelated classes of chemical compounds, the precise definitions of which have varied considerably with the development of chemistry. These changing definitions have led to frequent controversies, some of which are still unresolved. Acids initially were defined only by their common properties. They were substances which had a sour taste, which dissolved many metals, and which reacted with alkalies (or bases) to form salts. For a time, following the work of A. L. Lavoisier, it was believed that a common constituent of all acids was the element oxygen, but gradually it became clear that, if there were an essential element, it was hydrogen, not oxygen. In fact, the definition of an acid, formulated by J. von Liebig in 1840, as "a hydrogen-containing substance which will generate hydrogen gas on reaction with metals" proved to be satisfactory for about 50 years.

Bases initially were defined as those substances which reacted with acids to form salts (they were the "base" of the salt). The alkalies, soda and potash, were the best-known bases, but it soon became clear that there were other bases, notably ammonia and the amines.

Acids and bases are among the most important chemicals of commerce. The inorganic acids are often known as mineral acids, and among the most important are sulfuric, H_2SO_4; phosphoric, H_3PO_4; nitric, HNO_3; and hydrochloric, HCl (sometimes

called muriatic). Among the many important organic acids are acetic, CH_3COOH, and oxalic, $H_2C_2O_4$, acids, and phenol, C_6H_5OH. The important inorganic bases are ammonia, NH_3; sodium hydroxide or soda, NaOH; potassium hydroxide, KOH; calcium hydroxide or lime, $Ca(OH)_2$; and sodium carbonate, Na_2CO_3. There are also many organic bases, mostly derivatives of ammonia. Examples are pyridine, C_5H_5N, and ethylamine, $C_2H_5NH_2$.

Arrhenius-Ostwald theory. When the concept of ionization of chemical compounds in water solution became established, some considerably different definitions of acids and bases became popular. Acids were defined as substances which ionized in aqueous solution to give hydrogen ions, H^+, and bases were substances which reacted to give hydroxide ions, OH^-. These definitions are sometimes known as the Arrhenius-Ostwald theory of acids and bases and were proposed separately by S. Arrhenius and W. Ostwald. Their use makes it possible to discuss acid and base equilibria and also the strengths of individual acids and bases. The ionization of an acid in water can be written as Eq. (1). Qualitatively, an

$$HA = H^+ + A^- \tag{1}$$

acid is strong if this reaction goes extensively toward the ionic products and weak if the ionization is only slight. A quantitative treatment of this ionization of dissociation can be given by utilizing the equilibrium expression for the acid, as shown in Eq. (2), where

$$\frac{[H^+][A^-]}{[HA]} = K_{HA} \tag{2}$$

the brackets mean concentration in moles per liter and the constant K_{HA} is called the dissociation constant of the acid. This dissociation constant is a large number for a strong acid and a small number for a weak acid. For example, at 25°C (77°F) and with water as the solvent, K_{HA} has the value 1.8×10^{-5} for a typical weak acid, acetic acid (the acid of vinegar), and this value varies only slightly in dilute solutions as a function of concentration. Dissociation constants vary somewhat with temperature. They also change considerably with changes in the solvent, even to the extent that an acid which is fully ionized in water may, in some other less basic solvent, become decidedly weak. Almost all the available data on dissociation constants are for solutions in water, partly because of its ubiquitous character, and partly because it is both a good ionizing medium and a good solvent.

Acetic acid has only one ionizable hydrogen and is called monobasic. Some other acids have two or even three ionizable hydrogens and are called polybasic. An example is phosphoric acid, which ionizes in three steps, shown in Eqs. (3), (4), and (5), each with its own dissociation constant.

Ionization reaction	K_{HA}, *25°C (77°F)*	
$H_3PO_4 = H^+ + H_2PO_4^-$	1.5×10^{-3}	(3)
$H_2PO_4^- = H^+ + HPO_4^{2-}$	6.2×10^{-8}	(4)
$HPO_4^{2-} = H^+ + PO_4^{3-}$	4×10^{-13}	(5)

A similar discussion can be given for the ionization of bases in water. However, the concentrations of the species H^+ and OH^- in a water solution are not independently variable. This is because water itself is both a weak acid and a weak base, ionizing very slightly according to Eq. (6). For pure water,

$$H_2O = H^+ + OH^- \tag{6}$$

the concentration of H^+ and OH^- are equal. At ordinary temperatures, roughly 2×10^{-7}% of the water is present as ions. As a result of this ionization, the ionic concentrations are related through Eq. (7). At

$$\frac{[H^+][OH^-]}{[H_2O]} = K \tag{7}$$

25°C (77°F) and with concentrations in moles per liter, the product $[H^+][OH^-]$ is equal to 1×10^{-14}.

A major consequence of this interdependence is that measurement of the concentration of either H^+ or OH^- in a water solution permits immediate calculation of the other. This fact led S. P. L. Sørenson in 1909 to propose use of a logarithmic pH scale for the concentration of hydrogen ions in water. Although there are some difficulties in giving an exact definition of pH, it is very nearly correct for dilute solutions in water to write Eq. (8). It then turns out that pH

$$pH = -\log[H^+] \tag{8}$$

values of 0–14 cover the range from strongly acidic to strongly basic solutions. The pH of pure water at ordinary temperature is 7. *See* WATER.

For many situations, it is desirable to maintain the H^+ and OH^- concentration of a water solution at low and constant values. A useful device for this is a mixture of a weak acid and its anion (or of a weak base and its cation). Such a mixture is called a buffer. A typical example is a mixture of sodium acetate and acetic acid. From the treatment just given, it is evident that for this case Eq. (9) can be formed. In the

$$[H^+] = \frac{[CH_3COOH]}{[CH_3COO^-]} \times 1.8 \times 10^{-5} \tag{9}$$

equation $[CH_3COOH]$ and $[CH_3COO^-]$ represent the concentrations of acetic acid and acetate ion, respectively. Thus, if the concentrations of acetic acid and acetate ion are both 0.1 mole per liter, the H^+ concentration will be 1.8×10^{-5} mole per liter and OH^- will be 5.5×10^{-10} mole per liter. The pH of this solution will be about 4.7. Constant acidity is a most important aspect of blood and other life fluids; these invariably contain weak acids and bases to give the necessary buffering action.

Brönsted theory. The Arrhenius, or water, theory of acid and bases has many attractive features, but it has also presented some difficulties. A major difficulty was that solvents other than water can be used for acids and bases and thus need consideration. For many of the solvents of interest, the necessary extensions of the water theory are both obvious and plausible. For example, with liquid ammonia as the solvent, one can define NH_4^+ as the acid ion and NH_2^- as the base ion, and the former can be thought

of as a hydrogen ion combined with a molecule of the solvent. However, for a hydrogenless (aprotic) solvent, such as liquid sulfur dioxide, the extensions are less obvious. Consideration of such systems has led to some solvent-oriented theories of acids and bases to which the names of E. C. Franklin and A. F. D. Germann often are attached. The essence of these theories is to define acids and bases in terms of what they do to the solvent. Thus, one definition of an acid is that it gives rise to "a cation which is characteristic of the solvent," for example, SO^{2+} from sulfur dioxide. These theories have been useful in emphasizing the need to consider nonaqueous systems. However, they have not been widely adopted, at least partly because a powerful and wide-ranging protonic theory of acids and bases was introduced by J. N. Brönsted in 1923 and was rapidly accepted by many other scientists. Somewhat similar ideas were advanced almost simultaneously by T. M. Lowry, and the new theory is occasionally called the Brönsted-Lowry theory.

This theory gives a unique role to the hydrogen ion, and there appeared to be justification for this. One justification, of course, was the historically important role already given to hydrogen in defining acids. A rather different justification involved the unique structure of hydrogen ion. It is the only common ion which consists solely of a nucleus, the proton. As a consequence, it is only about 10^{-14} cm in diameter. All other ordinary ions have peripheral electron clouds and, as a result, are roughly 10^6 times larger than the proton. The small size of the latter makes it reasonable to postulate that protons are never found in a free state but rather always exist in combination with some base. The Brönsted theory emphasizes this by proposing that all acid-base reactions consist simply of the transfer of a proton from one base to another.

The Brönsted definitions of acids and bases are: An acid is a species which can act as a source of protons; a base is a species which can accept protons. Compared to the water theory, this represents only a slight change in the definition of an acid but a considerable extension of the term base. In addition to hydroxide ion, the bases now include a wide variety of uncharged species, such as ammonia and the amines, as well as numerous charged species, such as the anions of weak acids. In fact, every acid can generate a base by loss of a proton. Acids and bases which are related in this way are known as conjugate acid-base pairs, and the **table** lists several examples. By these definitions, such previously distinct chemical processes as ionization, hydrolysis, and neutralization become examples of the single class of proton transfer or protolytic reactions. The general reaction is expressed as Eq. (10). This equation can

$$\text{Acid}_1 + \text{base}_2 = \text{base}_1 + \text{acid}_2 \qquad (10)$$

be considered to be a combination of two conjugate acid-base pairs, and the pairs below can be used to construct a variety of typical acid-base reactions. For example, the ionization of acetic acid in water becomes Eq. (11). Water functions here as a base to

$$\underset{a_1}{CH_3COOH} + \underset{b_2}{H_2O} = \underset{b_1}{CH_3COO^-} + \underset{a_2}{H_3O^+} \qquad (11)$$

form the species H_3O^+, the oxonium ion (sometimes called the hydronium ion). However, water can also function as an acid to form the base OH^-, and this dual or amphoteric character of water is one reason why so many acid-base reactions occur in it.

As the table shows, strengths of acids and bases are not independent. A very strong Brönsted acid implies a very weak conjugate base and vice versa. A qualitative ordering of acid strength or base strength, as above, permits a rough prediction of the extent to which an acid-base reaction will go. The rule is that a strong acid and a strong base will react extensively with each other, whereas a weak acid and a weak base will react together only very slightly. More accurate calculations of acid-base equilibria can be made by using the ordinary formulation of the law of mass action. A point of some importance is that, for ionization in water, the equations reduce to the earlier Arrhenius-Ostwald type. Thus, for the ionization of acetic acid in water Eq. (11) leads to Eq. (12).

$$\frac{[H_3O^+][CH_3COO^-]}{[CH_3COOH][H_2O]} = K \qquad (12)$$

Remembering that the concentration of water will be almost constant since it is the solvent, this can be written as Eq. (13), where K_{HAc} is just the conventional dissociation constant for acetic acid in water.

$$\frac{[H_3O^+][CH_3COO^-]}{[CH_3COOH]} = K[H_2O] \equiv K_{HAc} \qquad (13)$$

One result of the Brönsted definitions is that for a given solvent, such as water, there is only a single scale of acid-base strength. Put another way, the relative strength of a set of acids will be the same for any base. Hence, the ordinary tabulation of ionization constants of acids in water permits quantitative calculation for a very large number of acid-base equilibria.

The Brönsted concepts can be applied without difficulty to other solvents which are amphoteric in the same sense as water, and data are available for many nonaqueous solvents, such as methyl alcohol, formic acid, and liquid ammonia. An important practical point is that relative acid (or base) strength turns out to be very nearly the same in these other solvents as it is in water. Brönsted acid-base reactions can also be studied in aprotic solvents (materials such as

Conjugate acid-base pairs

Strong acids	H_2SO_4	HSO_4^-	Weak bases
↑	HCl	Cl^-	↓
	H_3O^+	H_2O	
	HSO_4^-	SO_4^{2+}	
	$HF_{(aq)}$	F^-	
	CH_3COOH	CH_3COO^-	
	HN_4^+	NH_3	
	HCO_3^-	CO_3^{2+}	
Weak acids	H_2O	OH^-	Strong bases
	C_2H_5OH	$C_2H_5O^-$	

hexane or carbon tetrachloride which have virtually no tendency to gain or lose protons), but in this case, both the acid and the base must be added to the solvent.

A fact which merits consideration in any theory of acids and bases is that the speeds of large numbers of chemical reactions are greatly accelerated by acids and bases. This phenomenon is called acid-base catalysis, and a major reason for its wide prevalence is that most proton transfers are themselves exceedingly fast. Hence, reversible acid-base equilibria can usually be established very rapidly, and the resulting conjugate acids (or bases) then frequently offer favorable paths for the overall chemical reaction. The mechanisms of many of these catalyzed reactions are known. Some of them are specifically catalyzed by solvated protons (hydrogen ions); others, by hydroxide ions. Still others are catalyzed by acids or bases in the most general sense of the Brönsted definitions. The existence of this general acid and base catalysis constituted an important item in the wide acceptance of the Brönsted definitions.

Lewis theory. Studies of catalysis have, however, played a large role in the acceptance of a set of quite different definitions of acids and bases, those due to G. N. Lewis. These definitions were originally proposed at about the same time as those of Brönsted, but it was not until Lewis restated them in 1938 that they began to gain wide consideration. The Lewis definitions are: an acid is a substance which can accept an electron pair from a base; a base is a substance which can donate an electron pair. (These definitions are very similar to the terms popularized around 1927 by N. V. Sidgwick and others: electron donors, which are essentially Lewis bases, and electron acceptors, which are Lewis acids.) Bases under the Lewis definition are very similar to those defined by Brönsted, but the Lewis definition for acids is very much broader. For example, virtually every cation is an acid, as are such species as $AlCl_3$, BF_3, and SO_3. An acid-base reaction now typically becomes a combination of an acid with a base, rather than a proton transfer. Even so, many of the types of reactions which are characteristic of proton acids also will occur between Lewis acids and bases, for example, neutralization and color change of indicators as well as acid-base catalysis. Furthermore, these new definitions have been useful in suggesting new interrelations and in predicting new reactions, particularly for solid systems and for systems in nonaqueous solvents.

For several reasons, these definitions have not been universally accepted. One reason is that the terms electron donor and electron acceptor had been widely accepted and appear to serve similar predicting and classifying purposes. A more important reason is unwillingness to surrender certain advantages in precision and definiteness inherent in the narrower Brönsted definitions. It is a drawback of the Lewis definitions that the relative strengths of Lewis acids vary widely with choice of base and vice versa. For example, with the Brönsted definitions, hydroxide ion is always a stronger base than ammonia; with the Lewis definitions, hydroxide ion is a much weaker base than ammonia when reacting with silver ion but is stronger than ammonia when reacting with hydrogen ion. Another feature of the Lewis definitions is that some substances which have long been obvious examples of acid, for example, HCl and H_2SO_4, do not naturally fit the Lewis definition since they cannot plausibly accept electron pairs. Certain other substances, for example, carbon dioxide, are included by calling them secondary acids. These substances, too, tend to have electronic structures in which the ability to accept electron pairs is not obvious, but the more important distinction between them and primary acids is that their rates of neutralization by bases are measurably slow. However, in spite of these difficulties, the use of the Lewis definitions is increasing. Since there does not appear to be any simultaneous tendency to abandon the Brönsted definitions, chemistry seems to be entering a period when the term acid needs a qualifying adjective for clarity, for example, Lewis acid or proton acid.

Hard and soft acids and bases. As pointed out above, one of the drawbacks of such a broad definition as the Lewis one is that it is difficult to systematize the behavior of acids and bases toward each other. Attempts have been made to classify Lewis acids and bases into categories with respect to their mutual behavior. R. G. Pearson in 1963 proposed a simple and lucid classification scheme, based in part on earlier methods, that appears to be promising in its application to a wide variety of Lewis acid-base behavior.

Lewis bases (electron donors) are classified as soft if they have high polarizability, low electronegativity, are easily oxidized, or possess low-lying empty orbitals. They are classified as hard if they have the opposite tendencies. Some bases, spanning the range of hard to soft and listed in order of increasing softness, are H_2O, OH^-, OCH_3^-, F^-, NH_3, C_5H_5N, NO_2^-, NH_2OH, N_2H_4, C_6H_5SH, Br^-, I^-, SCN^-, SO_3^{2-}, $S_2O_3^{2-}$, and $(C_6H_5)_3P$. Acids are divided more or less distinctly into two categories, hard and soft, with a few intermediate cases. Hard acids are of low polarizability, small size, and high positive oxidation state, and do not have easily excitable outer electrons. Soft acids have several of the properties of high polarizability, low or zero positive charge, large size, and easily excited outer electrons, particularly *d* electrons in metals. Some hard acids are H^+, Li^+, Na^+, K^+, Be^{2+}, Mg^{2+}, Ca^{2+}, Sr^{2+}, Mn^{2+}, Al^{3+}, Sc^{3+}, Cr^{3+}, Co^{3+}, Fe^{3+}, As^{3+}, Ce^{3+}, Si^{4+}, Ti^{4+}, Zr^{4+}, Pu^{4+}, $BeMe_2$ (Me is the methyl group), BF_3, BCl_3, $B(OR)_3$, $Al(CH_3)_3$, AlH_3, SO_3, and CO_2. Examples of soft acids are Cu^+, Ag^+, Au^+, Tl^+, Hg^+, Cs^+, Pd^{2+}, Cd^{2+}, Pt^{2+}, Hg^{2+}, CH_3Hg^+, Tl^{3+}, BH_3, $CO(CN)_5^{2-}$, I_2, Br_2, ICN, chloranil, quinones, tetracyanoethylene, O, Cl, Br, I, N, metal atoms, and bulk metals. Intermediate acids are Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Pb^{2+}, Sn^{2+}, $B(CH_3)_3$, SO_2, NO^+, and R_3C^+.

The rule for correlating acid-base behavior is then stated as follows: hard acids prefer to associate with hard bases and soft acids with soft bases. Application, for example, to the problem of OH^- and NH_3, mentioned earlier (and recognizing that OH^- is hard

compared with NH_3), leads to reaction (14), which is unfavorable compared with reaction (15).

$$OH^-(\text{hard}) + Ag^+(\text{soft}) \rightarrow AgOH \quad (14)$$

$$NH_3(\text{soft}) + Ag^+(\text{soft}) \rightarrow Ag(NH_3)^+ \quad (15)$$

However, reaction (16) is unfavorable compared with reaction (17), which is in agreement with experiment.

$$NH_3(\text{soft}) + H^+(\text{hard}) \rightarrow NH_4^+ \quad (16)$$

$$OH^-(\text{hard}) + H^+(\text{hard}) \rightarrow H_2O \quad (17)$$

The rule is successful in correlating general acid-base behavior of a very wide variety of chemical systems, including metal-ligand interaction, charge-transfer complex formation, hydrogen bond formation, complex ion formation, carbonium ion formation, covalent bond formation, and ionic bond formation.

It is to be emphasized that the hard-and-soft acid-and-base concept is a means of classification and correlation and is not a theoretical explanation for acid-base behavior. The reasons why hard acids prefer hard bases and soft prefer soft are complex and varied. The already well-developed concepts of ionic and covalent bonds appear to be helpful, however. Hard acids and hard bases with small sizes and high charge would be held together with stable ionic bonds. Conversely, the conditions for soft acids and soft bases would be favorable for good covalent bonding. Existing theories of π-bonding also fit into the scheme.

Usanovich theory. Another comprehensive theory of acids and bases was proposed by M. Usanovich in 1939 and is sometimes known as the positive-negative theory. Acids are defined as substances which form salts with bases, give up cations, and add themselves to anions and to free electrons. Bases are similarly defined as substances which give up anions or electrons and add themselves to cations. Two examples of acid-base reactions under this scheme are reactions (18) and (19). In the first, SO_3 is an acid

$$Na_2O + SO_3 \rightarrow 2Na^+ + SO_4^{2-} \quad (18)$$

$$2Na + Cl_2 \rightarrow 2Na^+ + 2Cl^- \quad (19)$$

because it takes up an anion, O^{2-}, to form SO_4^{2-}. In the second example, Cl_2 is an acid because it takes up electrons to form Cl^-. Using conventional terminology, this second reaction is an obvious example of oxidation-reduction. The fact that oxidation-reduction can also be included in the Usanovich scheme is an illustration of the extensiveness of these definitions. So far, this theory has had little acceptance, quite possibly because the definitions are too broad to be very useful.

Generation of chemical species. The acidic or basic character of a solvent can be used to stabilize interesting chemical species, which would otherwise be difficult to obtain. For example, carbonium ions have been thought to be important intermediates in many organic reactions, but because of their fleeting existence as intermediates, their properties have been difficult to study. Most carbonium ions are very strong bases and would, for example, react, as shown by reaction (20). Accordingly, the equilibrium would

$$R_3C^+ + H_2O \rightarrow R_3COH + H^+ \quad (20)$$

lie far to the right. However, use of a very strongly acidic solvent reverses the reaction, and measurable amounts of carbonium ion are then found. Concentrated sulfuric acid has found use in this connection. The very high acidity of SbF_5 by itself, as a Lewis acid, and in mixtures with other Lewis acids, such as SO_2, or protonic acids, such as HF and FSO_3H, makes possible the study of many otherwise unstable carbonium ions. *See* SUPERACID.

Acidity functions. A very different approach to the definition of acids, or perhaps better, to the definition of acidity, is to base the definition on a particular method of measurement. (As one example, it is probably true that the most nearly exact definition of pH is in terms of the electromotive force of a particular kind of galvanic cell.) It is possible to define various acidity functions in this way, and several have been proposed. One of the earliest and also one of the most successful is the H_0 acidity function of L. P. Hammett. This defines an acidity in terms of the observed indicator ratio for a particular class of indicators, those which are uncharged in the basic form B. Suppose there is available a set of such indicators, and suppose further that the values of the dissociation constants of the acid forms BH^+ are known. Then the h_0 acidity of a solution is defined as Eq. (21),

$$h_0 = K_{BH^+} \frac{[BH^+]}{[B]} \quad (21)$$

where K_{BH^+} is the dissociation constant for the particular indicator employed, and where $[BH^+]/[B]$ is the experimentally observed ratio of concentrations of the conjugate acid and conjugate base forms of the indicator. To have a logarithmic scale (analogous to pH), the further definition is expressed in Eq. (22).

$$H_0 = -\log h_0 \quad (22)$$

The virtues of this scale are that measurements are relatively simple and can be made for concentrated solutions and for solutions in nonaqueous or mixed solvents, situations where the pH scale offers difficulties. A further point is that in dilute aqueous solutions this new acidity becomes identical to pH. Although it has been found that this measure of acidity is fairly consistent within a class of indicators used, different classes can give somewhat different measures of acidity. Hence, caution must be used in interpretation of acidity measured by this technique.

For a discussion of measurement of acidity *see* BASE (CHEMISTRY); BUFFERS (CHEMISTRY); HYDROGEN ION; IONIC EQUILIBRIUM; OXIDATION-REDUCTION; SOLUTION; SOLVENT. Franklin A. Long; Richard H. Boyd

Bibliography. J. F. Coetzee and C. D. Ritchie, *Solute-Solvent Interactions*, vol. 1, 1969, vol. 2, 1976; C. W.

Hand, *Acid-Base Chemistry*, 1986; H. F. Holtzclaw, Jr., W. R. Robinson, and J. D. Odom, *General Chemistry*, 9th ed., 1991.

Acid-base indicator

A substance that reveals, through characteristic color changes, the degree of acidity or basicity of solutions. Indicators are weak organic acids or bases which exist in more than one structural form (tautomers) of which at least one form is colored. Intense color is desirable so that very little indicator is needed; the indicator itself will thus not affect the acidity of the solution.

The equilibrium reaction of an indicator may be regarded typically by giving it the formula HIn. It dissociates into H^+ and In^- ions and is in equilibrium with a tautomer InH which is either a nonelectrolyte or at most ionizes very slightly. In the overall equilibrium shown as reaction (1), the simplifying assumption that the indicator exists only in forms (**1**) and (**2**) leads to no difficulty. The addition of acid will completely convert the indicator to form (**1**), which is therefore called the acidic form of the indicator although it is functioning as a base. A hydroxide base converts the indicator to form (**2**) with the formation of water; this is called the alkaline form. For the equilibrium between (**1**) and (**2**) the equilibrium constant is given by Eq. (2). In a manner similar to the pH designation of acidity, that is, $pH = -\log[H^+]$, the K_{In} is converted to pK_{In} with the result shown in Eq. (3). It is seen that the pK of an indicator has a numerical value approximately equal to that of a specific pH level.

$$\underset{(\mathbf{1})}{InH} \underset{\text{ioniz-ation}}{\rightleftharpoons} \underset{(\mathbf{2})}{HIn} \underset{\text{dissoci-ation}}{\rightleftharpoons} H^+ + In^- \qquad (1)$$

$$K_{In} = \frac{[H^+][In^-]}{[InH]} \qquad (2)$$

$$pK_{In} = pH - \log \frac{[In^-]}{[InH]} \qquad (3)$$

Use of indicators. Acid-base indicators are commonly employed to mark the end point of an acid-base titration or to measure the existing pH of a solution. For titration the indicator must be so chosen that its pK is approximately equal to the pH of the system at its equivalence point. For pH measurement, the indicator is added to the solution and also to several buffers. The pH of the solution is equal to the pH of that buffer which gives a color match. Care must be used to compare colors only within the indicator range. A color comparator may also be used, employing standard color filters instead of buffer solutions.

Indicator range. This is the pH interval of color change of the indicator. In this range there is competition between indicator and added base for the available protons; the color change, for example, yellow to red, is gradual rather than instantaneous. Observers may, therefore, differ in selecting the precise point of change. If one assumes arbitrarily that the indicator is in one color form when at least 90% of it is in that form, there will be uncertain color

Common acid-base indicators

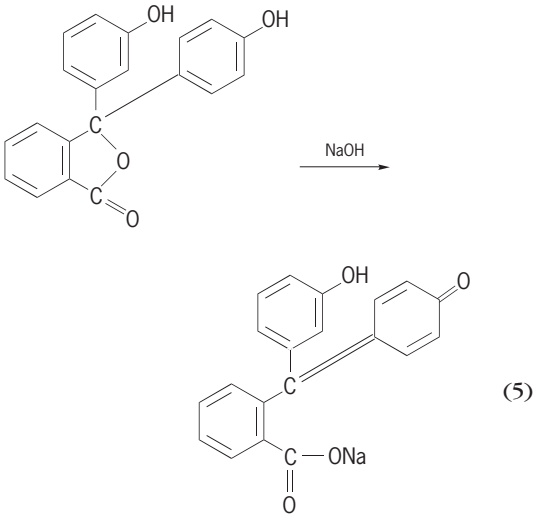
Common name	pH range	Color change (acid to base)	p*K*	Chemical name	Structure	Solution
Methyl violet	0–2, 5–6	Yellow to blue violet to violet		Pentamethylbenzyl-pararosaniline hydrochloride	Base	0.25% in water
Metacresol purple	1.2–2.8, 7.3–9.0	Red to yellow to purple	1.5	*m*-Cresolsulfonphthalein	Acid	0.73% in *N*/50 NaOH, dilute to 0.04%
Thymol blue	1.2–2.8, 8.0–9.6	Red to yellow to blue	1.7	Thymolsulfonphthalein	Acid	0.93% in *N*/50 NaOH, dilute to 0.04%
Tropeoline 00 (Orange IV)	1.4–3.0	Red to yellow		Sodium *p*-diphenyl-aminoazobenzene-sulfonate	Base	0.1% in water
Bromphenol blue	3.0–4.6	Yellow to blue	4.1	Tetrabromophenol-sulfonphthalein	Acid	1.39% in *N*/50 NaOH, dilute to 0.04%
Methyl orange	2.8–4.0	Orange to yellow	3.4	Sodium *p*-dimethyl-aminoazobenzene-sulfonate	Base	0.1% in water
Bromcresol green	3.8–5.4	Yellow to blue	4.9	Tetrabromo-*m*-cresolsulfonphthalein	Acid	0.1% in 20% alcohol
Methyl red	4.2–6.3	Red to yellow	5.0	Dimethylaminoazobenzene-*o*-carboxylic acid	Base	0.57% in *N*/50 NaOH, dilute to 0.04%
Chlorphenol red	5.0–6.8	Yellow to red	6.2	Dichlorophenolsulfonphthalein	Acid	0.85% in *N*/50 NaOH, dilute to 0.04%
Bromcresol purple	5.2–6.8	Yellow to purple	6.4	Dibromo-*o*-cresolsulfonphthalein	Acid	1.08% in *N*/50 NaOH dilute to 0.04%
Bromthymol blue	6.0–7.6	Yellow to blue	7.3	Dibromothymolsulfonphthalein	Acid	1.25% in *N*/50 NaOH, dilute to 0.04%
Phenol red	6.8–8.4	Yellow to red	8.0	Phenolsulfonphthalein	Acid	0.71% in *N*/50 NaOH, dilute to 0.04%
Cresol red	2.0–3.0, 7.2–8.8	Orange to amber to red	8.3	*o*-Cresolsulfonphthalein	Acid	0.76% in *N*/50 NaOH, dilute to 0.04%
Orthocresolphthalein	8.2–9.8	Colorless to red		—	Acid	0.04% in alcohol
Phenolphthalein	8.4–10.0	Colorless to pink	9.7	—	Acid	1% in 50% alcohol
Thymolphthalein	10.0–11.0	Colorless to red	9.9	—	Acid	0.1% in alcohol
Alizarin yellow GG	10.0–12.0	Yellow to lilac		Sodium *p*-nitrobenzeneazosalicylate	Acid	0.1% in warm water
Malachite green	11.4–13.0	Green to colorless		*p,p′*-Benzylidene-bis-*N,N*-dimethylaniline	Base	0.1% in water

in the range of 90–10% InH (that is, 10–90% In^-). When these arbitrary limits are substituted into the p*K* equation, the interval between definite colors is shown by Eqs. (4). Thus the pH uncertainty of the

$$\mathrm{pH} = \mathrm{p}K + \log\frac{10}{90} \quad \text{to} \quad \mathrm{pH} = \mathrm{p}K + \log\frac{90}{10} \qquad (4)$$

indicator is from pK + 1 to pK − 1 (approximately), and pK ± 1 is called the range of the indicator. The experimentally observed ranges may differ from this prediction somewhat because of variations in color intensity.

Examples. The **table** lists many of the common indicators, their chemical names, pK values, ranges of pH, and directions for making solutions. Many of the weak-acid indicators are first dissolved in N/50 NaOH to the concentration shown, then further diluted with water. The weak-base indicators show some temperature variation of pH range, following approximately the temperature change of the ionization constant of water. Weak-acid indicators are more stable toward temperature change. The colors are formed by the usual chromophoric groups, for example, quinoid and azo-. One of the most common indicators is phenolphthalein, obtained by condensing phthalic anhydride with phenol. The acid form is colorless. It is converted by OH^- ion to the red quinoid form, as shown by reaction (5).



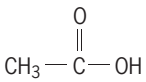
(5)

Other indiators such as methyl orange and methyl red are sodium salts of azobenzene derivatives containing sulfonic and carboxylic groups respectively. *See* ACID AND BASE; HYDROGEN ION; TITRATION.

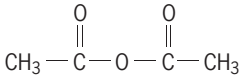
Allen L. Hanson

Acid anhydride

One of an important class of reactive organic compounds derived from acids via formal intermolecular dehydration; thus, acetic acid (**1**), on loss of water, forms acetic anhydride (**2**).



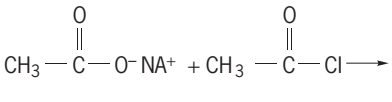
(**1**)



(**2**)

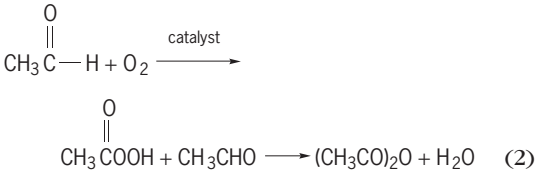
Anhydrides of straight-chain acids containing from 2 to 12 carbon atoms are liquids with boiling points higher than those of the parent acids. They are relatively insoluble in cold water and are soluble in alcohol, ether, and other common organic solvents. The lower members are pungent, corrosive, and weakly lacrimatory. Anhydrides from acids with more than 12 carbon atoms and cyclic anhydrides from dicarboxylic acids are crystalline solids.

Preparation. Because the direct intermolecular removal of water from organic acids is not practicable, anhydrides must be prepared by means of indirect processes. A general method involves interaction of an acid salt with an acid chloride, reaction (1).



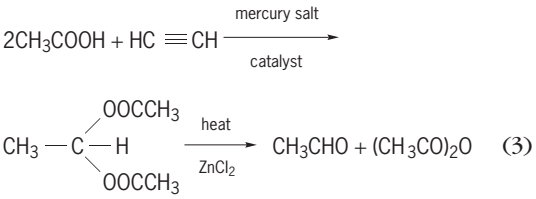
$$\mathrm{NaCl} + (\mathrm{CH_3CO})_2\mathrm{O} \qquad (1)$$

Acetic anhydride, the most important aliphatic anhydride, is manufactured by air oxidation of acetaldehyde, using as catalysts the acetates of copper and cobalt, shown in reaction (2); peracetic acid apparently is an intermediate. The anhydride is separated from the by-product water by vacuum distillation.



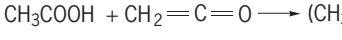
$$(\mathrm{CH_3CO})_2\mathrm{O} + \mathrm{H_2O} \qquad (2)$$

Another important process utilizes the thermal decomposition of ethylidene acetate (made from acetylene and acetic acid), reaction (3).



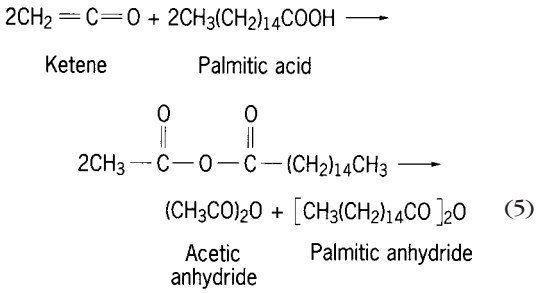
$$\mathrm{CH_3CHO} + (\mathrm{CH_3CO})_2\mathrm{O} \qquad (3)$$

Acetic anhydride has been made by the reaction of acetic acid with ketene, reaction (4).



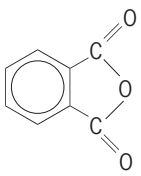
$$(\mathrm{CH_3CO})_2\mathrm{O} \qquad (4)$$

Mixed anhydrides composed of two different radicals are unstable, and disproportionate to give the two simple anhydrides. Direct use is made of this in the preparation of high-molecular-weight anhydrides, as seen in reaction (5). The two simple anhydrides are easily separable by distillation in a vacuum.

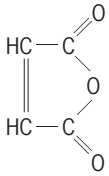


(5)

Cyclic anhydrides are obtained by warming succinic or glutaric acids, either alone, with acetic anhydride, or with acetyl chloride. Under these conditions, adipic acid first forms linear, polymeric anhydride mixtures, from which the monomer is obtained by slow, high-vacuum distillation. Cyclic anhydrides are also formed by simple heat treatment of cis-unsaturated dicarboxylic acids, for example, maleic and glutaconic acids; and of aromatic 1,2-dicarboxylic acids, for example, phthalic acid. Commercially, however, both phthalic (**3**) and maleic (**4**)



(**3**)



(**4**)

anhydrides are primary products of manufacture, being formed by vapor-phase, catalytic (vanadium pentoxide), air oxidation of naphthalene and benzene, respectively; at the reaction temperature, the anhydrides form directly.

Uses. Anhydrides are used in the preparation of esters. Ethyl acetate and butyl acetate (from butyl alcohol and acetic anhydride) are excellent solvents for cellulose nitrate lacquers. Acetates of high-molecular-weight alcohols are used as plasticizers for plastics and resins. Cellulose and acetic anhydride give cellulose acetate, used in acetate rayon and photographic film. The reaction of anhydrides with sodium peroxide forms peroxides (acetyl peroxide is violently explosive), used as catalysts for polymerization reactions and for addition of alkyl halides to alkenes. In Friedel-Crafts reactions, anhydrides react with aromatic compounds, forming ketones such as acetophenone.

Maleic anhydride reacts with many dienes to give hydroaromatics of various complexities (Diels-Alder reaction). Maleic anhydride is used commercially in the manufacture of alkyd resins from polyhydric alcohols. Soil conditioners are produced by basic hydrolysis of the copolymer of maleic anhydride with vinyl acetate.

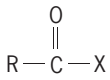
Phthalic anhydride and alcohols form esters (phthalates) used as plasticizers for plastics and resins. Condensed with phenols and sulfuric acid, phthalic anhydride yields phthaleins, such as phenolphthalein; with *m*-dihydroxybenzenes under the same conditions, xanthene dyes form, for example, fluorescein. Phthalic anhydride is used in manufacturing glyptal resins (from the anhydride and glycerol) and in manufacturing anthraquinone. Heating phthalic anhydride with ammonia gives phthalimide, used in Gabriel's synthesis of primary amines, amino acids, and anthranilic acid (*o*-aminobenzoic acid). With alkaline hydrogen peroxide, phthalic anhydride yields monoperoxyphthalic acid, used along with benzoyl peroxide as polymerization catalysts, and as bleaching agents for oils, fats, and other edibles.

Anhydrides react with water to form the parent acid, with alcohols to give esters, and with ammonia to yield amides; and with primary or secondary amines, they furnish *N*-substituted and *N*,*N*-disubstituted amides, respectively. *See* ACID HALIDE; ACYLATION; CARBOXYLIC ACID; DIELS-ALDER REACTION; ESTER; FRIEDEL-CRAFTS REACTION. Paul E. Fanta

Bibliography. F. Bettelheim and J. March, *Introduction to Organic and Biochemistry*, 1997; R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998.

Acid halide

One of a large group of organic substances possessing the halocarbonyl group



in which X stands for fluorine, chlorine, bromine, or iodine. The terms acyl and aroyl halides refer to aliphatic or aromatic derivatives, respectively.

The great inherent reactivity of acid halides precludes their free existence in nature; all are made by synthetic processes. In general, acid halides have low melting and boiling points and show little tendency toward molecular association. With the exception of the formyl halides (which do not exist), the lower members are pungent, corrosive, lacrimatory liquids that fume in moist air. The higher members are low-melting solids.

Preparation. Acid chlorides are prepared by replacement of carboxylic hydroxyl of organic acids by treatment with phosphorus trichloride, phosphorus pentachloride, or thionyl chloride [reactions (1)–(3)].

$$3RCOOH + PCl_3 \rightarrow 3RCOCl + H_3PO_3 \quad (1)$$

$$RCOOH + PCl_5 \rightarrow RCOCl + POCl_3 + HCl \quad (2)$$

$$RCOOH + SOCl_2 \rightarrow RCOCl + SO_2 + HCl \quad (3)$$

Although acid bromides may be prepared by the above methods (especially by use of PBr_3), acid iodides are best prepared from the acid chloride treatment with either CaI_2 or HI, and acid fluorides from the acid chloride by interaction with HF or antimony fluoride.

Reactions and uses. The reactivity of acid halides centers upon the halocarbonyl group, resulting in substitution ofthe halogen by appropriate

structures. Thus, as shown by reactions (4), with

$$\mathrm{RCOCl} + \begin{cases} \mathrm{HOH} \rightarrow \mathrm{HCl} + \mathrm{RCOOH} \\ \mathrm{C_2H_5OH} \rightarrow \mathrm{HCl} + \mathrm{RCOOC_2H_5} \\ \mathrm{2NH_3} \rightarrow \mathrm{NH_4Cl} + \mathrm{RCONH_2} \\ \mathrm{2HNR_2} \rightarrow \mathrm{NH_2R_2Cl} + \mathrm{RCONR_2} \end{cases} \quad (4)$$

substances containing active hydrogen atoms (for example, water, primary and secondary alcohols, ammonia, and primary and secondary amines), hydrogen chloride is formed together with acids, esters, amides, and *N*-substituted amides, respectively.

The industrially prepared acetyl and benzoyl chlorides are much used in reactions of the above type, particularly in the acetylating or benzoylating of amines and amino acids, and alcohols, especially polyalcohols such as glycerol, the sugars, and cellulose, to form amides, esters, and polyesters, respectively. In reactions (4) the by-product hydrogen chloride must be neutralized. With aliphatic acid halides, this must be done by using an excess of the amine (as shown); aromatic acid chlorides, being insoluble in water, can be used in the presence of aqueous sodium hydroxide (Schotten-Baumann reaction). An alternative technique uses the tertiary amine, pyridine, both as solvent and as neutralizing agent.

Interaction of acid chlorides with sodium salts of organic acids furnishes a general method for the preparation of acid anhydrides, as shown by reaction (5). Analogously, aromatic acyl peroxides, used

$$\mathrm{RCOCl} + \mathrm{RCOONa} \rightarrow \mathrm{NaCl} + (\mathrm{RCO})_2\mathrm{O} \quad (5)$$

as bleaching agents for flour, fats, and oils, and as polymerization catalysts, may be prepared from sodium peroxide and the acid halide [reaction (6)].

$$\mathrm{2C_6H_5COCl} + \mathrm{Na_2O_2} \rightarrow \underset{\text{Benzoyl peroxide}}{\mathrm{2NaCl} + \mathrm{C_6H_5COOOOCC_6H_5}} \quad (6)$$

See PEROXIDE.

Acid halides are used to prepare ketones, either by reaction with alkyl or aryl cadmium reagents [reaction (7)] or by the aluminum chloride catalyzed Friedel-Crafts reaction [reaction (8)].

$$\mathrm{RCOCl} + \mathrm{RCH_2CdCl} \rightarrow \mathrm{RCOCH_2R} + \mathrm{CdCl_2} \quad (7)$$

$$\mathrm{RCOCl} + \mathrm{C_6H_6} \xrightarrow{\mathrm{AlCl_3}} \mathrm{RCOC_6H_5} + \mathrm{HCl} \quad (8)$$

Reduction of acid halides is easily effected. In the Rosenmund method, used mainly for aromatic acid halides, hydrogen and a poisoned palladium catalyst are employed [reaction (9)]. The product here

$$\mathrm{RCOCl} + \mathrm{H_2} \xrightarrow[\text{poisoned}]{\mathrm{Pd}} \mathrm{HCl} + \mathrm{RCHO} \quad (9)$$

is the aldehyde. For reduction to the alcohol stage, the vigorous reagent lithium aluminum hydride reduces both aliphatic and aromatic acid halides [reaction (10)].

$$\mathrm{4RCOCl} + \mathrm{LiAlH_4} \rightarrow \mathrm{LiAlCl_4} + (\mathrm{RCH_2O})_4\mathrm{LiAl} \xrightarrow{\mathrm{4HCl}} \mathrm{4RCH_2OH} + \mathrm{LiAlCl_4} \quad (10)$$

Direct substitution of halogen (chlorine fastest, bromine more slowly) into acid aliphatic halides is relatively easy, compared with substitution in the parent acid, and takes place on the carbon next to the carbonyl group. The product is an α-halo acid halide, RCHXCOX. These compounds interact with carboxylic acids, via an equilibrium, to form an α-halo acid and an acid halide [reaction (11)].

$$\mathrm{RCHXCOX} + \mathrm{RCH_2COOH} \rightleftharpoons \mathrm{RCHXCOOH} + \mathrm{RCH_2COX} \quad (11)$$

Thus, if a small amount of acyl halide is added to a large amount of carboxylic acid, the latter can be chlorinated or brominated to completion by treatment with either chlorine or bromine (Hell-Volhard-Zelinski reaction). *See* ACYLATION; CARBOXYLIC ACID; FRIEDEL-CRAFTS REACTION.

Paul E. Fanta

Bibliography. N. L. Allinger et al., *Organic Chemistry*, 2d ed., 1976; W. H. Brown, *Introduction to Organic and Biochemistry*, 4th ed., 1987.

Acid rain

Precipitation that incorporates anthropogenic acids and acidic materials. The deposition of acidic materials on the Earth's surface occurs in both wet and dry forms as rain, snow, fog, dry particles, and gases. Although 30% or more of the total deposition may be dry, very little information that is specific to this dry form is available. In contrast, there is a large and expanding body of information related to the wet form: acid rain or acid precipitation. Acid precipitation, strictly defined, contains a greater concentration of hydrogen (H^+) than of hydroxyl (OH^-) ions, resulting in a solution pH less than 7. Under this definition, nearly all precipitation is acidic. The phenomenon of acid deposition, however, is generally regarded as resulting from human activity. *See* PRECIPITATION (METEOROLOGY).

Sources. Theoretically, the natural acidity of precipitation corresponds to a pH of 5.6, which represents the pH of pure water in equilibrium with atmospheric concentrations of carbon dioxide. Atmospheric moisture, however, is not pure, and its interaction with ammonia, oxides of nitrogen and sulfur, and windblown dust results in a pH between 4.9 and 6.5 for most "natural" precipitation. The distribution and magnitude of precipitation pH in the United States (**Fig. 1**) suggest the impact of anthropogenic rather than natural causes. The areas of highest precipitation acidity (lowest pH) correspond to areas within and downwind of heavy industrialization and urbanization where emissions of sulfur and nitrogen oxides are high. It is with these emissions that the most acidic precipitation is thought to originate. *See* ACID RAIN.

Atmospheric processes. The transport of acidic substances and their precursors, chemical reactions, and deposition are controlled by atmospheric processes. In general, it is convenient to distinguish between physical and chemical processes, but it must

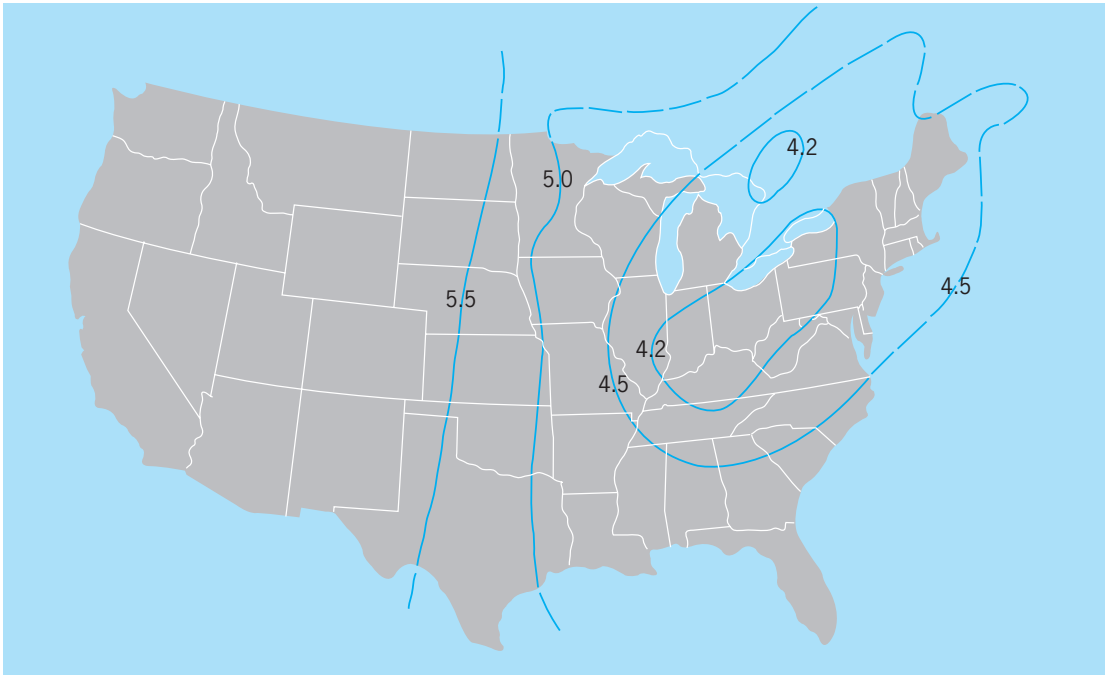


Fig. 1. Distribution of rainfall pH in the eastern United States.

be realized that both types may be operating simultaneously in complicated and interdependent ways. The physical processes of transport by atmospheric winds and the formation of clouds and precipitation strongly influence the patterns and rates of acidic deposition, while chemical reactions govern the forms of the compounds deposited.

In midlatitude continental regions, such as eastern North America, most precipitation arises from cyclonic storms. As shown by the schematic representation of a typical storm (**Fig. 2**), rain tends to form along the surface cold and warm fronts that define the so-called warm sector of the storm. This characteristic structure, arising from prevailing north-south temperature gradients, simultaneously sets the stage for chemical transformations of pollutants and for incorporation of the compounds into precipitation. The motion of the cold front toward the southeast forces the moist warm-sector air to stream along the cold front toward the low-pressure region (L), as shown by the broad arrow in Fig. 2. At the same time, the air is gradually lifted out of the surface layer and into the colder, upper parts of the storm, which allows the supply of water vapor to condense out, forming the cloud and precipitation. *See* CYCLONE; FRONT; STORM.

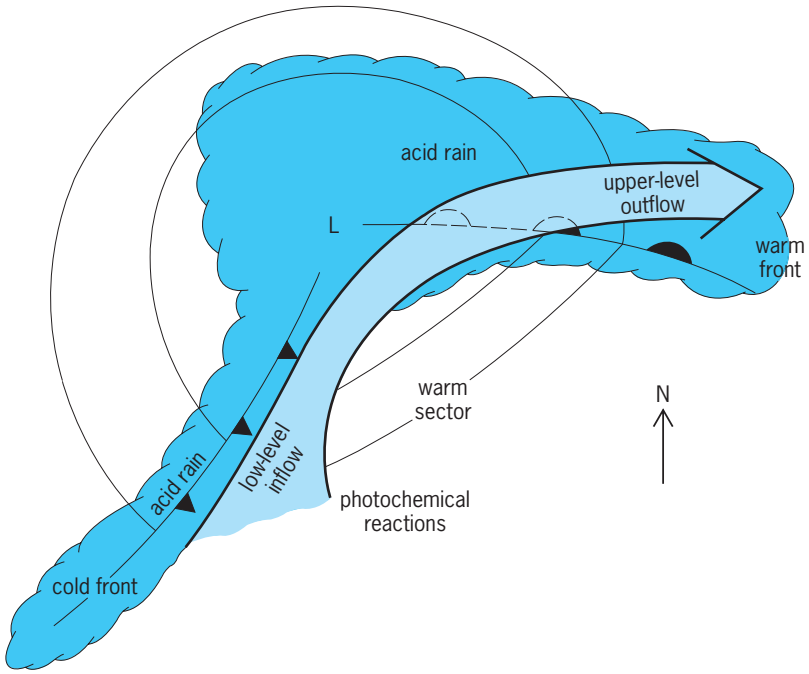


Fig. 2. Schematic view of a typical cyclonic storm in eastern North America with a low-pressure region (L).

There are a number of chemical pathways by which the primary pollutants, sulfur dioxide (SO_2) from industry, nitric oxide (NO) from both industry and automobiles, and reactive hydrocarbons mostly from trees, are transformed into acid-producing compounds. Some of these pathways exist solely in the gas phase, while others involve the aqueous phase afforded by the cloud and precipitation. As a general rule, the volatile primary pollutants must first be oxidized to more stable compounds before they are efficiently removed from the atmosphere. Ironically, the most effective oxidizing agents, hydrogen peroxide (H_2O_2) and ozone (O_3), arise from photochemical reactions involving the primary pollutants themselves. *See* AIR POLLUTION.

All of the ingredients needed to form the strong mineral acids of sulfur and nitrogen [sulfuric acid (H_2SO_4) and nitric acid (HNO_3)], which constitute most of the acids found in rain, exist in the warm sector. Especially in summertime, stagnant air conditions trap the pollutants under clear skies for several

days, permitting photochemical reactions to initiate other gas-phase reactions. Nitrogen oxides (NO_x) in combination with ultraviolet light from the Sun, reactive hydrocarbons, atmospheric oxygen, and water vapor simultaneously give rise to HNO_3 vapor, odd hydrogen radicals (particularly OH), and the strong oxidants H_2O_2 and O_3. Meanwhile, slow oxidation of sulfur dioxide (SO_2), initiated by reaction with the OH radical, leads to the gradual buildup of sulfate (SO^{-2}_4) aerosol, and hazy skies result from the highly condensable nature of H_2SO_4. While some dry deposition does occur under these conditions, the warm sector of midlatitude cyclones is a region conducive to the gradual accumulation of primary pollutants, oxidants, and acid-forming compounds in the lower atmosphere. *See* SMOG.

As the cold front approaches any given location in the warm sector, the airborne acids and their precursors are drawn into the circulations of the cyclone, particularly into the stream of moist air associated with the fronts. By this large-scale meteorological process, the pollutants become integrally associated with the frontal cloud systems, and attach to individual cloud particles by a variety of microphysical processes. Since most atmospheric sulfate particles are very soluble, they act as good centers (nuclei) for cloud drop formation. Thus, the process of cloud formation is itself a mechanism for scavenging particulate pollutants from the air. As the cloud drops grow by condensation, soluble gases will be absorbed by the cloud water. It is during this phase of the overall process that additional SO_2 will be oxidized to H_2SO_4 by the previously formed strong oxidants, particularly the H_2O_2.

The actual removal of pollutants from the atmosphere by wet deposition requires the formation of precipitation within the clouds. Without cloud elements greater than about 100 micrometers in diameter, the pollutant mass remains in the air, largely in association with the relatively small cloud drops, which have negligible rates of descent. Since most precipitation in midlatitude storms is initiated by the formation of ice particles in the cold, upper reaches of the clouds, pollutant fractionation between water phases inhibits the transfer of cloud water acidity to large particles, which do precipitate readily. Partly because of such microphysical phenomena and partly because of nonuniform distributions of pollutants within the clouds, the acidity of precipitation tends to be substantially less than that of the cloud water that remains aloft. The acidity of the precipitation appears to be acquired largely through collisions of the melted ice particles with the relatively concentrated cloud drops in the lower portions of the storm clouds. The rate of wet deposition is thus governed by a complicated set of meteorological and microscale processes working in harmony to effect a general cleansing of the atmosphere. *See* ATMOSPHERIC CHEMISTRY; CLOUD PHYSICS.

Terrestrial and aquatic effects. The effect of acid deposition on a particular ecosystem depends largely on its acid sensitivity, its acid neutralization capability, the concentration and composition of acid reaction products, and the amount of acid added to the system. As an example, the major factors influencing the impact of acidic deposition on lakes and streams are (1) the amount of acid deposited; (2) the pathway and travel time from the point of deposition to the lake or stream; (3) the buffering characteristics of the soil through which the acidic solution moves; (4) the nature and amount of acid reaction products in soil drainage and from sediments; and (5) the buffering capacity of the lake or stream.

In many ecosystems, except for foliar effects, the impacts of acid precipitation on aquatic and terrestrial ecosystems overlap. This is because soils are the key intermediate. They provide the root environment for terrestrial vegetation, and also control the water quality of runoff and soil drainage, which supplies most of the water to the aquatic system. A number of acid-consuming reactions occur in soil, which lessen the impact of acid additions on both the soil and soil drainage waters. In soils, indigenous or agriculturally amended carbonates (CO^{2-}_3) react with acid precipitation to raise the pH of soil drainage waters, while maintaining soil pH. Also, soils exhibit a cation exchange capacity that can serve to neutralize acid precipitation. At neutral soil pH (6.0–8.0), most cations on the exchange are calcium (Ca) and magnesium (Mg). When acids are added, H^+ from solution exchanges with the adsorbed Ca^{2+} and Mg^{2+}. Although this reduces the acidity of soil drainage waters, the soil acidity is increased. Eventually, when the neutralizing carbonates and exchangeable calcium and magnesium supplies are exhausted, soil minerals react with the acid. At a soil pH of about 5.2 or less, substantial aluminum (Al) can be solubilized from the dissolution of clay minerals and coatings on soil particles. *See* SOIL CHEMISTRY.

Acid deposition directly into lakes and streams causes chemical reactions analogous to those in the soil system. However, instead of soil materials, the carbonate-bicarbonate (HCO^-_3) system buffers the solution. As with soils, waters of low pH and buffering capacities, in this case bicarbonate concentrations, are most sensitive. At solution pH of 4.5, bicarbonate is essentially depleted, and subsequent acid additions that reduce pH also mobilize metals from suspended solids and the lake or stream bed. Generally, lakes and streams that drain acid-susceptible soils are most susceptible to direct acid deposition as well. Shallow soils with low pH, low cation-exchange capacity, and high permeability not only are least able to neutralize acid flows but also are poor sources of the bicarbonate waters needed to buffer the aquatic system. *See* BUFFERS (CHEMISTRY).

While the study of acid precipitation effects on terrestrial ecosystems is relatively new, soil acidification is a naturally occurring process in humid climates and has long been the subject of research, whose findings suggest acid precipitation effects. The generally accepted impact of soil acidification on the productivity of terrestrial plants is summarized as follows. As soil becomes more acidic, the basic cations (Ca, Mg) on the soil exchange are replaced by

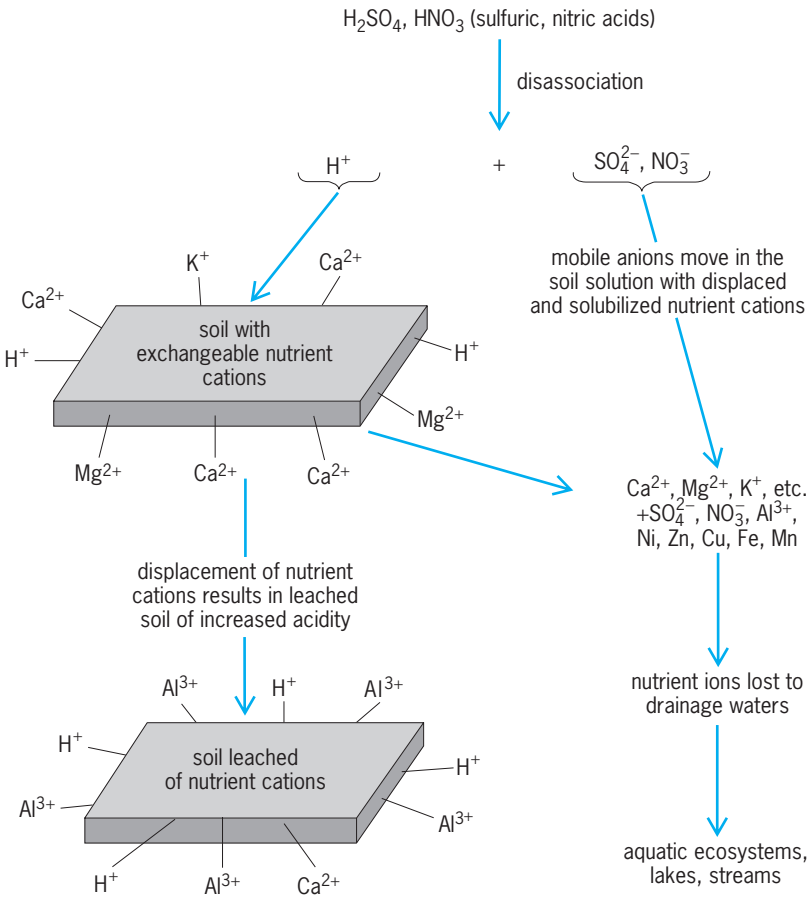


Fig. 3. Soil acidification and loss of soil nutrients following acid inputs.

hydrogen ions or solubilized metals. The basic cations, now in solution, can be leached through the soil (**Fig. 3**). As time progresses, the soil becomes less fertile and more acidic. Resultant decreases in soil pH cause reduced, less-active populations of soil microorganisms, which in turn slow decomposition of plant residues and cycling of essential plant nutrients. The plant availability of phosphorus, now composing mostly aluminum and iron (Fe) phosphates, decreases, while the plant availability of trace metals [aluminum, copper (Cu), iron, zinc (Zn), boron (B), manganese (Mn)] increases, sometimes to phytotoxic levels.

Acid precipitation may injure trees directly or indirectly through the soil. Foliar effects have been studied extensively, and it is generally accepted that visible damage occurs only after prolonged exposure to precipitation of pH 3 or less (for example, acid fog or clouds). Measurable effects on forest ecosystems will then more likely result indirectly through soil processes than directly through exposure of the forest canopy. The characteristics that make a soil sensitive to acidification are extremely variable from soil to soil. Accordingly, forest decline will not occur regionally across a deposition gradient, but it will occur in localized areas with sensitive soils. The occurrence of both acid-sensitive and acid-tolerant soils in areas receiving high levels of acid deposition and in areas receiving low levels of acid deposition makes cause-and-effect relationships between acid deposition and forest decline difficult to establish. For example, 10 years of monitoring indicated that hardwood forests in Ontario were on average healthy and stable. Still, tree condition was poorer in areas with both moderately high pollution deposition and acid-sensitive soils. The clearest cause-and-effect relationship exists for red spruce and white birch exposed to acid clouds and fog. Areas of declining red spruce and white birch coincide with areas of acid clouds and fog, and the extent of decline is related to the duration of exposure.

Many important declines in the condition of forest trees have been reported in Europe and North America during the period of increasing precipitation acidity. These cases include injury to white pine in the eastern United States, red spruce in the Appalachian Mountains of eastern North America, and many economically important species in central Europe. Since forest trees are continuously stressed by competition for light, water, and nutrients; by disease organisms; by extremes in climate; and by atmospheric pollutants, establishing acid deposition as the cause of these declines is made more difficult. Each of these sources of stress, singly or in combination, produces similar injury. However, a large body of information indicates that accelerated soil acidification resulting from acid deposition is an important predisposing stress that in combination with other stresses has resulted in increased decline and mortality of sensitive tree species and widespread reduction in tree growth. *See* FOREST ECOSYSTEM; TERRESTRIAL ECOSYSTEM.

Aquatic biology effects. Aquatic ecosystems are sustained by intricate exchanges of energy and material among organisms at all levels. Acidic deposition impacts aquatic ecosystems by harming individual organisms and by disrupting flows of energy and materials through the ecosystem. The effect of acid deposition is commonly assessed by studying aquatic invertebrates and fish. Aquatic invertebrates live in the sediments of lakes and streams and are vitally important to the cycling of energy and material in aquatic ecosystems. These small organisms, rarely larger than 20 mm (1 in.), break down large particulate organic matter for further degradation by microorganisms, and they are an important food source for fish, aquatic birds, and predatory invertebrates.

Individual aquatic organisms are harmed through direct physiological effects caused by low pH, high metal concentrations, and the interaction between pH and metals. Low pH causes lesions to form on gills and erodes gill tissue. Gill damage compromises respiration, excretion, and liver function, which in turn leaves organisms susceptible to further health problems. As pH decreases, the aluminum concentration of streams and lakes increases. Aluminum enhances the toxic effects of low pH. Together these phenomena result in severe acidosis and cause gills to secrete excessive amounts of mucus, further

interfering with respiration by reducing oxygen diffusion across gills. Additionally, gill membranes become more permeable, reducing sodium uptake and increasing sodium loss. The most likely physiological cause of fish death at pH 4–5 is the inability to maintain a proper salt balance (ion regulation).

Metals such as aluminum liberated from acidified soils may accumulate externally on the gill surfaces of invertebrates, and this accumulation also impairs respiration. Iron has been shown to accumulate on the gut membranes of invertebrates, inhibiting food absorption. It should be noted that the formation of metallo-organic complexes reduces the toxicity of metals in low-pH waters. Acid bogs and marshes support healthy populations of aquatic organisms because soluble metals and humic compounds form such complexes.

Currently, there are concerns that acid deposition is causing the loss of fish species, through the physiological damage discussed above and by reproductive impairment. In order to find water with a suitable pH, fish may be forced to spawn in areas that are less than optimal for egg development, with a consequent reduction in the number of, or total loss of, offspring. Female fish need high serum calcium levels for normal ovary formation and egg laying. Female fish in acidified waters maintain lower than normal serum calcium levels, and consequently egg laying may be inhibited, again resulting in fewer or no offspring. Early life stages of fish (eggs and fry) are usually more sensitive than adults to low pH, and acute mortality of these life stages may be high during periods of rapid declines in pH following rainfall. The death of fish embryos is thought to result from the corrosion of skin cells, which interferes with respiration and osmoregulation. While fish die from acidification, their numbers and diversity are more likely to decline from a failure to reproduce.

The effects of acid deposition on individuals described above in turn elicit changes in the composition and abundance of communities of aquatic organisms. The degree of change depends on the severity of acidification, and the interaction of other factors, such as metal concentrations and the buffering capacity of the water. The pattern most characteristic of aquatic communities in acidified waters is a loss of species diversity, and an increase in the abundance of a few, acid-tolerant taxa. At moderately affected sites, acid-sensitive species are frequently replaced by more acid-tolerant forms, while the total number or mass of organisms may not change. For example, ephemeropterans (mayflies) are quite sensitive to reductions in pH and their numbers may decline, while plecopterans (stone flies) and dipterans (true flies) become relatively more dominant following acidification. Invertebrates in the feeding group that shred leaves also become relatively more dominant than those that graze on diatoms in acid-impacted streams. As pH decreases further and toxic metals increase, both the number of species and the number of organisms decrease sharply (**Fig. 4**). Mollusks and crustaceans are particularly sensitive to acidification, since both groups require high levels of calcium to form their shells. Calcium is scarce in the poorly buffered systems that are most affected by acidic deposition.

Community-level effects may occur indirectly, as a result of changes in the food supply and in predator-prey relations. Reduction in the quality and amount of periphyton may decrease the number of herbivorous invertebrates, which may in turn reduce the number of organisms (predatory invertebrates and fish) that feed upon herbivorous invertebrates. The

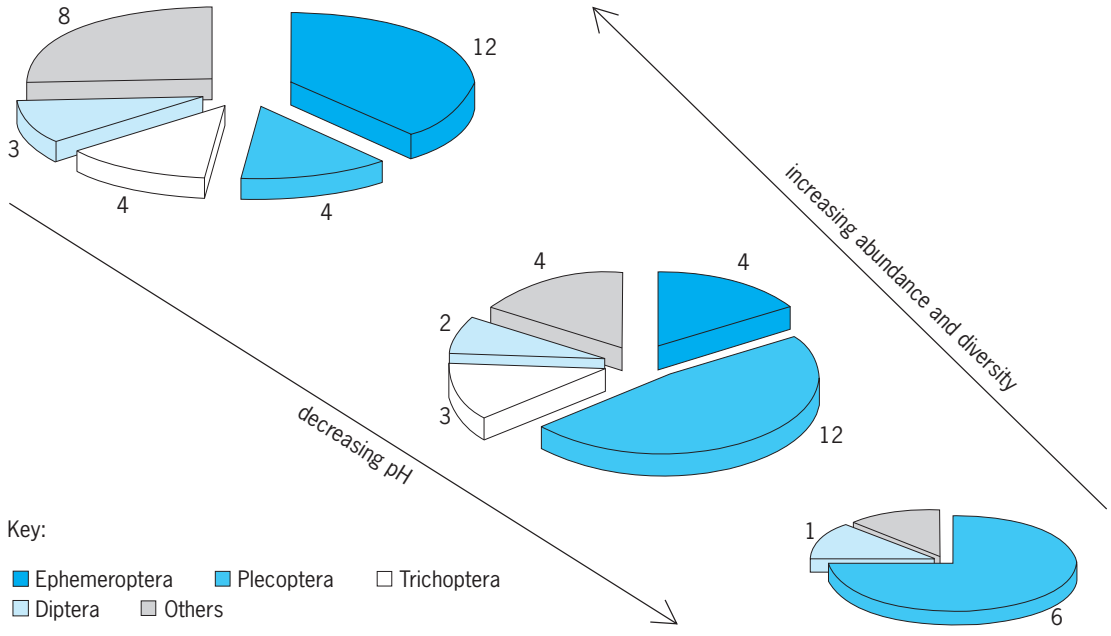


Fig. 4. General pattern of aquatic invertebrate abundance and diversity with a reduction in pH. Total abundance is indicated by disk sizes. Numbers of species within each order are indicated.

disappearance of fish may result in profound changes in plant and invertebrate communities. Dominant fish species function as keystone predators, controlling the size distribution, diversity, and numbers of invertebrates. Their reduction alters the interaction within and among different levels of the food web and the stability of the ecosystem as a whole.

Acid deposition affects the survival of individual organisms and the relationship of communities within aquatic ecosystems. Early life stages are generally more sensitive to low pH then adult organisms. Directs effects of pH and aluminum toxicity affect survival more than the indirect effects of food availability or predation.

The rate and magnitude of acidification of aquatic systems depend upon the factors discussed above. Predictable changes in water chemistry, as the pH of surface waters decreases, include decreased alkalinity, decreased buffering capacity, and increased concentrations of aluminum, magnesium, and iron, among other elements. Such changes in water chemistry generally result in decreased diversity of aquatic species and reduced productivity.

The impact of acid deposition on terrestrial and aquatic ecosystems is not uniform. While increases in acid deposition may stress some ecosystems and reduce their stability and productivity, others may be unaffected. The degree and nature of the impact depend on the acid input load, organismal susceptibility, and buffering capacity of the particular ecosystem. *See* BIOGEOCHEMISTRY; ECOSYSTEM.

Ronald R. Schnabel; Dennis Lamb; Harry B. Pionke; Dennis Genito

Bibliography. *Acid Deposition: Atmospheric Processes in Eastern North America*, National Academy Press, 1983; *Acid Rain and Transported Air Pollutants: Implications for Public Policy*, U.S. Congress, Office of Technology Assessment, 1984; R. J. Charlson and H. Rohde, Factors controlling the acidity of natural rainwater, *Nature*, 195:683–685, 1982; F. M. D'Itri (ed.), *Acid Precipitation: Effects on Ecological Systems*, 1982; D. L. Godbold and A. Hüttermann (eds.), *Effects of Acid Rain on Forest Processes*, Wiley-Liss, New York, 1994; R. M. Heard et al., Episodic acidification and changes in fish diversity in Pennsylvania headwater streams, *Trans. Amer. Fisheries Soc.*, 126:977–984, 1997; *Interim Assessment: The Causes and Effects of Acidic Deposition*, National Acid Precipitation Assessment Program, Washington, D.C., 1987; W. G. Kimmel et al., Macroinvertebrate community structure and detritus processing rates in two southwestern Pennsylvania streams acidified by atmospheric deposition, *Hydrobiologia*, 124:97–102, 1985; W. P. Robarge and D. W. Johnson, The effects of acidic deposition on forested soils, *Adv. Agron.*, 47:1–83, 1992; A. D. Rosemond et al., The effects of stream acidity on benthic invertebrate communities in the south-eastern United States, *Freshwater Biol.*, 27:193–209, 1992; J. N. Wordman and E. B. Cowling, Airborne chemicals and forest health, *Environ. Sci. Technol.*, 21:120–126, 1987.

Acipenseriformes

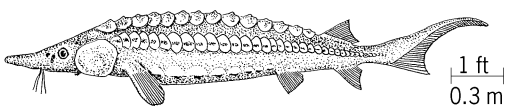
An order of actinopterygian fishes in the subclass Chondrostei, which probably originated from palaeonisciform ancestors during the Triassic or Jurassic periods. Acipenseriformes, the only extant order of about 10 orders of chondrosteans, are characterized by a largely cartilaginous skeleton, heterocercal tail, fins with more than one ray per pterygiophore, spiral valve intestine, myodome and preopercle reduced or absent, and gular plate absent. The cartilaginous skeleton of Acipenseriformes, long regarded as primitive, was taken as indicative of an alliance with the Chondrichthyes; however, cartilage is secondary and the true relationship is with bony fishes. *See* ACTINOPTERYGII; CHONDROSTEI; OSTEICHTHYES.

Acipenseriformes includes two suborders: the fossil Chondrosteoidei, which became extinct in the Mesozoic, and the Acipenseroidei, with two families, Acipenseridae and Polyodontidae, both with little apparent change from the Cretaceous.

Acipenseridae (sturgeons). Sturgeons are characterized by a robust body armed with five longitudinal rows of bony plates; a pronounced snout bearing sensitive barbels on the underside; an inferior mouth, which is protrusible and, in adults, toothless; spiracles present or absent; gill rakers fewer than 50; a large swim bladder; and pectoral fins, the leading edge of which consists of fused rays.

The family includes four extant genera and about 25 species in two subfamilies. The subfamily Acipenserinae comprises three genera: *Acipenser* (see **illustration**) is distinguished by having spiracles and a subconical snout, and is known from North America, Europe, and Asia. In *Pseudoscaphirhynchus*, which is endemic to the Aral Sea basin, spiracles are absent, the snout is depressed, and the caudal peduncle is short, depressed, and incompletely armored. *Scaphirhynchus* is similar to *Pseudoscaphirhynchus* but differs in having a long, depressed, and completely armored caudal peduncle; it is endemic to the Mississippi basin of North America.

The subfamily Husinae has two species: *Huso dauricus* (kaluga), which is native to the Amur River in Russia and China, and *Huso huso* (beluga), which occurs in the Black and Caspian seas and rarely in the Adriatic Sea. The beluga holds the distinction of being the largest freshwater fish in the world. Just how large is uncertain, as some records are in terms of length, others by weight. Hearsay has belugas reaching 6–8 m (20–26 ft) in length, but their weights are unknown. If such giants ever existed, it would have been well over 100 years ago. FishBase



Lake sturgeon (*Acipenser fulvescens*).

reports a 5-m (16.5-ft) 2072-kg (4558-lb) specimen estimated to be 118 years old.

Sturgeons are sluggish, slow-growing, long-lived fishes found in Eurasia and North America. All spawn in freshwater, but several species spend most of their lives in the sea. Valued commercially for their flesh and as the source of caviar, sturgeons have long been seriously depleted in North America and most of Europe, but a substantial fishery is maintained in Western and Central Asia. *See* STURGEON.

Family Polyodontidae (paddlefishes). Paddlefishes have a long, paddlelike snout. There are no large plates and scutes as in sturgeons, but patches of small ganoid scales; minute barbels on the underside of the snout; spiracles present; minute teeth; and greatly extended gill covers. There are several fossil genera known primarily from the Cretaceous Period in China and North America (Montana and Wyoming) and two living species, *Psephurus gladius* from China and *Polyodon spathula* from the United States.

Polyodon differs from *Psephurus* in having long gill rakers that number in the hundreds and are used as a filter in feeding on plankton, in contrast to gill rakers that are shorter and fewer in number in *Psephurus*; and scales mostly on the caudal pecuncle and caudal fin in contrast to scales mostly on the trunk. *Psephurus* is the larger, reaching 3 m (10 ft) in total length. The largest *Polyodon* of recent record, weighing 65 kg (143 lb), was caught in the Missouri River in Montana in 1973.

It is interesting to speculate on the utility of the paddle of the paddlefish. A common belief of nineteenth and early twentieth century fishers and professional naturalists was that the paddle is used to stir the bottom sediments to dislodge food, which then would be filtered from the water. Digging utility is not likely because the process would damage the numerous sensory and electroreceptors on the paddle. It is more likely that the paddle is used in detecting food. Another interesting hypothesis is that the paddle functions as a stabilizer to prevent dipping of the head that might occur as the fish cruises through water with its large mouth agape.

Herbert Boschung; Reeve M. Bailey

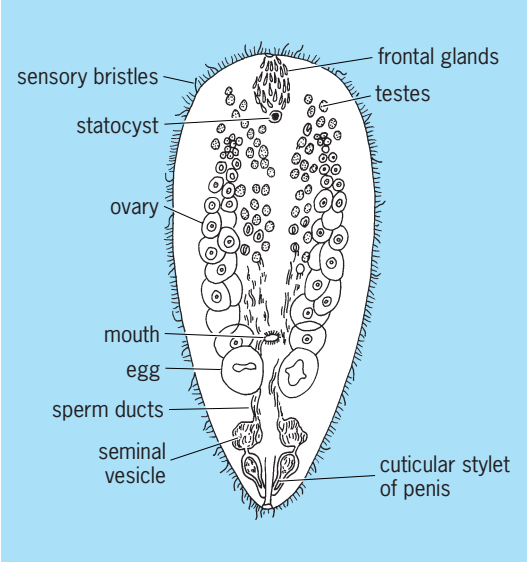
Bibliography. W. E. Bemis, E. K. Findeis, and L. Grande, An overview of Acipenseriformes, *Environ. Biol. Fishes*, 48:25–71, 1997; W. E. Bemis and B. Kynard, Sturgeon rivers: An introduction to Acipenseriformes biogeography and life history, *Environ. Biol. Fishes*, 48:167–183, 1997; V. J. Birstein and W. E. Bemis, How many species are there within the genus *Acipenser*?, *Environ. Biol. Fishes*, 48:157–163, 1997; V. J. Birstein, P. Doukakis, and R. DeSalle, Molecular phylogeny of Acipenseridae: Nonmonophyly of Scaphirhynchinae, *Copeia*, 2002(2):287–301, 2002; A. Choudhury and T. A. Dick, The historical biogeography of sturgeons (Osteichthyes: Acipenseridae): A synthesis of phylogenetics, paleontology and palaeogeography, *J. Biogeog.*, 25:623–640, 1998; R. Froese and D. Pauly (eds.), FishBase, version 05/2005; L. Grande and W. E. Bemis, Interrelationships of Acipenseriformes with comments on "Chondrostei," pp. 85–115 in M. L. J. Stassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; L. Grande and W. E. Bemis, Osteology and phylogenetic relationships of fossil and recent paddlefishes (Polyodontidae) with comments on the interrelationships of Acipenseriformes, *J. Vert. Paleontol.*, vol. 11, Mem. 1, 1991; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Acoela

An order of marine platyhelminthes of the class Turbellaria, generally regarded as primitive, and lacking protonephridia, oviducts, yolk glands, a permanent digestive cavity, and strictly delimited gonads. The nervous system is epidermal in the most primitive representatives and resembles that of cnidarians, but it is generally submuscular and consists of three to six pairs of longitudinal strands without anterior concentrations recognizable as a brain. The anterior or median ventral mouth opens into a simple pharynx which lacks elaborate musculature and leads directly to the syncytial vacuolated mass of endoderm forming the spongy intestinal tissue and lacking any true lumen.

Eyes are often absent; when present they are of the usual paired turbellarian pigment-cup type, with one to a few pigment and retinal cells. A single (sometimes double) statocyst and sensory bristles (modified cilia) are generally present anteriorly, and clusters of frontal glands open by one or more pores at the anterior tip. The epidermis, typically, is uniformly ciliated, and locomotion is accomplished by means of rapid, smooth ciliary gliding or swimming.

The acoels (see **illus.**) are mostly small (one to several millimeters in length), of simple contour, and plump and broad- to elongate-oval in form. Their



Childia spinosa.

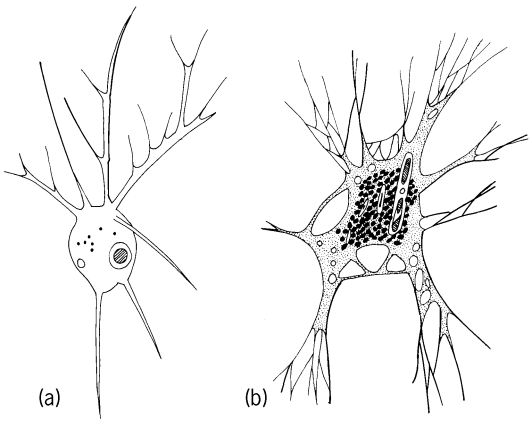
coloration is generally drab, or inconspicuous grayish-white, unless the body is colored brown or green by symbiotic algae. They are virtually worldwide in distribution, living beneath stones, among algae, on the bottom mud, or interstitially, from the littoral zone to deeper waters. A few are pelagic (in tropical and subtropical zones), and others have become symbiotic, living entocommensally in various invertebrates but mainly in echinoderms.

Free-living species feed on diatoms, protozoans, and other microscopic organisms by partially extruding the digestive syncytium through the mouth and engulfing the food, ameboid pseudopodial fashion, in temporary vacuoles in the syncytium. Larger organisms (crustacean larvae, adult copepods, and the like) are seized by the anterior body (which may be expanded or curved dorsoventrally); the grip is facilitated by sticky mucus from the frontal glands, and the prey is thrust through the mouth into a larger temporary digestive cavity. Digestion is rapid and effected by acid and alkaline proteases, carbohydrases, and lipases acting in sequence and secreted from the syncytium. Entocommensal species feed and digest in the same manner, the presence of appropriate cosymbiotic organisms permitting retention of the basic nutritional physiology characteristic of their free-living relatives.

The Acoela, as primitive Turbellaria, are of considerable interest phylogenetically since they show features reminiscent of the cnidarian grade of organization as well as others foreshadowing those of more elaborate turbellarians. In particular, their structure supports the hypothesis that ancient planula or planuloid larvae of hydrozoan cnidarians, by developing gonads and an oral aperture, gave rise to an acoeloid ancestral stock from which all modern flatworms, rhynchocoelans, and the coelomate Metazoa have subsequently evolved. *See* CNIDARIA; PLATYHELMINTHES; TURBELLARIA. J. B. Jennings

Aconchulinida

An order of the subclass Filosia comprising a small group of naked amebas which form filamentous pseudopodia (filopodia). The genus *Penardia*, possibly the only valid one in the order, includes filopodia-forming amebas with variable shape, an outer zone of clear ectoplasm, and inner cytoplasm often containing many vacuoles and sometimes zoochlorellae (see **illus.**). Pseudopodia are sometimes extended primarily at the poles of the organism, but also may fuse laterally to form small sheets of clear cytoplasm. A single nucleus and one water-elimination vesicle are characteristic of these amebas. Size of described species ranges from less than 10 to about 400 micrometers. *See* PROTOZOA; RHIZOPODEA; SARCODINA; SARCOMASTIGOPHORA. Richard P. Hall



Aconchulinida. (*a*) *Penardia cometa*. (*b*) *Penardia mutabilis*.

Acorales

An order of monocotyledonous angiosperms composed of a single genus, *Acorus* (sweet flag or sweet calamus), with two species widespread in the Northern Hemisphere. Formerly, *Acorus* was included in the aroid family, Araceae, but several lines of evidence, including deoxyribonucleic acid (DNA) sequences, have firmly established it as the first-branching lineage of the monocotyledons. These species are emergent aquatics with peculiar inflorescences, and they are unusual among the monocots in having characters otherwise confined to the magnoliid dicots, such as the anther formation and the presence of ethereal oils. These oils are the basis of the species' frequent use as a medicine to treat toothache and dysentery, and as an ointment in religious ceremonies. *See* MAGNOLIOPHYTA. Mark W. Chase

Acoustic emission

A method of nondestructive testing and materials characterization that uses mechanical waves moving through materials. It is similar to seismology, except in being concerned with the scale of engineering structures, such as aircraft, bridges, and chemical tanks. When a structure is subjected to external force (or stress), a defect (for example, a crack or welding flaw) on the structure is activated and enlarged dynamically, and thus generates waves, which spread through materials at a certain speed. Such waves, known as acoustic emission signals, are detected by sensors attached on the surfaces of the structure. Mechanical vibration due to acoustic emission signals is weak and requires high-sensitivity sensors and electronic amplification before it can be analyzed. *See* SEISMOLOGY.

In nondestructive testing of structures, acoustic emission signals are typically evaluated in order to know if the failure of a structure is imminent; if cracks and other defects, presumed to be present in any structure, are active; the positions of such active defects; and whether a structure with such defects can be safely operated. In evaluating materials behavior and quality, acoustic emission is used to assess how a material (such as a steel or an aluminum alloy) responds to mechanical stress, that is, when and how it changes shape permanently and how it proceeds to

eventual fracture; how an alloy withstands repeated application of stress (known as fatigue); the level of stress and corrosive environment that lead to failure of a material; and the types of microscopic failure processes that arise in a material under stress. *See* METAL, MECHANICAL PROPERTIES OF; PLASTIC DEFORMATION OF METAL; STRESS AND STRAIN.

Methodology. Acoustic emission signals emanating from their sources contain information about the source, such as the direction and speed of crack opening. For example, the high-speed cracking of brittle materials (such as high-strength steels and ceramics) produces short, fast-varying acoustic emission signals (**Fig. 1**), which are typically plotted against time measured in microseconds. In contrast, slow-growing defects in plastics result in longer, slowly varying signals (**Fig. 2**), which are typically plotted on a time scale of milliseconds. Because of the distortion of waves during the propagation through a complex structure and detection by a sensor, however, much of the information is lost. Thus, the presence of detectable acoustic emission signals is the most important clue in assessing the integrity of the structure.

By detecting one such signal at multiple sensor positions, the location of its source can be determined from the timing of signal arrivals. The basic principle of triangulation is the same as practiced in seismology, except that the differences in signal arrival times are of the order of microseconds to milliseconds. The speed of wave propagation is a material constant, determined by the stiffness and mass density of the propagating medium. Typically, the wave speed in solids is several kilometers per second. (By comparison, the sound speed in air is about 330 m or 1080 ft per second.) *See* SOUND.

An example of a sensor arrangement is shown in **Fig. 3** for a chemical tank. Here, the sensors are typically separated by several meters to 10 m (33 ft), and cover the entire surface of the tank with blocks of roughly triangular shape. The position of an acoustic emission source within each block is determined from the arrival times at the three surrounding sensors. A typical plot of acoustic emission source positions is shown in **Fig. 4**, indicating a concentration of sources.

The detection of acoustic emission signals requires a sensor possessing high sensitivity. A typical sensor uses a piezoelectric ceramic element, which converts mechanical vibration into an electrical signal, which can be amplified 1000 to 10,000 times. Various electrical measurement techniques are used to characterize and analyze the signals received. It is common to obtain and record several features of acoustic emission signals. These form the basis of real-time analysis and decision-making. In laboratory studies, the entire waveforms are also recorded for detailed analysis after testing. *See* PIEZOELECTRICITY.

Applications. By considering observed acoustic emission signal strength, correspondence to possible defective sites, and comparison to previous acoustic emission observations at the same position, the seriousness of the presumed defect is estimated, sometimes resulting in immediate shut-down of the entire plant. The sign of imminent failure is a rapid increase in the number of acoustic emission signals from a few concentrated sources. However, the value of the acoustic emission method lies in its pointing out potential trouble spots long before danger signs are generated. In this type of structural monitoring, the acoustic emission method can be applied without stopping normal operation (that is, without emptying the content of a tank or pipeline). Acoustic emission testing is applied to almost all tanks and vessels when new, and periodically during their service life.

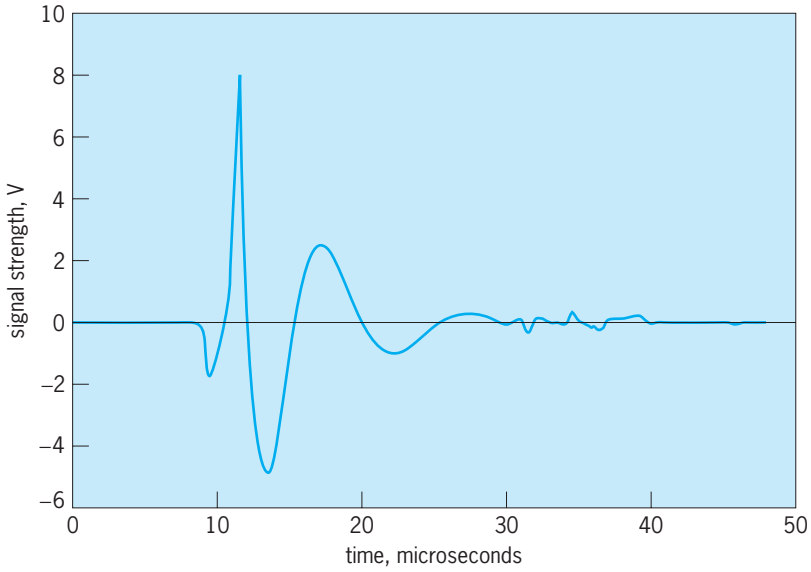


Fig. 1. Waveform of rapidly varying acoustic emission signal from a brittle material.

Other applications include detecting leakage of fluid and gas from pressurized tanks, pipes, and valves; monitoring key structural elements of bridges, dams, and tunnels; detecting damage in bearings and other rotating components, such as motors and compressors; detecting failure of cutting tools in automatic machining operation; and finding fissures and cracks in geothermal wells. In most

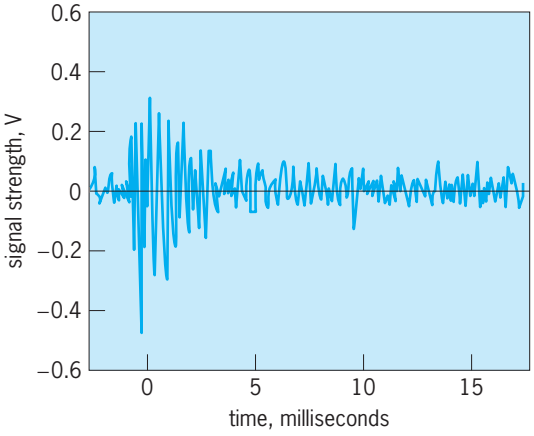


Fig. 2. Waveform of slowly varying acoustic emission signal from a plastic.

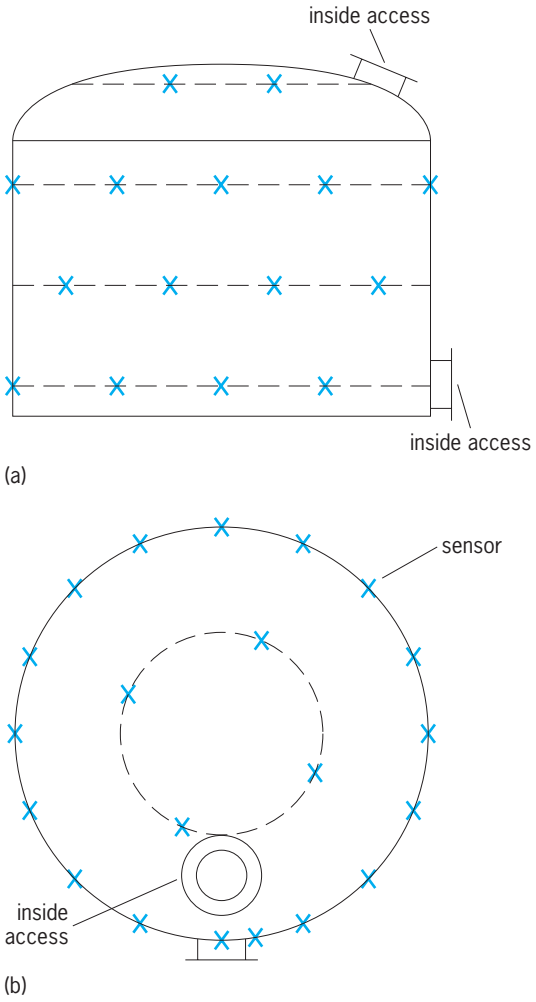


Fig. 3. Sensor arrangement for a chemical tank. (*a*) Side view. (*b*) Top view.

instances, the phenomena of acoustic emission generation are complex, and the evaluation of the results is largely learned from experience.

Acoustic emission methods are also valuable in the laboratory. Acoustic emission signals are emitted when materials are permanently deformed (stretched, pressed, bent, or twisted). Materials can exhibit several characteristic acoustic emission behaviors. One type of material emits many acoustic emission signals when the deformation begins, followed by a decreasing trend as more deformation occurs. Another type is characterized by the near absence of acoustic emission. This behavior is found when the material has been previously deformed heavily. These changes reflect the variation in the internal composition of materials, and acoustic emission provides key information in understanding deformation characteristics.

Acoustic emission signals are also produced when materials are stressed to a level nearing fracture. Material fracture results from the accumulation of microscopic internal damage. Such damage develops either gradually or rapidly. While acoustic emission cannot detect slow damage formation (such as forming voids), rapid damage processes that form cracks are serious defects and can be easily detected. One goal of materials engineering is to develop a strong and tough alloy with excellent fracture resistance, and acoustic emission provides valuable insights regarding the basic mechanical damage processes that are relevant to this endeavor.

An example of damage accumulation is shown in **Fig. 5**. Here, stress is applied to a high-strength steel specimen. As the applied stress is increased beyond a limit, the number of acoustic emission signals starts to rise and continues to rise quickly. Each such signal is produced in this steel by the fracture of a very hard and brittle carbide particle embedded in the steel matrix. Near the final fracture point, some acoustic emission signals result from microscopic cracks. In this study, carbide particles are shown to be the main source of damage. Consequently, the strength of the steel is improved by 20% or more by reducing the size of carbide particles. *See* STEEL.

In addition to the observation of deformation and fracture, acoustic emission is utilized in various areas of materials studies. These include abrasion, casting,

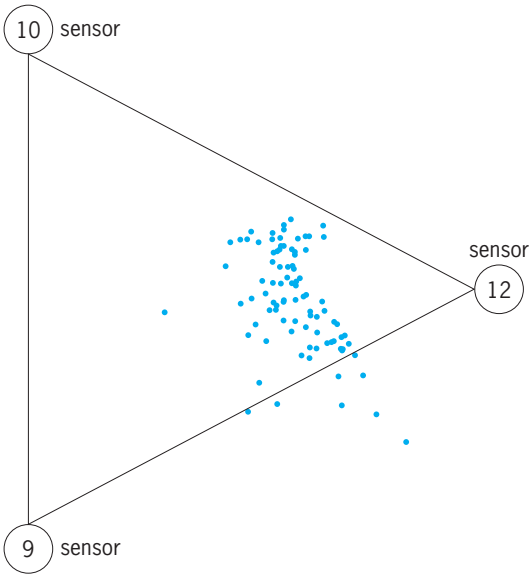


Fig. 4. Typical plot of acoustic emission source positions, indicating a concentration of sources.

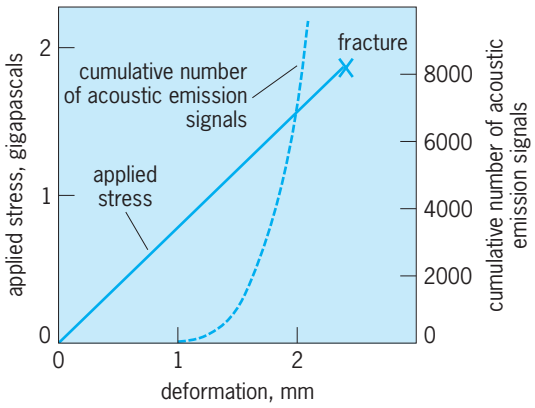


Fig. 5. Behavior of a high-strength tool steel specimen as stress is applied.

film coating and plating, corrosion, curing of plastics, friction, magnetization, oxidation, and hardening of steels. *See* NONDESTRUCTIVE EVALUATION.

Kanji Ono

Bibliography. Acoustic emission, in *Metals Handbook*, 9th ed., vol. 17: *Nondestructive Testing and Quality Control*, pp. 278–294, ASM International, Materials Park, OH, 1989; R. K. Miller and P. McIntire (eds.), Acoustic emission testing, *Nondestructive Testing Handbook*, 2d ed., vol. 5, American Society for Nondestructive Testing, Columbus, OH, 1987; K. Ono (ed.), *J. Acous. Emission*, 1982– ; K. Ono (comp.), *Progress in Acoustic Emission IX*, 1998 [Proceedings of the 14th International Acoustic Emission Symposium: Transitions in AE for the 21st century]; I. G. Scott, *Basic Acoustic Emission*, Gordon and Breach, New York, 1991.

Acoustic impedance

The ratio of the complex amplitude of spatially averaged acoustic pressure over a surface to the complex amplitude of volume velocity on that surface. The unit is the newton-second/meter^5, or the mks acoustic ohm. In the cgs system the unit is the dyne-second/centimeter^5. The volume velocity is the integral over the surface of the particle velocity normal to the surface. Acoustic impedance measures the degree to which an acoustic medium impedes the fluid motion created by an imposed pressure. It was originally defined by analogy to electrical systems. *See* ELECTRICAL IMPEDANCE; SOUND PRESSURE; UNITS OF MEASUREMENT.

In cases where the acoustic pressure p and volume velocity Q vary sinusoidally with time t, it is convenient to express them (as with quantities in alternating-current circuits) as the real parts of complex numbers that are rotating in the complex plane. This is done in Eqs. (1), where $\tilde{p}$ and $\tilde{Q}$ are complex

$$p(t) = \mathrm{Re}\{\tilde{p}\exp(i\omega t)\}$$
$$Q(t) = \mathrm{Re}\{\tilde{Q}\exp(i\omega t)\} \tag{1}$$

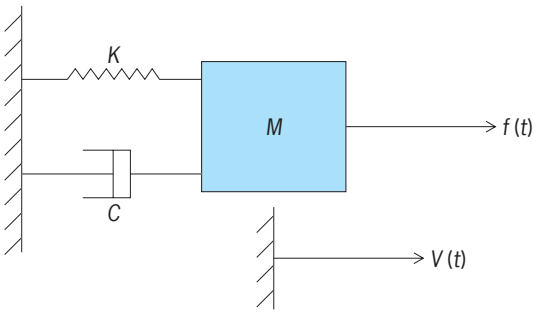
numbers, Re means "real part of," $i = \sqrt{-1}$, and ω is the angular frequency. The acoustic impedance is given by Eq. (2). It is generally complex-valued be-

$$Z = \frac{\tilde{p}}{\tilde{Q}} \tag{2}$$

cause the amplitudes $\tilde{p}$ and $\tilde{Q}$ are generally complex-valued. The phase of Z represents the phase difference between pressure and volume velocity. For a fixed pressure, a large volume velocity is associated with a small magnitude of acoustic impedance. *See* ALTERNATING-CURRENT CIRCUIT THEORY; COMPLEX NUMBERS AND COMPLEX VARIABLES.

Specific acoustic impedance z is defined as the acoustic impedance multiplied by the area, so given by Eq. (3), where $\tilde{V}$ is the complex amplitude of the

$$z = AZ = \frac{\tilde{p}}{\tilde{V}} \tag{3}$$



Mechanical system used to interpret acoustic impedance.

average velocity normal to the surface. The unit is the newton-second/meter^3, or the mks rayl. In the cgs system the unit is the dyne-second/centimeter^3, or the rayl. The characteristic acoustic impedance is the specific acoustic impedance for a plane wave propagating through an acoustic medium and is equal to ρc, where ρ is the mass density and c is the sound speed. The characteristic impedance of water is approximately 1.5×10^6 mks rayls, and the characteristic impedance of air at 0°C (32°F) is approximately 400 mks rayls.

Interpretation of real and imaginary parts. Acoustic impedance is often expressed as the sum of real and imaginary parts, as in Eq. (4), where R is the acous-

$$Z = R + iX \tag{4}$$

tic resistance and X is the acoustic reactance. These quantities may be interpreted by analogy to the mechanical system shown in the **illustration**, where the mechanical impedance is defined by Eq. (5). A

$$Z_{\mathrm{mech}} = \frac{\tilde{f}}{\tilde{V}} = C + i\left(\omega M - \frac{K}{\omega}\right) \tag{5}$$

comparison of Eqs. (4) and (5) indicates that the resistance is analogous to a mechanical dashpot (a damping element in which the applied force is proportional to velocity) and the reactance is analogous to either a mass or a spring, depending on its sign.

A more general interpretation of the resistance comes from Eq. (6) for the time-averaged power flow-

$$P_{\mathrm{ave}} = \tfrac{1}{2}R|\tilde{Q}|^2 \tag{6}$$

ing through the surface, which indicates that the power flow is proportional to resistance for a fixed volume velocity. The reactance is associated with an energy storage.

Circular aperture in planar baffle. One widely used example of acoustic impedance arises when the surface is chosen as a circular aperture in a planar rigid baffle of infinite extent with an acoustic fluid on one side. The velocity normal to the aperture is assumed constant over the aperture. The acoustic fluid has mass density ρ and sound speed c. When the radius of the circle is much less than an acoustic wavelength,

the resistance and reactance are given by Eqs. (7),

$$R_p = \rho c \frac{1}{2} \frac{(ka)^2}{\pi a^2} \quad \text{for} \quad (ka)^4 \ll 1$$
$$X_p = \rho c \frac{8}{3\pi} \frac{ka}{\pi a^2} \quad \text{for} \quad (ka)^2 \ll 1 \tag{7}$$

where $k = \omega/c$ is the acoustic wavenumber.

Acoustic reflection from a fluid interface. A common use of acoustic impedance arises in acoustic reflection problems. Reflected and transmitted waves are created when a plane wave is normally incident on a boundary between two fluids. If the incident wave propagates in a fluid denoted by 1 and the other fluid is denoted by 2, the complex-valued reflection coefficient is the ratio of reflected to incident wave amplitudes. It is related to the acoustic impedances of the fluids by Eq. (8).

$$R = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{8}$$

This result confirms the intuitive expectation that a relatively high impedance for fluid 2 will create a reflection coefficient similar to that of a rigid boundary ($R = 1$). Conversely, a relatively low impedance for fluid 2 will create a reflection coefficient similar to that of a pressure-release boundary ($R = -1$). *See* SOUND.

J. Gregory McDaniel

Bibliography. L. L. Beranek, *Acoustics*, 1948, reprint, 1989; M. C. Junger and D. Feit, *Sound, Structures, and Their Interaction*, 1972, reprint 1993; H. F. Olsen, *Dynamical Analogies*, 1958; A. D. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications*, 1981, reprint, 1989.

Acoustic interferometer

An instrument that is sensitive to the interference of two or more acoustic waves. It provides information on acoustic wavelengths that is useful in determining the velocity and absorption of sound in samples of gases, liquids, and materials, and it yields information on the nonlinear properties of solids.

Operating principles. In its simplest form, an acoustic interferometer for use in liquids has a fixed piezoelectric crystal (acting as a transmitter) tuned to the frequency of interest and a parallel reflector at a variable distance from it. Driven by an oscillating electrical voltage, the piezoelectric crystal generates a sound wave, which in turn is reflected by the reflector. The acoustic pressure amplitude on the front face of the crystal depends on the velocity amplitude at the face and the distance to the reflecting surface. The amplitude ratio (radiation impedance) of the acoustic pressure to the velocity and the relative phase shift between the two oscillating quantities depend solely on the distance to the reflecting surface. If the reflector acts as a rigid surface, this amplitude ratio is ideally zero whenever the net round-trip distance between the crystal and the reflector is an odd number of half-wavelengths because the reflected wave is then exactly out of phase with the incident wave at the crystal's location. The crystal then draws the maximum current since the oscillations are unimpeded. *See* ACOUSTIC IMPEDANCE; PIEZOELECTRICITY; WAVE MOTION; WAVELENGTH.

Wavelength and sound-speed measurement. During operation, the current drawn by the crystal is monitored as the reflector is gradually moved away from the crystal. Whenever the reflector position is such that the crystal is at a pressure antinode (place of maximum pressure in a standing wave), there is a strong dip in the current drawn due to the relatively high radiation impedance presented by the standing wave to the crystal face. Consecutive antinodes are a half-wavelength apart. For a given frequency f, a measured distance L between the location of any one antinode and that of its nth successor yields the wavelength $2L/n$ and the speed of sound $c = 2Lf/n$. An acoustic interferometer based on this principle can achieve a precision of 0.01%. Since the current drawn by the crystal is relatively insensitive to the frequency for a given radiation impedance, the sound speed can also be determined by keeping the distance between the crystal and the reflector fixed and gradually sweeping the frequency. With electronic feedback to lock the interferometer on a particular mode, signal measurement is enhanced.

Absorption measurement. The pressure nodes and antinodes correspond to the local maxima and minima, respectively, in the current drawn. The peak of the current amplitude decreases with the distance traversed by the reflector. If the separation distance is sufficiently large that the exponential decrease associated with absorption dominates any spreading losses, the absorption coefficient for the medium can be derived by measurement of the ratios of current amplitudes at two successive points where the current drawn is a local maximum. *See* SOUND; SOUND ABSORPTION; ULTRASONICS.

George S. K. Wong; Allan D. Pierce; Sameer I. Madanshetty

Bibliography. E. Carr Everbach and R. E. Apfel, An interferometric technique for B/A measurement, *J. Acous. Soc. Amer.*, 98(6):3428–3438, 1995; L. Cuscó and J. P. M. Trusler, Identification of environmentally acceptable low-sound speed liquids, *Int. J. Thermophys.*, 16(3):675–685, 1995; J. Tapson, Stochastic resonance in a mode locked acoustic interferometer, *Ultrasonics*, 36:415–419, 1998; M. S. Zhu et al., Sound velocity and ideal-gas specific heat gaseous 1,1,1,2-tetrafluoroethane (R134a), *Int. J. Thermophys.*, 4(5):1039–1050, 1993.

Acoustic levitation

The use of intense acoustic waves to hold a body that is immersed in a fluid medium against the force of gravity without obvious mechanical support.

In general, levitation involves the use of a fundamental (noncontact) force to balance gravity, and it is usually employed to determine some basic characteristic associated with the force in question. The near levitation of charged oil drops was used by

Waterdrop acoustically levitated in air between a vibrating surface (below the drop) and a reflecting surface (above the drop). The drop is located slightly below the pressure antinode of a 21-kHz standing wave. The equilibrium shape of the drop is oblate because of the balance between acoustic radiation pressure and capillary stresses.

R. A. Millikan to measure the quantum of electrical charge, and this work has been extended by using electromagnetically levitated, superconducting niobium spheres to search for the hypothetical fractionally charged quarks of elementary particle theory. The levitation of particles by light beams has also been demonstrated. *See* ELECTRON; QUARKS; RADIATION PRESSURE.

Levitation is not restricted to electromagnetic and optical means, but can also occur in the presence of fluid flow, including the back-and-forth fluid flow produced by the passage of an acoustic wave. Such acoustically generated forces are extremely small in common experience. But intense acoustic waves are nonlinear in their basic character and, therefore, may exert a net acoustic radiation pressure on an object sufficient to balance the gravitational force and thus levitate it (see **illustration**). Since sound is a mechanical wave, the force on the object will depend on the degree to which the mechanical properties (density and elasticity) of the levitated object differ from those same properties of the surrounding medium. *See* ACOUSTIC RADIATION PRESSURE.

The applications of acoustic levitation in air or other gas include an acoustic positioning module that has been carried in the space shuttle and used in fundamental studies of the oscillation and fission of spinning drops. An acoustic levitation furnace has been designed to study the possibility of containerless solidification of molten materials. This could result in materials of commercial interest, and lead to the bulk processing of materials in space. *See* SPACE PROCESSING.

Levitation in the normal gravitational field of the Earth can occur in the presence of sufficiently intense acoustic fields. Objects can also be levitated by ultrasound in a liquid. Not only is the required force smaller in a liquid than in air because of buoyancy, but also much stronger acoustic standing waves can be established in a liquid than in a gas on account of the liquid's cohesive strength.

Applications of the levitation of objects in liquids have included measurements of the ultimate tensile strengths of liquids, mechanical characterization of superheated and supercooled liquids, the measurement of properties of biological materials (including human red blood cells). The shape oscillations and interfacial properties of drops levitated in air and in liquids have been studied. The evaporation of charged arrays of drops levitated electroacoustically has been characterized.

Closely related to acoustic levitation is the trapping of cold and heavy gases near the pressure nodes of stationary ultrasonic fields. Aerosol particles in air can also be localized and concentrated by stationary ultrasonic fields. Ultrasonic fields in liquids are also used to trap single isolated bubbles in such a way that a light pulse is emitted during each acoustic oscillation. *See* SONOLUMINESCENCE; SOUND; ULTRASONICS.

Robert E. Apfel; Philip L. Marston

Acoustic microscope

An instrument that utilizes focused acoustic waves to produce images of surface and subsurface features in materials, and to measure elastic properties on a microscopic scale. It has been used to image and measure local elastic properties in metals, ceramics, semiconductor integrated circuits, polymeric materials, and biological materials including individual cells.

Scanning acoustic microscope. The most important component of a scanning acoustic microscope is the acoustic lens (**Fig. 1**). It is usually made from a cylinder of single-crystal sapphire. At one end of the cylinder a spherical cavity is ground and polished, while to the other end a high-frequency (100–2000 MHz) piezoelectric transducer is attached. The sample is placed below the spherical cavity near the focal point with a drop of water between the lens and sample. An electrical signal consisting of a narrow-width tone burst at a very high radio frequency is applied to the transducer. The transducer converts the electrical signal into plane-wave acoustic pulses, which propagate through the cylindrical rod and impinge

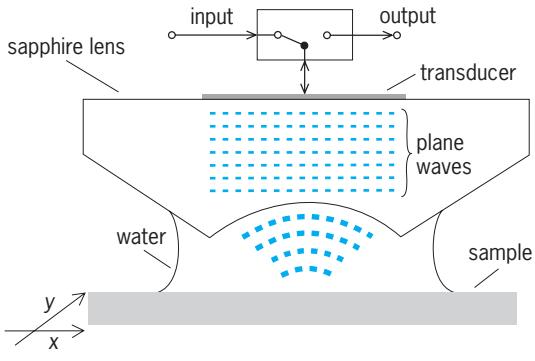


Fig. 1. Acoustic lens.

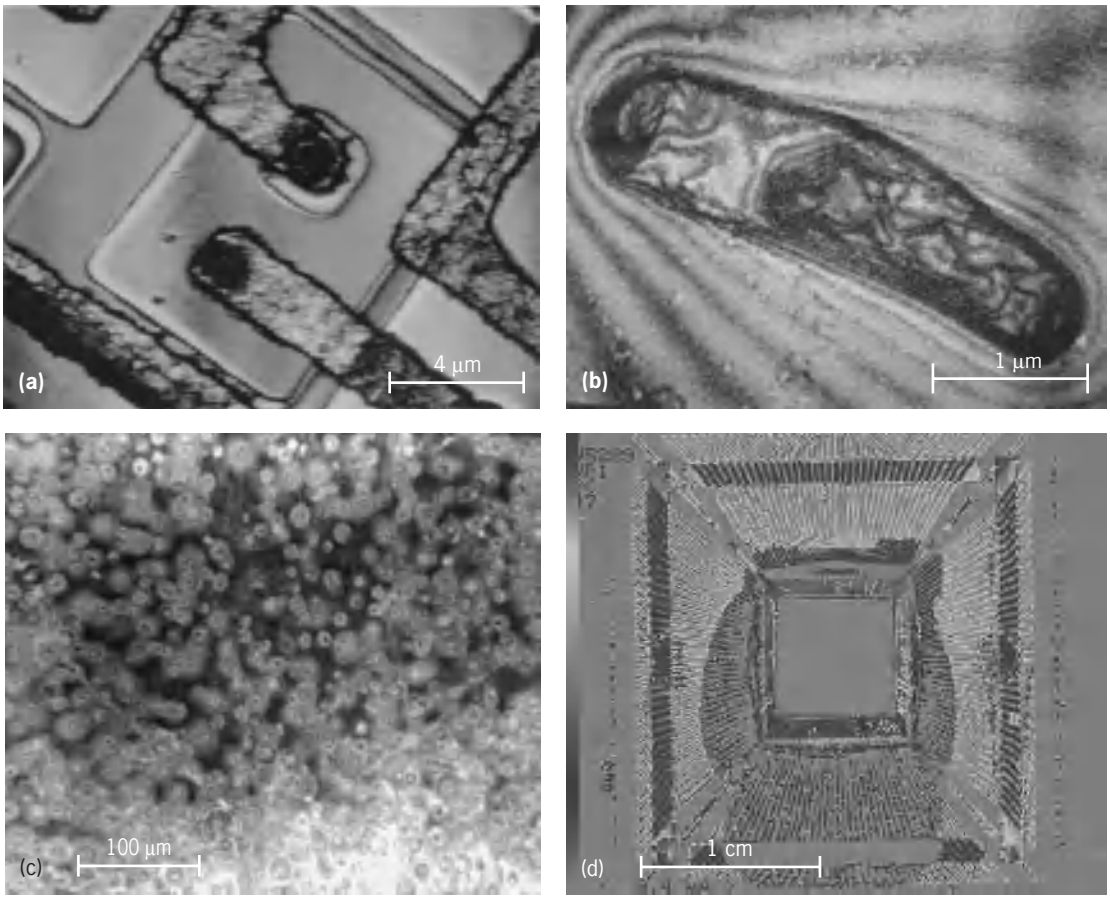


Fig. 2. Acoustic images. (*a*) Bipolar transistor on silicon integrated circuit and (*b*) myxobacterium, both images taken in superfluid helium at 0.2 K (*from J. S. Foster and D. Rugar, Low-temperature acoustic microscopy, IEEE Trans., SU-32, 139–151, 1985*). (*c*) Delaminations at a polymer-metal interface (*courtesy of S. Sathish*). (*d*) Delamination in an electronic ceramic package (*courtesy of G. Pfannschmidt*).

on the spherical cavity of the lens. At the interface between lens and water, the acoustic waves converge and focus to a diffraction-limited spot on the surface of the sample. A part of the incident acoustic energy is reflected by the sample, while the rest propagates into the sample. The reflected signal propagates back into the water and reaches the transducer after propagating through the lens. The transducer converts the acoustic signal into an electrical signal, which is electronically amplified and digitized, and the amplitude is stored in a computer. The lens is mechanically raster-scanned across the surface of the sample, and the reflected amplitude at each location is acquired and displayed on the monitor of a computer as an acoustic image. Since the acoustic image is built with the help of reflected amplitude of the signal, the microscope is called a reflection scanning acoustic microscope. *See* PIEZOELECTRICITY; TRANSDUCER.

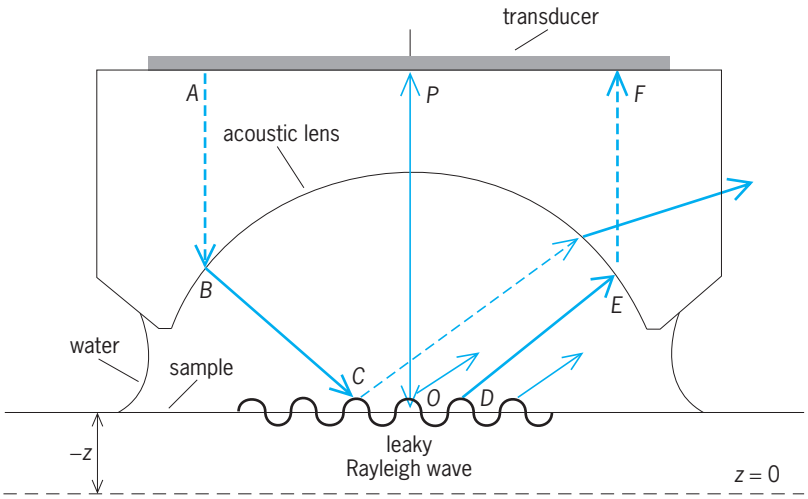


Fig. 3. Acoustic rays interacting to produce contrast enhancement in imaging at defocus and to generate *V(z)* curve.

Resolution and contrast. The diffraction-limited acoustic spot size and hence the resolution of an acoustic microscope increases with the increasing frequency of the acoustic waves. A resolution of 0.5 micrometer can be routinely achieved at a frequency of 2 GHz with water as coupling fluid. This is comparable to the resolution of an optical microscope. By employing liquid helium as the coupling fluid, a resolution of 20 nanometers has been achieved. The contrast in acoustic images arises from the local variation in the elastic properties of the materials and is related to the local acoustic refractive index, which is defined as the ratio of the velocity of sound in the material to the velocity of sound in the coupling fluid. *See* ELASTICITY; LOW-TEMPERATURE ACOUSTICS.

Defocused operation. Acoustic images can be obtained not only with lens focused on the surface of the sample but also with the lens defocused so that the focal point is inside the material (a mode of operation known as imaging at a negative defocus). Acoustic waves focused on the surface produce images of the surface features (**Fig. 2***a*, *b*). However, when the lens is defocused, acoustic waves penetrating into the sample are brought to focus in the interior, and this produces images of the interior features of materials. This feature of an acoustic microscope has been used to image materials that are opaque to light. It has found applications in the analysis of flaws and defects in several industrial materials, including electronic components (Fig. 2*c*, *d*). *See* NONDESTRUCTIVE EVALUATION.

When a large-aperture acoustic lens is used for imaging at a negative defocus, the contrast and the resolution do not degrade as observed in other types of microscopes. Instead, spectacular enhancement in the contrast occurs. The enhancement of the contrast is due to the ability of solids to support surface acoustic waves, and the capability of an acoustic lens to transmit and receive them. The origin of the contrast enhancement can be explained by the interaction of waves inside an acoustic lens (**Fig. 3**). Using a ray picture of acoustic waves, in an acoustic lens, as the angle of incidence increases, a critical angle is reached when total internal reflection occurs. At the critical angle, Rayleigh surface acoustic waves are generated on the surface of the sample, and propagate along the interface between water and the sample. The surface acoustic waves, while propagating at the interface, reradiate energy back into the water, adding an extra contribution to the signal reaching the transducer. These surface acoustic waves are very sensitive to surface and near-surface defects and to the anisotropic elastic properties of materials. This mode of operation of the acoustic microscope is helpful for observing microstructure in polished and unetched metallic materials, and for the detection of surface microcracks. Acoustic images obtained at several different defocus distances show reversal of contrast in many grains.

Quantitative acoustic microscopy. Recording the variation of the received signal amplitude as the lens-to-sample distance is reduced produces a $V(z)$ curve, which can be used to calculate the elastic properties in a microscopic region of the sample. The most important feature of the curve is a periodic oscillation in the amplitude. The origin of this oscillatory behavior is explained using the acoustic rays propagating between the transducer, the acoustic lens, the water, and the sample (Fig. 3). The directly reflected ray (*PO*) and the Rayleigh ray (*AB-BC-CD-DE-EF*) interfere at the face of the transducer and produce a periodic variation in the amplitude. The period of the oscillations can be used to calculate the Rayleigh wave velocity, which, in turn, provides information about elastic properties.

Measuring elastic anisotropy. In an acoustic microscope operating with a spherical lens, the signal returning to the transducer is an average over the diffraction-limited spot. Hence, it is not sensitive to the elastic anisotropy of materials. Acoustic microscopes sensitive to elastic anisotropy have been developed using a cylindrical lens that focuses the acoustic beam along a line. With a cylindrical lens, the Rayleigh surface acoustic waves propagate perpendicular to the cylindrical axis. A measurement of the $V(z)$ curve with such a lens will provide the Rayleigh wave velocity along a line. Such lenses have been used to investigate the elastic anisotropy of materials.

Ultrasonic force microscope. The development of scanning probe microscopy has led to a new type of acoustic microscope capable of imaging elastic properties at nanometer resolution. In the ultrasonic force microscope, an acoustic wave is generated by a piezoelectric transducer attached to one face of the sample, and the tip of an atomic force microscope is used to measure the amplitude of this wave when it has propagated through the sample to the other side. By scanning the tip in contact with the sample, an acoustic image with elastic stiffness information can be produced. The resolution of the microscope depends on the diameter of the tip rather than the acoustic wavelength in the material. *See* MICROSCOPE; SCANNING ELECTRON MICROSCOPE.

Shamachary Sathish

Bibliography. A. Briggs, *Acoustic Microscopy*, Clarendon Press, Oxford, 1992; A. Briggs, *Advances in Acoustic Microscopy*, vol. 1, Plenum Press, New York, 1995; A. Briggs and W. Arnold, *Advances in Acoustic Microscopy*, vol. 2, Plenum Press, New York, 1996; B. T. Khuri-Yakub and C. F. Quate (eds.), *Selected Papers in Scanning Acoustic Microscopy*, SPIE, Bellingham, WA, 1992.

Acoustic mine

A mine that either passively listens to a target's sound noises, or periodically interrogates its environment by actively emitting acoustic pulses that may return echoes if prospective targets come within range. A mine is an underwater weapon consisting of a shell case which contains high explosives. Mines can be planted by airplanes, surface ships, or submarines. Submarines are used for deployment when secrecy is important. There are three types of mines: drifting, moored, and bottom. Drifting mines float freely at the sea surface. Moored mines are positively buoyant and are held at a given depth by a cable anchored to the bottom. Bottom mines rest on the sea floor. A mine can be activated by various means, such as by actually contacting a target, by sensing a target's magnetic field, by listening to the acoustic noises that emanate from a target, by sensing the excess pressure field that may be induced on the mine's sensor by a target passing above it, by the reception of acoustic echoes that a target may return to the mine's sonar after it has sent out its interrogating pings, or by a combination of several of these. *See* SONAR.

Classification mechanisms. Some mines have sophisticated classification mechanisms that allow them to distinguish between various types of nearby

objects. Only when a true target (as opposed to a false alarm) is identified is the mine actually triggered. Simpler mines detonate when they sense any large object nearby, which is always assumed to be enemy. More advanced mines consist of cases with a torpedo inside, and they rest near the sea bottom. The acoustic sensor in the outer capsule constantly explores its environment with its sonar. When a true target is detected and identified, the inner torpedo is fired. It then tracks the prospective target with its own sensor and guidance system until it hits the target. *See* ACOUSTIC TORPEDO.

Any mine can be set so that it will allow the passage of one or more ships and then explode when the desired next one is detected as it comes within range. The acoustic classification phase of any mine mechanism requires considerably more advanced signal processing techniques, and more favorable environmental conditions, than the initial detection phase; hence, it is less reliable.

Countermeasures. Closely related to acoustic mines is the area of acoustic mine countermeasures. Minesweeping vessels also try to find and neutralize mines by means of sonar. Bottom mines are the hardest to find. Often they become buried beneath the bottom sediment, which dampens their possible vibrations, making them nearly impossible to detect. The acoustic detection, classification, and neutralization of mines lying close to—not necessarily beneath—the sea bottom is the most challenging problem in mine countermeasures, and the research required to adequately address these problems is in its infancy.

Guillermo C. Gaunaurd

Bibliography. G. K. Hartmann, *Weapons That Wait*, U.S. Naval Institute, 1979.

Acoustic noise

Unwanted sound. Noise control is the process of obtaining an acceptable noise environment for people in different situations. These definitions and the words "unwanted" and "acceptable" suggest that criteria need to be established to determine when noise from different sources is unwanted and that these criteria could or should be used to decide on acceptable noise limits. Understanding noise and its control, then, requires a knowledge of the major sources of noise, sound propagation, human response to noise, and the physics of methods of controlling noise. The continuing increase in noise levels from many different human activities in industrialized societies led to the term noise pollution. Different governments have passed legislation and created regulations to control noise.

Noise as an unwanted by-product of an industrialized society affects not only the operators of machines and vehicles, but also other occupants of buildings in which machines are installed, passengers of vehicles, and most importantly the communities in which machines, factories, and vehicles are operated.

Propagation of Sound

Sound is a three-dimensional wave motion in the atmosphere. The acoustic waves travel from the source with a speed independent of their amplitude (unless the latter is very large). This speed is then dependent only on the acoustic medium and is proportional to the square root of the absolute temperature for any given medium. For air at 68°F (20°C), the speed of sound c is 1117 ft/s (343 m/s). *See* ACOUSTICS; WAVE MOTION.

Pure tones. The sources of sound may consist either of vibrating solid bodies (such as loudspeakers or vibrating metal panels on machines) or of random motion of air particles (such as when an air jet mixes with the atmosphere). If a loudspeaker cone is made to vibrate with simple harmonic motion at a given frequency f (in hertz or cycles per second), it gives rise to a sinusoidal disturbance in the atmosphere. This sound is known as a pure tone, of frequency f. At any instant there will be a sinusoidal variation of the atmospheric pressure with distance away from the source. The sound pressure p is defined as the difference in the pressure from the undisturbed pressure. For a pure tone, wave crests or wave troughs (maximum and minimum values of pressure) are separated by a distance called the wavelength λ. The relationship between the speed of sound c, wavelength λ, and frequency f is given by Eq. (1). Of course, at

$$\lambda = c/f \tag{1}$$

some point in space the sound pressure p also varies sinusoidally with time. The period T between pressure maxima is related to the frequency by Eq. (2) for a pure tone (**Fig. 1**).

$$T = 1/f \tag{2}$$

Many sound sources, such as a singing voice or a musical instrument, contain strong pure-tone components. These sounds contain several simultaneous pure tones with a fundamental (or lowest-frequency tone) which is usually the strongest. Examples of noise sources which contain pure tones include fans, engine exhausts, pumps, compressors, gears, bearings, and electric motors.

Random noise. If a sound source is made to vibrate with a random motion, it also creates sound waves. However, the sound pressure–time history at some point in space then resembles **Fig. 2** rather than

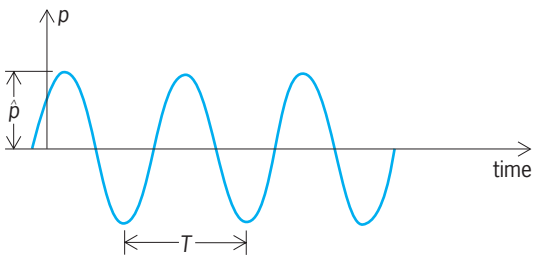


Fig. 1. Sound pressure p at one point in space, varying sinusoidally with time, in response to a source vibrating with simple harmonic motion.

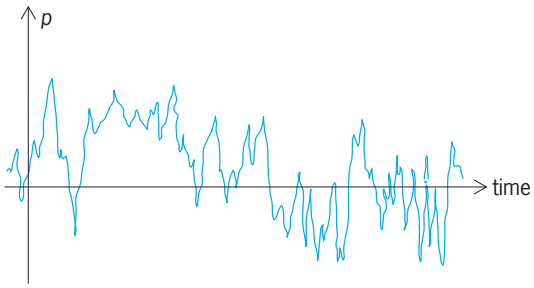


Fig. 2. Sound pressure *b* at one point in space, varying at random with time (random noise), in response to a source vibrating with random motion.

Fig. 1. This sound is called random noise, because it contains all frequencies instead of just one. Many noise sources, for example, jet engine exhausts, fans, flow noise in pipes, waterfalls, and wind, contain mostlyrandom noise.

Although the strength of a pure tone can be described by its amplitude or peak pressure $\hat{p}$ (Fig. 1), a random noise cannot, because the amplitude is never constant. Thus, noise signals are normally described by their time-averaged (effective or root-mean-square) values, P_{rms}, given by Eq. (3), where T

$$p_{\text{rms}} = (1/T)\int_0^T p^2 dt \qquad (3)$$

is the averaging time, which is allowed to approach infinity. Only in the case of pure tone does $p_{\text{rms}} = \hat{p}/\sqrt{2}$.

Sound pressure level. The value of effective sound pressure p_{rms} increases about 10^6 times from a pin drop to a thunderclap. Because of these large variations in sound magnitudes, and because the human hearing sensation seems to vary in a logarithmic way, logarithms are used in measurement of sound. The sound pressure level L_p of a sound of effective pressure p_{rms} is given by Eq. (4), where the units of L_p

$$L_p = 10\log(p_{\text{rms}}^2/p_{\text{ref}}^2) \qquad (4)$$

are decibels (dB). The reference pressure is internationally accepted to be $p_{\text{ref}} = 2 \times 10^{-4}$ dyne/cm.

Sound power level and intensity. If a source is constantly pulsating, it will radiate acoustic energy at a rate that is called the acoustic power and is expressed as W, in watts (joules/s). Again, because acoustic powers vary from very small to very large values, logarithms are used. The sound power level L_w (in decibels) of such a source is given by Eq. (5), where

$$L_w = 10\log(W/W_{\text{ref}}) \qquad (5)$$

the reference sound power is $W_{\text{ref}} = 10^{-12}$ W. In an ideal source, if it is assumed that there is no energy dissipation as the sound energy radiates away from the source, then the intensity I (power/unit area) is

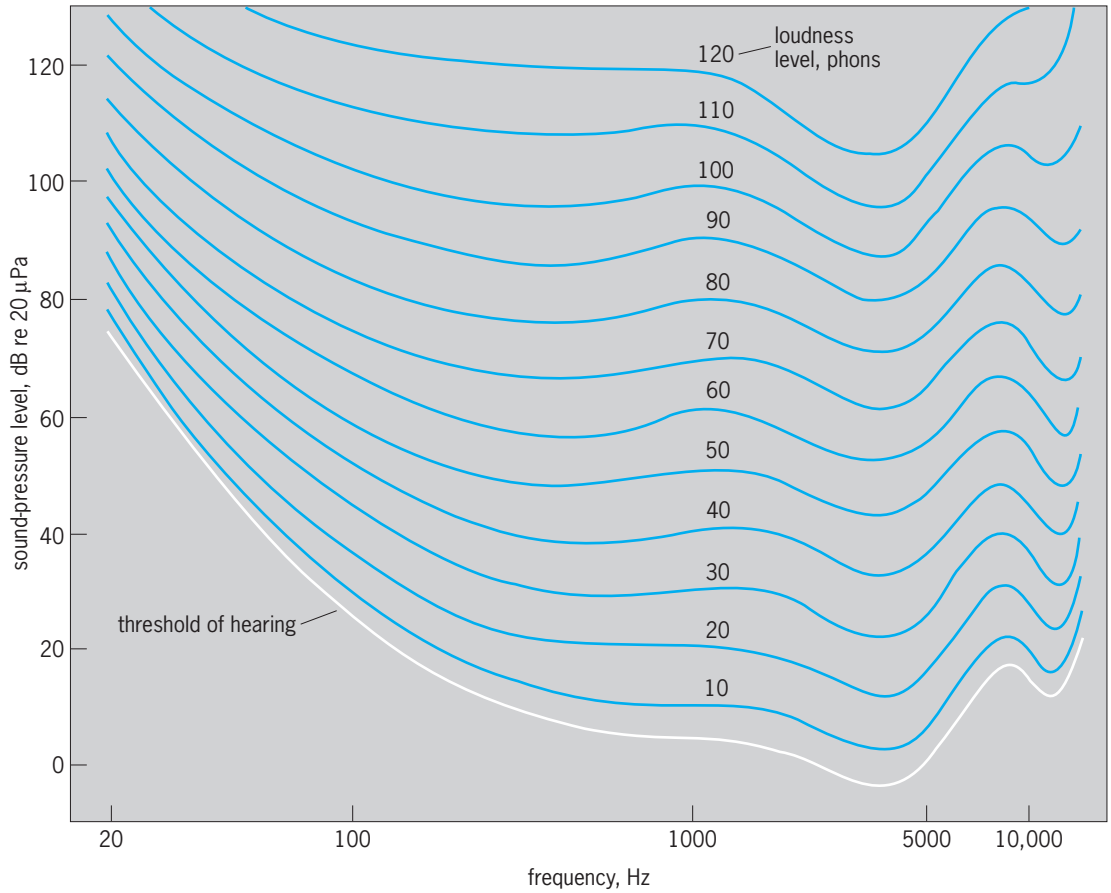


Fig. 3. Equal-loudness contours for pure tones. (*After D. W. Robinson and R. S. Dadson, A redetermination of the equal loudness relations for pure tones, Brit. J. Appl. Phys., 7:166, 1956*)

reduced by a factor of $(1/r^2)$ or $^1/_4$ every time the distance r from the source is doubled. This is the inverse-square law, Eq. (6). Because it is easily shown

$$I \propto 1/r^2 \tag{6}$$

that the intensity I is proportional to p^2_{rms}, doubling the distance from the source reduces the sound pressure level L_p [Eq. (4)] by 10 log $(2)^2 = 20(0.30) \approx 6$ dB, provided there are no reflections. One of the simplest noise control measures thus becomes evident; noise sources should be placed as far away from receivers as possible.

Frequency analysis. In order to determine the frequency distribution of a noise, the intensity or sound pressure level in different frequency bands is measured. The frequency analysis may be done by using electrical filters or fast Fourier transform (FFT) analyzers. The bands most commonly used are constant percentage bands of width one octave or one-third of one octave. (An octave is a doubling in frequency.) Constant-bandwidth bands such as 5, 20, or 50 Hz are also used, to obtain information on pure tones contained in a noise signal. The constant-percentage filter rapidly increases in bandwidth as the center frequency is raised. The range of hearing for most people is from about 20 to 16,000 Hz, and can be covered with 10 octave band filters with center frequencies: 31.5, 63, 125, 250, 500, 1000, 2000, 4000, 8000, and 16,000 Hz. The one-octave band has a bandwidth of 73% of the center frequency, while the one-third-octave band has a bandwidth of 23% of the center frequency. *See* HARMONIC ANALYZER; NOISE MEASUREMENT.

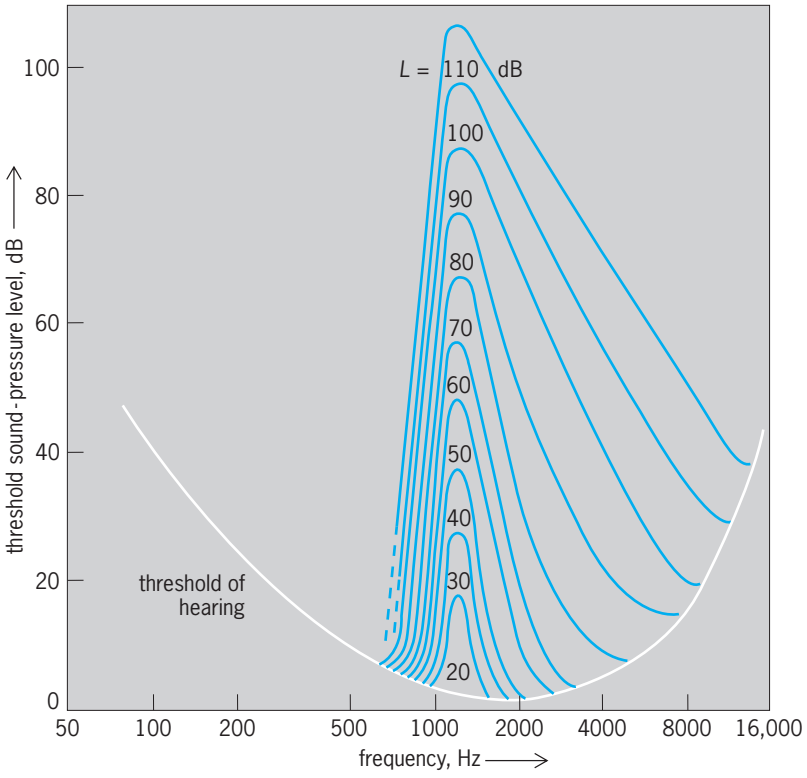


Fig. 4. Masking effect of narrow-band noise with a center frequency of 1200 Hz. Parameter *L* is root-mean-square value of sound pressure level of noise band. Curves show sound pressure level at threshold of hearing in presence of noise band. (*Data from E. Zwicker*)

Human Response to Noise

Noise interferes with some human activities and if sufficiently intense can permanently damage the ear.

Hearing mechanism. The ear is a complicated transducer which converts acoustical energy into mechanical vibrations of the eardrum, the three auditory ossicles (bones), and the cochlear membrane, and eventually into electrical energy which triggers nerve impulses in the auditory nerve. The outer ear canal and the eustachian tube are filled with air, while the inner ear (the cochlea) is filled with fluid. *See* EAR (VERTEBRATE); HEARING (HUMAN); PHYSIOLOGICAL ACOUSTICS.

Loudness. Individual responses to pure tones of different frequencies vary a little. Equal-loudness level contours (**Fig. 3**) represent pure tones which appear equally loud to most people at various frequencies. The lowest curve in Fig. 3 represents the threshold of hearing or the softest sounds that can be heard. The top curve approximately represents the threshold of "in" or "feeling." The equal-loudness level curves are given units of phons P, and are numerically equal to their sound pressure level value at 1000 Hz. Sounds must be increased about 10 phons in level before they double in loudness, and because of this nonlinear subjective response of the ear, a linear loudness scale with units of sones S, defined by Eq. (7), is used. Thus a sound level of 50 phons has

$$S = 2^{(P-40)/10} \tag{7}$$

a loudness of 2 sones; a sound of 60 phons, 4 sones; and so on.

Equal-loudness curves for bands of noise resemble those shown in **Fig. 4** for pure tones, but with some differences. It follows from Fig. 4 that the low-frequency sounds must be much more intense than mid-frequency and high-frequency sounds in order to seem equally loud. The greatest hearing sensitivity is in the range of about 1000 to 4000 Hz, which corresponds to the middle of the speech range.

Equal-loudness level contours have been used in acoustical instrumentation to produce useful single-number measures of the loudness or disturbing effect of noise. A-, B-, and C-weighting filters (**Fig. 5**), corresponding approximately to the inverse of the 40-, 70-, and 100-phon curves respectively, have been built into most sound level meters. The sound readings obtained using these filters are known as the A-weighted, B-weighted, or C-weighted sound levels. Although originally intended for low-level sounds, the A-weighted sound level is used for monitoring both low-level and intense sounds from almost all machine and vehicle noise sources. A-weighted levels are sometimes abbreviated as dB(A). *See* LOUDNESS; PSYCHOACOUSTICS.

Hearing damage. Immediate permanent hearing damage, normally to the eardrum or ossicles, can result from very intense sounds, above about 140

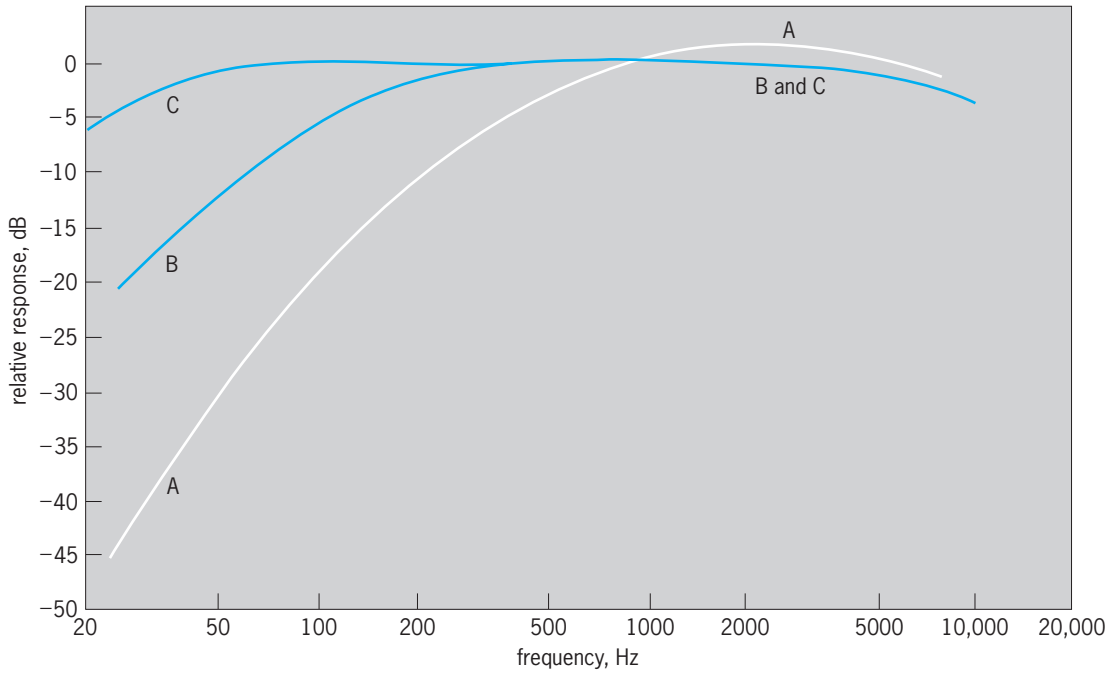


Fig. 5. A-, B-, and C-frequency weightings used with many sound-level meters.

or 150 dB, such as those associated with nearby gunfire or explosions. Lower-level continuous intense noises about 90 to 110 dB(A) can cause temporary hearing loss, from which a person recovers after a period of rest in a quiet environment. However, if such lower-level intense noise is experienced every day over a period of years, permanent hearing damage occurs. The amount of hearing loss depends upon frequency and sound pressure level of noise, bandwidth of noise, duration of exposure each day, and number of years of exposures. A general criterion for specifying tolerable exposures to noise is that the noise should not exceed levels or durations that will cause the average person (after 10 years of exposure) a measurable loss in understanding normal conversation. The damage risk criteria given in **Fig. 6** are based primarily on industrial surveys, and many people can be exposed to greater intensities and durations than these criteria, particularly if the total duration is less than 8 hours per day. *See* HEARING IMPAIRMENT.

Masking and speech interference. When two sounds contain the same frequency components, the more intense sound can normally be heard and it is difficult or impossible to detect the other. This phenomenon is known as masking. If the sounds are of different frequencies, the effect is more complicated and depends on the bandwidth and frequency separation of the two sounds. Figure 4 shows how the hearing threshold is changed for different levels of narrow band of noise centered at 1200 Hz. High-frequency sounds are masked easily by lower-frequency noise, while low-frequency sounds are not masked very easily by high-frequency noise.

Masking noise, sometimes called acoustic perfume, is supplied by loudspeakers in some buildings such as open-plan offices to try to mask unwanted sound such as conversation between other individuals at a distance. Noise in offices or workshops can become loud enough to mask wanted speech sounds, which normally have a sound pressure level of about 65 dB at the listener's ear. People do not become used to hearing speech above masking noises. One measure of the interference with speech of a steady broadband background noise is the preferred speech interference level (PSIL), the arithmetic average of the sound pressure levels of the background noise in the three octave bands centered at 500, 1000, and 2000 Hz (**Table 1**). *See* MASKING OF SOUND.

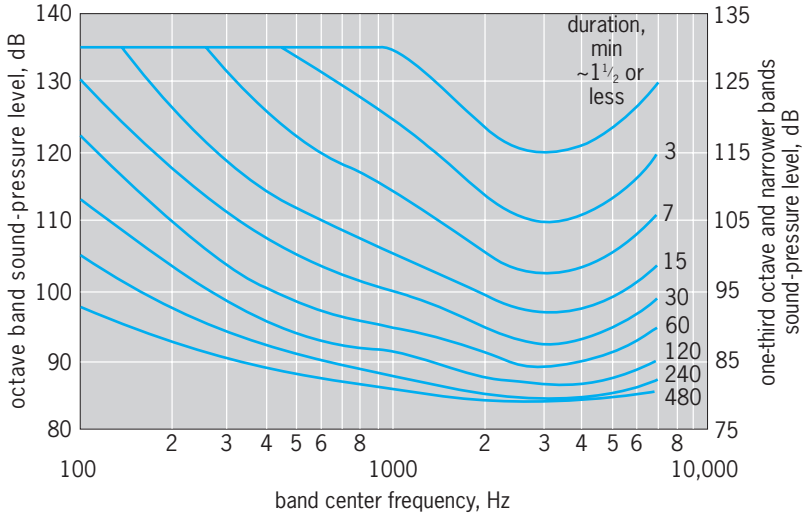


Fig. 6. Damage-risk contours for one exposure per day to one-octave (left-hand ordinate) and one-third-octave or narrower (right-hand ordinate) hands of noise. This graph can be applied to individual band levels which are present in broadband noise. (*After K. D. Kryter et al., Hazardous exposures to intermittent and steady state noise, J. Acous. Soc. Amer., 39:451–463, 1966*)

TABLE 1. Preferred speech-interference levels (500–2000 Hz) that barely permit reliable conversation

Distance between talker and listener, ft (m)	Speech-interference level, dB			
	Normal vocal effort	Raised vocal effort	Very loud vocal effort	Shouting
0.5(0.15)	72	77	83	89
1 (0.3)	66	71	77	83
2 (0.6)	60	65	71	77
3 (0.9)	56	61	67	73
4 (1.2)	54	59	65	71
5 (1.5)	52	57	63	69
6 (1.8)	50	55	61	67
12 (3.7)	44	49	55	61

Annoyance. The louder the noise, the more annoying it tends to be. With continuing exposure to a noise, adaptation occurs as long as the noise is accepted as a part of the environment. Because of adaptation and the difficulty of separating noise annoyance from the effects of other environmental factors, it has not been possible to determine an acceptable annoyance criterion for noise.

Sleep interference. Sufficiently intense noise will awaken a person; less intense noise usually arouses a person from deep sleep to more shallow sleep. However, as with annoyance, people tend to adapt to noise during sleep, making it difficult to specify a noise criterion for sleep interference.

Work performance. Studies to find the effects (if any) of noise on work productivity, efficiency, concentration, incidents of errors and accidents, and so forth, have been inconclusive. Some researchers state that after a period of adaptation noise has little or no effect on work performance, provided it is not sufficiently intense to interfere with speech communication, while others claim that noise can interfere particularly with those engaged in intellectual tasks. *See* INDUSTRIAL HEALTH AND SAFETY.

Community reaction. The effect of noise on whole communities rather than individuals or relatively small groups has also been studied (**Fig. 7**). Several physical measures have been used to characterize the noise environment, including effective perceived noise level (EPNL), composite noise rating (CNR), noise and number index (NNI), and noise exposure forecast (NEF). Most of these measures were originally created to rate aircraft noise while others, such as equivalent sound level L_{eq} and day-night level L_{dn}, were developed to characterize other sources such as traffic, factory, and construction noise.

Noise Reduction Methods

Most noise problems can be modeled as source–path–receiver systems. It is most desirable to reduce the strength or number of the sources. For example, the noise from the impact of two metal machine parts in a punch press might be reduced by replacing one or both of the metal contacts with softer material such as nylon or strong durable plastic. However, it is sometimes difficult to reduce the noise at a source without extensive redesign. For example, changes in some metal-cutting machines may affect the cutting process adversely. In such cases, it may be possible to reduce the source strength by substituting a quieter machine or using a different process.

When all possible ways of reducing the source strength have been tried, ways of reducing noise propagation along paths from the sources to the receivers should be investigated. As a last resort, the receiver, which in most cases is the human ear, can be protected by using earplugs or earmuffs or by enclosure in a booth. However, the ear devices can interfere with communication and be uncomfortable, and booths can be inconvenient and expensive. In cases of community noise problems caused by traffic or aircraft, receiver noise control usually becomes socially unacceptable, extremely expensive, or even virtually impossible. It is usually necessary to control path noise propagation.

Planning. It is often possible to use distance and source directivity with advantage in noise reduction. Equation (6) shows that outdoors the sound pressure level theoretically decreases by 6 dB for each doubling of distance. Indoor sound behavior is described by Eq. (8), sometimes known as the room

$$L_p = L_w + 10\log\left(\frac{Q}{4\pi r^2} + \frac{4}{R}\right) \qquad (8)$$

equation. Here, L_w is the sound power level of the noise source [Eq. (5)], Q is the directivity index of the noise source, r is the distance from source to receiver in meters, and R is the room constant defined as $R = S\overline{\alpha}/(1 - \overline{\alpha})$, where $\overline{\alpha}$ is the average absorption coefficient of the room walls of area S, in m^2. The sound pressure level L_p (in decibels) theoretically decreases at 6 dB per doubling of distance near to the source (where the term $Q/4\pi r^2$ dominates) and at a lower rate in the reverberant field (where the $4/R$ term and reflections dominate). Outdoors, one may assume no reflections, so that $\overline{\alpha} = 1$, $4/R = 0$, and Eq. (8) reduces to Eq. (6). If a source is omnidirectional (radiates equal sound energy to all directions), $Q = 1$. However, many noise sources are directional, making it advantageous in outdoor situations to position the receiver at locations relative to the source where the directivity index Q is small (directions in which the noise source radiates little energy). The same procedure is effective indoors but only close to the source, since the $4/R$ term dominates Eq. (8) in the reverberant field. *See* ARCHITECTURAL ACOUSTICS; DIRECTIVITY; REVERBERATION.

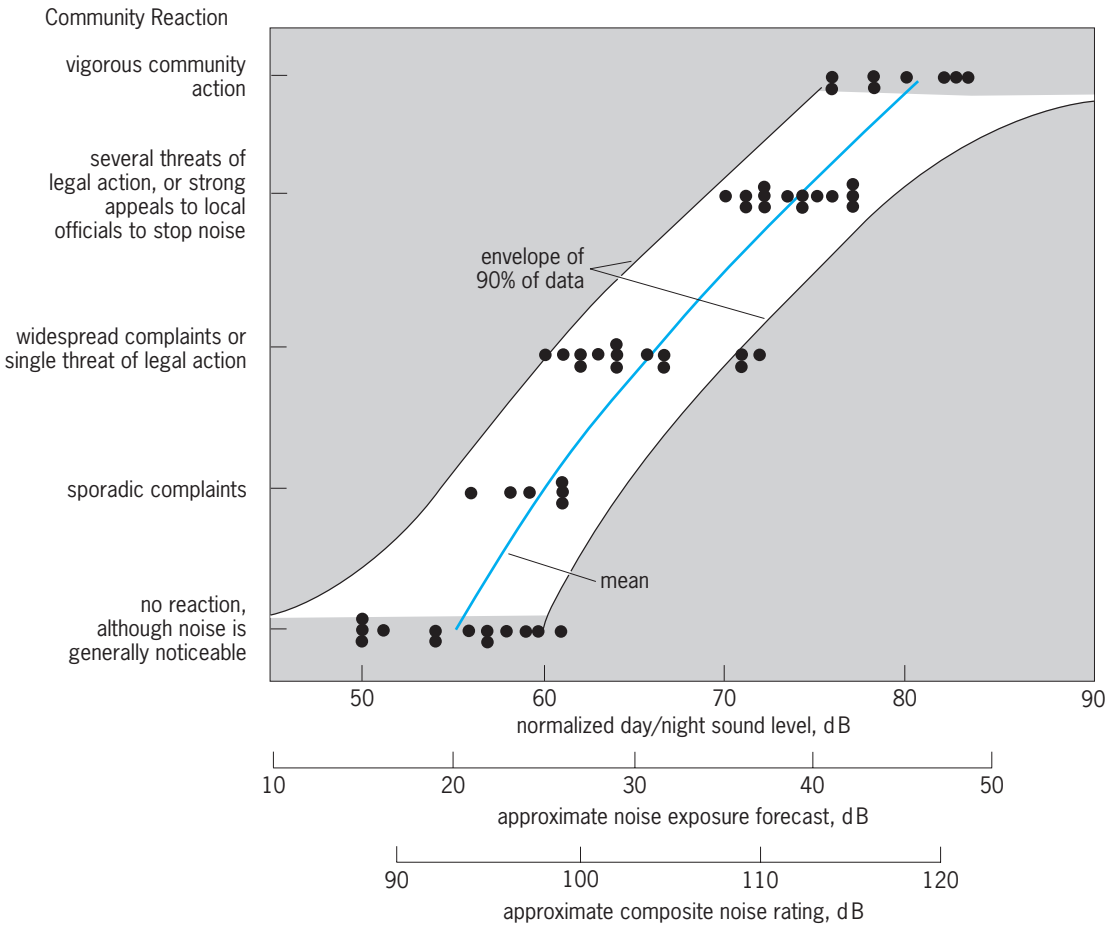


Fig. 7. Community reaction to many types of instrusive noise. Data normalized to residential urban noise, some prior exposure, windows partially open, no pure tones or impulses. (*After K. M. Eldred, Assessment of community noise, Noise Control Eng., 3(2):88–95, 1974*)

Absorbing materials. Although Eq. (8) is not accurate in rooms with low ceilings (such as factories), it can still be used for qualitative guidance with noise problems in irregularly shaped rooms. For example, in close proximity to a noise source, there is no reduction in sound pressure level L_p if the absorption coefficient of the materials on the walls is increased since the term $Q/4\pi r^2$ dominates; thus, it is not feasible to help the operator of a machine by using absorbing materials. However, far from a noise source in a room, a reduction of the reverberant noise level is achieved by increasing the absorption of the room walls. The absorption coefficient of a material α is defined as the fraction of incident acoustic intensity which is absorbed. Acoustic absorbing materials are usually made of porous materials such as fiberglass or open-celled foams. In environments where oil, water, or dirt can clog the pores of an absorbing material, a very thin impervious sheet of plastic may be placed over the absorbing material without substantially altering its sound-absorbing properties. The sheet (**Fig. 8**) increases the absorption at low frequencies but reduces it athigh frequencies.

Enclosures. Noise may be reduced by enclosure of the source. If it is essential to have continuous access to a noise source such as a machine used in mass production, or if cooling is necessary for a machine, a partial enclosure must be used. However, in the latter case noise leakage can be minimized by providing cooling vents, built from bent ducts, lined with absorbing material, and supplied with air from cooling fans, if necessary.

Transmission loss. The transmission less TL in dB of a partition is defined by Eq. (9), where τ is the fraction of incident acoustic intensity transmitted by the

$$\mathrm{TL} \approx 10 \log(1/\tau) \tag{9}$$

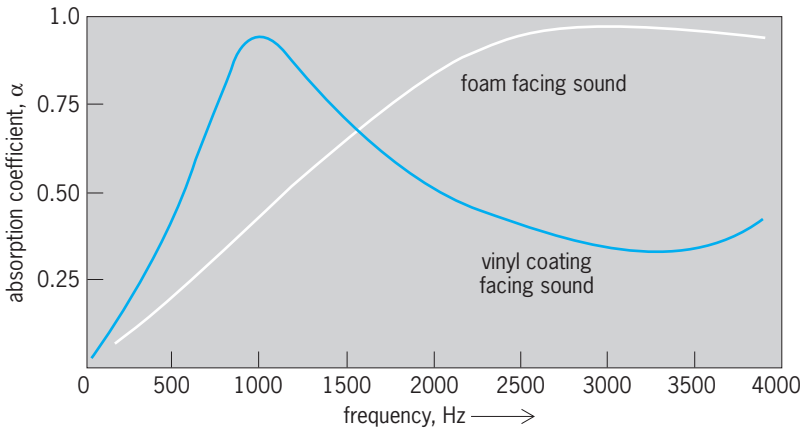


Fig. 8. Normal incidence absorption coefficient of a $^3/_4$-in.-thick (19-mm) porous foam material with vinyl coating facing sound and with foam facing sound.

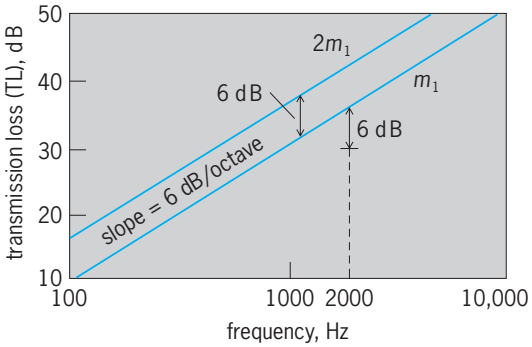


Fig. 9. Random incidence transmission loss for walls with mass per unit area m_1 and $2m_1$.

wall. For random incidence sound, the transmission loss in dB of an enclosure wall, a wall in a building, or even an airplane cabin wall is given empirically by Eq. (10), where m is the wall mass per unit area

$$\mathrm{TL} \approx 20 \log mf - 48 \tag{10}$$

in kg/m^2. **Figure 9** shows a plot of Eq. (10). Enclosure walls are almost "transparent" to low-frequency sound, but for every doubling of frequency the transmission loss increases by 6 dB. Transmission loss also increases by 6 dB at any given frequency if mass per unit area is doubled. These increases are not always observed in practice; an increase of about 4 to 5 dB per octave is often found. If massive walls are used in buildings to obtain a high transmission loss, care must be used to prevent leaks since these will seriously reduce the sound insulation of the wall. Two walls with an air gap usually have a greater transmission loss than one wall of the same total mass.

Other measures of performance. Sometimes measures other than transmission loss are used to rate the effectiveness of enclosures, walls, or even mufflers. The noise reduction (NR) of an enclosure is the sound pressure level L_{p1} measured inside minus the sound pressure level L_{p2} at some point outside the enclosure, Eq. (11). The insertion loss (IL) of the enclosure

$$\mathrm{NR} = L_{p1} - L_{p2} \tag{11}$$

is the difference in sound pressure levels at some point with $[L_p(w)]$ and without $[L_{p(wo)}]$ the presence of the enclosure, Eq. (12).

$$\mathrm{IL} = L_{p(wo)} - L_{p(w)} \tag{12}$$

Although Eqs. (10), (11), and (12) have been defined for enclosures, very similar definitions can be used for walls in buildings and mufflers. In general, transmission loss, noise reduction, and insertion loss are not equal. Sometimes transmission loss is averaged throughout the important audible frequency range 125–4000 Hz to give a single-number rating of a wall. A more complicated scheme in which the transmission loss–against–frequency curve is compared with a set of standard curves yields a single number rating called sound transmission class (STC).

Vibration isolation. If a machine is rigidly attached to the floor or supporting structure of a building or vehicle, the structure or floor can act in a manner similar to the sounding board of a musical instrument, and radiate large amounts of noise. This problem can be overcome by placing vibration isolators between machines and their supports. Such isolators may be metal or elastomeric springs.

Forces on machines at discrete frequencies can be caused by magnetic forces (usually at 120 Hz and integer multiples in the United States), and out-of-balance forces in rotating machines such as engines, motors, pumps, and fans. Vibration isolators should be designed so that the natural frequency of the machine on the isolators is very much less than any such forcing frequency. Such isolators will transmit much smaller forces to a support than those which act on the machine itself. However, isolators with this property generally must be very soft, and the machine may undergo static deflections when placed on them, resulting in misalignment of parts.

When a machine is started up, the forcing frequency increases and coincides briefly with the natural frequency before exceeding it. When the frequencies coincide, the machine will have a large vibration amplitude and transmit a large force to its support. To reduce this problem, during start-up, and during stopping, some damping is usually provided in isolators. Vibration and consequently noise can also be reduced by the application of viscous damping materials to the structures. *See* MECHANICAL VIBRATION; VIBRATION; VIBRATION DAMPING; VIBRATION ISOLATION.

Barriers. Barriers are widely used in industry and alongside roads and railways to shield receivers from noise sources. For a given geometry, the insertion loss of a barrier depends on frequency, and can be calculated approximately from **Fig. 10** by first determining by Fresnel number N, given by Eq. (13). Here λ = sound wavelength [Eq. (1)], $a + b$ = shortest path length of the sound wave over the barrier (Fig. 10), and d = straight-line distance between source S and receiver R. The positive sign is for the receiver in the shadow zone (receiver able to see source) and the negative sign for the receiver in the bright zone (receiver unable to see source). Figure 10 shows the theoretical insertion loss of a barrier as a function of N for a point source, under the assumption that the barrier is infinitely long and impervious to sound waves. If the sound source is composed of a line of sources (for example, closely spaced road vehicles), then the insertion loss is changed somewhat.

$$N = \pm(2/\lambda)(a + b - d) \tag{13}$$

Obviously barriers for any given geometry are more effective at high frequencies, when λ is small, and are ineffective at low frequencies, when λ is large, and thus N is small. For any particular frequency, the insertion loss of the barrier can be made larger by increasing $a + b - d$, or equivalently by placing the barrier as near the source or receiver R as possible. If barriers are used in low factory-type buildings, the barrier effectiveness can be reduced by sound reflections from the ceiling and walls near the barrier, but this can be mitigated by adding absorbing materials to such ceilings and walls. Barriers

will not reduce the noise on the receiver side, but will increase it, unless the barrier is also covered in absorbing material. If the source must be seen, a transparent barrier (made of Plexiglas) can be placed between a machine and operator.

Mufflers. Mufflers are used to reduce the sound from systems containing a noise source connected to a pipe or duct system such as air-conditioning systems, fans, and industrial blowers, gasoline and diesel engines, compressors, and jet engine inlets and exhausts. There are two main types of mufflers, reactive and dissipative. Reactive mufflers are usually composed of several chambers of different volumes and shapes connected together with pipes, and tend to reflect the sound energy back to the source. They are essentially sound filters. Dissipative mufflers are usually composed of ducts or chambers which are lined with acoustic absorbing materials that absorb the acoustic energy and turn it into heat. Some mufflers are a combination of reactive and dissipative types. The type of muffler selected for any particular application will depend upon the noise source to be silenced and several environmental factors.

Reactive mufflers. Reactive types (**Fig. 11**) are most useful when the noise source to be reduced contains pure tones at fixed frequencies or when there is a hot, dirty, high-speed gas flow. Reactive mufflers for such purposes can be made quite inexpensively and require little maintenance. Such mufflers lose their effectiveness when used with large-diameter ducts and at high frequencies (that is, if 0.8 of the sound wavelength is less than the greatest diameter or lateral dimension of the muffler) due to the formation of lateral waves or cross modes in the muffler.

Dissipative mufflers. Dissipative types are useful when the source produces noise in a broad frequency band. They are particularly effective at high frequencies, but special precautions must be taken if the gas stream has a high speed and temperature and if it contains particles or is corrosive. If the speed of the gas stream is above about 15 m/s, the absorbing material (fiberglass or rock wool) should have surface bonding to prevent damage. At high speeds (up to 300 ft/s or 100 m/s), facing materials such as wire screens or perforated metal sheets are needed to prevent erosion of the absorbing material. If the gas stream has a high temperature (up to 1000°F or 550°C), materials such as Corten with special paints or stainless steel may be used for the facing material. Contamination of the absorbing material with oil, water, and dirt may be prevented by using a very thin surface sheet of impervious plastic material such as Mylar. The thin sheet causes a slight loss of acoustic absorption at high frequency but an increase at low frequency. Both parallel baffle and blocked line-of-sight mufflers are used. The latter tend to be better particularly at high frequency, although construction is more complicated.

Measures of performance. As with enclosures, the insertion loss, transmission loss, and noise reduction of a muffler can be defined and measured. The first two quantities are usually more useful. In general, the three quantities are not necessarily equal, although, in special cases, insertion loss can equal transmission loss. Insertion loss and transmission loss are sometimes loosely described interchangeably as attenuation. *See* MUFFLER.

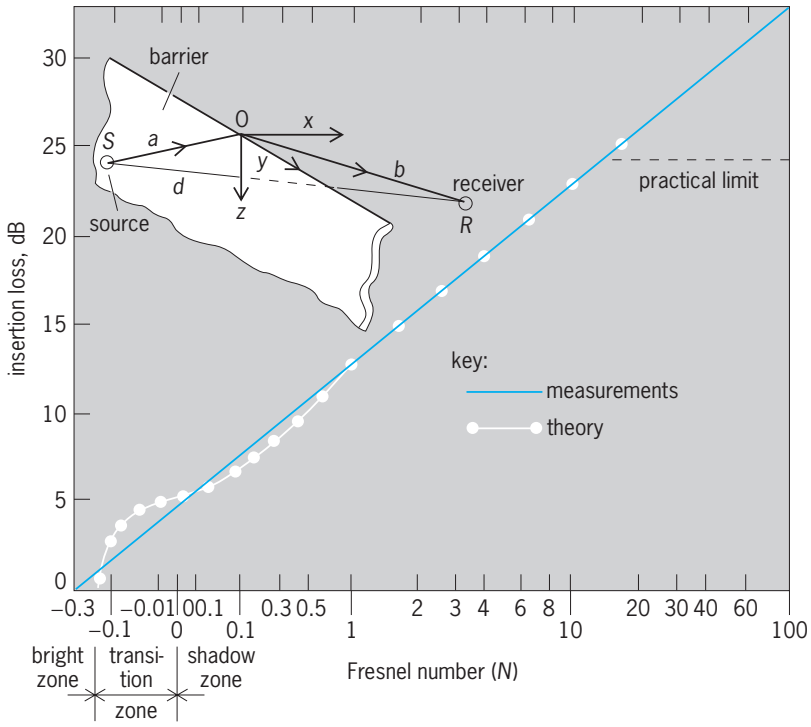


Fig. 10. Insertion loss of sound from a point source by a rigid barrier as a function of Fresnel number *N*. Negative *N* refers to the case where the receiver is able to see the source. (*After L. L. Beranek, Noise and Vibration Control, McGraw-Hill, 1971*)

Ear protectors. When all possible ways of reducing noise source strength and propagation along paths have been exhausted, the last resort is to protect the ear (the receiver). If a limited number of people are involved, this may sometimes be an attractive and economic approach (for example, to protect the sole operator of a large noisy machine). In many cases it is undesirable or impossible (for example, to protect thousands of residents living around airports).

There are three main types of ear protectors: earplugs that fit into the ear canal; earmuffs that fit over the external ear; and rigid helmets under which earmuffs or earplugs or both may be worn. There is some advantage in using more than one of these types simultaneously, although usually the attenuation is not strictly additive. Fortunately, both continuous unwanted sound and wanted sound are reduced by the same amount so that the ability to

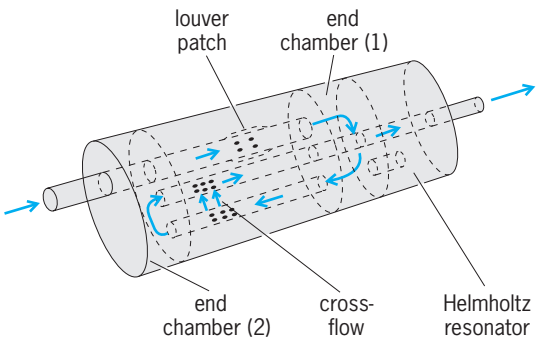


Fig. 11. Typical reverse flow automobile muffler.

hear warning signals and speech is not adversely affected. This is because the signal-to-noise ratio is not changed.

In cases where only one or two people are involved, as in monitoring instrumentation near a large noisy machine, it may be desirable to build a special acoustic enclosure around the operators. The booth should be built as far as possible from the machine, vibration-isolated and acoustically sealed, with forced ventilation if necessary. *See* EAR PROTECTORS.

Noise Sources

The principal sources of noise may be classified as surface transportation noise, aircraft noise, industrial noise, noise in the community from industrial and construction sites, and noise at home.

Surface transportation noise. There are several sources of this noise: road traffic, railroads, off-road recreational vehicles, ships, and hovercraft.

Road traffic has become the dominant source of noise annoyance in most industrialized countries. The power/weight ratio of road vehicles has been continually increased to permit higher payloads, greater acceleration, and higher cruising speeds, resulting in more powerful engines which are usually more noisy than lower-power ones. The main effect of traffic noise is annoyance caused by interference with speech sounds. Traffic noise can also interfere with sleep, although (except for heavy truck traffic) it tends to decrease at night and people tend to adapt to it. Attempts have been made to devise special noise measures such as the traffic noise index and noise pollution level to account for the annoying effects of fluctuations in level, in addition to the overall level, but these measures are not universally accepted.

In many countries, there are government regulations for car and truck noise. In the United States there are federal regulations for the maximum sound levels permitted from trucks. In addition, some cities and states have regulations for maximum sound levels for automobiles. For new vehicles, peak A-weighted sound level limits at a distance of 50 ft (15 m) from the center of the road during maximum acceleration are normally specified. These limits have been progressively reduced.

The major sources of community noise from vehicles are exhaust, cooling fan, engine, and, at high speed [above about 50 mi/h (80 km/h)], tires. Exhaust and fan noise can be reduced fairly easily; engine and tire noise are more difficult to reduce. Heavy diesel engine trucks have become the major source of traffic noise in the United States, producing about four times as much acoustic energy each day as the total fleet of automobiles. In planning new cities or new highways, care should be taken to route heavy traffic far from residential areas and to locate light and heavy industry nearer to such highways. The double glazing of windows and "soundproofing" homes in some European cities are probably equally as cost-effective as reducing the vehicle noise. Barriers can also shield residential areas from highway and railroad noise.

The interior noise for passengers of cars and rail systems is not normally a major problem. However, some truck drivers are subject to noise which is potentially hazardous if experienced for long periods.

Aircraft noise. Aircraft noise is a much more localized problem than surface transportation noise, since it occurs only around major airports. In the United States and many other industrialized countries, only about 25% as many people are seriously disturbed by aircraft noise as by road traffic noise. However, because it is localized and more easily identifiable, it seems to receive more attention. Most of the noise is produced by scheduled airlines; the contribution from the large numbers of light general-aviation aircraft is relatively small.

Jet airliners produce more community noise than did the early propeller airliners, particularly at high frequency. However, jet airliners using fanjets, which bypass a considerable proportion of the inlet air past the main compressor and combustion chambers, are about 10 or 15 dB quieter than the earlier pure-jet airliners because their design results in a lower exhaust velocity. Exhaust noise is proportional to the eighth power of velocity, so that reducing this velocity results in much lower exhaust noise. Radiation through the fan inlet and noise from the exhaust of the fan and the compressor are still problems in fanjet airliners. However, extensive use of sound absorptive liners in the fan intake and exhaust ducts has effectively decreased this noise also.

Measures such as perceived noise level (PNL) and A-weighted sound level [dB(A)] are usually used to monitor the noise of individual aircraft, and since 1969 the United States has had federal regulations which set noise limits for new passenger airliners. In order to describe the effect of aircraft noise near an airport, more complicated measures have been created to allow not only for the noise of each event (takeoff and landing) but also for the number of aircraft movements and the time of day when each occurs. In the United States the noise exposure forecast (NEF) is widely used; in Britain, the noise and number index (NNI) is used. The International Civil Aviation Organization (ICAO) has recommended the use of the equivalent continuous perceived noise level (ECPNL).

Interior noise levels in early jet aircraft were sometimes very high. These resulted from transmission of the engine noise exhaust and from fuselage wall excitation by boundary-layer pressure fluctuations. Modern fanjet airliners have much lower interior noise levels. However, interior noise levels of commuter propeller-driven passenger airliners remain intolerably high in many cases. These interior levels can be reduced by using larger slower-rotating propellers and locating them further from the fuselage.

Industrial noise. Industrial noise is a widespread problem in the United States and other industrialized countries. However, unlike traffic and aircraft noise which are mainly annoyance problems, exposure to industrial noise each day or over a period of years can cause permanent hearing damage. Probably about 5 million people in the United States have varying degrees of such damage.

TABLE 2. Permissible noise exposures for occupational noise in the United States

Duration per day, h	Sound level, dB(A)
8	90
6	92
4	95
3	97
2	100
1½	102
1	105
½	110
¼ or less	115

In 1969 the U.S. government created industrial noise regulations under the Walsh-Healey Act, and in 1971 these were extended to cover almost all workers under the Occupational Safety and Health Act (OSHA). For every 5-dB increase in A-weighted sound level, a halving in exposure time is allowed (**Table 2**). In most other countries, using energy considerations, the halving in exposure time occurs for a 3-dB(A) increase. In the United States a maximum peak overall level of 140 dB is allowed for impulsive noise.

Most metal-cutting, metal-forming, and woodcutting machines produce intense noise; many manufacturing industries are noisy. Noise reduction methods of enclosure, absorption, and vibration isolation and damping described previously can be used in many different cases; however, the cost of achieving significant noise reduction through such engineering means is very high.

Community site noise. In countries where cities are compact, factories have been built close to residential communities. Although community industrial noise can be reduced by some of the engineering methods already described, careful planning and zoning of new factories and cities constitute the best solution to such problems. There are no federal regulations governing exterior industrial noise in the United States, although some cities and states do have regulations. Some other countries have such regulations at the national level.

Construction noise is dissimilar to exterior industrial noise in that it is more temporary in nature and sometimes continues evenings, nights, and weekends. Several phases of construction are usually involved, including site clearing; demolition; excavation; placing of foundations; erection of floors, frames, walls, windows, and pipes; and finishing, filling, and paving. Often the initial and final phases of construction are the noisiest. Two different methods have been proposed to control construction noise: setting up acceptable noise limits at site or nearest residential boundaries; and specifying acceptable noise limits for each piece of equipment used. Both approaches are used in different countries, and a mixed approach is probably desirable in the control of both exterior industrial noise and construction noise.

Noise at home. Appliances used in and around houses can cause annoyance not only to the users but to others in adjoining rooms, apartments, and even separate houses. A few of the appliances that can cause noise include furnaces, plumbing, air conditioners, fans, water heaters, pumps, dishwashers, refrigerators, vacuum cleaners, blenders, mixers, electric razors, hair driers, saws, drills, sanders, typewriters, and lawnmowers. The problem can be reduced by manufacturers producing, and consumers purchasing, quieter appliances; by more carefully constructed houses which provide better sound insulation between walls and apartments; and by the use of acoustic tiles, drapes, and carpets.

Malcolm J. Crocker

Flow Noise

It is the ear's sensitivity to pressure variation that causes the sensation of sound. When those variations propagate through the air at the speed of sound, they are acoustical waves. But when the pressures heard by the ear are unsteady only because the ear is near to and buffeted by rough flow, then the ears sense flow noise, distinct from sound in its inability to propagate away from its turbulent source.

Unsteady flow always involves pressure fluctuations whose gradients produce the forces that accelerate the fluid particles. At large scale the motions tend to evolve slowly, as in the weather, but in smaller flows they have a shorter life, particularly when the flow is fast. The ratio of flow speed to the length scale of the eddying flow is the characteristic frequency of flow noise. The flow in a round jet of speed u and diameter L is very rough and unsteady, the associated pressure variations being flow noise. In fact, that noise is most energetic at frequency $0.3u/L$, and the buffeting effects on nearby parts of an aircraft structure can cause structural failure by acoustical fatigue.

A body moving through the air carries with it a thin boundary layer where the adjacent air tends to move with the body. Often that boundary layer is turbulent, and the rough unsteady pressures within and near that turbulence are another instance of flow noise. Such unsteady pressures cause vibration of the body structure, and that in turn causes noise in the interior, which can be a problem for passengers close to the fuselage surface in an aircraft; they hear the flow noise of the boundary layer. Similarly, the surface of a submarine is buffeted by the flow noise of its surrounding boundary layer, and that interferes with the sensitive sonar equipment used for underwater navigation. The same kind of flow noise is heard inside an automobile when the windows are open, or even on a rapidly moving bicycle as the wind causes flow noise at the ear. *See* BOUNDARY-LAYER FLOW; TURBULENT FLOW.

The term flow noise is sometimes used for all noises generated by flow processes, even when that noise is organized in sound waves that propagate away from the flow. Indeed it is often very difficult to distinguish between the local field where unsteady pressure is simply a reflection of the local unsteady flow, and the acoustic elements that escape. Sometimes the term pseudosound is used for flow noise, on account of its inability to propagate as sound. Whenever the flow speed u is comparable with the sound speed c, the distinction is very hard to

maintain; but in many common flows of small characteristic Mach number, that is, where u is much smaller than the speed of sound, the coupling between flow and sound is very weak. The sound proper is then very small, with most of the buffeting noise being flow noise whose characteristics are only weakly influenced by acoustic properties. *See* MACH NUMBER.

A microphone in a wind senses flow noise that is localized to the windy region. The compressibility of air is unimportant to that noise, and flow noise does not signify any energy or power transmission. Acoustic motions are ones that depend crucially on compressibility, and this is possibly the clearest distinction that can be made between the two kinds of noise. *See* SOUND. John E. Ffowcs-Williams

Bibliography. J. Anderson and M. Brates-Anderson, *Noise: Its Measurement, Analysis, Rating, and Control*, 1993; A. Barber (ed.), *Handbook of Noise and Vibration Control*, 6th ed., 1992; L. L. Beranek (ed.), *Noise and Vibration Control Engineering*, 1992; M. J. Crocker, A. J. Price, and F. M. Kessler (eds.), *Noise and Noise Control*, vol. 1, 1975, vol. 2, 1982; A. P. Dowling and J. E. Ffowcs Williams, *Sound and Sources of Sound*, 1983; C. M. Harris (ed.), *Handbook of Acoustical Measurements and Noise Control*, 4th ed., 1998; R. S. Jones, *Noise and Vibration Control in Buildings*, 1984; M. P. Norton, *Fundamentals of Noise and Vibration Analysis for Engineers*, 1990; H. K. Pelton, *Noise Control Management*, 1992.

Acoustic phonetics

The discipline of acoustic phonetics can be narrowly defined as the study of the acoustic output of the vocal tract for speech, but ultimately it encompasses much more. Acoustic phonetics makes direct contact with, and in many cases is the foundation of, areas of study such as speech synthesis, automatic (computer) recognition of speech, speech perception, phonological analysis and theory, and speech pathology.

Source-filter theory. There is a well-articulated theory of acoustic phonetics, the source-filter theory, where the source is the input to the system and the filter is the resonator. It is easiest to introduce the theory by first discussing vowels, the kinds of speech sound for which the theory is best known and for which it makes the most precise predictions.

Source. The source is the energy produced by the vibrating vocal folds (**Fig. 1**), which are set into motion by a combination of muscular forces that close the vocal folds as well as raise the pressure in the trachea relative to the pressure above the vocal folds. When the tracheal pressure overcomes the resistance offered by the closed vocal folds, the latter are blown apart and set into oscillatory motion that will continue, by virtue of a delicate interplay of aerodynamic and mechanical forces, as long as an adequate pressure differential exists. The nearly periodic vibration of the vocal folds can be characterized in either the time domain (as a waveform) or frequency domain (as a spectrum); most models of the vibration focus on parameters of the time-domain representation and their relationship to the spectral composition of vibratory energy. Vibration of the vocal folds generates a consecutive-integer harmonic spectrum, with greatest energy at the fundamental frequency (F0), or lowest-frequency component, and decreasing energy across the higher harmonics at the frequencies 2F0, 3F0, 4F0, 5F0, . . . ,nF0. The time-domain models show how subtle changes in the temporal parameters result in systematic frequency-domain changes. In certain cases, the periodic output of the vibratory motion may be mixed with variable degrees of aperiodic energy, as in more breathy types of voice. Thus the frequency-domain representation, also called the glottal source spectrum, can be associated very directly with the perception of voice quality (encompassing harshness, breathiness, and so forth). *See* HARMONIC (PERIODIC PHENOMENA); VIBRATION.

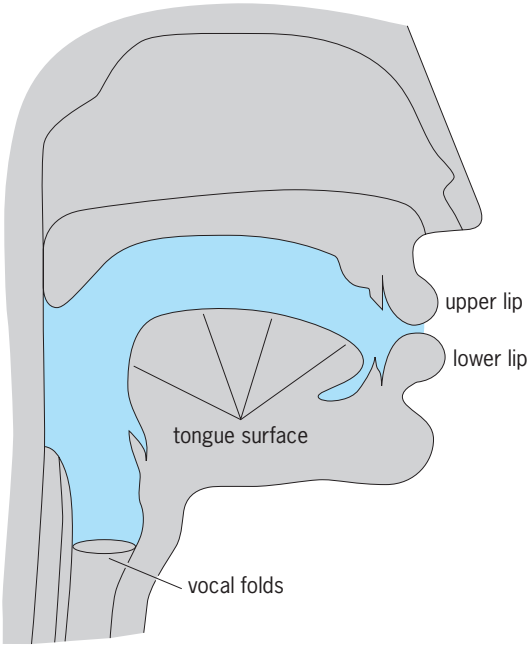


Fig. 1. Midsagittal schematic view of vocal folds (source) and vocal tract (filter, colored area).

Filter. The glottal source spectrum can be considered as the input to the filter, which in anatomical terms is the vocal tract. The vocal tract is the length of airway running from the vocal folds, in the larynx, to the lips (Fig. 1). The vocal-tract filter has the characteristics of a resonating tube, with one end closed. The closed end is at the vocal folds because the acoustic energy generated there excites the vocal-tract filter each time the vocal folds snap shut during consecutive vibratory cycles. The open end is at the lips because vowel sounds are produced with the mouth open to varying degrees, depending on the vowel. A tube resonator with a closed end and uniform cross-sectional area from end to end has multiple resonant frequencies determined by the tube's length. For such a tube having a length of 17 cm (6.7 in.),

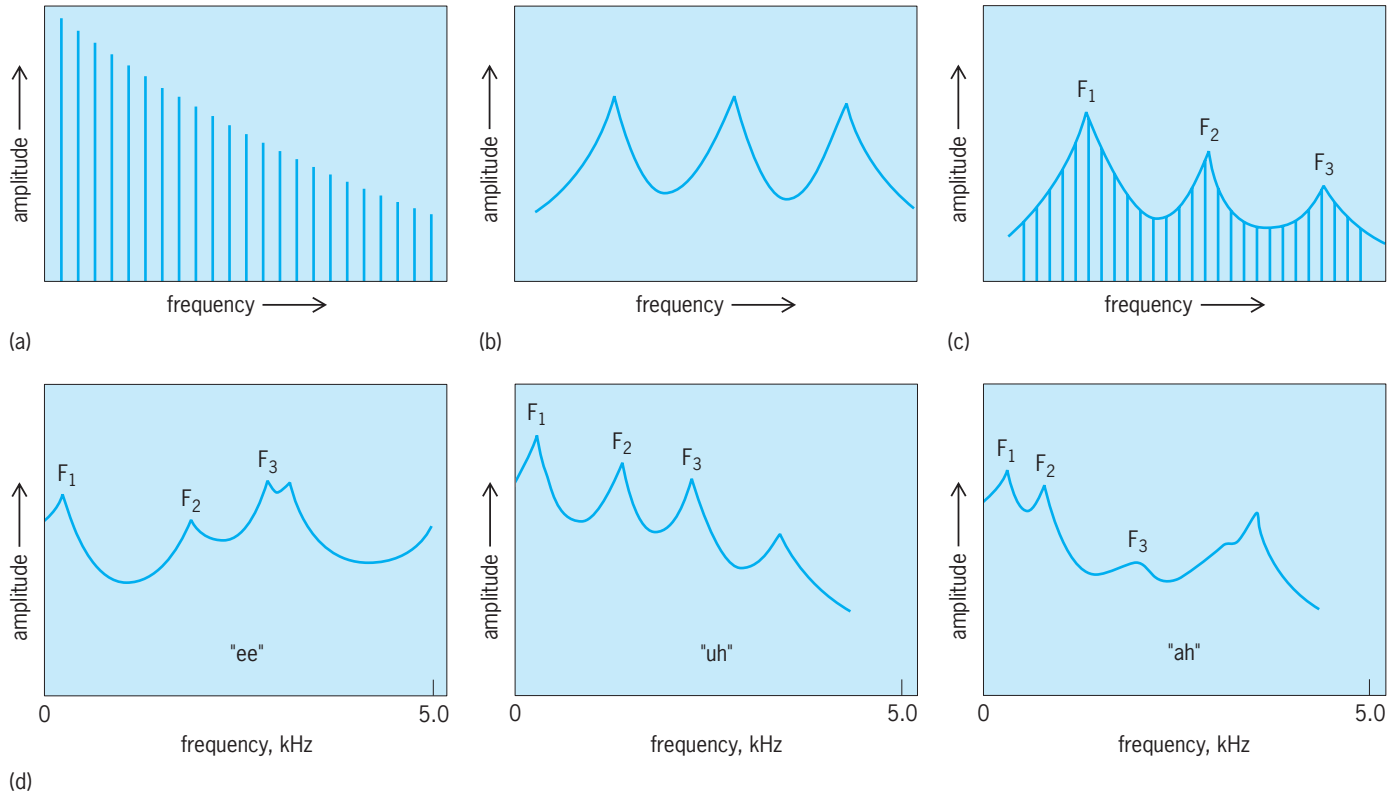


Fig. 2. Production of output spectrum of the vocal tract. (*a*) Source spectrum produced by vibrating vocal folds. (*b*) Resonance curve (filter) of vocal-tract tube having uniform cross-sectional area. (*c*) Output spectrum from combined source and filter. (*d*) Output spectra for vowels "ee," "uh," and "ah."

a typical length for the adult male vocal tract, the resonant frequencies are determined by the quarter-wavelength rule, given by the equation below,

$$f_r = \frac{(2n-1)\,c}{4l}$$

where f_r is a resonant frequency, $n = 1, 2, 3, \ldots$, c is the constant speed of sound in air (35,400 cm/s), and l is the length of the tube. This formula gives an odd-integer series of resonant frequencies at approximately 520 Hz, 1560 Hz, 2600 Hz, and so forth. These resonant frequencies can be thought of as peaks in a theoretical spectrum (where "theoretical" means calculated from theory, rather than being measured empirically), which have noninfinite amplitude as a result of both energy loss within the vocal tract and sound radiating from the lips. The overall shape of the theoretical spectrum can be thought of as the resonance curve or filter function of the vocal tract.

Output signal. The filtering effect of the vocal tract can be thought of in another way, that of the vocal-tract filter function shaping the source harmonic energy generated by the input, or vibrating vocal folds. The output signal thus represents the combined effect of the source and filter (**Fig. 2***a–c*). The peaks in the output spectrum are referred to as formants, and the frequency locations of the first three formants (F1, F2, and F3, the first three resonances of the tube) are a function of the configuration of the vocal-tract tube, which is to say of the vowel articulation. The resonant frequencies predicted by the quarter-wavelength rule—when the tube has a uniform cross-sectional area from end to end—are changed in systematic ways that depend on the location of constrictions relative to the standing-wave patterns of pressure and velocity within the tube. Thus a vowel such as the "ee" in the word "heat," for which the vocal tract tube shape is constricted toward the front (that is, toward the lips) and wide open toward the back (that is, toward the vocal folds) is associated with a resonance curve, and therefore with an output spectrum, substantially different from the case of the tube without any constrictions (as for the first vowel "uh" in the word "about"); on the other hand, a vowel like the "ah" in "father" constricts the tube toward the back and leaves the front wide open, yielding a different kind of deviation from the quarter-wavelength resonances of an unconstricted tube (Fig. 2*d*).

Although a tube resonator has, in theory, an infinite series of resonance peaks, only the first three formant frequencies have critical phonetic importance in the sense of distinguishing acoustically among the vowels of a language. Not surprisingly, these same three formant frequencies have been shown to be of critical importance in the perception of vowel categories, although in some cases the auditory system may not treat the individual formant frequencies as separate pieces of acoustic information. *See* SPEECH PERCEPTION.

Consonants and nasals. The model of a source spectrum shaped by a vocal-tract filter or resonator also

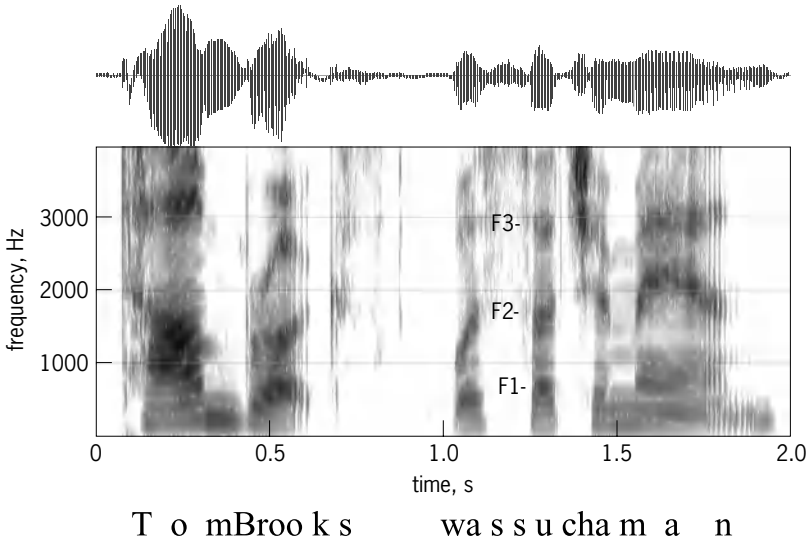


Fig. 3. Spectrogram of the utterance, "Tom Brooks was such a man." Intensity is coded by the gray scale. The speech waveform (amplitude over time) is at the top of the plot. Letters beneath the plot show the approximate locations of the sounds. First three formants, F1, F2, and F3, are indicated for the vowel "uh" in "such."

applies to consonants such as stops (such as p, b), fricatives (such as s, z), and affricates (such as ch, j), but in these cases the sources are likely to involve aperiodic (inharmonic) spectra generated by impulse or turbulent air flows. For some sounds (such as b, z, j) aperiodic sources within the vocal tract are combined with the periodic vocal fold source described above. Another class of sounds, the nasals (such as m, n), require a theory for parallel resonating tubes where one of the tubes has a dead end and therefore traps energy at its resonant frequencies. In this case the theoretical and empirical spectrum will have reverse peaks or antiresonances, indicating frequencies at which the combination of source and filter results in little or no output energy. Antiresonances are also found in the spectra of the stops, fricatives, and affricates.

Properties of speech sounds. The last half-century has seen an enormous amount of research on the temporal and spectral properties of speech sounds. This work has been critical to the design of speech synthesizers and algorithms for computer recognition of speech, and in some cases to the design of hearing aids and signal-processing schemes for cochlear implants; the data have also proven useful in understanding speech disorders. The standard representation of speech sounds in acoustic phonetics research is the spectrogram (**Fig. 3**), which shows the formant frequencies of the vocal-tract signal as a function of time, with intensity coded in a gray or color scale. Measurement conventions have been established for the correspondence between temporal intervals and sound categories, as well as for the spectral characteristics of speech sounds. A large amount of data has been generated for naturally spoken vowel and consonant durations and spectra, as well as the characteristics of the vocal fold vibratory source, and these have been used to design high-quality speech synthesizers. The very high intelligibility of sophisticated speech synthesizers, whose algorithms are based on these measurements, is strong evidence for the high quality of acoustic-phonetics research. *See* ACOUSTIC SIGNAL PROCESSING; HEARING IMPAIRMENT; SPEECH RECOGNITION; VOICE RESPONSE.

Factors affecting speech sounds. A predictable characteristic of any spectral measurement of vocal-tract output is its dependence on the age and gender of the speaker. Because the vocal tract resonates like a tube closed at one end with tube length being a scaling factor for the resonant frequencies, the formant frequencies of a particular sound spoken by men, women, and children will be quite different as a result of their general size differences, which are mirrored by differences in vocal-tract length. More generally, the acoustic characteristics of any speech sound, including their durations and spectra, are affected by many factors including (but not limited to) age, gender, speaking rate, speaking style, dialect, linguistic stress (for example, *re*bel versus re*bel*), and phonetic context (the sounds surrounding the sound of interest). Thus, the acoustic characteristics of a given sound cannot be considered as fixed or templatelike. One challenge of research in acoustic phonetics is to account for the factors that can affect the acoustic representation of speech sounds, and to relate these factors to speech synthesis, automatic speech recognition, speech perception, phonological theory, and clinical application in the diagnosis and understanding of speech disorders. Sophisticated models of the vocal tract are available to complement models of vocal vibration, allowing careful experimentation on the specific factors that contribute to the acoustic variation of speech sounds in natural speech. *See* SOUND; SPEECH.

Gary Weismer

Bibliography. G. Fant, *Acoustic Theory of Speech Production*, rev. ed., Mouton de Gruyter, The Hague, 1970; K. Forrest et al., Statistical analysis of word-initial voiceless obstruents: Preliminary data, *J. Acous. Soc. Amer.*, 84:115–123, 1988; J. Hillenbrand et al., Acoustic characteristics of American English vowels, *J. Acous. Soc. Amer.*, 97:3099–3111, 1995; D. H. Klatt and L. C. Klatt, Analysis, synthesis, and perception of voice quality variation among female and male talkers, *J. Acous. Soc. Amer.*, 87:820–857, 1990; K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, 1998; B. H. Story, A parametric model of the vocal tract area function for vowel and consonant simulation, *J. Acous. Soc. Amer.*, 117:3231–3254, 2005.

Acoustic radiation pressure

The net pressure exerted on a surface or interface by an acoustic wave. One might presume that the back-and-forth oscillation of fluid caused by the passage of an acoustic wave will not exert any net force on an object, and this is true for sound waves normally encountered. (The sound power of a normal speaking voice is less than one-millionth of the electric power

of a 100-W light.) Intense sound waves, however, can exert net forces in one direction of sufficient magnitude (proportional to the sound intensity) to balance gravitational forces and thus levitate an object in air. These acoustic forces, generated by acoustic radiation pressure, are generally not of the magnitude to lift a table off the floor, although such is not inconceivable.

The nature of the acoustic radiation pressure forces was clarified in a simple experiment by G. Hertz and H. Mende in which an acoustic transducer in water sends a beam of sound at a water–carbon tetrachloride interface. These mutually immiscible liquids are chosen because their properties are matched in such a way as to permit the passage of the sound beam across the interface with minimal reflection. But as the sound intensity is increased, an observer will notice that the interface statically deforms toward the water. Moreover, if the transducer is immersed in the carbon tetrachloride and pointed at the interface, the interface will again deform toward the water, suggesting that the difference in acoustic energy densities between the two liquids—a direction-independent quantity—is the relevant factor.

The problem is somewhat more complex when considering the force on a small object rather than an interface, not only because the problem is three-dimensional, but also because the scattering from the object and the oscillating and translatory motion must be considered. Such a force is relatively easy to generate by using a standing acoustic wave field, as is found when sound of a certain resonance frequency bounces back and forth in an organ pipe. An acoustic standing wave field is characterized by regions of acoustic pressure maxima and minima. For rigid objects in air or liquid, the sound field pushes the object toward minima. But objects that have some significant compressibility as compared to the surrounding medium (such as some liquid drops, or small bubbles held in liquid) are pushed toward maxima.

Forces due to acoustic radiation pressure have been used to calibrate acoustic transmitters, to deform and break up liquids, to collect like objects or to separate particles (including biological cells) based on mechanical properties, and to position objects in a sound field, sometimes levitating the sample so that independent studies of the object's properties can be performed. Single bubble sonoluminescence phenomena depend on acoustic radiation forces to maintain a bubble in a zone while its substantial radial oscillations take place.

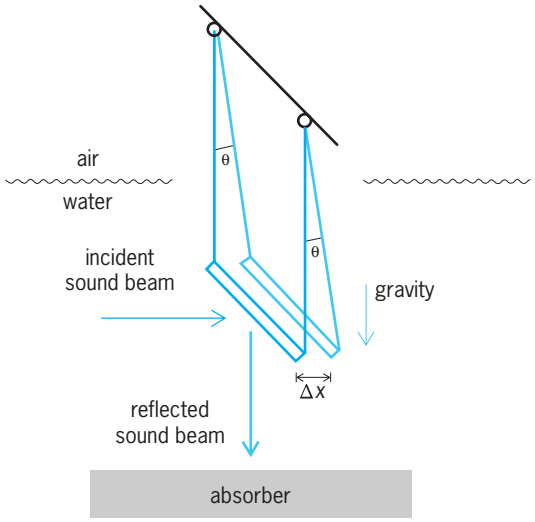
Radiation pressure of sound in air and in liquids has been used to suppress the capillary instability of cylindrical liquid surfaces that ordinarily causes long liquid columns to break into drops. Experiments in the 1970s demonstrated that low-frequency oscillations in fluid systems could be generated by modulating acoustic radiation pressure. This technique has been applied to imaging the low-frequency elastic response of tissue in the human body. *See* ACOUSTIC LEVITATION; SONOLUMINESCENCE; SOUND; ULTRASONICS.

Robert E. Apfel; Philip L. Marston

Acoustic radiometer

A device to measure the acoustic power or intensity of a sound beam by means of the force or torque that the beam exerts on an inserted object or interface. The underlying theory involves the concept of radiation pressure, which is the time-independent part of the pressure associated with a nominally sinusoidal acoustic disturbance. Such pressure occurs, for example, when a plane sound wave is partially reflected at an interface between two materials, with the nonlinear interaction between the incident and reflected waves giving rise to a steady pressure on the interface. If a narrow beam is incident on the interface and the transmitted wave is fully absorbed by the second material, the magnitude of the radiation force F (area integral of radiation pressure) equals a constant times W/c, where W is the power of the sound beam and c is the sound speed. The multiplicative constant is of the order of unity and depends on the thermodynamic properties of the fluid. The force acts in the vector direction of propagation of the sound wave. When the inserted object is of small size, the force is not given by as simple an expression, but it can nevertheless be predicted from basic principles. Thus, acoustic radiometers do not require external calibration and are themselves sometimes used in the calibration of transducers.

Vane radiometer. A modern design of an acoustic radiometer, used to measure the total power of an ultrasonic sound beam in water and other liquids, employs a vane suspended in the fluid in such a manner that its displacement in a direction normal to its face is proportional to the net force pushing on its front face. The vane is ideally of dimensions somewhat larger than the incident beam's diameter, so that the encountered force is associated with the entire incident beam. To eliminate the possibility of sound being reflected toward the transmitting transducer, the vane is oriented at 45° to the incident sound



Acoustic radiometer, used to measure the acoustic power of a sound beam in water. The vane is suspended from a pendulum whose rotation is opposed by gravity. Deflection Δx of the vane is proportional to the force on the vane.

beam. The vane's horizontal displacement is made to be proportional to the imposed force by fastening the vane at one end of a long pendulum (see **illus.**) whose rotation from the vertical is opposed by the effect of gravity, such that the apparent spring constant for displacement in a direction at 45° to the face is approximately Mg/L, where M is the apparent mass of the vane (corrected for the presence of water), g is the acceleration of gravity, and L is the length of the pendulum. A nonlinear acoustics theory for such a circumstance yields a proportionality relation between the net horizontal radiation force on the vane and the acoustic power associated with the incident beam. Because the deflection of the vane is proportional to the radiation force, the acoustic power can be determined. *See* BUOYANCY; PENDULUM.

Rayleigh disk. The concept of the vane device evolved from that of the Rayleigh disk, which was a circular disk that could rotate about its diameter and whose deflection from a nominal 45° orientation was opposed by a torsional spring. The Rayleigh disk was taken to have a radius much smaller than the wavelength, and its use ideally yielded a measurement of the local acoustic intensity that would have existed at the center of the disk were the disk not present. This is in contrast to the vane device, which is intended for use in circumstances when the vane's surface dimensions are large compared to a wavelength and the incident beam is entirely captured by the vane's surface. *See* ACOUSTIC RADIATION PRESSURE; SOUND; SOUND INTENSITY; SOUND PRESSURE.

Allan D. Pierce; Sameer I. Madanshetty

Acoustic signal processing

A discipline that deals generally with the extraction of information from acoustic signals in the presence of noise and uncertainty. Acoustic signal processing has expanded from the improvement of music and speech sounds and a tool to search for oil and submarines to include medical instrumentation; techniques for efficient transmission, storage, and presentation of music and speech; and machine speech recognition. Undersea processing has expanded to studying underwater weather and long-term global ocean temperature changes, mammal tracking at long ranges, and monitoring of hot vents. These techniques stem from the rapid advances in computer science, especially the development of large, inexpensive memories and ever-increasing processing speeds.

Sound can be said to be low and slow compared to computers. The audio frequencies are low, spanning roughly the range from 15 to 25,000 hertz, with seismic frequencies below this range and frequencies relevant to medical ultrasonics above. All of these frequencies are well within the limits of analog-to-digital input samplers and digital-to-analog output converters that form the bridges between continuous-time physical waveforms and the streams of digits handled by a computer. The speed of sound is slow, about 335 m/s (1100 ft/s) in air and 1500 m/s (5000 ft/s) in water, compared to light and radio waves at 3×10^8 m/s (10^9 ft/s), the upper limit for electrical signals in computer circuits. A computer can carry out thousands of elementary computations between input samples. *See* ANALOG-TO-DIGITAL CONVERTER; DIGITAL-TO-ANALOG CONVERTER; SOUND.

Time and frequency domains. Signal processors, both human and computer, "think" in two domains, two pictures of the signal world. One domain is the pictures of waveforms, the sound pressure as it changes in time. This time-domain picture is $s(t)$; t is time, and s might be the sound pressure or a microphone voltage. The other picture is the complex magnitude at every important frequency during an interval of time. This frequency-domain or spectrum picture is $S(f \mid T_n)$; f is the frequency in hertz, and T_n is the nth time interval. J. Fourier (1768–1830) formalized his series version of this picture and showed that $s(t)$ could be calculated from $S(f)$, and vice versa. Mathematicians have developed other versions of spectral transforms, each with its special area of application. *See* FOURIER SERIES AND TRANSFORMS; INTEGRAL TRANSFORM.

Each domain picture provides the same information, but sometimes it is easier to think, or compute, using one domain rather than the other. The computer version used is the DFT (discrete Fourier transform). Transforms were formerly time-consuming computations; for N points in one domain the time needed was proportional to N squared. In the 1960s a layered algorithm was developed to speed up the DFT. The emphasis is on sample sizes which are integer powers of 2. For example, 1024 is the tenth power of 2; it uses 10 layers, and the time factor dropped from 1024∗1024 to 10∗1024, over 100 times faster. These fast Fourier transforms have had a major impact on spectral processing.

Medical applications. Ultrasonic imaging of the unborn fetus is commonly used to observe growth and gender, as well as the positioning of multiple fetuses, and to warn of delivery problems. Similar techniques are used to view the heart in action, and to carry out breast examinations for cancerous nodules. Ultrasonic imaging devices typically use a pulse transmitter surrounded by a two-dimensional matrix of microphones. The array can be focused about a given depth by selecting the travel time (the time it takes the pulse to go from the transmitter to a specific location and back to a receiver). Tiny changes in travel time are manifested as signal phase changes, and at each depth the processor focuses on each point in the plane by adjusting the received signal phases to add constructively. The Doppler effect is used to observe blood flow: blood flowing toward the sensors causes the reflected frequency to be higher than transmitted, and blood flowing away lowers the frequency. Doppler imaging needs long pulses to measure slight changes in frequency, so wideband frequency-modulated signals may be used to achieve the time resolution of short broadband pulses and the frequency resolution of longer-duration signals. *See* BIOMEDICAL ULTRASONICS; MEDICAL ULTRASONIC TOMOGRAPHY; ULTRASONICS.

The cochlea in the inner ear is nature's time-domain–to–frequency-domain transformer, exciting

a line of nerve endings in response to something like the short-time spectrum of the input. For people in whom this natural mechanism is dysfunctional but the cochlear nerve is intact, a technology for restoring hearing is developing based on modest DFT analysis of sound near an ear, feeding perhaps 32 channels to an implanted microconnector in the cochlear nerve bundle. *See* HEARING (HUMAN); HEARING AID; PHYSIOLOGICAL ACOUSTICS.

Hearing aids can make use of miniaturized electronics to provide tailored spectral gain and overall dynamic gain adjustment to keep the total signal amplitude at a suitable level. *See* HEARING AID.

A binaural "eyeglasses sonar" for blind persons who have normal hearing was developed in New Zealand in the 1970s. With a transmitter in the bridge and receivers on either side, the beam moves naturally with head motion. The signal is a periodic, octave-wide, ultrasonic frequency sweep. The signal processing removes the non-information-bearing sounds due to the periodicity of the signal, and frequency translates the signal to the normal hearing frequencies. It is a sensitive motion detector. The intelligent signal processing is done by the wearer, whose head movement and walking pace cause the auditory sensations that contain the information about the wearer's surroundings that would be provided by vision for a sighted person.

Applications to rooms and halls. Audio filters are the hardware that adjust the gain across the spectrum to emphasize some frequencies and deemphasize (notch) others. Analog filters required heavy bulky components for low frequencies. Digital filters are programs in computer chips of negligible weight. Digital "equalizers" allow the user to adjust the spectral gain in 3 to 30 adjacent frequency bands to suit the listener. In churches and lecture halls that echo and emphasize resonant frequencies, fixed-frequency notch filters may be used; a dynamic system may use a fast Fourier transform to watch for frequencies that are too loud and apply appropriately deep notches, then remove the notches when the bothersome frequencies are no longer excited. For halls with distributed speakers, the signal can be written into a short memory and read out at desired delay times to different speakers; this is a natural solution for long narrow churches. *See* DIGITAL FILTER; ELECTRIC FILTER; EQUALIZER; SOUND FIELD ENHANCEMENT.

Transmission and storage applications. Part of information theory deals with removing all the redundancy in a processor input so that only essential, information-bearing bits remain, just enough to restore the original signal. Lossless algorithms yielding 2:1 compression for text and data are in common use. Lossy compression tolerates minor errors in reconstruction and may further reduce the number of needed bits. With the interest in sending speech, music, and video as well as written text over the Internet, algorithms were developed yielding 12:1 compression on such material. *See* DATA COMPRESSION; INFORMATION THEORY.

Speech recognition. One of the most difficult signal processing areas is machine recognition of spoken speech. Vocabulary size and number of speakers are key parameters. The fundamental problems are: (1) words are not spoken separately, but in streams of connected sound, (2) a phrase is seldom said the same way every time, (3) the vocabulary is enormous, and (4) there is a huge variety of speaker accents and rhythms. The signal processing must segment an utterance into phonemes, words, or phrases; pick out key features; or compare the whole segment with a library for likely matches.

Many telephone information systems now accept single digits spoken in English and perhaps Spanish. This avoids the first three problems. Also in use are user-specific algorithms that are trained separately with each user, with a specific library of commands. User-specific dictation algorithms with large vocabularies are under development. *See* SPEECH RECOGNITION.

Sonar. The ocean is transparent to sound and nearly opaque to light and radio waves. Even for sound, the attenuation increases exponentially as the frequency squared, so low audio frequencies must be used to transmit signals as distances increase. The speed of sound varies with temperature and pressure, causing sound to travel in paths that oscillate from near the surface to kilometers deep in the oceans. In most underwater situations the propagation is by multiple paths or by multiple modes, with many interfering sources. *See* SOUND ABSORPTION; UNDERWATER SOUND.

Input processing. The receiving array is the initial signal processor. A large number of hydrophones in a three-, two-, or one-dimensional array is used to focus a beam in both vertical and horizontal angles. Each beam former adds the inputs from every hydrophone after each is time-delayed to synchronize all inputs from the desired angle, and after differences in individual gains are compensated to form a clean beam. The gain in average signal-to-noise ratio is proportional to the number of hydrophones; however, ships, biological organisms, and storms are localized sources of acoustic noise, and therefore may be suppressed more than average by the beam pattern.

Postinput processing. Signal detection and estimation theory guide the processing of noise, interference, and their power levels as well as the uncertainty as to their exact description (stochastic and mechanistic) and the uncertainty in the received signal (which may also be part stochastic and part mechanistic). The ultimate goal in processing is to include the total reception as input, together with the propagation possibilities, noise, and interference, and to determine how likely each combination of these would be to cause the reception just observed. Signal detection, parameter estimation, and propagation learning proceed together, not sequentially. *See* ESTIMATION THEORY; STOCHASTIC PROCESS.

This approach was impractical in the past unless one met or assumed specific statistics. Now the speed, lower cost, and increased memory of computers make this viewpoint ever more applicable; what had been considered "random" may now be considered "complicated but manageable." This opens the

way to vastly improved performance at low input-signal–to–noise ratios. *See* SONAR.

Theodore G. Birdsall

Acoustic torpedo

An autonomous undersea vehicle that can be launched from submarines, surface ships, or aircraft to attack enemy submarines and surface ships. An acoustic torpedo is a sophisticated weapon. Its main components are a guidance and control system, a power plant to provide propulsive and electrical energy, a propulsor to control speed and direction, and a warhead. The launching platform performs the function of determining the approximate location of the target and launching the torpedo in the proper direction. The torpedo typically utilizes an acoustic sensor in its nose and is controlled by an on-board computer. During its operation, the torpedo searches the volume of the ocean determined by the launch platform. It progresses through the following phases: detection (an object is present), classification (the object is a target of interest), homing (steer at the object), and detonation of the warhead. *See* GUIDANCE SYSTEMS; HOMING.

Acoustics is the primary means of propagation in the undersea environment. The acoustic sensor enables the detection, classification, and homing functions. An electromechanical device converts acoustic (pressure) waves in the ocean to electrical signals that are processed. Similarly, electrical signals produced by the torpedo are converted to acoustic waves for active sonar transmissions. Because of their small size, torpedoes utilize high frequencies of operation. This results in short detection ranges, perhaps several hundred meters. *See* ACOUSTIC MINE; TRANSDUCER.

Acoustic torpedoes have two modes of sonar operation: passive and active. In passive sonar the torpedo listens for sounds emitted by the target. If the emissions are detected and classified relative to other sounds in the ocean, the torpedo enters into homing on the target. In active sonar the torpedo transmits acoustic pulses that reflect off objects in the environment. These reflections are processed on board the torpedo for detection, classification, and homing. *See* SONAR.

Limitations on torpedo performance are provided by three mechanisms: background noise provided by the flow of water over the sensor face; reverberation (the backscattering of acoustic energy from the ocean surface, bottom, and particles in the water column); and acoustic devices that are designed to deliberately produce noise to mask the target or to provide a decoy to divert the torpedo. These limitations are addressed through a combination of processing algorithms, strategies for tactical employment, and logic commands that are inserted into the on-board computer. *See* ACOUSTIC SIGNAL PROCESSING; ANTISUBMARINE WARFARE; ELECTRONIC WARFARE; REVERBERATION; UNDERWATER SOUND.

Frank W. Symons

Acoustical holography

The recording of sound waves in a two-dimensional pattern (the hologram) and the use of the hologram to reconstruct the entire sound field throughout a three-dimensional region of space. Acoustical holography, which first appeared in the 1960s in studies in ultrasonics, is an outgrowth of optical holography, invented by Dennis Gabor in 1948. The wave nature of both light and sound make holography possible. The objective of optical holography is to observe (reconstruct) three-dimensional images of the sources of reflected light (visible or nonvisible). Acoustical holography involves reconstruction of the sound field that arises due to radiation of sound at a boundary, such as the vibrating body of a violin, the fuselage of an aircraft, or the surface of a submarine. This reconstruction represents a solution to an inverse wave propagation problem explained heuristically using Huygens' principle. Both acoustical holography and optical holography rely on the acquisition of an interferogram, a two-dimensional recording at a single frequency of the phase and amplitude of an acoustic or electromagnetic field, usually in a plane. Gabor called this interferogram a hologram. *See* HOLOGRAPHY; HUYGENS' PRINCIPLE; INVERSE SCATTERING THEORY.

Farfield and nearfield holography. Two distinct forms of acoustical holography exist. In farfield acoustical holography (FAH), the hologram is recorded far removed from the source (in the Fresnel or Fraunhofer zones). This form of acoustical holography is characterized by the fact that the resolution of the reconstruction is limited to a half-wavelength. This resolution restriction is removed, however, when the hologram is recorded in the acoustic nearfield, an important characteristic of nearfield acoustical holography (NAH), invented by E. G. Williams and J. D. Maynard in 1980.

In all forms of holography the reconstruction of the field is based theoretically on a mathematically rigorous inversion of the Helmholtz integral equation. In practice, however, most applications compromise much of this mathematical rigor, especially in farfield acoustical holography. Nearfield acoustical holography treats the inverse problem much more rigorously than farfield acoustical holography, due to the inclusion of evanescent waves in the reconstruction process. Evanescent waves decay exponentially from the source surface, where they are generated. In order to capture them the hologram must be recorded very close to the source surface, that is, in the acoustic nearfield. The reconstruction provided by the solution of the inverse problem can be thought of as a backtracking of the sound field from the measurement surface to the source. This backtracking process is carried out in a computer which is able to increase exponentially the amplitude of the evanescent waves until their initial value at the source surface is reached. It is these waves that contain the high-resolution spatial information about the source. The most accurate reconstructions are obtained when the two-dimensional hologram is

larger than the object (**Fig. 1**) or when the hologram is a closed surface conformal to the surface of the vibrating object.

Using multiple reconstructions, a three-dimensional image of the sound field is obtained (**Fig. 2**). With this volume mapping, it is possible to derive the particle velocity, acoustic intensity, energy densities, and farfield directivity patterns. Besides the reconstruction of the pressure, nearfield acoustical holography is able to reconstruct the normal velocity on the surface of the object, providing detailed information about its vibrations—a key to unraveling its vibration and radiation mechanisms. Nearfield acoustical holography can also reconstruct sound fields within boundaries (for example, the noise inside an aircraft). In that case, the hologram is acquired in the interior, and multiple reconstructions provide the complete acoustic field in the interior as well as the velocity on the fuselage.

Nearfield acoustical holography has been used in the automotive industry to study interior noise and tire noise, in musical acoustics to study vibration and radiation of violin-family instruments, and in the aircraft industry to study interior cabin noise and fuselage vibrations. Applications are also found in underwater acoustics, especially in studies of vibration, radiation, and scattering from ships and submarines. *See* ACOUSTIC NOISE; MUSICAL ACOUSTICS; UNDERWATER SOUND.

Data acquisition and reconstruction. Typically, temporal acoustic data are acquired by measurement of the acoustic pressure with a single microphone or hydrophone, which scans an imaginary two-dimensional surface. This surface is most often a planar or cylindrical one, but surfaces of varied shapes have been used. In some cases, an array of microphones is used and the pressure is measured instantaneously by the array.

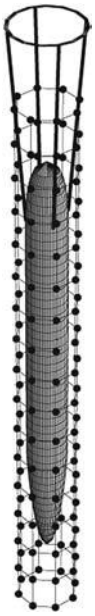


Fig. 1. Underwater hologram measurement system for the study of vibration, radiation, and scattering from air-filled steel shells. Diagram shows vibrating source and array of points marking measurement locations in the hologram. Hydrophone and and axial scanner are not shown.

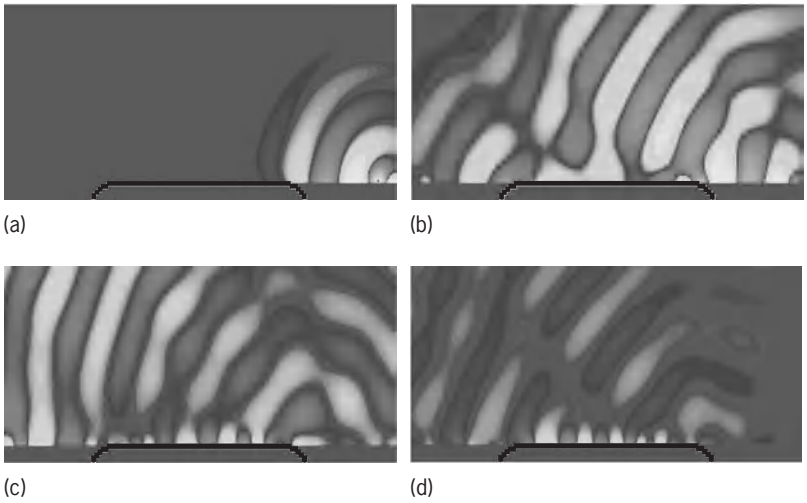


Fig. 2. Reconstruction, by means of nearfield acoustical holography, of the pressure field at four instants in time during the scattering of a wave pulse (traveling from right to left). Different grays indicate positive pressure, negative pressure, and ambient pressure. (*a*) 50 μs. (*b*) 350 μs. (*c*) 650 μs. (*d*) 950 μs.

Figure 1 shows an example of the measurement of a cylindrical hologram used for a coherent vibrator and scatterer underwater. In this case, the pressure field is sensed by a single hydrophone attached to an axial scanning device. In the circumferential direction, data are obtained by rotation of the vibrating object about its axis. At each data point in the scan, the pressure is recorded in a computer as it varies with time, along with a reference signal which provides a phase reference and amplitude calibration. Typical reference signals might be the force applied by a force generator exciting the object, the acceleration at a fixed point on the body, or the incident field pressure (for a scattering hologram).

The measured data are processed in a computer to reconstruct the pressure at the surface of the object as well as the vibration of the surface. In this processing, the measured time data is Fourier-transformed into the frequency domain, creating a set of holograms, one for each frequency bin in the transform. In the inversion process, each hologram is broken up into a set of waves or modes whose propagation characteristics are known from basic principles. Each wave or mode is then back-propagated to the source surface by multiplication by the known inverse propagator, and the field is then recomposed by addition of all these waves or modes. Evanescent modes beyond the dynamic range of the system are filtered out and are not included in this recomposition. *See* FOURIER SERIES AND TRANSFORMS.

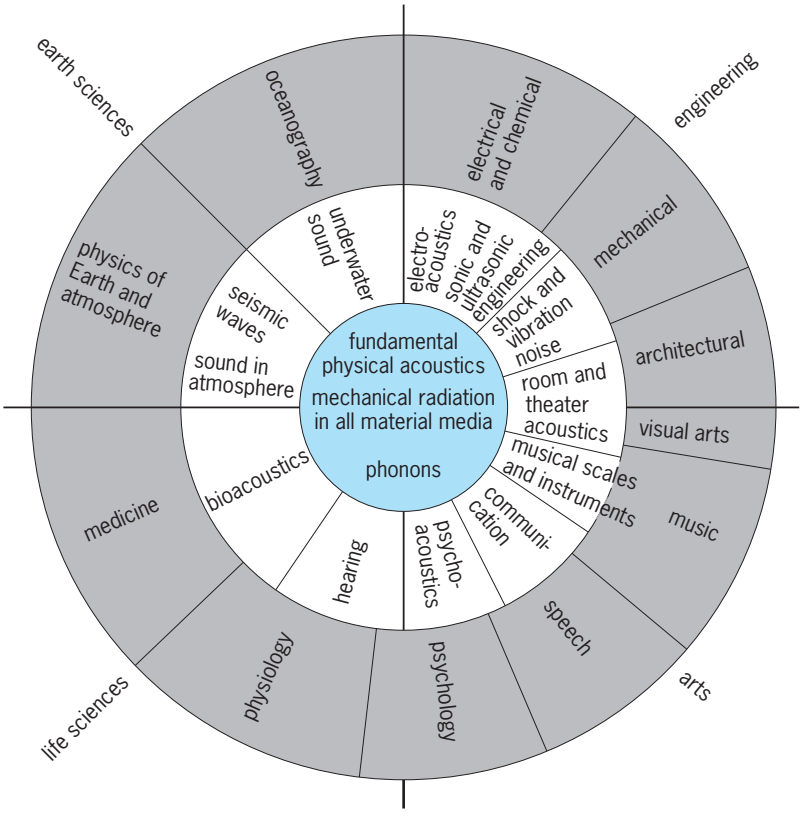
Two kinds of inverse propagators exist, one which reconstructs velocity and the other which reconstructs pressure. This whole process is carried out for each hologram (each frequency). If desired, the reconstructed holograms are inverse-transformed from frequency back to time. An example is shown in Fig. 2 for a hologram generated by the scattering

of a wave pulse (traveling from right to left) off a cylindrical, air-filled, steel shell submerged in a water tank. The four time snapshots show the reconstructed pressure field not only at the surface of the shell (drawn at the bottom of each snapshot) but also at 32 concentric cylindrical surfaces of increasing radius (32 separate reconstructions) put together to generate a planar view. Earl G. Williams

Bibliography. J. W. Goodman, *Introduction to Fourier Optics*, 2d ed., McGraw-Hill, New York, 1996; E. G. Williams, *Fourier Acoustics, Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, 1999.

Acoustics

The science of sound, which in its most general form endeavors to describe and interpret the phenomena associated with motional disturbances from equilibrium of elastic media. An elastic medium is one such that if any part of it is displaced from its original position with respect to the rest, as for example by an impact, it will return to its original state when the disturbing influence is removed. Acoustics was originally limited to the human experience produced by the stimulation of the human ear by sound incident from the surrounding air. Modern acoustics, however, deals with all sorts of sounds which have no relation to the human ear, for example, seismological disturbances and ultrasonics.



Relations among the various branches of acoustics and related fields of science and technology.

Basic acoustics may be divided into three branches, namely, production, transmission, and detection of sound.

Sound production. Any change of stress or pressure producing a local change in density or a local displacement from equilibrium in an elastic medium can serve as a source of sound. The human vocal mechanism is an obvious example. Other illustrations are provided by struck solids (as a drum, or violin or piano string), flow of air in a jet, and underwater explosions.

Modern acoustics as a science and technology has experienced enormous development by the invention of sources of sound which can be precisely controlled with respect to both frequency and intensity. The most important of these are called acoustic transducers, namely, devices by which any form of energy can be transformed into sound energy and vice versa. The most useful type is the electroacoustic transducer, in which electrical energy is transformed into the mechanical energy of sound. An example is the electrodynamic loudspeaker used in sound-reproducing systems, public address systems, and radio and television. Piezoelectric and magnetostrictive transducers are widely used in scientific and industrial applications of acoustics. An example of a nonelectric acoustic transducer is the siren, in which interrupted fluid flow results in the production of sound.

Sound transmission. Transmission of sound takes place through an elastic medium by means of wave motion. A wave is the motion through the medium of a disturbance as distinguished from the motion of the medium as a whole. An obvious example is a surface wave on water. Most sound waves of importance are transmitted compressional disturbances, that is, disturbances in which the pressure or density at any point in the medium is caused to vary from its equilibrium value. The change is then propagated through the medium with a velocity which depends on the type of wave in question, the nature of the medium, and the temperature. For example, the velocity of sound in still air of normal atmospheric composition at 0°C (32°F) is 331.45 m/s (1087.43 ft/s). Another important property of sound transmission is intensity, measured by the average rate of flow of energy in the wave per unit of time and per unit area perpendicular to the direction of propagation. The intensity of all practical sound waves diminishes with distance from the source, a property known as attenuation. This is due either to the spreading of the sound-wave energy over larger and larger surfaces or to actual absorption of the energy by the medium with transformation into heat.

The most important sound waves are harmonic waves, defined as waves for which the propagated disturbance at any point in its path varies sinusoidally with time with a definite frequency or number of complete cycles per second (the unit being the hertz). For a harmonic wave the disturbance at any instant repeats itself in magnitude and phase at intervals along the direction of propagation equal to

the so-called wavelength, which is, therefore, equal to the sound velocity divided by the frequency.

Acoustics deals with waves of all frequencies, but not all frequencies are audible by human beings, for whom the average range of audibility extends from 20 to 20,000 Hz. Sound below 20 Hz is referred to as infrasonic, and that above 20,000 Hz is called ultrasonic.

Sound detection. The detection of sound is made possible by the incidence of transmitted sound energy or an appropriate acoustic transducer. For human beings with so-called normal hearing, the most important transducer is the ear, a remarkably sensitive organ able to detect a sound intensity as low as 10^{-16} W/cm^2. For modern applied acoustics, transducers such as the microphone, based on the piezoelectric effect, are widely used. Generally speaking, any transducer used as a source of sound is also available as a detector, though the sensitivity varies considerably with the type.

Applications. The practical applications of acoustics are multifarious. They include architectural acoustics, or the study of sound waves in closed rooms arranged for the satisfactory production and reception of speech and music; underwater acoustics, dealing specifically with all aspects of sound in the sea and its use for the detection of vessels and the exploration of the sea bed; engineering acoustics, including the whole technology of sound reproduction and recording, sound motion pictures, and radio and television, as well as the study of vibrations of solids and their control and the use of high-intensity ultrasonics in industrial processing (for example, in metallurgy). Noise control is also an important case of engineering acoustics. Further examples are provided by physiological and psychological acoustics, dealing with hearing in humans and animals. Communication acoustics considers the production and transmission of speech. Musical acoustics deals with the physics of musical instruments. Bioacoustics and medical acoustics take up the use of sound in medical diagnosis and therapy as well as its use in the study of the overall behavior of animals in general.

All applications of acoustics are based on the fundamental physical properties of sound waves. The extensive inderdisciplinary is exhibited in the **illustration**. *See* ACOUSTIC NOISE; ARCHITECTURAL ACOUSTICS; ATMOSPHERIC ACOUSTICS; LOUDSPEAKER; MECHANICAL VIBRATION; MICROPHONE; MUSICAL ACOUSTICS; PHYSIOLOGICAL ACOUSTICS; PSYCHOACOUSTICS; SOUND; SOUND ABSORPTION; SOUND RECORDING; SOUND-REINFORCEMENT SYSTEM; SOUND-REPRODUCING SYSTEMS; ULTRASONICS; UNDERWATER SOUND; VIBRATION; VIBRATION DAMPING; VIBRATION ISOLATION. R. Bruce Lindsay

Bibliography. D. T. Blackstock, *Fundamentals of Physical Acoustics*, 2000; F. Fahy, *Foundations of Engineering Acoustics*, 2001; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., Wiley, 2000; A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, 1989.

Acoustooptics

The field of science and technology that is concerned with the diffraction of visible or infrared light (usually from a laser) by high-frequency sound in the frequency range of 50–2000 MHz. The term "acousto" is a historical misnomer; sound in this frequency range should properly be called ultrasonic. Such sound cannot be supported by air, but propagates as a mechanical wave disturbance in amorphous or crystalline solids, with a sound velocity ranging from 0.6 to 6 km/s (0.4 to 4 mi/s) and a wavelength from 3 to 100 μm. *See* LASER; ULTRASONICS.

The sound wave causes a displacement of the solid's molecules either in the direction of propagation (longitudinal wave) or perpendicular to it (shear wave). In either case, it sets up a corresponding wave of refractive-index variation through local dilatation or distortion of the solid medium. It is this wave that diffracts the light by acting as a three-dimensional grating, analogous to x-diffraction in crystals. The fact that the grating is moving is responsible for shifting the frequency of the diffracted light through the Doppler effect. *See* DIFFRACTION GRATING; DOPPLER EFFECT; REFRACTION OF WAVES; WAVE MOTION; X-RAY DIFFRACTION.

Since the 1960s, acoustooptics has moved from a scientific curiosity to a relevant technology. This evolution was initially driven by the need for fast modulation and deflection of light beams, and later by demands for more general optical processing. It was made possible by the invention of lasers, the development of efficient ultrasonic transducers, and the formulation of realistic models of sound-light interaction. *See* TRANSDUCER.

Bragg diffraction. When the sound propagates in a very wide, parallel beam that approximates a plane wave, the analogy with x-ray diffraction is most pronounced. The phenomenon is then called Bragg diffraction and is conventionally illustrated by the wave-vector diagram (**Fig. 1**).

The diagram shows an isosceles triangle, the sides of which are formed by the wave vectors $\mathbf{k}_0$ and $\mathbf{k}_1$ of the incident and diffracted light, and $\mathbf{K}$ of the sound. The directions of the wave vectors show which way sound and light propagate, while their lengths are proportional to the momenta of the corresponding field quanta, that is, photons and phonons. In quantum-mechanical terms, the wave-vector diagram represents conservation of momentum in a photon-phonon collision. The incident light photon ($\mathbf{k}_0$) collides with a sound phonon ($\mathbf{K}$) to create a

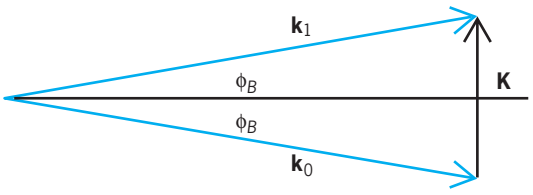


Fig. 1. Wave-vector diagram of acoustooptic interaction.

diffracted photon ($\mathbf{k}_1$). *See* CONSERVATION OF MOMENTUM; PHONON; PHOTON; QUANTUM MECHANICS.

Conservation of momentum in the collision requires that the arrows line up to form a closed triangle. This determines the incident angle ϕ_B, the Bragg angle. Its magnitude equals approximately $\lambda/2\Lambda$, where λ and Λ are the wavelengths of light and sound. Typical Bragg angles are small, ranging from 0.1 to 10°.

In addition to momentum, energy must be conserved in the collision. The energy of the colliding phonon is added to that of the incident photon to create the diffracted photon. As the energy of a quantum is proportional to its frequency, this process results in the frequency of the diffracted photon being upshifted by the sound frequency. This is the quantum-mechanical interpretation of the Doppler effect. *See* CONSERVATION OF ENERGY.

Bragg cell. In the wave-vector diagram, each arrow represents a plane wave carrying photons or phonons. In actuality, however, both sound and light propagate as beams of finite width rather than as infinitely wide plane waves. The finite size of the sound beam relaxes the requirement that the incident angle precisely equal the Bragg angle. In this way a tolerance about the Bragg angle is provided that makes practical acoustooptic devices possible, such as the Bragg cell (**Fig. 2**).

The acoustic beam of width L is generated by applying an electrical signal at frequency F to the acoustic transducer. The transducer is made of a thin layer of piezoelectric material, often crystalline, that generates mechanical vibrations when activated electrically. *See* PIEZOELECTRICITY.

Frequency shifting. The incident light beam (called the zeroth order) has a frequency f, a wavelength λ, and a width d. It is incident at an angle ϕ_B as required by the wave-vector diagram. The diffracted beam (called the first order) propagates at the deflection angle $\phi_d = 2\phi_B$. This beam has a frequency $f + F$, where F is the frequency of the acoustic signal. Thus the first obvious application of a Bragg cell is as a frequency shifter. The frequency F can be varied within limits without having to change the incident angle, because of the tolerance about the Bragg angle. *See* OPTICAL ISOLATOR.

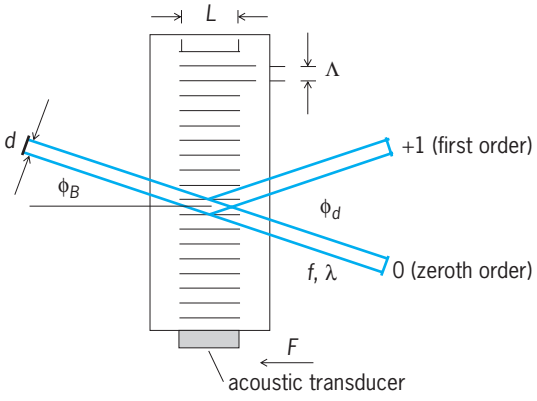


Fig. 2. Basic acoustooptic Bragg cell.

Light modulation. A second application is a light modulator. Here use is made of the fact that the intensity of the diffracted light is proportional to the intensity of the sound wave. Thus, if the electrical signal at frequency F is modulated in amplitude, the result is a similar modulation of the diffracted light. *See* OPTICAL MODULATORS.

Light deflection. Another application of the Bragg cell is as an electronic light deflector. Varying the sound frequency changes the sound wavelength and hence the value of the Bragg angle. This in turn changes the deflection angle in proportion. (Again, because of the tolerance about the Bragg angle, the input angle need not be changed.) Thus, by sweeping the acoustic frequency, the Bragg cell can be made to produce a scanning, diffracted beam of light. For random-access applications the frequency change is discontinuous so as to direct the diffracted beam quickly in any desired direction.

It is customary to focus the diffracted beam upon a focal plane by use of a suitable lens. When the beam scans, the focused spot traces out a line. The number of resolvable spots, N, that can be addressed along this line is given by the product of the maximum frequency swing $\Delta F_{\max}$ and the transit time τ of the sound through the light beam. Typical values are $\Delta F_{\max} = 40$ MHz, $\tau = 10$ μs, $N = 400$. The transit time τ also determines the shortest time in which the beam can change direction.

Spectrum analysis. If a spectrum of discrete frequencies, F_1, F_2, $F_3, \ldots$, is applied simultaneously to the transducer, each frequency produces its own diffracted beam and hence its own spot in the focal plane. Thus, in yet another application, the Bragg cell acts as a spectrum analyzer by making the spectral distribution of the sound signal visible in the focal plane. The obtainable frequency resolution equals $1/\tau$. *See* SPECTRUM ANALYZER.

Imaging. If the light beam is wide enough to fill the sound cell, a substantial portion of the sound signal is "addressed" by the incident light. The cell then generates in diffracted light a "moving image" of the amplitude modulation of the sound wave. Such an image may be further optically processed to enhance key features or to extract specific information. Alternatively, if the amplitude modulation represents successive lines of a television image, that image may be projected upon a screen, using appropriate dynamic optics to immobilize it. Such an acoustooptic television display was in fact developed in 1936 in Great Britain, long before the invention of the laser.

Width of sound beam. The efficiency of a Bragg cell increases by increasing the width L of the sound beam. Unfortunately the tolerance about the Bragg angle (and therefore the acoustic frequency range and the acceptance angle) are inversely proportional to this length. The design of a Bragg cell is thus a compromise between efficiency and bandwidth. Furthermore, if the width of the sound beam is too small, the beam does not resemble a plane wave, and undesirable higher-order light beams are generated by multiple diffraction. The cell is then said to work in

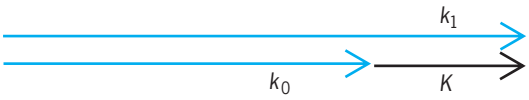


Fig. 3. Wave-vector diagram showing collinear interaction in an anisotropic crystal.

the Raman-Nath region (named after two of the first investigators of the phenomenon), rather than the Bragg region.

Optical filters. An important application of the Bragg cell is to electronically tunable optical filters. Here a narrow tolerance about the Bragg angle is required to minimize the window of allowable incident wavelengths clustered around a central one. When the sound frequency is changed, a different central wavelength will satisfy the Bragg angle condition, and the window moves along the optical spectrum.

Use of anisotropic crystals. In isotropic materials a tunable acoustooptic filter requires a wide sound beam, which is not always feasible. In practice, anisotropic crystals are used for this application. In such crystals the light velocity depends upon direction of propagation and polarization. The wave-vector diagram (Fig. 2) must then be modified to account for different polarizations of incident and diffracted light. This results in the light wave vectors $\mathbf{k}_0$ and $\mathbf{k}_1$ having unequal lengths. By collapsing the wave-vector triangle to three parallel lines (**Fig. 3**), a configuration is obtained in which the sound and the light propagate in the same direction. This collinear interaction maximizes the interaction length and obviates the need for a wide sound beam. *See* CRYSTAL OPTICS; POLARIZED LIGHT.

Another advantage of using anisotropic crystals is that some of them have sound velocities an order of magnitude smaller than isotropic materials such as glass or fused quartz. (In tellurium dioxide, TeO_2, for example, one of the possible shear waves has a sound velocity of 617 m/s or 2024 ft/s). Since the efficiency of a Bragg cell—often summarized by a material figure of merit called M_2—is inversely proportional to the cube of the sound velocity, it is clear that crystals such as tellurium dioxide offer potential design advantages.

Two-dimensional devices. It has proven possible to construct two-dimensional acoustooptic configurations. Here, acoustic surface waves interact with light propagating in optical waveguides made of thin films, deposited on a suitable substrate. Such devices are of obvious importance when combined with integrated electronic circuits. *See* INTEGRATED CIRCUITS; INTEGRATED OPTICS; LIGHT; SOUND.

Adrian Korpel

Bibliography. A. P. Goutzoulis and D. R. Pape (eds.), *Design and Fabrication of Acousto-optic Devices*, Marcel Dekker, New York, 1994; A. Korpel, *Acousto-optics*, 2d ed., Marcel Dekker, New York, 1997; A. Korpel (ed.), *Selected Papers on Acousto-optics*, SPIE, Bellingham, 1990; A. Sliwinski, B. B. J. Linde, and P. Kwiek (eds.), Acousto-optics and applications, *Proc. SPIE*, vol. 3581, 1998.

Acquired immune deficiency syndrome (AIDS)

A viral disease of humans caused by the human immunodeficiency virus (HIV), which attacks and compromises the body's immune system. Individuals infected with HIV proceed through a spectrum of stages that ultimately lead to the critical end point, acquired immune deficiency syndrome. The disease is characterized by a profound progressive irreversible depletion of T-helper-inducer lymphocytes (CD4+ lymphocytes), which leads to the onset of multiple and recurrent opportunistic infections by other viruses, fungi, bacteria, and protozoa (see **table**), as well as various tumors (Kaposi's sarcoma, lymphomas). HIV infection is transmitted by sexual intercourse (heterosexual and homosexual), by blood and blood products, and perinatally from infected mother to child (prepartum, intrapartum, and postpartum via breast milk).

HIV infection. In 1983, a virus recovered from the blood of a patient with persistent generalized lymphadenopathy attacked and killed T lymphocytes but could not be well characterized because of the inability to grow it in sufficient amounts. In 1984, T cells were identified that supported growth of the virus and were able to at least partially withstand its cytopathic effects. This retrovirus was variously named LAV (lymphadenopathy-associated virus) and HTLV-III (human T-cell lymphotropic virus, type III), but is now universally named HIV-1 (human immunodeficiency virus, type 1). *See* RETROVIRUS.

The availability of large quantities of virus and the resultant diagnostic reagents led to the finding that HIV-1 was the primary etiological agent of this syndrome. A majority of those infected will ultimately develop AIDS. Virtually all individuals with AIDS, as well as those with the generalized

Agents of major opportunistic infections in patients with AIDS

Type	Agent
Viruses	*Cytomegalovirus*
	Herpes simplex
	Herpes zoster
	Papovavirus (progressive multifocal leukoencephalopathy)
Fungi	*Candida* species
	Cryptococcus neoformans
	Histoplasma capsulatum
	Coccidioides immitis
Bacteria	*Mycobacterium avium* complex
	Mycobacterium tuberculosis
	Streptococcus pneumoniae
	Haemophilus influenzae
	Salmonella typhimurium
	Shigella flexneri
	Campylobacter species
	Treponema pallidum
	Rochalimaea henselae (bacillary angiomatosis)
Protozoa	*Pneumocystis carinii*
	Toxoplasma gondii
	Isospora belli
	Cryptosporidia
	Microsporidia

lymphadenopathy syndrome, tested seropositive for HIV-1. This important breakthrough also permitted the introduction of serological testing of blood donors, vastly decreasing the incidence of transfusion-associated HIV infection. The time interval from initial HIV infection until HIV antibodies can be detected by using commercially available assays ranges from 3 weeks to 6 months; however, HIV infection can precede the appearance of clinical disease by 10 years or more. Serological testing allows for the detection of viral infection before any clinical manifestations appear.

Since retroviruses such as HIV-1 integrate their genetic material into that of the host cell, infection is generally lifelong and cannot be eliminated easily. Therefore, medical efforts have been directed toward preventing the spread of virus from infected individuals. The best method for preventing transmission of HIV infection is to avoid high-risk behaviors, such as unprotected sexual intercourse or intravenous drug abuse.

Clinical disease. Approximately 50–70% of individuals with HIV infection experience an acute mononucleosis-like syndrome approximately 3–6 weeks following primary infection. In the acute HIV syndrome, symptoms include fever, pharyngitis, lymphadenopathy, headache, arthralgias, myalgias, lethargy, anorexia, nausea, and erythematous maculopapular rash. These symptoms usually persist for 1–2 weeks and gradually subside as an immune response to HIV is generated.

Asymptomatic stage. Although the length of time from initial infection to development of the clinical disease varies greatly from individual to individual, a median time of approximately 10 years has been documented for homosexual or bisexual men, depending somewhat on the mode of infection. Intravenous drug users experience a more aggressive course than homosexual men and hemophiliacs because their immune systems have already been compromised and the virus is acting on a weakened system.

Early symptomatic disease. As HIV replication continues, the immunologic function of the HIV-infected individual declines throughout the period of clinical latency. At some point during that decline (usually after the CD4+ lymphocyte count has fallen below 500 cells per microliter), the individual begins to develop signs and symptoms of clinical illness, and sometimes may demonstrate generalized symptoms of lymphadenopathy, oral lesions (thrush, hairy leukoplakia, aphthous ulcers), herpes zoster (shingles), and thrombocytopenia.

Secondary opportunistic infections. Secondary opportunistic infections are a late complication of HIV infection, usually occurring in individuals with less than 200 CD4+ lymphocytes per microliter. They are characteristically caused by opportunistic organisms such as *Pneumocystis carinii* and cytomegalovirus that do not ordinarily cause disease in individuals with a normally functioning immune system. However, the spectrum of serious secondary infections that may be associated with HIV infection also includes common bacterial pathogens, such as *Streptococcus pneumoniae*. Secondary opportunistic infections are the leading cause of morbidity and mortality in persons with HIV infection. Therefore, HIV-infected individuals are administered protective vaccines (pneumococcal) as well as prophylactic regimens for the prevention of infections with *P. carinii, Mycobacterium tuberculosis*, and *M. avium* complex. *See* MYCOBACTERIAL DISEASES; PNEUMOCOCCUS; STREPTOCOCCUS.

Tuberculosis has become a major problem for HIV-infected individuals. Those with pulmonary tuberculosis cannot be distinguished from those with community-acquired pneumonias, as well as *P. carinii* pneumonia. Any HIV-infected individual with pneumonia should be isolated immediately from other persons and health care staff members until a diagnosis of tuberculosis is ruled out. Health care settings should implement the use of negative-pressure sputum induction booths, as well as negative-pressure isolation rooms. Further complicating the dual epidemic of HIV disease and tuberculosis has been the emergence of multidrug-resistant tuberculosis, with its therapeutic and prophylactic dilemmas. *See* DRUG RESISTANCE; OPPORTUNISTIC INFECTIONS; TUBERCULOSIS.

Treatments. Antiretroviral treatment with deoxyribonucleic acid (DNA) precursor analogs—for example, azidothymidine (AZT), dideoxyinosine (ddI), and dideoxycytidine (ddC)—has been shown to inhibit HIV infection by misincorporating the DNA precursor analogs into viral DNA by the viral DNA polymerase. Nevertheless, these agents are not curative and do not completely eradicate the HIV infection.

Abe M. Macher; Eric P. Goosby

Acquired immunological tolerance

An induced state in which antigens originally regarded as foreign become regarded as self by the immune system. Tolerance can be induced (tolerization) in all of the cells of the immune system, including T cells (also known as T lymphocytes), the antibody-forming B cells (also known as B lymphocytes), and natural killer cells. Artificially induced immunological tolerance can be helpful in a number of clinical settings.

Autoimmune disease. The major function of the normal mammalian immune system is distinguishing foreign antigens (usually derived from microorganisms) from self-antigens so that only foreign antigens will be the targets of immune attack. To ensure the ability to make this distinction, developing T cells and B cells normally undergo selection processes based on the antigen receptors they express. Thereby, only cells with potential reactivity to foreign antigens, and not those with the ability to recognize self-antigens (autoantigens), contribute to the lymphocyte repertoire. However, this selection is incomplete, and additional mechanisms are required to help prevent autoreactive lymphocytes from mounting harmful antiself immune responses (termed autoimmunity). Regulatory processes, however, sometimes go awry, resulting in autoimmune disease, in which a variety of tissues (for example,

the insulin-producing beta cells of the pancreas in diabetes and the thyroid gland in autoimmune thyroiditis) may be targeted. *See* ANTIBODY; ANTIGEN; CELLULAR IMMUNOLOGY; IMMUNITY.

Inducing self-tolerance in the immune system could be an approach to curing autoimmune diseases. Some autoimmune diseases are believed to be initiated by autoreactive T lymphocytes. These may also be associated with autoantibody production by B cells that are dependent on help from autoreactive T cells. For these diseases, tolerization of T cells should, in theory, be sufficient to correct the disease. Other autoimmune diseases may be B cell–autonomous, so tolerization of autoreactive B cells would be essential for an effective cure. *See* AUTOIMMUNITY; IMMUNOLOGICAL DEFICIENCY.

Organ transplantation. Tolerization can also be used to facilate organ transplantation. Immunosuppressive drug therapy, which globally suppresses immune responses, is normally taken continuously by organ transplant recipients in an effort to prevent rejection in the graft. Despite improvements in immunosuppressive drug therapy, teaching the immune system to regard a set of foreign antigens presented by the organ graft as self (that is, tolerance induction) has become an important goal for several reasons: (1) It would eliminate the need for chronic immunosuppressive therapy, which is associated with lifelong increased risks of infection and malignancy, and other side effects. (2) It would prevent chronic rejection (a major problem even with immunosuppressive therapy), which often leads to late graft loss. (3) It presents a less toxic alternative to the unacceptably high levels of nonspecific immunosuppressive therapy that would likely be required to prevent rejection of xenografts (grafts from a donor of another species).

With the exception of blood group antigens, which are usually matched before transplanting organs between allogeneic donors and recipients (donors and recipients of the same species), most individuals do not have preexisting natural antibodies against organs from other humans. Since natural killer cells do not appear to make a significant contribution to the rejection of allogeneic organs (other than bone marrow), induction of T cell tolerance would be sufficient to overcome the problems of allotransplantation. In contrast, T cell–, natural killer cell–, and B cell–mediated immune responses to xenografts are generally greater than those to allografts, so in such cases the induction of B cell– and natural killer cell tolerance is of interest in addition to T cell tolerance. For example, although pigs are considered the most promising xenogeneic species for organ donation to humans, they express a carbohydrate antigen [Galα1,3Gal (Gal)] on most of their tissues that is recognized by natural antibodies present in the sera of all humans. These antibodies can bind to the Gal antigen expressed on the cells lining the blood vessels of pig organs and initiate a very rapid rejection process called hyperacute rejection. Although studies in primates indicate that hyperacute rejection can be prevented by removing the antibodies from the recipient serum prior to transplant, the antibodies rapidly return to the circulation and cause other, more delayed forms of rejection. Tolerization of the B cells that produce these natural antibodies presents a possible solution to this problem. Natural killer cells seem to respond more strongly to xenografts than to allografts, perhaps because xenogeneic tissues do not express molecules that can be recognized by inhibitory receptors on natural killer cells that normally prevent the natural killer cells from killing autologous cells. Although little is known about the process by which natural killer cells are educated to distinguish self from nonself, bone marrow transplantation studies in mice suggest that acquired tolerance can be induced among natural killer cells. *See* TRANSPLANTATION (BIOLOGY).

Central tolerance. Many strategies for inducing immunological tolerance involve reproducing the mechanisms involved in natural central tolerance—the phenomenon by which self-tolerance is maintained among immature lymphocytes developing in the central lymphoid organs. For developing T cells, tolerance occurs in the thymus, the central organ for T cell development. For B cells, development occurs in the bone marrow, and an encounter with self antigens can induce tolerance there among the immature cells. *See* IMMUNOLOGICAL ONTOGENY.

Clonal deletion. Clonal deletion (or negative selection) is a major natural mechanism of central tolerance. Developing lymphocytes with uniquely rearranged receptors (T cell receptor in the case of T cells; immunoglobulin receptor in the case of B cells) that recognize self antigens may be selected against if they encounter these self antigens on an appropriate cell in the thymic (for T cells) or marrow (for B cells) environment. Developing T cells encountering such antigens will undergo apoptotic death (programmed cell death). Developing B cells in the marrow or immature B cells in the spleen that recognize self-antigens may die or undergo additional gene rearrangements that produce new immunoglobulin receptors with different recognition specificities (receptor editing).

Hematopoietic cell transplantation. The transplantation of bone marrow or other sources of hematopoietic (blood cell–producing) stem cells provides a very powerful means of inducing T cell central tolerance. One reason is that cells derived from the donor stem cells (for example, dendritic cells) enter the thymus and present donor antigens to developing thymocytes, which are killed if they bind strongly to an antigen. If a constant supply of these donor cells is provided by a successful hematopoietic stem cell graft, T cells developing subsequently will be tolerized to the marrow donor by the same mechanism through which self-tolerance is normally achieved (death of alloreactive cells). The clinical potential of bone marrow or other types of hematopoietic cell transplantation for the induction of transplantation tolerance in humans has not yet been realized, however, because the methods traditionally used to prepare a recipient for a hematopoietic cell transplantation (for the treatment of leukemia, for example) are far too toxic to be appropriate in this setting.

In addition, major problems due to host-versus-graft and graft-versus-host immune reactivity occur with extensively mismatched marrow transplants.

More recently, animal studies have indicated that less toxic but more specific treatments to target the cells responsible for this immune reactivity can allow hematopoietic cell transplantation to be performed with minimal toxicity. Since the host marrow is not subjected to ablative treatments (that is, treatments that eliminate host hematopoietic cells), a state of mixed chimerism (the coexistence of blood cells derived from both the donor and the host) results when donor hematopoietic stem cell engraftment is achieved. Thus, hematopoietic cell transplantation may have clinical applicability for the induction of allograft tolerance. *See* CHIMERA.

Hematopoietic cell transplantation has been shown in animal models to induce T cell central tolerance even across xenogeneic barriers when the donor and host species are very closely related. T cells are made tolerant to the donor by similar mechanisms to those described above for allogeneic marrow. Natural antibody-producing B cells (both preexisting and newly developing cells) are made tolerant to the donor antigens when mixed chimerism is achieved. Therefore, successful achievement of marrow engraftment across highly disparate species barriers has the potential to prevent both T cell– and natural antibody–mediated rejection of xenografts. In the most disparate species combinations, however, nonimmunological barriers such as species differences in the proteins important for the function of donor hematopoietic stem cell grafts make it difficult to achieve engraftment. Donor cells may also be destroyed by cells of the innate immune system (for example, macrophages).

Peripheral tolerance. Peripheral tolerance comprises mechanisms to prevent immune responses among mature lymphocytes in the peripheral tissues.

Anergy. One major mechanism of T cell and B cell peripheral tolerance is anergy, in which the cells cannot be fully activated by encounter with the antigens that their receptors recognize. Numerous methods of inducing T and B cell anergy have been described. For B cells, the induction of anergy versus activation might be dependent on antigen concentration, with exposure to low antigen levels leading to an anergic state. However, T cells may be anergized by an encounter with antigen in the absence of costimulatory signals from the antigen-presenting cell that are required for T cell activation, or by encounter with an antigenic variant for which it has low affinity. Many experimental approaches to inducing transplantation tolerance, such as the use of costimulatory blockade, may induce a state of anergy among donor-reactive T cells. Costimulatory blockade may also lead to peripheral deletion of T cells encountering alloantigens (antigens that stimulate responses in another member of the some species) of a graft.

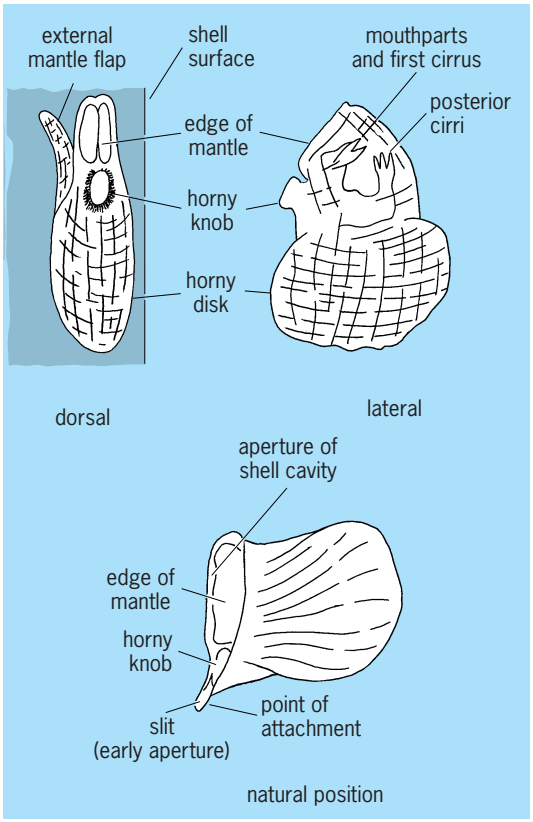
Suppression. Another major mechanism of peripheral tolerance is suppression, in which both B cells and T cells may be rendered tolerant of a specific antigen through the activity of substances or cells that actively suppress the lymphocyte's function. Numerous means of inducing B cell and T cell suppression have been described, and convincing evidence implicates cells with the ability to suppress T cell alloresponses (immune responses to alloantigens) in transplantation models. Recently, T cells expressing the molecules CD4 and CD25 have been implicated in the tolerance induced with costimulatory blockade and other strategies used in rodent models. However, although a variety of mechanisms resulting in suppression have been described, the mechanisms operative in vivo are incompletely understood. A better understanding of these complex phenomena may lead to the ability to apply some of these strategies from animal models to clinical transplantation. *See* IMMUNOSUPPRESSION.

Megan Sykes

Bibliography. W. E. Paul (ed.), *Fundamental Immunology*, Lippincott-Raven, Philadelphia, 1999; M. Sykes and S. Strober, Mechanisms of tolerance, in E. D. Thomas, K. G. Blume, and S. J. Forman (eds.), *Hematopoietic Cell Transplantation*, pp. 264–286, Blackwell Science, Malden, 1999.

Acrothoracica

An order of the subclass Cirripedia. All members burrow into shells of mollusks and thoracican barnacles, echinoderm tests, polyzoans, dead coral, and limestone. The entrance to the burrow is slitlike. In thin shells the burrow and animal are strongly compressed (see **illus.**). The mantle lacks calcareous



***Trypetesa lateralis* Tomlinson. (*After J. T. Tomlinson, A burrowing barnacle of the genus Trypetesa, order Acrothoracica, J. Wash. Acad. Sci., 43(11):373–381, 1953*)**

plates and is attached to the burrow by cement from glands in the dorsal part of the mantle. The attached area becomes thickened by the buildup of successive layers of chitin which cannot be shed at molting. The normal three pairs of cirriped mouthparts are present: mandibles, maxillules, and maxillae. Four to six pairs of cirri occur, often greatly reduced in size, the first pair close to the oral appendages and the remainder packed together at the posterior end of the body. The sexes are separate; dwarf males are found on the mantle or wall of the burrow of the female. Naupliar larval stages may be omitted, but a cypris larva always ccurs in the life cycle. Three families, about eight genera, and 40 species are recognized. Detailed examination of shell debris, especially from warmer seas, has revealed many new species. Doubtless many more exist. Fossils, known from burrows only, occur from the Carboniferous to Recent. *See* CIRRIPEDIA. H. G. Stubbings

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; J. T. Tomlinson, The burrowing barnacles (Cirripedia: Order Acrothoracica), *Bull. U.S. Nat. Mus.*, 296:1–162, 1969.

Acrotretida

An order of inarticulated brachiopods consisting of a group of sessile, suspension-feeding, marine, benthic, epifaunal bivalves, with representatives occurring throughout the Phanerozoic Era (from the Early Cambrian to the present). Extant members of this group originated in the Ordovician Period, when the order achieved maximum diversity.

Most shells are circular or subcircular in outline. The dorsal (brachial) valve is usually larger than the ventral (pedicle) valve, and is convex and often conical. The ventral valve is flattened and disklike. The valves are separate and do not articulate about a hinge but possess a complex arrangement of muscles. Acrotretids have two pairs of anterior and posterior adductor muscles, two pairs of oblique muscles, an elevator, and three pairs of minor muscles: the lophophore protractors, retractors, and elevators.

The tentacular feeding organ (lophophore) occupies the mantle cavity. The digestive system consists of a mouth, pharynx, esophagus, stomach, digestive diverticula, pylorus, and intestine, terminating in a functional anus. Four digestive diverticula in *Discinisca* and two in *Neocrania* open through ducts to the stomach. The excretory system consists of a pair of ciliated funnels (metanephridia) which during spawning act as gonoducts and allow the discharge of gametes from the coelom into the mantle cavity. Some solid waste may also be ejected through the nephridiopores, enmeshed in mucus, while the main excretory product, ammonia, is voided through the tissues of the mantle and lophophore.

Acrotretids possess an open circulatory system of blood vessels and coelomic canals, containing a fluid that is coagulable and has free cellular inclusions consisting of blood cells and coelomocytes. Acrotretids have a central nervous system containing unsheathed nerves, but no differentiated sense organs. Members of this order have been reported only as gonochoristic.

Two suborders are recognized: (1) Acrotretidina have a short pedicle emerging through a foramen in the ventral valve, chitinophosphatic shells, and planktotrophic larval development. (2) Craniidina lack a pedicle and are cemented to the substrate by the ventral valve, have shells composed of calcite and scleroproteins, and produce lecithotrophic larvae. *See* BRACHIOPODA; INARTICULATA. Mark A. James

Bibliography. G. A. Cooper, *Jurassic Brachiopods of Saudi Arabia*, 1989; M. J. S. Rudwick, *Living and Fossil Brachiopods*, 1970.

Acrylonitrile

An explosive, poisonous, flammable liquid, boiling at 77.3°C (171°F), partly soluble in water. The formula $CH_2{=}CH{-}C{\equiv}N$ indicates it may be regarded as vinyl cyanide, and its systematic name is 2-propenonitrile. Acrylonitrile is prepared by ammoxidation of propylene, according to the reaction below, over various sorts of catalysts, chiefly metallic

$$2CH_3{=}CHCH_3 + 2NH_3 + 3O_2 \longrightarrow 2CH_2{=}CHC{\equiv}N + 6H_2O$$

oxides. Older processes, such as catalytic addition of hydrogen cyanide to acetylene, and the catalytic reaction of propylene with nitric oxide, are little used today.

Most of the acrylonitrile produced is consumed in the manufacture of acrylic and modacrylic fibers. Substantial quantities are used in acrylonitrile-butadiene-styrene (ABS) resins, in nitrile elastomers, and in the synthesis of adiponitrile by electrodimerization. The adiponitrile is subsequently hydrogenated to hexamethylenediamine, a constituent of nylon. Smaller amounts of acrylonitrile are used in cyanoethylation reactions, in the synthesis of drugs, dyestuffs, and pesticides, and as co-monomers with vinyl acetate, vinylpyridine, and similar monomers.

Acrylonitrile undergoes spontaneous polymerization, often with explosive force. It polymerizes violently in the presence of suitable alkaline substances. *See* NITRILE; POLYMERIZATION. Frank Wagner

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; P. A. Smith, *Open-Chain Nitrogen Compounds*; *Chemistry of Non-Cyclic Nitrogen-Containing Organic Functional Groups*, 2d ed., 1982.

Actiniaria

An order of the cnidarian subclass Zoantharia (Hexacorallia) known as the sea anemones, the most widely distributed of the anthozoans. They have even been discovered in frigid waters. Usually they are solitary animals which live under the low-tide mark attached to some solid object by a basal expansion

or pedal disk. They feed on prey such as copepods, mollusks, annelids, crustaceans, and fishes. The burrowing species, like *Edwardsia*, *Halcampella*, and *Harenactis*, lack a pedal disk and bury their elongated bodies in the soft sediment of oceans. The actinians move rather actively. *Gonactinia* and *Boloceroides* can swim by using their tentacles.

Morphology. The freely retractile, skeletonless polyp has a cylindrical body, with a thick, tough, rough column wall often bearing rugae, verrucae, tubercles, or suckers (**Fig. 1**). The body is often encrusted with sand grains, pebbles, and other detritus. *Actinia*, *Diadumene*, *Metridium*, and other species have smooth, thin walls. The acrorhagi or marginal sphaerules, small rounded bodies covered with nematocysts, are arranged in a circlet on the margin which is the junction between the oral disk and column. When lacking nematocysts, they are termed fronds or pseudoacrorhagi. Nematocysts discharge a toxic substance, tetramethylammonium hydroxide or tetramine; however, the human skin is seldom affected by this. The junction between the pedal disk and column is termed the limbus. In *Tealia*, *Anemonia*, *Phellia*, *Ilyanthus*, and many other species, the upper part of the column is folded to form a collar or parapet, which divides the body into an upper

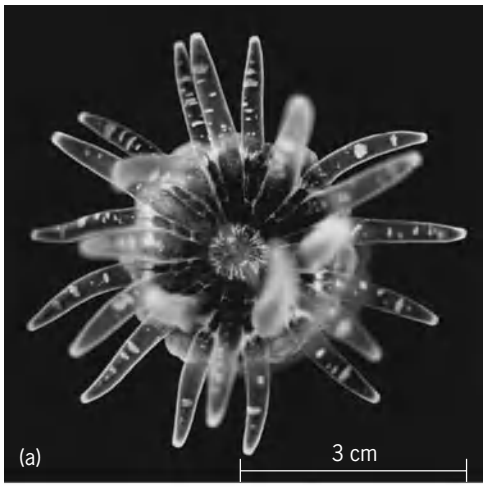
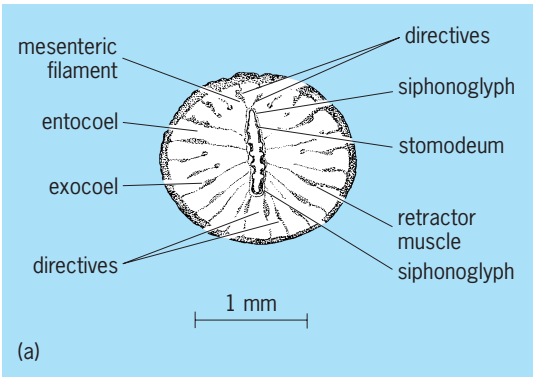


Fig. 1. Sea anemones, oral view. *Anthopleura* sp. (*a*) Young. (*b*) Adult.



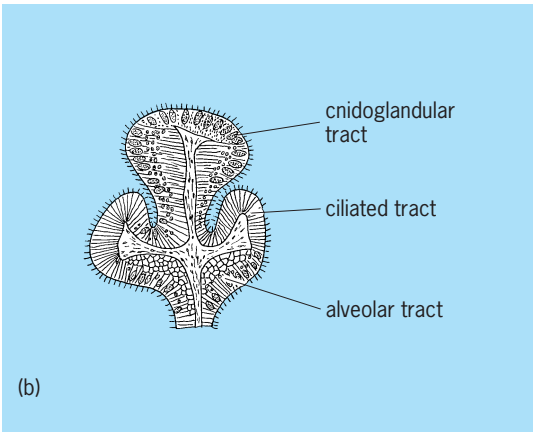


Fig. 2. Some of the morphological features of the actinians. (*a*) Cross-sectional view of *Anthopleura* sp. (*b*) Three-tracted mesenteric filament of *Phymanthus crucifer*.

capitulum and a lower scapus. Normally anemones have two siphonologlyphs, the sulcus and sulculus, and are termed diglyphic (**Fig. 2***a*); however, monoglyphic and occasionally tri- or tetraglyphic species such as *Diadumene* and *Metridium* occur. The colors vary with species and many variations occur even among the same species. *Diadumene luciae* has been divided into four races according to differences in color.

The tentacles increase in number regularly and are arranged in several cycles. There are 6 primary, 6 secondary, 12 tertiary, 24 quaternary, and so forth in the hexameroustype (Fig. 1). The paired mesenteries appearing first in couples show a bilateral arrangement. Their retractor muscles face each other, except the sulcal and asulcal pairs or directives which face away from each other. The space between each set of the paired mesenteries is an entocoele while its neighboring space is an exocoele, in which additional pairs of mesenteries are added (Fig. 2*a*). All the spaces communicate with each other either by oral or labial stomata; marginal or parietal ones, or both, may be present. The mesenteric filament is composed of three tracts, the cnidoglandular, ciliated, and alveolar (Fig. 2*b*). In Diadumenidae, Metridiidae, and Sagartiidae whitish threads or acontia protrude from small pores or cinclides of the column. These acontia are continuations of the septal filaments.

The musculature is the most highly developed in

the coelenterates. Generally, the longitudinal muscle is ectodermal, and the circular muscle, endodermal. The principal muscles are the longitudinal retractors, which form characteristic muscle bands; circular muscles of the column; marginal and tentacular sphincters, which serve to contract the body, oral disk, and tentacles respectively; the basilar muscle in the pedal disk; and, near this, the parietobasilar.

Reproduction. Most actinians are dioecious. The egg is comparatively large and contains a great amount of yolk. It is covered with numerous short processes (**Fig. 3***a*). The spermatozoon has a long flagellum (Fig. 3*b*). Developing larvae (**Fig. 4**) pass through the *Edwardsia* stage with eight complete mesenteries, then the *Halcampoides* stage with six pairs of protocnemes (primary mesenteries). Longitudinal fission frequently occurs as well as budding (**Fig. 5***a*, *b*, and *c*). Sometimes new individuals result from laceration (Fig. 5*d* and *e*). *See* CNIDARIA; REPRODUCTION (ANIMAL); HEXACORALLIA. Kenji Atoda

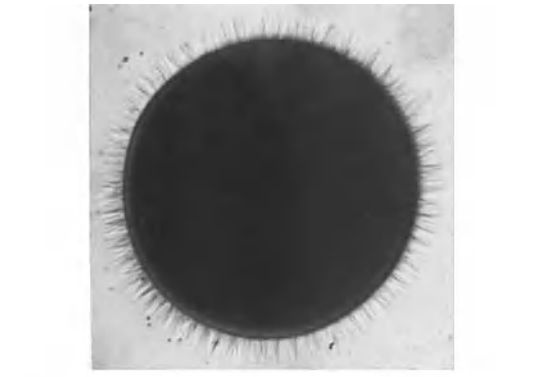


Fig. 3. Gametes of actinians. (*a*) Egg of *Anthopleura stella*. (*b*) Spermatozoa of *A. xanthogrammica*.

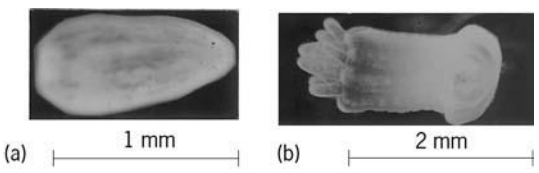


Fig. 4. Larval stages of *Anthopleura* sp. (*a*) Ciliated larva. (*b*) Larva just after extrusion.

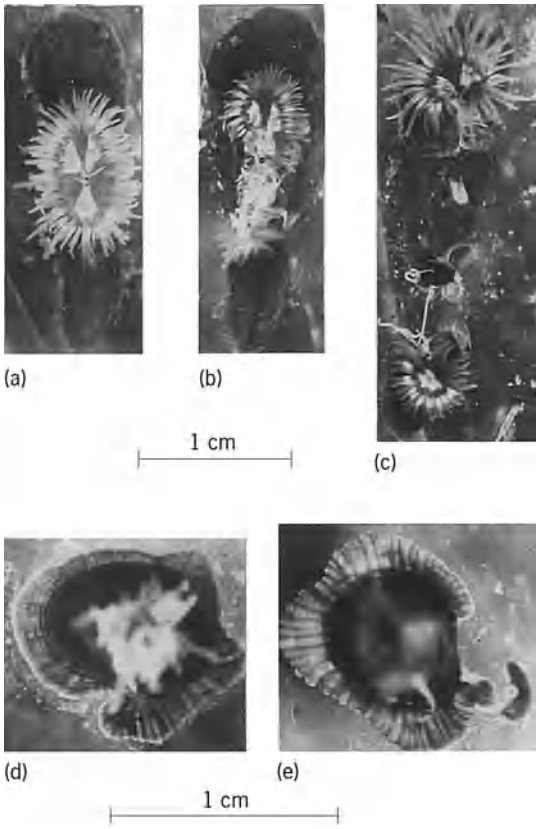


Fig. 5. Division of the pedal disk in *Diodumene luciae*. (*a*) Beginning of fission. (*b*) After 9½ h. (*c*) Polyp dividing into three unequal portions. (*d*) Beginning of laceration. (*e*) A small piece separated.

Bibliography. J. M. Shick, *A Functional Biology of Sea Anemones*, Chapman and Hall, 1991.

Actinide elements

The series of elements beginning with actinium (atomic number 89) and including thorium, protactinium, uranium, and the transuranium elements through the element lawrencium (atomic number 103). These elements, chemically similar, have a strong chemical resemblance to the lanthanide, or rare-earth, elements of atomic numbers 57 to 71. Their atomic numbers, names, and chemical symbols are 89, actinium (Ac), the prototype element, sometimes not included as an actual member of the actinide series; 90, thorium (Th); 91, protactinium (Pa); 92, uranium (U); 93, neptunium (Np); 94, plutonium (Pu); 95, americium (Am); 96, curium (Cm); 97, berkelium (Bk); 98, californium (Cf); 99, einsteinium (Es); 100, fermium (Fm); 101, mendelevium (Md); 102, nobelium (No); 103, lawrencium (Lr).

Studies of the chemical and physical properties of actinide and lanthanide elements and their compounds have indicated that their electronic structure must be similar; an inner electron shell of fourteen 5*f* electrons in the case of the actinides, and fourteen

Oxidation states of actinide elements in aqueous solution

Oxidation state		Elements*
II		Cf, Es, Fm, Md, No
III		Ac, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr
IV		Th, Pa, U, Np, Pu, Bk
V		Pa
V	(as MO_2^{+} ions)	U, Np, Pu, Am
VI	(as MO_2^{2+} ions)	U, Np, Pu, Am
VII	(as MO_5^{3-} ions)	Np, Pu

*The most stable states are underscored.

4f electrons in the case of the lanthanides, is filled in progressing across the series. Except for thorium and uranium, the actinide elements are not present in nature in appreciable quantities. The transuranium elements were discovered and investigated as a result of their synthesis in nuclear reactions. All are radioactive, and except for thorium and uranium, weighable amounts must be handled with special precautions.

The uranium isotopes ^{233}U and ^{235}U and the plutonium isotope ^{239}Pu undergo nuclear fission with slow neutrons with the liberation of large amounts of energy. Thorium can be converted to ^{233}U, and the isotope ^{238}U to ^{239}Pu by neutron irradiation; hence thorium and natural uranium can be used indirectly as nuclear fuels in breeder reactors.

Ion-exchange chromatography has been an important experimental technique in the study of the chemistry of the actinide elements. This method together with the analogy between corresponding actinides and lanthanides was the key to the discovery of the transcurium elements.

The actinide elements are very similar chemically. Most have the following in common: trivalent cations which form complex ions and organic chelates; soluble sulfates, nitrates, halides, perchlorates, and sulfides; and acid-insoluble fluorides and oxalates.

The characteristic oxidation state in aqueous solution is the III state, with higher oxidation states prominent in the early and the lower II state appearing in the latter part of the series (see **table**).

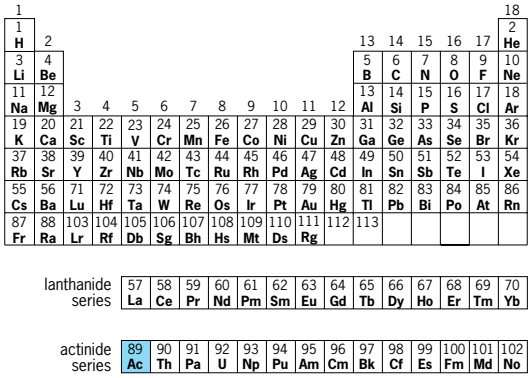
Many solid compounds including hydrides, oxides, and halides have been prepared by dry chemical methods. Binary compounds with carbon, nitrogen, silicon, and sulfur are of interest because of their stability at high temperatures. *See* ACTINIUM; AMERICIUM; ATOMIC STRUCTURE AND SPECTRA; BERKELIUM; CALIFORNIUM; CURIUM; EINSTEINIUM; FERMIUM; LAWRENCIUM; MENDELEVIUM; NEPTUNIUM; NOBELIUM; PERIODIC TABLE; PLUTONIUM; PROTACTINIUM; THORIUM; TRANSURANIUM ELEMENTS; URANIUM.

Glenn T. Seaborg

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; S. Cotton, *Lanthanide and Actinide Chemistry*, 2d ed., 2006; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; J. J. Katz, G. T. Seaborg, and L. R. Morse (eds.), *The Chemistry of the Actinide Elements*, 2 vols., 2d ed., 1986; G. Meyer and L. R. Morse, *Synthesis of Lanthanide and Actinide Compounds*, 1991.

Actinium

A chemical element, Ac, atomic number 89, and atomic weight 227.0. Actinium was discovered by A. Debierne in 1899. Milligram quantities of the element are available by irradiation of radium in a nuclear reactor. Actinium-227 is a beta-emitting element whose half-life is 22 years. Six other radioisotopes with half-lives ranging from 10 days to less than 1 minute have been identified.



The relationship of actinium to the element lanthanum, the prototype rare earth, is striking. In every case, the actnium compound can be prepared by the method used to form the corresponding lanthanum compound with which it is isomorphous in the solid, anhydrous state. *See* ACTINIDE ELEMENTS; LANTHANUM; NUCLEAR REACTION; RADIOACTIVITY.

Sherman Fried

Bibliography. S. Cotton, *Lanthanide and Actinide Chemistry*, 2d ed., 2006; J. J. Katz, G. T. Seaborg, and L. R. Morse (eds.), *The Chemistry of the Actinide Elements*, 2 vols., 2d ed., 1986; G. Meyer and L. R. Morse, *Synthesis of Lanthanide and Actinide Compounds*, 1991.

Actinobacillus

A genus of gram-negative, immotile and nonspore-forming, oval to rod-shaped, often pleomorphic bacteria which occur as parasites or pathogens in mammals (including humans), birds, and reptiles. They are facultatively aerobic, capable of fermenting carbohydrates (without production of gas) and of reducing nitrates. Most species are oxidase- and catalase-positive. Some cultures tend to stick on the surface of agar media, particularly on primary isolation. The genome deoxyribonucleic acid contains between 40 and 47 mol % guanine plus cytosine. The actinobacillus group shares many biological properties with the genus *Pasteurella*. At least two of the following features or combinations of features differentiate members of the *Actinobacillus* group

from *Pasteurella*: hemolysis, delayed or lacking fermentation of D-galactose or D-mannose, fermentation of inositol, positive reactions for both urease and β-galactosidase, hydrolysis of salicin or esculin, and fermentation of maltose together with negative reactions for trehalose fermentation and ornithine decarboxylase. *See* PASTEURELLA.

Actinobacillus lignieresii, the type species, causes granulomatous lesions predominantly in the upper alimentary tract of cattle (in particular, "wooden tongue"), and putrid skin and lung lesions in sheep. *Actinobacillus equuli* causes suppurative lesions, mainly in the kidneys and joints of foals. *Actinobacillus suis* is the etiologic agent of various lesions in piglets, and *A. pleuropneumoniae* that of epizootic necrotizing pleuropneumoniae in swine. *Actinobacillus capsulatus* was isolated from joint disease in rabbits. *Actinobacillus* (*Pasteurella*) *ureae* and *A. hominis* occur in the respiratory tract of healthy humans and may be involved in the pathogenesis of sinusitis, bronchopneumonia, pleural empyema, and meningitis.

Actinobacillus actinomycetemcomitans occurs in the human oral microflora, and together with anaerobic or capnophilic organisms may cause endocarditis and suppurative lesions in the upper alimentary tract. A specific role of *A. actinomycetemcomitans* in the pathogenesis of localized juvenile periodontitis has been suggested. *Pasteurella pneumotropica, P. haemolytica, Haemophilus paragallinarum*, and a variety of other groups of organisms that occur in healthy or diseased mammals, birds, and reptiles have been located with, or in the close vicinity of, the genus *Actinobacillus* on the basis of genetic relatedness.

Actinobacilli are susceptible to most antibiotics of the β-lactam family, aminoglycosides, tetracyclines, chloramphenicol, and many other antibacterial chemotherapeutics. *See* ANTIBIOTIC; MEDICAL BACTERIOLOGY.

Walter Mannheim

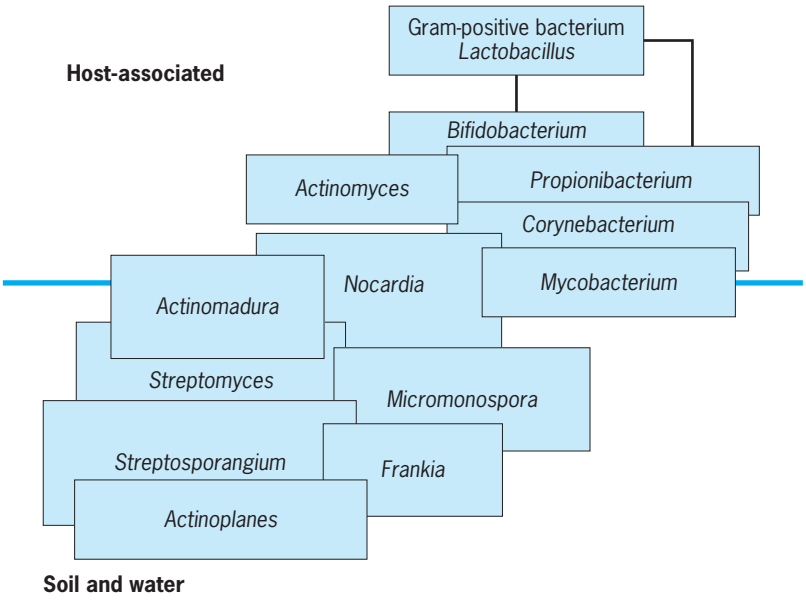
Bibliography. R. G. E. Murray, N. R. Krieg, and J. G. Holt (eds.), *Bergey's Manual of Systematic Bacteriology,* vol. 1, 1984; C. C. Tsai et al., Serum neutralizing activity against *Actinobacillus actinomycetemcomitans* leukotoxin in juvenile periodontitis, *J. Clin. Periodont.,* 8:338–348, 1981.

Actinomycetes

A heterogeneous collection of bacteria with diverse ecologies and physiologies. The initial definition was based on their branching filamentous cellular morphology during a stage of the growth cycle. The name actinomycete means "ray fungus," and for many years the actinomycetes were erroneously considered to be fungi, or at least closely related to the fungi. True fungi, such as the common bread mold, also grow as branching filaments (hyphae) that are several micrometers in diameter, whereas those of the actinomycetes are about 1 μm in diameter. The genetic material of the eukaryotic true fungi consists of DNA associated with proteins, and is enclosed in a membrane (the nucleus), whereas the genetic material of the prokaryotic actinomycetes consists of free DNA in direct contact with the cell sap (cytoplasm). In general, actinomycetes do not have membrane-bound organelles, whereas the fungi have an array of organelles such as mitochondria and vacuoles. The cell walls of fungi are made of chitin or cellulose, whereas the cell walls of the actinomycetes are made of a cross-linked polymer containing short chains of amino acids and long chains of amino sugars. Actinomycetes are susceptible to a wide range of antibiotics used to treat bacterial diseases, such as penicillin and tetracycline. *See* AMINO ACIDS; AMINO SUGAR; ANTIBIOTIC; DEOXYRIBONUCLEIC ACID (DNA); FUNGI.

Diversity and distribution. The actinomycetes occupy virtually all environmental niches, with some general tendency for members of one genus to have a more limited distribution. Not unexpected for such a diverse collection of prokaryotic microorganisms, the boundaries of the actinomycetes are not sharply defined. To add to the difficulty in selecting delineating characteristics, the genus *Actinomyces* is, in most respects, an exceptional taxon within the actinomycetes, or Actinomycetales. The actinomycetes include plant pathogens, animal pathogens, commensals in the intestines of invertebrates, and free-living bacteria in soil and in fresh and marine water. Only a few actinomycetes are pathogenic for plants, but *Streptomyces scabies* is one of the exceptions. For the most part, the aerobic *Streptomyces* are found in soil with a pH range of 5–6.5 with 10% organic matter. *Micromonospora* and *Streptosporangium* are found in soil with a pH of 4–5, with an organic content of 4–7%. *Nocardia* and *Actinomadura* are best isolated from soil with a pH of 4–5 and an organic content of 10%. The unique *Actinoplanes* with motile spores are found in aquatic habitats. The microaerophilic *Actinomyces* tend to colonize mammals and are often found in the mouth and gastrointestinal tract. When displaced from their normal sites within the mouth or gastrointestinal tract, *Actinomyces* may cause diseases in humans such as lung abscesses and appendicitis, and lumpy jaw in cattle. The animal pathogens are distributed throughout several genera, including *Mycobacterium*, *Corynebacterium*, *Nocardia*, *Actinomadura*, and *Actinomyces* (see **illustration**). *See* BACTERIA; BACTERIAL GROWTH; MEDICAL BACTERIOLOGY; PATHOGEN; SOIL MICROBIOLOGY.

Metabolism. The actinomycetes display a wide range of metabolic activities. Physiologically the actinomycetes have extraordinary ability to break down polysaccharides and hydrocarbons. In general, actinomycetes grow more slowly than "true" bacteria such as *Bacillus*, enteric bacteria, and *Pseudomonas*, in part because they are best adapted to the temperature of soil and water (70°F or 21°C). Members of the genus *Actinomyces* do not require oxygen for growth and are sometimes referred to as anaerobic bacteria. It is actually a requirement for elevated levels of carbon dioxide rather than a



Relationships among some better-known actinomycete genera. The actinomycetes are gram-positive bacteria that bridge through *Bifidiobacterium* and *Propionibacterium* to *Lactobacillus*. The genera above the colored line tend to live in association with hosts, but infrequently as pathogens. The genera below the colored line are found in soil and water.

negative effect of oxygen. *Actinomyces* is cultivated in the laboratory at body temperature (98°F or 37°C) in a rich nutrient medium, and frequently in an atmosphere that is enriched with carbon dioxide. *Actinomyces* is moderately slow-growing, requiring 1–2 weeks to achieve visible growth. Members of the genus *Nocardia* require oxygen for growth and have the ability to digest and transform many organic chemicals, including petroleum products and industrial products, intermediates, and by-products. *Nocardia* strains have been used commercially to modify chemical compounds related to cholesterol to make biologically active steroids. *Nocardia* inhaled into the lungs from soil may cause a disease in humans resembling tuberculosis. The streptomycetes are aerobic bacteria that have the ability to thrive with a wide range of nitrogen and carbon sources, including cellulose. The aroma of fresh soil and newly dug potatoes is actually due to streptomycetes. Streptomycetes do not produce disease in humans and are best known for their prolific production of antibiotics. Members of the genus *Frankia* symbiotically fix atmospheric nitrogen in the root nodules of the alder tree (*Alnus rugosa*) and the bayberry shrub (*Myrica pennsylvanica*). *See* TUBERCULOSIS.

Identification and classification. The actinomycetes were brought together as a group based upon morphological considerations, and morphological and developmental characteristics remain important in identification and classification. Members of the genus *Actinomyces* initially grow as branching filaments and then fragment into short rods. *Micromonospora* produces solitary spores in the vegetative hyphae. Streptomycetes persist as filamentous vegetative hyphae and form chains of spores in aerial hyphae. *Nocardia* does not produce aerial hyphae. The hyphae of some *Nocardia* species form branching filaments only for a brief time, whereas other species have persistent long, branching vegetative hyphae. There are differences in the chemical composition of actinomycete cells, especially the cell walls, which can be characterized by staining or chemical analysis. *Mycobacterium* cells are difficult to stain, and the films on microscope slides must be heated for the dye to penetrate. Once stained, the mycobacterium cells tenaciously retain the dye, and even treatment with dilute acid does not remove the dye; hence the name acid-fast bacteria. The actinomycete genera differ in the amounts of meso-diaminopimelic acid, arabinose, galactose, and long-chain mycolic acids in their cell walls. Streptomycete cell walls contain LL-diaminopimelic acid, whereas those of *Nocardia* contain meso- and/or DD-diaminopimelic acid. Analysis of menaquinones and lipid profiles of cell extracts have been used to demonstrate relatedness between actinomycete strains. For example, dihydromenaquinones are found in *Corynebacterium* and *Mycobacterium*, whereas tetrahydromenaquinones are found in *Nocardia*. Analysis of various nucleic acid preparations has become a widely used and accepted means for identifying and demonstrating relationships among actinomycetes. Initially, the guanine and cytosine (G+C) content of DNA was used to establish unrelatedness but could not establish relatedness. The ability of denatured DNA to reanneal became the standard method for measuring relatedness. Currently one of the popular nucleic acid techniques involves the analysis of the 16S fraction of ribosomes. As the capability to sequence the nucleotides in DNA samples has improved, direct comparison of genome sequences has become the definitive methodology. *See* NUCLEIC ACID; NUCLEOTIDE.

Molecular biology. Genetically the actinomycetes have DNA with high proportions of the purine guanine and the pyrimidine cytosine (55–72% G+C), and have rather large genome sizes (three to eight megabases). The high G+C content of the DNA affords some degree of protection of the genome for those species living in the soil and water. The smaller genomes of *Corynebacterium*, *Mycobacterium*, and *Rhodococcus* are organized into circular chromosomes, whereas the larger chromosomes of *Streptomyces* and *Nocardia asteroids* are usually linear. The *Streptomyces coelicolor* genome contains 8,667,507 base pairs, organized into a linear chromosome containing an estimated 7825 genes. The central core of the *S. coelicolor* chromosome contains considerable similarity to the whole chromosome of *Mycobacterium tuberculosis* and *Corynebacterium diphtheriae*. The actinomycetes have the ability to transfer genetic material among closely related strains by several different mechanisms, including conjugation and plasmid transfer. There are limited data indicating that plasmid transfer of genetic material occurs in nature as well as the laboratory. Methodology for the transfer of genes to or from actinomycetes to taxonomically distant bacteria, such as *Escherichia coli*, has been accomplished. There is growing evidence

that limited gene transfer has occurred between actinomycetes and eukaryotic organisms. *See* CHROMOSOME.

S. Gaylen Bradley

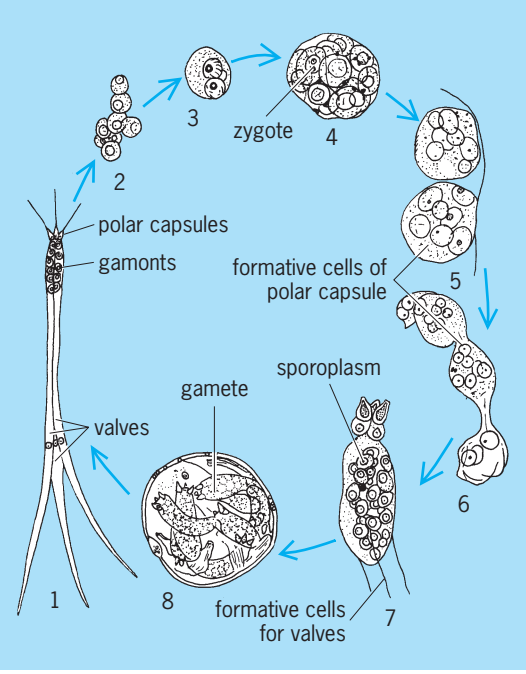
Bibliography. S. Donadio et al., Microbial technologies for the discovery of novel bioactive metabolites, *J. Biotech.*, 99:187–198, 2002; T. Kieser et al., *Practical Streptomyces Genetics*, John Innes Foundation, Norfolk, UK, 2000; E. M. Miguelez, C. Hardisson, and M. B. Manzanal, Streptomycetes: A new model to study cell death, *Int. Microbiol.*, 3:153–158, 2000.

Actinomyxida

An order of the protozoan class Myxosporidea (subphylum Cnidospora). It is characterized by the production of trivalved spores with three polar capsules and one to many sporoplasms. The spore membrane may be extended into anchor-shaped processes, which may have bifurcate tips. These protozoan parasites are found in the body cavity or in the intestinal lining of fresh-water annelids and in marine worms of the phylum Sipunculoidea. *See* INVERTEBRATE PATHOLOGY.

The life cycle for most species is not well known. Uninucleate sporoplasms may pair and fuse (plasmogamy) to form binucleate amebulas. Each amebula may produce a number of small and large cells through repeated binary fission. A large and small cell will pair and fuse (anisogamy) into a single cell (zygote) which becomes the sporoblast. The sporoblast divides repeatedly by binary fission, producing cells that will form the spore. Thus, in *Triactinomyxon legeri* (see **illus.**) parasitic in the gut of tubificid annelids, three cells form the polar capsules, three form the valves of the membrane, and one cell becomes the sporoplasm. The nucleus of the sporoplasm divides repeatedly by mitosis until 27 nuclei are formed, of which 24 become the sporoplasmic nuclei and 3 become the residual somatic nuclei. *See* CNIDOSPORA; MYXOSPORIDEA; OLIGOCHAETA; PROTOZOA.

Ross F. Nigrelli



Life cycle of *Triactinomyxon:* 1, mature spore; 2, liberated gamonts; 3, gamonts pairing; 4, zygote formation; 5, spore formation from zygote; 6–7, later stages in spore formation; 8, cyst in host tissue filled with young spores.

Actinophage

Any of a number of bacteriophages that infect and lyse certain bacteria (as originally classified in the order Actinomycetales; see **table**). Actinophages of particular interest are those that include in their host range any organisms of the genus *Streptomyces,* since most of the therapeutically useful antibiotics require culture of the particular *Streptomyces* species on a large scale. Contamination of a culture with a specific actinophage may result in lysis and destruction of the bacteria, which obviously halts antibiotic production. The elimination of actinophages from the environment is sometimes difficult, and the problem of phage contamination is often solved by the isolation and use of a phase-resistant mutant of the antibiotic-producing bacterium. *See* ANTIBIOTIC; BACTERIOPHAGE; LYSOGENY.

Lane Barksdale

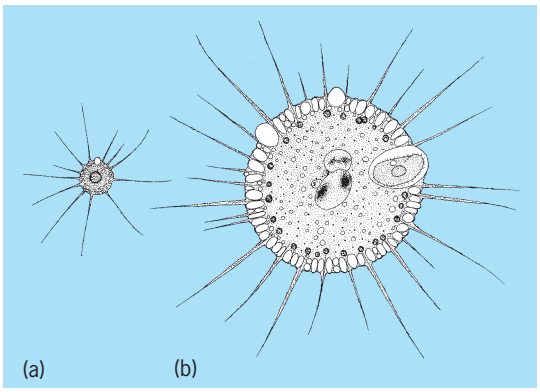
Host ranges of some actinophages*

Actinophage	Sensitive organism†		
	Streptomyces griseus	*Nocardia braziliensis*	*Mycobacterium rhodocrous*
MNP 6	+	+	−
MSP 1	−	+	−
MNP 8	−	−	+

*After L. A. Jones and S. G. Bradley, *Mycologia*, 56:505–513, 1964.
†Here + = lysed by bacteriophage, − = not lysed.

Actinophryida

An order of Heliozoia. These protozoans are characterized by stiff arms radiating from the central cell mass. Each arm is supported internally by an array of microtubules which terminate on the surface of nuclei. The organisms are found in fresh-water, marine, and soil habitats. They feed on other protozoans which collide with and stick to the arms. Well-studied genera are *Actinophrys* (**illus.** *a*), with a single central nucleus, *Actinosphaerium*, and *Echinosphaerium* (illus. *b*); the latter two have many nuclei lying just internal to a peripheral layer of vacuoles. The body size ranges from 50 to several hundred micrometers. Actinophryids form cysts within which a type of uniparental sexual reproduction



Actinophryida. (*a*) *Actinophrys sol*. (*b*) The much larger *Echinosphaerium nucleofilum*.

occurs. *See* ACTINOPODEA; HELIOZOIA; PROTOZOA; SARCODINA. David J. Patterson

Bibliography. J. J. Lee, S. H. Hutner, and E. C. Bovee, *Illustrated Guide to the Protozoa*, Society of Protozoologists, 1984; D. J. Patterson, On the organization and classification of the protozoon *Actinophrys sol* Ehrenberg, 1830, *Microbios*, 26:165–208, 1979.

Actinopodea

A class of Sarcodina; or in some modern schemes a superclass (Actinopoda) within the phylum Sarcomastigophora. Some of these protozoans have more or less permanent pseudopodia, composed of a central shaft of microtubules surrounded by a thin cytoplasmic envelope (axopodia); others have delicate and often radially arranged filopodia (filamentous, flexible pseudopodia) or filoreticulopodia (filamentous, branched pseudopodia) without axial microtubules. Although some are stalked and attached to the substratum (sessile), most are floating types. There are four subclasses: Radiolaria (Polycystinea and Phaeodarea), Acantharia, Heliozoia, and Proteomyxidia.

Radiolaria are all marine planktonic protozoans with a characteristic morphology consisting of a central capsule surrounded by a perforated capsular wall enclosed by the frothy or weblike mass of extracapsular cytoplasm. Solitary and colonial (multicellular) forms occur in some groups, and many species possess a characteristic, often highly ornamented, siliceous skeleton that persists after death of the protozoon and contributes to the sedimentary deposits in the ocean floor. The skeleton may be spherical, composed of concentric lattice shells, dome-shaped, or consist of a variety of solid geometric shapes, often with an esthetically pleasing architecture. Some species are surface-dwelling, and others occur from near surface to great depths in the ocean. *See* RADIOLARIA.

Acantharia, marine planktonic organisms resembling the Radiolaria, possess a central capsule, but the skeleton is composed of strontium sulfate. The skeleton is composed of rods arranged in a strict dimensional pattern, usually following the plan of a set of cartesian axes. In some species the skeleton is augmented with latticelike plates. The skeletal rods originate typically from a central point of attachment within the capsule and radiate outward through the capsular wall. Upon death of the acantharian, the strontium sulfate skeleton dissolves; hence no record is deposited in the ocean sediments. Algae, presumably symbionts, have been observed in the extracapsular cytoplasm or sometimes within the capsule of some species. Axopodia are characteristic. An apparently contractile apparatus (myoneme) attaches the cytoplasmic sheath to the skeleton and may regulate the shape of the peripheral cytoplasm, changing the hydrostatic properties of the organism and adjusting its buoyancy. *See* ACANTHAREA; ACANTHOMETRIDA.

Heliozoia include both fresh-water and marine organisms. The pseudopodia may be either filopodia or axopodia radially arranged around the central cytoplasm. Skeletal elements are either absent or simpler than observed in Acantharia or Radiolaria, and may consist of scales, spines, or in certain genera a perforate continuous exoskeleton (test). A centroplast (microtubular organizing center) is present in some genera, but lacking in others. Reproduction as exemplified by *Actinophrys* and *Actinosphaerium* is by fusion of gametes produced by the mother cell (autogamy). A zygocyst is formed and gives rise to the mature floating radiate form bearing radial pseudopodia. In *Actinophrys sol*, the sexual process begins with formation of a cyst followed by division within the cyst and formation of the haploid gametes that fuse to yield the zygote. *See* HELIOZOIA.

Proteomyxidia lack skeletons, but may gather detritus or mineral matter around the central cell body to form a loose protective coat. Filopodia or filoreticulopodia are characteristic and emerge either from one or both poles of the central body, which in some forms becomes spindle-shaped, especially during fission. Food of some species in the form of bacteria or small detrital particles is gathered by pseudopodial streaming from the surrounding environment. A number of species invade algae and, rarely, higher plants. During periods unfavorable for growth, some species may form resting cysts which resume active growth when the environment improves. *See* PROTEOMYXIDIA.

Little is known about the reproduction of many Actinopodea. *See* PROTOZOA; SARCODINA; SARCOMASTIGOPHORA. O. Roger Anderson

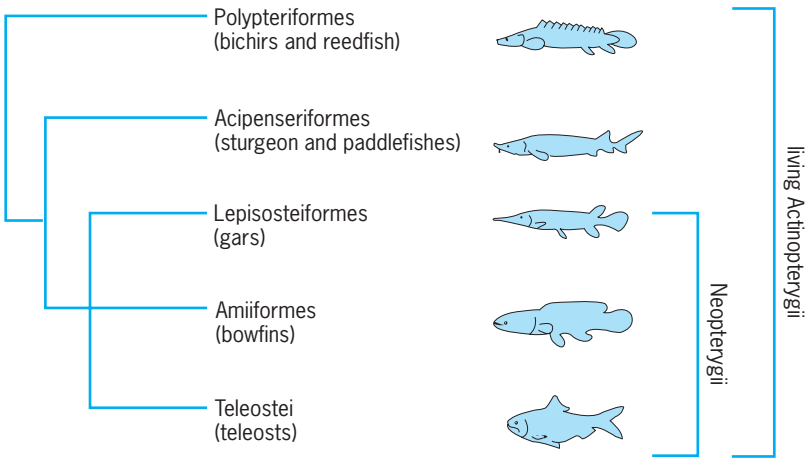
Actinopterygii

A class of teleostom fishes commonly known as the ray-finned fishes comprising the subclasses Cladistia (Polypteriformes and fossil orders), Chondrostei (Acipenseriformes and fossil orders), and Neopterygii (the remaining actinopterygian orders). The Neopterygii, minus the Lepisosteiformes,

Amiiformes, and several fossil taxa, comprise the Teleostei (see **illustration**). The Actinopterygii comprise about half of all vertebrate species and about 96% of all currently existing fishes. It is considered a nonmonophyletic group that is derived from more than one lineage when tetrapods are excluded. The Actinopterygii and the Sarcopterygii, minus the Tetrapoda, comprise what was previously called Osteichthyes (the bony fishes). The class includes 44 orders, 453 families, 4289 genera, and about 27,000 described extant species. Many species are known to science but are not yet described; further, in regions such as the Amazon and Congo basins, species are probably becoming extinct before they are discovered and described in the scientific literature.

Characteristics. Actinopterygii, being an extremely large and heterogenous group of fishes, are not identifiable by a unique character common to all taxa; suffice it to characterize the class as follows: The scales are dermal. Ganoid-type scales (scales covered by a hard mineral substance called ganoin or ganoine) are present in the Cladistia, Chondrostei, and holosteans (Lepisosteiformes, Amiiformes) and in several fossil orders. The Neopterygii (less the holosteans) have bony-ridge scales, which lack dense enamel or ganoin and are marked with bony ridges (circuli) that alternate with valleylike depressions. Bony-ridge scales are thin and overlapping and are typically cycloid or ctenoid, or highly modified as scutes, semirigid body rings (pipefishes and seahorses), rigid encasements resembling a turtle shell (trunkfishes and cowfishes), armature of bony plates (for example, the catfish families, Callichthyidae and Loricariidae; and poachers, Agonidae), and lancets (bladelike structures, one on each side of the caudal peduncle) of surgeonfishes (Acanthuridae). Some neopterygians lack scales entirely (for example, most catfishes and most eels). A sensory canal is present in the dentary bone, which is usually enclosed. Dermal bones of the skull are completely ossified in most groups. Radial bones of the pectoral girdle are attached to the scapulocoracoid complex. An interopercular bone and branchiostegal rays are usually present. Typically there is on each side one otolith in each of the three membranous sacs of the labyrinth of the inner ear. External nares (nostrils) are usually paired and end in a blind sac; internal nares are absent; and spiracles are usually absent.

About 40% of living actinopterygian species live exclusively or almost exclusively in freshwater. The rest inhabit mostly marine, brackish, or combination environments. The group has diversified through time to occupy an enormous range of habitats, from oceanic regions over 11,000 m (36,000 ft) deep (numerous species) to mountain streams up to 5200 m (17,000 ft) above sea level (homalopterid river loaches of Tibet), and from subfreezing water at −1.98°C (28.4°F) [Antarctic icefishes, Notothenioidei] to hot springs at 43.8°C (110.8°F) [pupfishes, *Cyprinodon*, and livebearers, *Gambusia*].



Tree illustrating some basal groups of living Actinopterygii. The arrangement reflects current understanding of their relationships, and the lack of clear resolution among certain neopterygian groups.

Some species move up on land to feed (mudskippers, *Periophthalmodon*), while others may take to the air by gliding on wind currents (flying fishes, Exocoetidae) or fly short distances using their large, well-muscled pectoral fins as wings (Gastropelecidae). While most species are herbivores (plant eaters) or piscivores (fish eaters), some are parasitic (candiru, Trichomycteridae) of South America. Some actinopterygians are ram-filter feeders (paddlefish, *Polyodon*), and some live within the internal organs of other organisms (pearlfishes, Carapidae) in possibly symbiotic, parasitic, or mutualistic relations.

Actinopterygians are extremely diverse morphologically and include the smallest known adult vertebrates [for example, a goby called the stout infant fish, *Schindleria brevipinguis*, from the Great Barrier Reef of Australia reaching only 7 mm (0.28 in.) in length]. The longest actinopterygian is the giant oarfish (*Regalecus*), which reaches reported lengths of over 8 m (26 ft).

Fossil record. The fossil record indicates that actinopterygians are at least as old as the Late Silurian (about 420 million years before present). Fossil actinopterygians are species-rich and extremely abundant, making up the majority of vertebrate fossils that are known by complete skeletons. Many major radiations of early actinopterygians, such as pycnodonts, semionotiforms, and palaeonisciforms, have been extinct for tens of millions of years. Other early actinopterygian groups, such as the Cheirolepiformes, have been extinct for hundreds of millions of years. Based on the fossil record, the most major differentiation of the group began in the late Mesozoic. *See* SEMIONOTIFORMS.

Living forms. Living Actinopterygii (see illustration) comprise Polypteriformes (bichirs and reedfishes, represented by at least 10 species, all in freshwaters of Africa); Acipenseriformes (sturgeons and paddlefishes, represented by at least 24 anadromous or freshwater species,

widespread within the Northern Hemisphere); Lepisosteiformes (gars, represented by at least seven species, mostly in freshwater but also rarely in brackish or even marine water, found in eastern North America, Central America, and Cuba); Amiiformes (bowfins, represented by one living species in the freshwaters of eastern North America); and Teleostei (teleosts, represented by about 24,000 species, widespread in nearly all aquatic environments, and including about 99.8% of all living actinopterygians). The gars, bowfins, and teleosts are grouped together in the Neopterygii. *See* ACIPENSERIFORMES; AMIIFORMES; POLYPTERIFORMES.

Phylogeny. There are two major theories to explain how the three basal extant lineages of Neopterygii are related. The older theory recognizes Holostei (a group containing gars and bowfins but excluding teleosts) as a natural (monophyletic) group. The other theory, resulting from detailed anatomical studies of the 1970s and 1980s, recognizes Halecostomi (a group containing bowfins and teleosts but excluding gars) as the natural group. Most systematic ichthyologists since the mid-1970s have favored recognition of Halecostomi over Holostei. However, with the advent of molecular evolutionary studies in the 1990s, the controversy has become even more complicated, because the most detailed morphological data now indicate Halecostomi as the natural group but molecular data favor Holostei. Thus, gars, bowfins, and teleosts are presented in the illustration as an unresolved trichotomous branch with inconclusive evidence for unambiguous phylogenetic resolution. *See* CHONDROSTEI; HOLOSTEI; OSTEICHTHYES; TELEOSTEI.

Herbert Boschung; Lance Grande

Bibliography. L. Grande and W. E. Bemis, A Comprehensive Phylogenetic Study of Amiid Fishes (Amiidae) Based on Comparative Skeletal Anatomy: An Empirical Search for Interconnected Patterns of Natural History, *Soc. Vert. Paleontol. Mem.*, no. 4, suppl. to *J. Vert. Paleontol.*, vol. 18, no. 1, 1998; G. V. Lauder and K. Liem, The evolution and interrelationships of the actinopterygian fishes, *Bull. Mus. Comp. Zool.*, 150:95–107, 1983; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; J. R. Paxton and W. N. Eschmeyer (eds.), *Encyclopedia of Fishes*, University of New South Wales Press, 1994; M. Stiassny, L. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996.

Action

Any one of a number of related integral quantities which serve as the basis for general formulations of the dynamics of both classical and quantum-mechanical systems. The term has been associated with four quantities: the fundamental action S, for general paths of a dynamical system; the classical action S_C, for the actual path; the modified action S', for paths restricted to a particular energy; and action variables, for periodic motions.

Least-action principles. A dynamical system can be described in terms of some number N of coordinate degrees of freedom that specify its configuration. As the vector q whose components are the degrees of freedom $q_1, q_2, \ldots, q_N$ varies with time t, it traces a path $q(t)$ in an N-dimensional space. The fundamental action S is the integral of the lagrangian of the system taken along any path $q(t)$, actual or virtual, starting from a specified configuration q_1 at a specified time t_1, and ending similarly at configuration q_2 and time t_2. The value of this action $S[q(t)]$ depends on the particular path $q(t)$. The actual path $q_C(t)$ which is traversed when the system moves according to newtonian classical mechanics gives an extremum value of S, usually a minimum, relative to the other paths. This is Hamilton's least-action principle. The extremum value depends only on the end points and is called the classical action $S_C(q_1, q_2; t_1, t_2)$. *See* DEGREE OF FREEDOM (MECHANICS); LAGRANGE'S EQUATIONS.

An important variant of Hamilton's principle applies when the virtual paths $q(t)$ are restricted to motions all of the same energy E, but no longer to a specific time interval, $t_1 - t_2$. The modified action $S' = S - E(t_1 - t_2)$ obeys a modified least-action principle, usually called Maupertuis' principle, namely, that the classical path gives again an extremal value of S' relative to all paths of that energy. Maupertuis' principle is closely related to Fermat's principle of least, time in classical optics for the path of light rays of a definite frequency through a region of inhomogeneous refractive index. *See* HAMILTON'S PRINCIPLE; LEAST-ACTION PRINCIPLE; MINIMAL PRINCIPLES.

Action variables. When some degree of freedom q_i executes a periodic or oscillatory motion at energy E, its contribution S_i to the action S' is given by Eq. (1),

$$S_i = \int p_i(q_i, E)dq_i \tag{1}$$

where p_i is the momentum conjugate to q_i. A new quantity, the action variable $J_i(E)$, can be defined for that degree of freedom, as the amount that the integral S_i increases for each cycle of motion of the coordinate q_i, that is, $J_i(E)$ is given by Eq. (2), where

$$J_i(E) = \oint p_i(q_i, E)dq_i \tag{2}$$

$\oint$ represents integration over one cycle of motion. Action variables have the important property of adiabatic invariance, namely, that they do not change during a process in which parameters λ, representing either external or internal features of the system, change sufficiently slowly, but not necessarily by a small amount, from λ to λ'. Although quantities, including the energy, change in general (say, from E to E'), J_i remains unchanged, so that $J_i(E; \lambda) = J_i(E'; \lambda')$. Two examples of such adiabatic invariants are those associated with a simple mechanical oscillator, and with a charged particle moving in a magnetic field.

Simple mechanical oscillator. For this case, λ may represent internal parameters such as the mass m or the

stiffness constant k. The single action variable $J(E)$ is the ratio $2\pi E/\omega$, where $\omega =$ is the angular frequency. The adiabatic invariance of J means that if the parameters change slowly enough that the relative change of frequency during each cycle is small ($\Delta\omega/\omega \ll 1$), the energy will change in proportion to the angular frequency. *See* HARMONIC OSCILLATOR.

Charged particle in magnetic field. For this case, there are several adiabatic invariants in situations where the field changes sufficiently slowly in space and time as to be nearly constant over distances of the order of the cyclotron orbit radius of the motion around the field lines. The principal adiabatic invariants are $J_\perp$ and $J_{//}$, which are associated with cyclotron-orbit motion perpendicular to the field line and the motion of the orbit "guiding center" along the field line, respectively. Their adiabatic invariance underlies the basic features of confinement of charged particles in magnetic-bottle field configurations such as the Van Allen belt in the Earth's magnetosphere. The value of $J_\perp$ is $(2\pi mc/e)\mu_z$, where μ_z is the orbital magnetic moment parallel to the field. This moment is diamagnetic, thus tending to confine the particles to weaker fields, say, away from the Earth's polar regions, where they then execute periodic motion along the field lines. Under time-varying magnetic fields, induction effects will accelerate the particles and change their energy. For slow enough rates of variation, the energy changes are determined by the condition of near-invariance of the longitudinal action variable $J_{//}[E(t); \lambda(t)]$. *See* DIAMAGNETISM; MAGNETIC MOMENT; MAGNETOSPHERE; PLASMA (PHYSICS); VAN ALLEN RADIATION.

Use in old quantum theory. In the early period of development of quantum theory, action variables were recognized as being candidates from within classical mechanics which could possibly be assigned quantized values. Their invariance under slowly changing external fields would then maintain the observed discreteness of energy levels in atoms. For example, the invariance of E/ω led to Eq. (3) for the energy levels

$$E_n = n\hbar\omega \tag{3}$$

of an oscillator in the so-called old quantum theory, where $\hbar$ is Planck's constant divided by 2π, and n is any positive integer.

Significance in quantum mechanics. In quantum mechanics, as originally formulated by E. Schrödinger, the state of particles is described by wave functions which obey the Schrödinger wave equation. States of definite energy in, say, atoms are described by stationary wave functions, which do not move in space. Nonstationary wave functions describe transitory processes such as the scattering of particles, in which the state changes. Both stationary and nonstationary state wave functions are determined, in principle, once the Schrödinger wave propagator (also called the Green function) between any two points q_1 and q_2 is known. In a fundamental restatement of quantum mechanics, R. Feynman showed that all paths from q_1 to q_2, including the virtual paths, contribute to the wave propagator. Each path contributes a complex phase-term $\exp i(\phi[q(t)])$, where the phase ϕ is proportional to the action, $S[q(t)]$, for that path, being given by Eq. (4). The resulting sum over paths, appropriately

$$\phi[q(t)] = \frac{S[q(t)]}{\hbar} \tag{4}$$

defined, is the path integral (or functional integral) representation of the Schrödinger wave propagator. The path integral has become the general starting point for most formulations of quantum theories of particles and fields. The classical path $q_C(t)$ of least action now plays the role in the wave function as being the path of stationary phase $\phi_{\text{stationary}}$, which is equal to $S_C(q_1, q_2; t_1, t_2)/\hbar$. *See* GREEN'S THEOREM; PROPAGATOR (FIELD THEORY).

Semiclassical approximation. The functional integration can be performed explicitly for only very few systems. An important case where it can be carried out explicitly, but approximately, is when large quantum fluctuations Δq of the paths $q(t)$ away from the extremal path $q_C(t)$, on the scale of $\hbar^{1/2}$, are neglected. This results in the very useful semiclassical approximation. The only surviving phase factor is now the classical stationary phase, so that the semiclassical approximation also establishes the classical least-action principle in the limit that $\hbar$ approaches 0, that is, the short-wavelength limit of the quantum theory.

WKB approximation. The semiclassical approximation for the propagator at a definite energy is similar, but is based on the phase $S'/\hbar$ and Maupertuis' least-action principle. Its most common use is for approximating stationary states of motions that are restricted to one coordinate dimension. Here it is known as the Wentzel-Kramers-Brillouin (WKB) approximation, and bound-state energies E_n are determined by the condition $J_i\ (E_n) = 2\pi(n + \delta)$. The factor δ represents a correction to the old quantum theory due to wavelike behavior at the end points of the motion. For the simple oscillator discussed above, δ has the value $^1/_2$ and changes the energies to those given by Eq. (5).

$$E_n = \left(n + {}^1/_2\right)\hbar\omega \tag{5}$$

See HAMILTON-JACOBI THEORY; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; WENTZEL-KRAMERS-BRILLOUIN METHOD.

Bernard Goodman

Bibliography. W. Dittrich and M. Reuter, *Classical and Quantum Dynamics: From Classical Paths to Path Integrals*, 3d ed., 2001; H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; R. J. Goldston and P. H. Rutherford, *Introduction to Plasma Physics*, 1995; L. D. Landau and E. M. Lifshitz, *Quantum Mechanics, Nonrelativistic Theory,* 3d ed., 1977; L. S. Schulman, *Techniques and Applications of Path Integration,* 1981, reprint 2005.

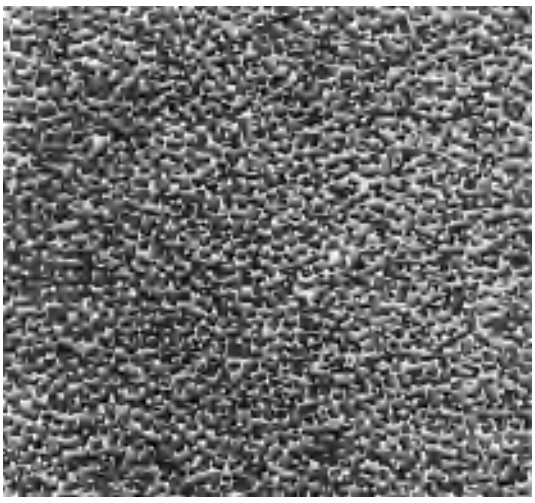
Activated carbon

A powdered, granular, or pelleted form of amorphous carbon characterized by very large surface area per unit volume because of an enormous number of fine pores. Activated carbon is capable of collecting gases, liquids, or dissolved substances on the surface of its pores. For many gases and liquids, the weight of adsorbed material approaches the weight of the carbon. *See* ADSORPTION.

Adsorption on activated carbon is selective, favoring nonpolar over polar substances and, in a homologous series, generally improving with increasing boiling point. Adsorption is also improved with increased pressure and reduced temperature. Reversal of the physical adsorptive conditions (temperature, pressure, or concentration) more or less completely regenerates the carbon's activity, and frequently allows recovery of both the carrier fluid and adsorbate. Compared with other commercial adsorbents, activated carbon has a broad spectrum of adsorptive activity, excellent physical and chemical stability, and ease of production from readily available, frequently waste materials.

Large-pore decolorizing carbons are used in liquid-phase work. Applications include improving the color of manufactured chemicals, oils, and fats, as well as controlling odor, taste, and color in potable water supplies, beverages, and some foods. Gas-adsorbent carbons are generally harder, higher-density, finer-pore types useful in gas separations, recovering solvent vapors, air conditioning, gas masks, respirators, carbon canisters in automobiles, and supporting metal salt catalysts, particularly in the production of vinyl-resin monomers.

Almost any carbonaceous raw material can be used for the manufacture of activated carbon. Wood, peat, and lignite are commonly used for the decolorizing materials. Bone char made by calcining bones is used in large quantity for sugar refining. Nut shells (particularly coconut), coal, petroleum coke, and other residues in either granular, briqueted, or pelleted form (see **illus.**) are used for adsorbent products.



Activated carbon pellets, 4- to 6-mesh.

Activation. This is the process of treating the carbon to open an enormous number of pores in the 1.2- to 20-nanometer-diameter range (gas-adsorbent carbon) or up to 100-nanometer-diameter range (decolorizing carbons). After activation, the carbon has the large surface area (500–1500 m^2/g) responsible for the adsorption phenomena. Carbons that have not been subjected previously to high temperatures are easiest to activate. Selective oxidation of the base carbon with steam, carbon dioxide, flue gas, or air is one method of developing the pore structure. Other methods require the mixing of chemicals, such as metal chlorides (particularly zinc chloride) or sulfides or phosphates, potassium sulfide, potassium thiocyanate, or phosphoric acid, with the carbonaceous matter, followed by calcining and washing the residue. The economics of the latter process requires recovery of the chemical agent.

Tests. Tests to describe activated carbon's ability to perform are designed to simulate operating conditions. Carbon tetrachloride activity shows a gas-adsorbent carbon's capacity for vapors, and is the percentage by weight of carbon tetrachloride adsorbed at 25°C (77°F) from dry air saturated at 0°C (32°F). Retentivity, which correlates with a carbon's ability to remove low concentrations of vapor from a gas stream, is then determined by blowing dry air through the saturated carbon. Iodine activity indicates a carbon's ability to remove iodine from a standard stock solution and is used in specifications for liquid purification carbon. Quantitative liquid decolorizing evaluations are frequently determined by adding varying amounts of carbon to a standard series of solution aliquots and plotting (Freundlich isotherm) on logarithmic paper the concentration of adsorbate remaining in solution against the ratio of adsorbed material to weight of carbon. Minute service is applied to gas mask carbons, and represents the length of time during which a thin bed of activated carbon will completely adsorb chloropicrin gas. The break point occurs when penetration of gas through the bed activates a detection device downstream. Hardness, or strength, of the coarser activated carbons is calculated from the change in screen analysis experienced after mechanically abrading the carbon. Other quality tests, such as moisture and ash content, bulk or apparent density, and screen analysis, are also used. *See* CARBON; CHARCOAL; DESTRUCTIVE DISTILLATION; RESPIRATOR.

H. Burnham Allport

Bibliography. American Water Works Assoc., *Design and Use of Granular Activated Carbon*, 1989; H. Janowska, A. Swiatowski, and J. Choma, *Active Carbon*, 1991.

Activation analysis

A technique in which a neutron, charged particle, or gamma photon is captured by a stable nuclide to produce a different, radioactive nuclide which is then measured. The technique is specific, highly sensitive, and applicable to almost every element

in the periodic table. Because of these advantages, activation analysis has been applied to chemical identification problems in many fields of interest.

Neutron method. Neutron activation analysis (NAA) is the most widely used form of activation analysis. In neutron activation analysis the sample to be analyzed is placed in a nuclear reactor where it is exposed to a flux of thermal neutrons. Some of these neutrons are captured by isotopes of elements in the sample; this results in the formation of a nuclide with the same atomic number, but with one more mass unit of weight. A prompt gamma ray is immediately emitted by the new nuclide, hence the term (n,γ) reaction, which can be expressed as reaction (1), where z refers to the atomic number and

$$ {}^{z}_{m}A(n,\gamma) \rightarrow {}^{z}_{m+1}A + \gamma \qquad (1) $$

m the atomic weight. Usually the product nuclide (A^{z}_{m+1}) is radioactive, and by measuring its decay products one can identify and quantify the amount of target element in the sample. The basic activation equation is given by Eq. (2), and enables one

$$ A = Nf\sigma\left(1 - e\frac{-0.693t}{t_{1/2}}\right) \qquad (2) $$

where A = activity of product nuclide (disintegrations per second)
N = atoms of target element
f = flux of neutrons (neutrons per cm^2-s)
σ = cross section of the target nuclide (cm^2)
$t_{1/2}$ = half-life of induced radioactive nuclide
t = time of irradiation

to calculate the number of atoms of an unknown target element by measuring the radioactivity of the product. These radioactive products usually decay by emission of a beta particle (negative electron) followed by a gamma ray (uncharged). Cross sections and half-lives are well known, and neutron fluxes can be measured by irradiation and measurement of known materials. When such techniques are used, the method is known as absolute activation analysis. A simple (and older) method is to simultaneously irradiate and compare with the unknown sample a standard containing known amounts of the elements in question. This is called the comparator technique.

Measurement techniques. Measurement of the induced radioactivities is the key to activation analysis. This is usually obtained from the gamma-ray spectra of the induced radionuclides. Gamma rays from radioactive isotopes have unique, discrete energies, and a device that converts such rays into electronic signals that can be amplified and displayed as a function of energy is a gamma-ray spectrometer. It consists of a detector [germanium doped with lithium, GeLi, or sodium iodide doped with thallium, NaI(Tl)] and associated electronics. Gamma rays interact in the detector to form photoelectrons, and these are then amplified and sorted according to energy. Peaks in the resulting gamma-ray spectra are called gamma-ray photo-peaks. By taking advantage of the different half-lives and different gamma-ray energies of the induced radionuclides, positive identification of many elements can be made. A computer, or provision for recording the spectrometric data in computer-compatible form, is almost a necessity. Calibration of counting conditions with a particular detector enables the activation analyst to relate the area under each gamma-ray photo-peak to an absolute disintegration rate; this supplies the A in Eq. (2). Such techniques are multielement, instrumental, and absolute. Where the sought element captures a neutron to produce a non-gamma-emitting nuclide, the analyst must make chemical separations and then beta-count the sample, or must go to another technique. *See* GAMMA-RAY DETECTORS.

Charged-particle method. Activation analysis can also be performed with charged particles (protons or He^{3+} ions, for example), but because fluxes of such particles are usually lower than reactor neutron fluxes and cross sections are much smaller, charged-particle methods are usually reserved for special samples. Charged particles penetrate only a short distance into samples, which is another disadvantage. A variant called proton-induced x-ray emission (PIXE) has been highly successful in analyzing air particulates on filters. Here the samples are all similar, and are low in total mass, and many of the elements of interest such as sulfur, calcium, iron, zinc, and lead are not easy to determine by neutron activation analysis. The protons excite prompt x-rays characteristic of the element, and these are measured. Prompt gamma rays have been used for measurement in some neutron activation analysis studies, but that method has had only limited success. Photon activation, using photons produced by electron bombardment of high-z targets such as tungsten, is another special variant of rather minor interest. Neutron sources other than reactors are sometimes used: Cf^{252} is an element that emits neutrons as it decays, and can be used for neutron activation analysis; and small accelerators called 14-MeV neutron generators produce a low flux of high–energy neutrons which are used primarily for determination of oxygen and nitrogen.

Applications. Activation analysis has been applied to a variety of samples. It is particularly useful for small (1 mg or less) samples, and one irradiation can provide information on 30 or more elements. Samples such as small amounts of pollutants, fly ash, very pure experimental alloys, and biological tissue have been successfully studied by neutron activation analysis. Of particular interest has been its use in forensic studies; paint, glass, tape, and other specimens of physical evidence have been assayed for legal purposes. In addition, the method has been used for authentication of art objects and paintings where only a small sample is available. Activation analysis services are available in numerous university, private, commercial, and United States government laboratories, and while it is not an inexpensive method, for many special samples it is the best and cheapest method of acquiring necessary data. *See* FORENSIC CHEMISTRY;

NUCLEAR REACTION; PARTICLE DETECTOR; RADIOISOTOPE. W. S. Lyon

Bibliography. Z. B. Alfassi (ed.), *Activation Analysis*, vols. 1 and 2, 1990; W. S. Lyon and H. H. Ross, Nucleonics, *Anal. Chem. (Annu. Rev.)*, 54(4):227R, 1982; 1976 International Conference on Modern Trends in Activation Analysis, *J. Radioanal. Chem.*, vols. 37–39, 1977; 1981 International Conference on Modern Trends in Activation Analysis, *J. Radioanal. Chem.*, vols. 69–71, 1982.

Active sound control

The modification of sound fields by using additional sound transmitted from loudspeakers. The signals from the loudspeakers are controlled electronically by using digital signal processing techniques.

Physical principles. The physical principle underlying the active control of sound is interference. That is, the sound pressure fluctuations produced at a given point in space by two sources of sound simply add together at each point in time. Thus if a secondary source of sound is made to produce the opposite pressure fluctuation to a given primary source at a given point in space, the net result will be silence at that point. *See* INTERFERENCE OF WAVES.

In 1877 Lord Rayleigh described an experiment in which he used electromagnets driven from the same electrical source to synchronize the vibrations of two tuning forks, placed about 9 m (30 ft) apart, at a frequency of 250 Hz. With one ear closed, he was able to delineate regions of silence, and he observed that moving his head about 2.5 cm (1 in.) was "sufficient to produce a marked revival of sound." **Figure 1** shows the results of a computer simulation of Rayleigh's experiment. The two sources of sound radiate spherical waves at the frequency of 250 Hz, and the fields superpose to give regions of both destructive and constructive interference. The region of complete silence detected by Rayleigh lies along the vertical line bisecting that connecting the two sources. Although it is indeed possible to cancel the field produced by a primary source by introducing a secondary source, the quiet zone produced is restricted in dimensions: the size of the zone is proportional to the wavelength of the sound. It is therefore possible, in general, to produce larger quiet zones at lower frequencies, where the wavelength of the sound is longer.

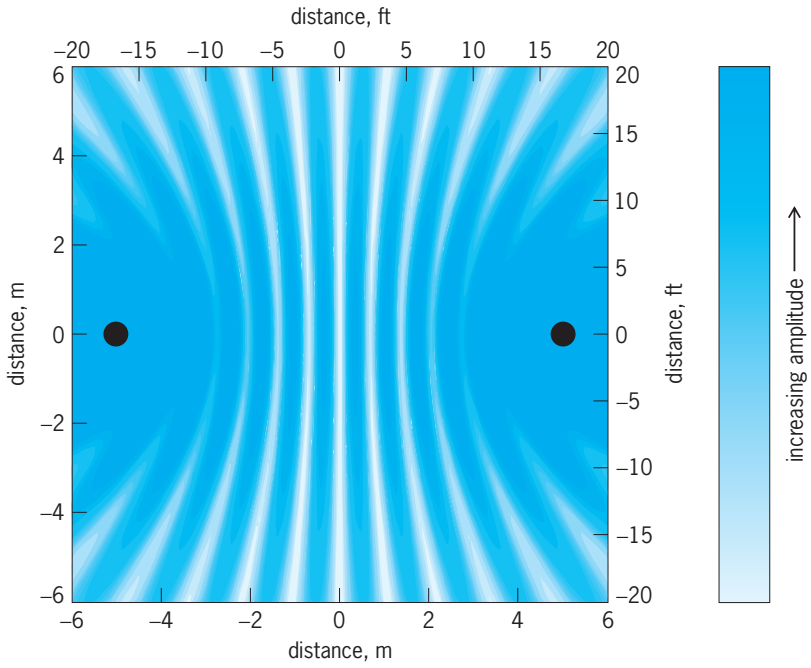


Fig. 1. Results of a computer simulation of the sound field produced by two sources of spherical sound waves in air that have identical source strengths and are separated by a distance of 10 m. The sources are out of phase and radiate sound at a frequency of 250 Hz. The amplitude of the resulting pressure fluctuation is shown.

It is also possible to interfere destructively with the sound field radiated by a primary source of single-frequency sound by placing a secondary source of the same frequency at a distance from the primary source which is very much less than the wavelength of the sound. In this case, if the secondary source is driven out of phase with the primary source, the amplitude of the sound is very much reduced over all space. The secondary source does not "destroy the energy" of the sound radiated by the primary source but merely prevents its radiation in the first place. If the secondary source is placed within one-tenth of a wavelength of the primary source, the amplitude of the sound is reduced everywhere (except very close to the sources) by about 10 decibels on average. This amounts roughly to a halving of the loudness of the sound. Since the wavelength of sound in air is about 3.4 m at 100 Hz (and 0.34 m at 1000 Hz), it is clear that the application of this approach is limited to localized primary sources radiating sound at the lowest frequencies of practical interest.

Application of digital signal processing. The developments in modern electronics during the latter part of the twentieth century enabled the practical implementation of the active control of sound with much greater ease than was possible in Lord Rayleigh's time (**Fig. 2**). A detection microphone is first used to sense the primary sound. The electrical signal from this microphone is passed through an analog low-pass filter, then sampled and converted into digital format. This resulting sequence of numbers is then passed through a digital filter prior to being converted back into an analog signal and fed to the loudspeaker comprising the secondary source. The value of the output from the digital filter at a given time is typically calculated from both the current and a number of previous values of the input sequence. These values are multiplied by a series of numbers comprising the filter coefficients before being added together to produce the output value at that particular time. These arithmetic operations can be carried out extremely fast on a special-purpose microprocessor that is used to implement the digital filter. An error microphone is then used to sense the degree of interference between the primary and secondary sound. *See* DIGITAL FILTER; MICROPROCESSOR.

The coefficients of the digital filter are chosen to ensure that the waveform of the sound radiated from the loudspeaker is aligned in time in order to

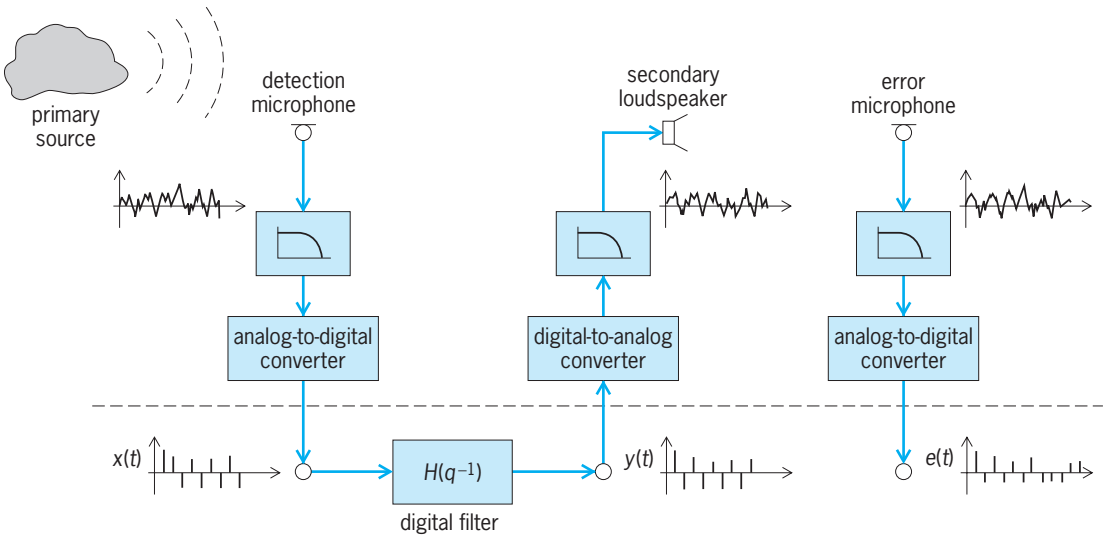


Fig. 2. Feed-forward active noise control system. Elements of the system below the broken line are entirely digital, the filter *H* operating on the input sequence *x* to produce the output sequence *y*. The characteristics of the filter are designed by using the error sequence *e*.

be (as far as possible) opposite to the waveform produced at the error microphone by the primary source. When the waveform of the primary sound is unpredictable, it is important to ensure that the time taken to process and transmit the secondary sound is sufficiently short that it arrives soon enough at the error microphone to cancel the primary sound. This basic approach to the problem is often referred to as feed-forward active control, since it makes use of the advanced warning of the arrival of unwanted sound. This is provided by the finite time required for sound to propagate from the detection microphone to the error microphone.

The design of the filter is often based on an approach that was developed during the 1940s by Norbert Wiener. This technique can be used to design a filter that, for example, minimizes the average over time of the squared value of the signal from the error microphone. Furthermore, the digital filter can be made adaptive so that the coefficients of the filter are rapidly updated at every time sample to ensure that the time-averaged value of the error signal is minimized. An algorithm for accomplishing this is a variant of the LMS algorithm developed by Bernard Widrow and Marcian (Ted) Hoff in the 1960s. Such a strategy can be very useful when, for example, the characteristics of transmission path from the primary source to the error microphone change with time.

Sound from the secondary source may also feed back into the detection microphone. Such a problem often arises when active control systems are used to reduce the amplitude of sound propagating in tubes or ducts (such as those in air-conditioning systems). It is possible to compensate for this effect by modeling the feedback path electronically with a further filter whose output is then subtracted from that of the detection microphone. *See* ADAPTIVE SIGNAL PROCESSING.

Multichannel systems. Some physical limitations of the technique can be overcome simply by increasing the number of secondary sources used to control the field. For example, if 10 loudspeakers are used to control a pure-tone sound field, it is possible in principle to produce 10 points of silence in the field. Of course, depending upon the geometrical arrangement and the acoustical environment of the primary and secondary sources, increases in level may well be produced at other positions. Another strategy is to drive the secondary sources in order to minimize the sound level at a number of error microphones which is larger than the number of secondary sources. An example of the practical application of this approach is in the control of propeller noise inside aircraft (**Fig. 3**). In this case the dominant source of unwanted noise is at the blade passage frequency of the propellers, typically below about 100 Hz. In addition, the interiors of such aircraft are usually relatively confined and the sound field produced by the rotation of the propellers excites a relatively small number of resonant modes of the aircraft cabin. Under these circumstances it is possible to produce widespread reductions in sound level. **Figure 4** shows the sound field both before and after when 16 secondary loudspeaker sources are used to minimize the sum of

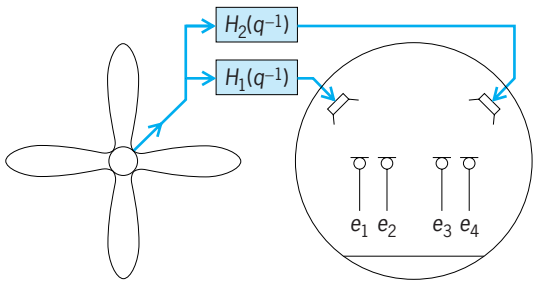


Fig. 3. Active control system for the suppression of propeller-induced aircraft cabin noise. The digital filters H_1 and H_2 are designed to minimize the sum of the squared error signals e_1–e_4.

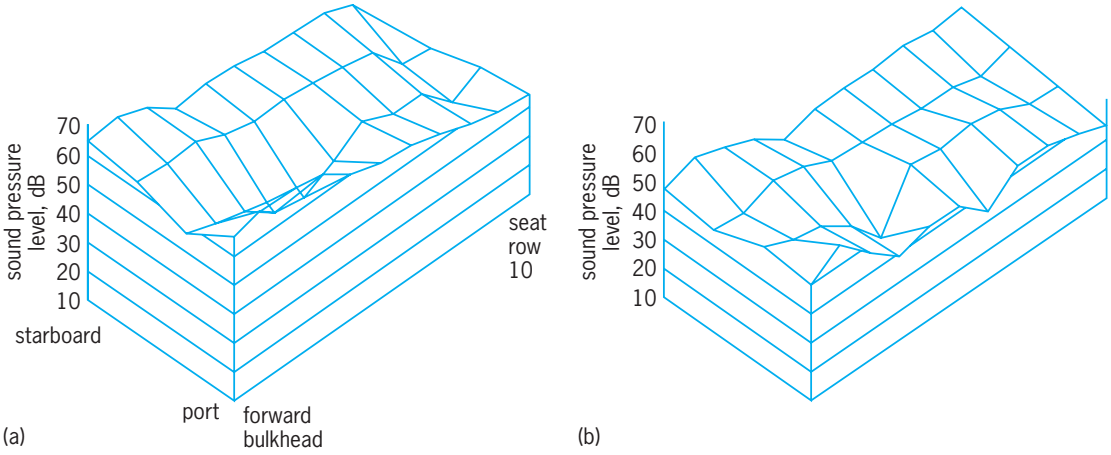


Fig. 4. Distribution of sound pressure level (in decibels on an arbitrary scale) at head height in a 48-seat propeller aircraft (*a*) before and (*b*) after the use of a multichannel active noise control system consisting of 16 loudspeakers and 32 microphones. The results are at a frequency of 88 Hz and show the reduction in the contribution due to the port propeller only. (*After S. J. Elliott et al., In-flight experiments on the active control of propeller-induced cabin noise, J. Sound Vib., 140:219–238, 1990*)

the squared pressure amplitudes at 32 error microphones distributed at head height through the cabin of a 48-seat passenger aircraft. Since the basic source of sound is due to the rotation of the propellers, all that is required is the detection of the frequency of the primary source (88 Hz in this example), and this is readily provided by filtering a signal from a tachometer on one of the engines. This reference signal is then sampled and passed through a digital filter associated with each secondary source. (This signal is uncorrupted by feedback from the secondary source outputs.) The coefficients of these filters are then adjusted adaptively in response to the signals sampled at the error microphones by using a multichannel generalization of the LMS algorithm. This ensures the minimization of the sum of the squared outputs from the error microphones and thus the suppression of the sound field.

Many acoustic waveforms of practical interest are highly unpredictable (unlike the propeller noise referred to above). Examples include the waveforms of the noise generated inside an automobile due to the airflow past the passenger cabin and the vibrations generated by the contact of the tires with the road. In addition, there may be multiple primary sources of such unwanted sound. In these cases it becomes more difficult to find reference signals that give a prior indication of the acoustic pressure fluctuations well before their arrival at the ears of the occupants. **Figure 5** shows a matrix of digital filters whose inputs are from a number of vibration sensors placed

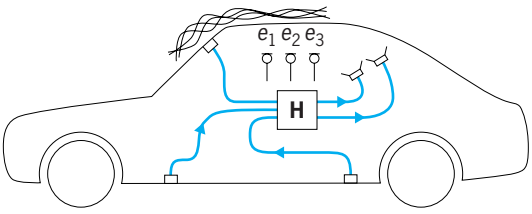


Fig. 5. Multichannel feed forward active control system for suppressing noise from multiple primary sources.

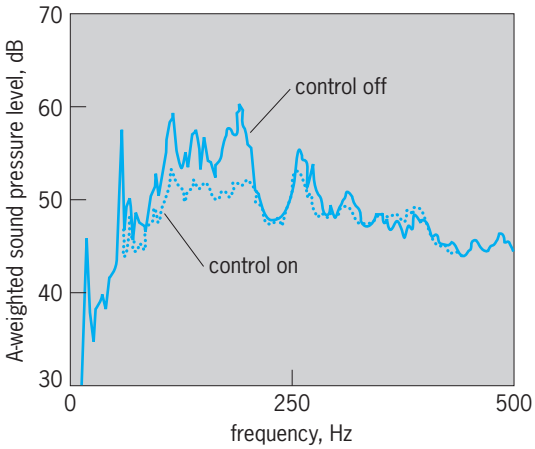


Fig. 6. Performance of a multichannel feed forward system for the active suppression ofroad-induced noise in a small passenger car at a speed of 60 km/h (37 mi/h). Reference signals were taken from six vibration sensors on the vehicle suspension. The A-weighted sound pressure level is shown before and after the application of active control. (*After T. J. Sutton et al., Active control of road noise inside vehicles, Noise Control Eng. J., 42:137–148, 1994*)

on the body of the automobile and whose outputs are used to drive the secondary sources. It is possible to find the optimal matrix of digital filters that ensures the minimization of the sum of the time-averaged signals from a number of error microphones. **Figure 6** shows the sound pressure level (in decibels) at a position close to the ear of a driver of a small passenger car both before and after the implementation of active control. The filter inputs were provided by reference signals from six vibration sensors placed at carefully chosen positions on the car body. Two secondary loudspeakers were used, each being fed by a combination of all the digitally filtered reference signals in order to minimize the sum of the time-averaged signals from two error microphones. It is clear that some reductions in sound level can be produced, but these are again restricted to frequencies below about 250 Hz. *See* ACOUSTIC NOISE; CONTROL SYSTEMS; SOUND.

Philip A. Nelson

Bibliography. S. Elliott, *Signal Processing for Active Control*, Acadmic Press, 2000; P. A. Nelson and S. J. Elliott, *Active Control of Sound*, Academic Press, 1992; A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Practical Applications*, McGraw-Hill, 1981, reprint, Acoustical Society of America, 1989; L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice Hall, 1975; B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.

Activity (thermodynamics)

The activity of a substance is a thermodynamic property that is related to the chemical potential of that substance. Activities are closely related to measures of concentration, such as partial pressures and mole fractions, and are more convenient to use than chemical potentials. The conditions that hold in chemical reaction equilibrium and in phase equilibrium can be expressed in terms of activities of the species involved.

The activity a_i of chemical species i in a phase (a homogeneous portion) of a thermodynamic system is defined by Eq. (1), where μ_i is the chemical po-

$$a_i \equiv \exp[(\mu_i - \mu_i^\circ)/RT] \tag{1}$$

tential of i in that phase, μ_i° is the chemical potential of i in its standard state, R is the gas constant, and T is the absolute temperature. Equation (1) shows that the activity a_i depends on the choice of standard state for species i. Since μ_i is an intensive quantity that depends on the temperature, pressure, and composition of the phase, a_i is an intensive function of these variables. a_i is dimensionless. When $\mu_i = \mu_i^\circ$, then $a_i = 1$. The degree of departure of a_i from 1 measures the degree of departure of the chemical potential from its standard-state value. The chemical potential is defined by $\mu_i \equiv \partial G/\partial n_i$, where G is the Gibbs energy of the phase, n_i is the number of moles of substance i in the phase, and the partial derivative is taken at constant temperature, pressure, and amounts of all substances except i. *See* CHEMICAL THERMODYNAMICS; FREE ENERGY.

From Eq. (1), it follows that μ_i is given by Eq. (2).

$$\mu_i = \mu_i^\circ + RT \ln a_i \tag{2}$$

The chemical potentials μ_i are key thermodynamic quantities of a phase, since all thermodynamic properties of the phase can be found if the chemical potentials in the phase are known as functions of temperature, pressure, and composition. Activities are more convenient to work with than chemical potentials, because the chemical potential of a substance in a phase goes to minus infinity as the amount of that substance in the phase goes to zero; also, one can determine only a value of μ_i relative to its value in some other state, whereas one can determine the actual value of each a_i.

Gases. For a pure gas i or a component i of any gas mixture, the standard state is defined as pure ideal gas i at the temperature T of the pure gas or mixture and at a standard pressure P° of 1 bar (100 kilopascals). Formerly, 1 atm was used as the standard-state pressure. Since real gases do not exhibit ideal behavior at 1 bar, the standard state of a gas is a hypothetical state.

For a component of an ideal gas mixture (one whose density is low enough that intermolecular forces are negligible), the chemical potential can be shown to be given by Eq. (3), where μ_i° is a function

$$\mu_i = \mu_i^\circ + RT \ln(P_i/P^\circ) \tag{3}$$

of T, and the partial pressure P_i of gas i is defined by $P_i \equiv x_i P$, where x_i is the mole fraction of i in the mixture and P is the mixture's pressure. Comparison with Eq. (2) shows that $a_i = P_i/P^\circ$ for a component of an ideal gas mixture. The motivation for the Eq. (1) definition of activity is to produce an expression for the chemical potential that closely resembles the expression for μ_i in an ideal system, so that the degree of departure from ideal behavior can be readily assessed. Thus, the general expression Eq. (2) for μ_i resembles Eq. (3) for an ideal gas mixture. For a component i of a real gas mixture, the fugacity f_i is defined as $f_i \equiv a_i P^\circ$, so that $a_i = f_i/P^\circ$ in a real gas mixture, as compared with $a_i = P_i/P^\circ$ for an ideal gas mixture. The fugacity coefficient ϕ_i of component i of a gas mixture is defined by $\phi_i \equiv f_i/P_i$; so $a_i = \phi_i P_i/P^\circ$. In an ideal gas mixture, each ϕ_i equals 1.

For a gas mixture at temperature and pressure T and P', the activity a_i of component i can be found by using Eq. (4) to find the fugacity coefficient ϕ_i, where

$$\ln \phi_i = \int_0^{P'} \left(\frac{\overline{V}_i}{RT} - \frac{1}{P} \right) dP \tag{4}$$

the integration is at constant T and composition and $\overline{V}_i$ is the partial molar volume of component i in the mixture. A reliable equation of state is often used for the mixture to find an approximate expression for $\overline{V}_i$ to use in Eq. (4).

Pure solids and pure liquids. For a pure solid or liquid, the standard state is defined as the pure substance at a pressure of 1 bar (100 kPa) and at the temperature T of the substance. For most pure solids or liquids at pressures below 20 bar (2000 kPa), a_i is quite close to 1.

Solutions. Several possible choices exist for standard states of the components of a liquid or solid solution.

One choice (called the symmetrical convention) is to treat all components on the same footing and take the standard state of each component i as pure substance i at the temperature and pressure of the solution. This choice is commonly made for solutions of miscible liquids. With this choice, the composition of the solution is expressed using mole fractions.

For an ideal solution (one where the components of the solution resemble each other so closely in molecular structure that the differences in molecular

size and shape and in intermolecular forces are negligible), one can show that the chemical potentials are given by Eq. (5), where μ_i^* is the chemical potential

$$\mu_i = \mu_i^* + RT \ln x_i = \mu_i^\circ + RT \ln x_i \qquad (5)$$

of pure liquid i at the temperature and pressure of the solution and x_i is the mole fraction of i in the solution. Thus, $a_i = x_i$ for an ideal-solution component.

For a component of a nonideal solution where the symmetrical convention is used, the activity is defined by Eq. (1) and the activity coefficient f_i of component i (not to be confused with the symbol used for fugacity in gases) is defined by $f_i \equiv a_i/x_i$, so $a_i = f_i x_i$. For an ideal solution, each f_i equals 1. In a nonideal solution, f_i measures the degree of departure from ideality and f_i goes to 1 as x_i goes to 1.

In the unsymmetrical convention, one component of the solution is designated as the solvent and the other components are called the solutes. The standard state of the solvent is chosen as pure solvent at the temperature and pressure of the solution. If the solution composition is specified using mole fractions, the standard state of solute i is chosen as the hypothetical state with $x_i = 1$ and solute i behaving as it would in an ideally dilute solution. An ideally dilute solution is one so extremely dilute that molecules of i interact only with solvent molecules. If the solution composition is specified using molalities m_i for solutes, then the standard state of solute i is the hypothetical state with $m_i = m^\circ \equiv 1$ mol/kg and i behaving as if in an ideally dilute solution. If molar concentrations c_i are used, the standard state has $c_i = c^\circ \equiv 1$ mol/L and i exhibiting ideally dilute behavior. These three choices and Eq. (1) give three different solute activities, $a_{x,i}$, $a_{m,i}$, and $a_{c,i}$. Three different solute activity coefficients, $\gamma_{x,i}$, $\gamma_{m,i}$, and $\gamma_{c,i}$, are defined to satisfy $a_{x,i} = \gamma_{x,i} x_i$, $a_{m,i} = \gamma_{m,i} m_i/m^\circ$, and $a_{c,i} = \gamma_{c,i} c_i/c^\circ$. All these solute-$i$ activity coefficients go to 1 as the solvent mole fraction goes to 1. *See* CONCENTRATION SCALES.

For an electrolyte solute C_cA_a, where c and a are the numbers of cations and anions, respectively, in the formula, Eq. (2) can be applied to the solute as a whole and also to each individual ion. The molality scale is usually used for the solute. The activity a_i of the electrolyte solute is related to the activities a_+ and a_- of its cation and anion, respectively, by $a_i = a_+^c a_-^a$. Since properties of individual ions are not experimentally determinable, a mean molal activity coefficient $\gamma_\pm$ is defined by the relation $(\gamma_\pm)^{c+a} = (\gamma_+)^c(\gamma_-)^a$. Values of $\gamma_\pm$ can be found from vapor-pressure measurements or galvanic-cell data.

Some workers take the standard-state pressure for each solution component as the fixed pressure of 1 bar.

Reaction equilibrium. The equilibrium constant for a chemical reaction can be expressed in terms of the equilibrium values of the activities of the reactants and products. *See* CHEMICAL EQUILIBRIUM.

Phase equilibrium. For a solution in equilibrium with its vapor (assumed to be ideal), the equilibrium condition that the chemical potential of substance i must be equal in each phase leads to the relation $P_i = f_i x_i P_i$, where P_i is the partial pressure of substance i in the vapor in equilibrium with the solution, P_i° is the vapor pressure of pure i at the temperature of the solution, x_i is the mole fraction of i in the solution, and f_i is the symmetrical-convention activity coefficient of i in the solution. This relation allows the activity coefficient and hence the activity of i in the solution to be found. If the vapor departs significantly from ideality, pressures are replaced by fugacities in this relation.

Ira N. Levine

Bibliography. K. Denbigh, *The Principles of Chemical Equilibrium*, Cambridge University Press, 4th ed., 1981; I. N. Levine, *Physical Chemistry*, 4th ed., McGraw-Hill, 1995.

Activity-based costing

A specific system or procedure for determining accurate manufacturing costs in order to achieve profitability goals. Traditional accounting systems allocate overhead costs by using a volume-oriented base, such as direct labor hours or direct material dollars. The cost allocation bases of direct labor or production quantity were designed primarily for inventory valuation. As a consequence, traditional cost accounting methods are fully effective only when direct labor (or direct materials) is the dominant cause of cost.

Cost models. While traditional standard cost systems were effective in the past, changes in manufacturing technologies, such as the just-in-time philosophy, robotics, and flexible manufacturing systems, made traditional cost models somewhat obsolete. Rapid technological advancement has resulted in the restructuring of manufacturing cost patterns (for example, the direct labor and inventory costs are decreasing, while those of technology depreciation, engineering, and data processing are increasing). Because of the changing nature of these types of costs, existing (that is, traditional) cost accounting systems often do not adequately support the objectives of advanced manufacturing. *See* FLEXIBLE MANUFACTURING SYSTEM.

In automated manufacturing facilities, distortion in product costs results from volume-based allocations such as those based on direct labor, because the bases used to allocate overhead do not cause the costs. As a result, product cost distortion occurs because of high overhead rates that are inflated by many costs that should be directly traceable to the product rather than arbitrarily allocated on the basis of direct labor or direct materials content. Several components of a product's cost that should be traced to the product include hidden overhead costs, such as for material movement, order processing, process planning, rework, warranty maintenance, production planning and control, and quality assurance. *See* MATERIALS HANDLING.

Key concepts. Activity-based costing systems track direct and indirect costs to the specific activities that

TABLE 1. Manufacturing costs for Models A and B (current year)

Category	Model A	Model B
Direct material cost	$2,800,600	$1,500,000
Direct labor cost	350,000	250,000
Direct labor hours	35,000	25,000
Total manufacturing overhead costs		$12,000,000

cause them and thus attempt to provide a more reliable estimate of product cost. In this approach, four key concepts differentiate activity-based costing from traditional costing systems: activity accounting, cost drivers, direct traceability, and non-value-added costs. These four basic concepts are embodied in activity-based costing systems and lead to more accurate costing information. In addition, activity-based costing systems provide more flexibility than conventional (for example, absorption-based) costing systems because they produce a variety of information useful for technology accounting, product costing, and design-to-cost.

Activity accounting. In an activity-based system, the cost of the product is the sum of all costs required to manufacture and deliver the product. The activities that a firm pursues consume its resources; and resource availability and usage create costs. Activity accounting decomposes an organization into its basic business processes and their associated activities in order to provide a cause-and-effect rationale for how fundamental objectives create costs and result in outputs.

Cost drivers. A cost driver is an activity that causes costs to be incurred. Familiar cost drivers include the number of machine setups, the number of engineering change notices, and the number of purchase orders. By controlling the cost driver, unnecessary costs can be eliminated, resulting in improved product cost.

Direct traceability. Direct traceability involves attributing costs to those products or processes that consume resources. Many hidden overhead costs can be effectively traced to products, thus providing more accurate information upon which to base decisions such as product costing, make-versus-buy, and investment justification.

Non-value-added costs. In manufactured products, customers may perceive that certain features add no value to the product. Through identification of cost drivers, a firm can pinpoint these sources of unnecessary cost. In fact, an activity-based cost system attempts to identify and place a cost on the manufacturing processes performed (either value-adding or non-value-adding) so that management can determine desirable changes in resource requirements for these activities. This approach is in contrast to traditional cost systems that accumulate costs by budgetary line items and by functions.

Application. An example will illustrate the key concepts of activity-based costing. A hypothetical firm, the XYZ Company, makes pagers for the telecommunications industry. Currently, the company is making two models of pagers with equal annual production volume of 50,000 each. Ever since its inception, the company has been using an absorption-based traditional costing system for product costing. Thus, the company uses direct labor hours as the allocation base for all its manufacturing overhead costs (**Table 1**).

Recently, the company installed an activity-based costing system and subsequently identified six main activities to be responsible for most of the manufacturing overhead costs. The overhead costs were distributed to these six activities through activity accounting. Also, six corresponding cost drivers and their budget consequences for the current year were established (**Table 2**). Besides establishing the activity-based costing data, the activity-based costing system also measured the levels of each activity (or cost driver rate) required by the two products (**Table 3**). For postaudit purposes the company ran a cost comparison between the two costing systems to arrive at product costs for two models, A and B. The two sets of computations involved the traditional (absorption-based) costing system and the activity-based costing system.

For the cost computation using the traditional method, the computation involved determining the overall manufacturing overhead cost rate, that is, total overhead costs divided by total direct hours. When the total cost per unit for each model was computed, the value for Model A was $203 and the value for Model B was $135.

A very different cost per unit resulted from cost computation using activity-based costing. In this method, the first step was to establish the applied activity rate of the six activities; the applied activity rate equals the activity cost divided by the budgeted cost driver rate in Table 2. In the second step, the activity cost is traced to each pager model, that is, the cost driver rate from Table 3 multiplied by the

TABLE 2. Activity-based costing data

Activity	Costs	Cost driver	Budgeted rate
Production	$8,000,000	Number of machine hours	200,000 hours
Engineering	1,000,000	Number of engineering change orders	40,000 orders
Material handling	1,000,000	Number of material moves	60,000 moves
Receiving	800,000	Number of batches	500 batches
Quality assurance	800,000	Number of inspections	20,000 inspections
Packing and shipping	400,000	Number of products	100,000 products

TABLE 3. Cost driver rates for Models A and B

Activity	Model A	Model B
Production	50,000 hours	150,000 hours
Engineering	15,000 orders	25,000 orders
Materials handling	20,000 moves	40,000 moves
Receiving	150 batches	350 batches
Quality assurance	6,000 inspections	14,000 inspections
Packing and shipping	50,000 products	50,000 products

TABLE 4. Product cost comparison

Costing system	Model A	Model B
Traditional	$203.00	$135.00
Activity-based	130.78	207.24

applied activity rate from Table 1. When the total cost per unit model was determined by this method, the cost per unit was $130 for Model A and $207 for Model B. There is a significant difference in the cost per unit as computed by these two different methods.

Thus the two costing systems produced different cost estimates (**Table 4**). Using absorption-based (traditional) costing, Model A costs more to make than Model B. However, based on activity-based costing, Model A is less expensive to make than Model B.

Traditional costing bases its cost estimates on the assumption that Model A is responsible for more overhead costs than Model B, using direct labor hours as an allocation base. However, because of differences in the number of transactions that affect indirect (overhead costs), the reality is that Model B actually incurs more overhead costs than Model A. This is shown by the analysis carried out with activity-based costing. As this example demonstrates, a traditional costing method generates distorted product costs that do not accurately reflect the actual amount of overhead costs incurred by the product. Companies making decisions based on distorted costs may unknowingly price some of their products out of the market while selling others at a loss. Or they may make key design decisions, based on competitor target costs, that do not reflect the true costs of manufacturing their new or revised products. *See* INDUSTRIAL ENGINEERING.

William G. Sullivan

Bibliography. J. A. Brimson, *Activity Accounting: An Activity-Based Costing Approach*, 1997; R. Cooper and R. S. Kaplan, Measure costs right: Make the right decisions, *Harvard Bus. Rev.*, September-October 1988; R. Cooper and R. S. Kaplan, *The Design of Cost Management Systems*, 2d ed., 1998; H. T. Johnson, Activity-based management: Past, present, and future, *Eng. Econ.*, 36(3):219–238, Spring 1991; R. S. Kaplan, New approaches to measurement and control, *Eng. Econ.*, 36(3):201–218, Spring 1991; J. K. Shank, Strategic cost management: New wine, or just new bottles, *Manag. Acc. Res.*, 1:47–65, Fall 1989.

Acylation

A process in which a hydrogen atom in an organic compound is replaced by an acyl group (R—CO where R = an organic group). The reaction involves substitution by a nucleophile (electron donor) at the electrophilic carbonyl group (C═O) of a carboxylic acid derivative. The substitution usually proceeds by an addition-elimination sequence [reaction]. Two common reagents, with the general formula RCOX, that bring about acylation are acid halides (X = Cl,Br) and anhydrides (X = OCOR). There are also other acylating reagents. The carboxylic acid (X = OH) itself can function as an acylating agent when it is protonated by a strong acid catalyst as in the direct esterification of an alcohol. Typical nucleophiles in the acylation reaction are alcohols (ROH) or phenols (ArOH), both of which give rise to esters, and ammonia or amines (RNH_2), which give amides. *See* ACID ANHYDRIDE; ACID HALIDE; AMIDE; AMINE; ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.

Carbon acylation. Acylation at a carbon atom can take several forms. Reaction of an acid halide with an organometallic compound such as a Grignard reagent (RMgX) is a rapid and efficient route to ketones. A widely used reaction is acylation of an aromatic ring with an acid chloride or anhydride in a Friedel-Crafts reaction. In acylation of the α-CH_2 group of a ketone or ester, hydrogen usually is first removed with a base to give the enolate, and the acyl group is then introduced. An acid chloride can be used, or since the enolate is a strong nucleophile, another ester can serve. *See* FRIEDEL-CRAFTS REACTION; GRIGNARD REACTION.

Esters are not generally useful as acylating agents unless the nucleophile is a strong base such as an enolate or the reaction is catalyzed by strong acid. However, thioesters are significantly more reactive acylating agents. In living systems, acylation occurs in several metabolic pathways; a thioester, acetyl coenzyme A, is the key intermediate in such reactions. *See* ESTER.

Another type of acylation is the reaction of one ester with a different alcohol to give a new ester. In this case the original ester is the acylating agent; the process is called transesterification and is exemplified by methanolysis of a fat. This reaction, which is usually carried out with an acid catalyst, is the first step in analysis of the fatty acid composition of fats.

Other carboxyl activation methods. Acylation can be thought of as the transfer of an acyl group from some reactive form of a carboxylic acid to a nucleophilic acceptor. As organic synthesis has become more refined and targets more complex, improved ways of activating carboxylic acids have been developed. These enable acylation to be carried out with a variety of acids, more selectively and under milder

conditions, than is possible with traditional acid chlorides or anhydrides.

One approach is to convert an acid to a mixed carboxylic-carbonic anhydride. These derivatives are prepared from the acid, ethyl chloroformate, and a tertiary amine at 0°C (32°F). They are more reactive than acid chlorides for acylation of amines or diazomethane.

Formation of an amide bond between two protected amino acids is an important process. This acylation is called peptide coupling, and it is the crucial reaction in the synthesis of peptides and proteins. Many methods have been developed for this purpose. One of the most enduring is the use of carbodiimide reagents, which effect activation of the acid and reaction with the amine in one operation. *See* PEPTIDE; PROTEIN.

James A. Moore

Bibliography. W. Carruthers and I. Coldham, *Modern Methods of Organic Synthesis*, 4th ed., 2004; R. J. Sundberg and F. A. Carey, *Advanced Organic Chemistry*, *Part B: Reaction and Synthesis*, 4th ed., 2001; B. M. Trost and I. Fleming (eds.), *Comprehensive Organic Synthesis*, vol. 6, 1991.

Adaptation (biology)

A characteristic of an organism that makes it fit for its environment or for its particular way of life. For example, the Arctic fox (*Alopex lagopus*) is well adapted for living in a very cold climate. Appropriately, it has much thicker fur than similar-sized mammals from warmer places; measurement of heat flow through fur samples demonstrates that the Arctic fox and other arctic mammals have much better heat insulation than tropical species. Consequently, Arctic foxes do not have to raise their metabolic rates as much as tropical mammals do at low temperatures. This is demonstrated by the coati (*Nasua narica*), which lives in Panama and has a body mass similar to Arctic foxes and about the same metabolic rate at comfortable temperatures. When both animals are cooled, however, the coati's metabolic rate starts to rise steeply as soon as the temperature falls below 68°F (20°C), while that of the Arctic fox begins to rise only below −22°F (−30°C). The insulation is so effective that Arctic foxes can maintain their normal deep-body temperatures of 100°F (38°C) even when the temperature of the environment falls to −112°F (−80°C). Thus, thick fur is obviously an adaptation to life in a cold environment. *See* THERMOREGULATION.

In contrast to the clear example above, it is often hard to be sure of the effectiveness of what seems to be an adaptation. For example, tunas seem to be adapted to fast, economical swimming. Their swimming muscles are kept warmer than the environment. The body has an almost ideal streamlined shape, rounded in front and tapering behind to a very narrow caudal peduncle immediately in front of the tail fin. The tail fin itself is tall and narrow, the shape that should enable it to propel the fish at least energy cost. Experiments with young tunas, however, have failed to show that they are faster or more economical than apparently less well adapted relatives.

A structure that evolved as an adaptation for one function may later come to serve a different function. This phenomenon is known as exaptation. For example, feathers seem to have evolved in the ancestors of birds as an adaptation for heat insulation, but in the evolution of wings feathers were exapted for flight. The wings in turn became an exaptation for swimming, when the penguins evolved.

Evolution by natural selection tends to increase fitness, making organisms better adapted to their environment and way of life. It might be inferred that this would ultimately lead to perfect adaptation, but this is not so. It must be remembered that evolution proceeds by small steps. For example, squids do not swim as well as fish. Fish swim by beating their tails, and squids by jet propulsion. Comparison of the swimming performance of a trout and a similar-sized squid showed that the fish could swim faster and with less energy cost. The squid would be better adapted for swimming if it evolved a fish-like tail instead of its jet propulsion mechanism, but evolution cannot make that change because it would involve moving down from the lesser adaptive summit before climbing the higher one. *See* ANIMAL EVOLUTION; ORGANIC EVOLUTION.

R. McNeill Alexander

Bibliography. R. McNeill Alexander, *Optima for Animals*, 2d ed., 1996; U. Dieckmann et al., *Adaptive Speciation*, 2004; S. H. Orzack and E. Sober, *Adaptationism and Optimality*, 2001; D. Schluter, *The Ecology of Adaptive Radiation*, 2000; T. Shanahan, *The Evolution of Darwinism*, 2004; E. R. Weibel, C. R. Taylor, and L. Bolis (eds.), *Principles of Animal Design*, 1998.

Adaptive control

A special type of nonlinear control system which can alter its parameters to adapt to a changing environment. The changes in environment can represent variations in process dynamics or changes in the characteristics of the disturbances. *See* NONLINEAR CONTROL THEORY.

A normal feedback control system can handle moderate variations in process dynamics. The presence of such variations is, in fact, one reason for introducing feedback. There are, however, many situations where the changes in process dynamics are so large that a constant linear feedback controller will not work satisfactorily. Control of a supersonic aircraft is a typical example. The dynamics of the airplane changes drastically with Mach number and dynamic pressure. A flight control system with constant parameters will not work well for an aircraft which operates over wide ranges of speeds and altitudes. *See* FLIGHT CONTROLS.

Adaptive control is also useful for industrial process control. In a given operating condition, most

processes can be controlled well with regulators with fixed parameters. Since delay and holdup times depend on the production, it would, however, be desirable to retune the regulators when there is a change in production. Adaptive control can also be used to compensate for changes due to aging and wear. Typical examples are variations in catalyst activity in chemical reactions and slow changes in heat transfer due to sediments. Wear in valves and mechanical systems are other examples. *See* PROCESS CONTROL.

Gain scheduling. It is sometimes possible to find auxiliary variables in a system which correlate well with the changes in process dynamics. It is then possible to eliminate the influences of parameter variations by changing the parameters of the regulator as functions of the auxiliary variables (**Fig. 1**). This method of eliminating variations in process dynamics is called gain scheduling. Gain scheduling could be considered an extension of feed-forward compensation. It can be seen from Fig. 1 that there is no way to correct for an incorrect schedule.

There is a controversy in nomenclature as to whether or not a system with gain scheduling should be considered an adaptive system. Gain scheduling is, nevertheless, a very useful technique for eliminating parameter variations. It is, in fact, the predominant method used to handle parameter variations in flight control systems. In that case, the Mach number and the dynamic pressure are measured by air data sensors and used as scheduling variables. The parameters of the flight control system are then determined by table look-up and interpolation.

One drawback of systems based on gain scheduling is that their design is time-consuming. The controllers must be designed for each operating condition. The interpolation method and the safe operation of the system must also be verified by extensive simulations. It is sometimes possible to obtain the gain scheduling by using normalized dimensionfree parameters. The auxiliary measurements are used together with the process measurements to obtain the normalized variables. The normalized control variable is calculated as the output of a linear constant-coefficient system driven by the normalized measurements. The control variable is retransformed before it is applied to the process.

Model reference adaptive systems. In model reference adaptive systems (MRAS) the dynamic specifications are given in terms of a reference model which tells how the process output ideally should respond to the command signal. The reference model is part of the control system (**Fig. 2**). The regulator can be thought of as consisting of two loops. The inner loop is an ordinary control loop composed of the process and a regulator. The parameters of the regulator can be adjusted, and the adjustments are made in the outer loop, which attempts to drive the regulator parameters in such a way that the error between the model output y_m and the process output y becomes small. The outer loop thus also looks like a regulator loop. The key problem is to determine the adjustment mechanism so that a stable system which brings the error to zero is obtained. This problem is nontrivial. It is easy to show that it cannot be solved with a simple linear feedback from the error to the controller parameters.

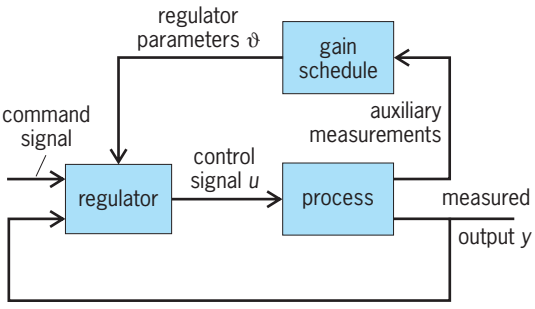


Fig. 1. Block diagram of a system where parameter variations are eliminated by gain scheduling.

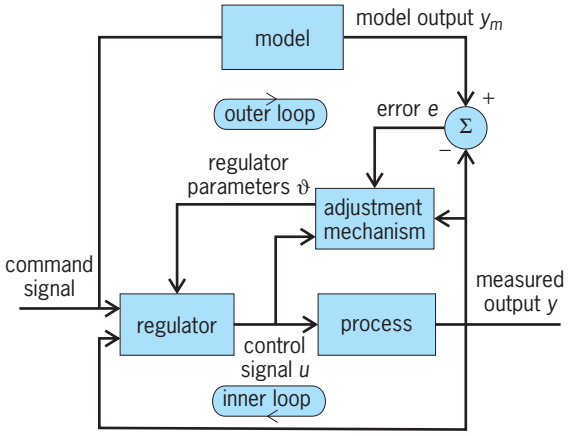


Fig. 2. Block diagram of model reference adaptive system (MRAS).

The parameter adjustment mechanism of Eq. (1),

$$\frac{d\vartheta_i}{dt} = -k\frac{\partial e}{\partial \vartheta_i} \cdot e \quad i = 1, \ldots, n \tag{1}$$

called the MIT rule, was used in the original MRAS. The variables $\vartheta_1, \ldots, \vartheta_n$ are the adjustable regulator parameters, $e = y_m - y$ is the error, $\partial e/\partial \vartheta_i$ are the sensitivity derivatives, and k is a parameter which determines the adaptation rate. Equation (1) represents a parameter adjustment mechanism which is composed of three parts: a linear filter for computing the sensitivity derivatives from process inputs and outputs, a multiplier, and an integrator. This configuration is typical for many adaptive systems.

The MIT rule can unfortunately give an unstable closed-loop system. The rule can be modified using Lyapunov or Popov stability theory. But it was only in the late 1970s that real progress was made in the theory of stability for adaptive systems.

Self-tuning regulators. The self-tuning regulator (STR; **Fig. 3**) can be thought of as composed of two loops. The inner loop consists of the process and an ordinary linear feedback regulator. The parameters of the regulator are adjusted by the outer loop, which is composed of a recursive parameter estimator and

a design calculation. There are many variants of the self-tuners because there are many combinations of design and parameter estimation schemes.

The regulator shown in Fig. 3 is called a regulator based on identification of an explicit process model. It is sometimes possible to reparametrize the process in such a way that the process is expressed in terms of the regulator parameters. The self-tuning regulator is then considerably simplified, because the design calculations are eliminated. Such a self-tuner is called an algorithm, based on estimation of an implicit process model. It is very similar to a model reference adaptive system because the parameter estimator can be interpreted as an adjustment mechanism for the regulator parameters, as can be seen by comparison of Figs. 2 and 3.

The self-tuning regulators can be used in several different ways. Since the regulator becomes an ordinary constant-gain feedback if the parameter estimates are kept constant, it can be used as a tuner whose purpose is to adjust the parameters of a control loop. In this case, the self-tuner is connected to the process, which is run until satisfactory performance is obtained. The self-tuner is then disconnected, and the system is left with the constant regulator parameters obtained. Since the tuning is done automatically, it is possible to use control algorithms with many adjustable parameters. The self-tuner can also be used to build up a gain schedule. The system is then run with the self-tuner at different operating points. The controller parameters obtained are stored. In this way, it is possible to obtain suitable regulator settings for different operating conditions. The self-tuner can, of course, also be used as a true adaptive controller for systems with varying parameters.

Stochastic adaptive control. Regulator structures such as MRAS and STR are based on heuristic arguments. It would be appealing to obtain the regulators from a unified theoretical framework. This can, in principle, be done by using nonlinear stochastic control theory. The system and its environment are then described by a stochastic model. The criterion is formulated so as to minimize the expected value of a loss function which is a scalar function of states and controls. *See* STOCHASTIC CONTROL THEORY.

The problem of finding a control which minimizes the expected loss function is difficult. Useful explicit conditions for existence are not known in general.

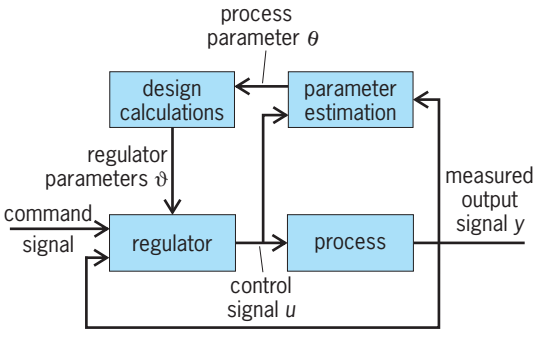


Fig. 3. Block diagram of a self-tuning regulator (STR).

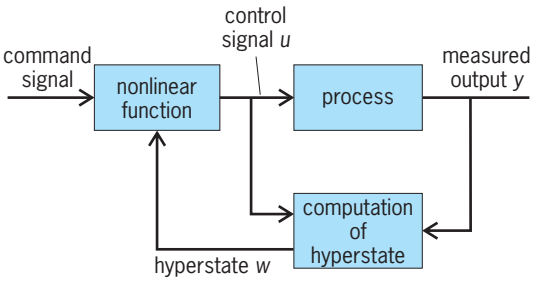


Fig. 4. Block diagram of an optimal nonlinear stochastic controller. The hyperstate is generated from a dynamical system using *u* and *y* as inputs. The regulator is a static nonlinear function which gives the control variable as a function of the hyperstate and the command signal.

Under the assumption that a solution exists, a functional equation for the optimal loss function can be derived by using dynamic programming. This equation, which is called the Bellman equation, can be solved numerically only in very simple cases. Nevertheless, the approach is of interest because it gives an insight into the structure of the optimal controller (**Fig. 4**). The controller can be thought of as composed of two parts, an estimator and a feedback regulator. The estimator generates the conditional probability distribution of the state from the measurements. This distribution is called the hyperstate of the problem. The feedback regulator is a nonlinear function which maps the hyperstate into the space of control variables. To solve a nonlinear stochastic control problem, it is necessary to determine the estimator, that is, the formula for updating the hyperstate, and to solve the Bellman equation. The structural simplicity of the solution is obtained at the price of introducing the hyperstate, which is a quantity of very high dimension. For a problem where the state space is R^4 (four-dimensional euclidean space), the hyperstate is, for example, a distribution over R^4. *See* ESTIMATION THEORY; OPTIMAL CONTROL THEORY.

Since it is difficult to solve the Bellman equation, approximative solutions are of considerable interest. A simple example will be used to illustrate some common approximations. Consider a process described by Eq. (2), where u is the control, y the

$$y(t+1) = y(t) + bu(t) + e(t) \tag{2}$$

output, e white noise, and b a constant parameter. Equation (2) can be interpreted as a sampled-data model of an integrator with unknown gain. Let the criterion be to minimize expression (3), where E de-

$$\lim_{N\to\infty} E\frac{1}{N}\sum_{1}^{N} y^2(t) \tag{3}$$

notes the mathematical expectation.

If the parameter b is known, the control law which minimizes (3) is given by Eq. (4). If the parameter

$$u(t) = \frac{1}{b}y(t) \tag{4}$$

b has a gaussian prior distribution, it follows that the conditional distribution of b, given inputs and outputs up to time t, is gaussian with mean $\hat{b}(t)$

and variance $P(t)$. The hyperstate of the problem can then be characterized by the triple $[y(t), \hat{b}(t), P(t)]$. In this simple case, the Bellman equation can be solved numerically. The control law obtained cannot be characterized in a simple way. It has, however, a very interesting property. Not only will the control signal attempt to bring the output close to zero; but also, when the parameters are uncertain, the regulator will inject signals into the system to reduce the uncertainty of the parameter estimates. The optimal control law will give the right balance between keeping the control errors and the estimation errors small. This is called dual control.

The control law given by Eq. (5) is an approxi-

$$u(t) = -\frac{1}{\hat{b}(t)}y(t) \tag{5}$$

mative solution. This control is called the certainty equivalence control. It is obtained simply by solving the control problem in the case of known parameters and substituting the known parameters with their estimates. The self-tuning regulator can be interpreted as a certainty-equivalence control.

The control law given by Eq. (6) is another

$$\begin{aligned} u(t) &= -\frac{\hat{b}(t)}{\hat{b}^2(t) + P(t)}y(t) \\ &= -\frac{1}{\hat{b}(t)} \cdot \frac{\hat{b}^2(t)}{\hat{b}^2(t) + P(t)}y(t) \end{aligned} \tag{6}$$

approximation, called cautious control because it hedges and uses low gain when estimates are uncertain. The cautious control law minimizes the criterion given by expression (7), but it is not optimal for Eq. (4).

$$E[y^2(t)/y(t-1), y(t-2), \ldots] \tag{7}$$

Status and prospects. The phrase adaptive control is unfortunately used in many different ways and its precise meaning is subject to controversy. The field is nevertheless in a state of rapid development. The availability of microprocessors, which make it possible to implement the controllers economically, has been a strong driving force.

The theory of adaptive control is still in its infancy. Major steps have been taken toward an understanding of the stability problem, but much research is needed to develop the appropriate conceptual framework and the key theoretical problems. *See* CONTROL SYSTEMS.

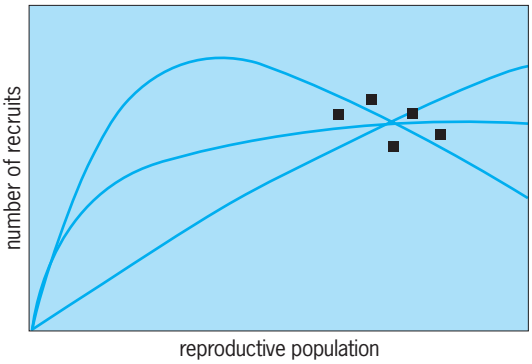
Karl Johan Åström

Bibliography. K. J. Åström and B. Wittenmark, *Adaptive Control*, 2d ed., 1994; C. C. Hang, T. H. Lee, and W. K. Ho, *Adaptive Control*, 1993; R. Isermann et al., *Adaptive Digital Control Systems*, 1991; P. P. Kokotovic, *Foundations of Adaptive Control*, 1991; K. S. Narendra, R. Ortega, and P. Dorato (eds.), *Advances in Adaptive Control*, 1991.

Adaptive management

An approach to management of natural resources that emphasizes how little is known about the dynamics of ecosystems and that as more is learned management will evolve and improve. Natural systems are very complex and dynamic, and human observations about natural processes are fragmentary and inaccurate. As a result, the best way to use the available resources in a sustainable manner remains to be determined. Furthermore, much of the variability that affects natural populations is unpredictable and beyond human control. This combination of ignorance and unpredictability means that the ways in which ecosystems respond to human interventions are unknown and can be described only in probabilistic terms. Nonetheless, management decisions need to be made. Adaptive management proceeds despite this uncertainty by treating human interventions in natural systems as large-scale experiments from which more may be learned, leading to improved management in the future.

Experimental work. For example, consider the case of a newly developed fishery. Managers must decide how much can be harvested from a given fish population in order to maximize profits without jeopardizing the viability of the population and the sustainability of the fishery. The adequate rate of harvest will depend on the effect that reducing the size of the reproductive population has on the subsequent number of juveniles recruiting annually into the fishery. All fisheries management hinges on identifying the stock size (number of reproductives in the population) that can produce a maximum sustained harvest and also ensure against overexploitation or chance extinction. In order to identify this target stock size, how the number of recruits (new fish added to the population by reproduction each year) varies with stock size must be determined, which requires collecting additional data (see **illus.**). Using adaptive management allows heavy harvesting, thereby lowering the stock size. Doing this management experiment should allow the identification of the appropriate recruitment function, and the progression toward a better-informed management



Plausible curves relating number of recruits (squares) to stock size. The level of harvest that a fish population can sustain depends on the relationship between the number of reproductive individuals and the subsequent number of recruits. Several hypotheses are consistent with the available data, each of which would lead to a different management policy. A large-scale management experiment involving harvesting down the population and observing recruitment responses would help to discriminate among the alternative models.

policy. Of course, in the real world this effort is further clouded by the fact that recruitment varies enormously from year to year due to the vagaries of the environment, quite independent of stock size. Delaying action to collect more information will not help, because in order to acquire more information the system must be moved to states at which the different models predict different responses.

This example illustrates a general problem that managers and environmental policy makers frequently face: The information available about natural systems is normally consistent with several competing hypotheses, each of which may dictate a different best course of action. Experimental work and basic science may help to postulate alternative hypotheses about possible consequences of management actions. However, it is generally difficult to extrapolate results from experiments conducted at small temporal and spatial scales directly to the actual ecosystem. Intervention in the real world and conducting large-scale management experiments are ways to learn about methods of management. But to be effective, adaptive management has to go far beyond a simple trial-and-error approach: it must involve the design and implementation of a formal plan for action, a plan by which management decisions are determined based on available knowledge and relevant information gaps, population responses are monitored, and future management is refined based on what is learned from the effects of previous management actions.

Stages. A key first step in the development of an adaptive management program is the assessment of the problem. During this stage, existing knowledge and interdisciplinary experience is synthesized and formally integrated by developing a dynamic model of the system. This modeling exercise helps to identify key information gaps and to postulate hypotheses about possible system responses to human intervention consistent with available information. Different management policies have to be screened in order to narrow down the alternatives to a few plausible candidates. Possible outcomes, including biological, economical, and societal variables of interest, have to be anticipated as much as possible based on the information gathered. The assessment of the problem can be accomplished through a series of workshops with participation of the different stakeholders together with the scientists and managers. Because effective management requires cooperation from stakeholders, their involvement during the initial planning stages is essential.

The second stage involves the formal design of a management and monitoring program. This usually involves the use of models to evaluate several alternative management policies in terms of their ability to meet management goals. To the extent that new information can result in improved future management, adaptive management programs may include large-scale experiments deliberately designed to accelerate learning. Some management actions may be more effective than others at filling the relevant information gaps. In cases where spatial replication is possible (such as small lakes, patches of forest, and reefs), policies that provide contrasts between different management units will be much more informative about the system dynamics than those that apply the same rule everywhere. For example, to control a pest population a predator is released into the field. To learn what level of control is most effective, a number of spatially discrete populations may be selected and different numbers of predators may be released into these units. Without experimental management (such as a single level of control) and without monitoring, little or no information will be gained to improve future management. This works much in the same way as treatment and control work in experimental science. In small-scale research experiments, however, it is much easier to control most sources of variability and to replicate the treatments so that results from the experiments can be unequivocally attributed to the factors tested. In the real world, most sources of variability cannot be controlled, and the effects of changes in environmental conditions tend to be confounded with those of the management actions. In the fishery problem discussed earlier, recruitment might improve as the reproductive population is harvested down; but if the policy is applied to a single population without controls, it will not be possible to determine if that is a consequence of management and not the effect of a more favorable environment. The lack of replicates can be a major impediment to learning through experimental management.

There are other barriers to the implementation of large-scale management experiments. Experiments usually have associated costs; thus, in order to be worthwhile, benefits derived from learning must overcompensate short-term sacrifices. Choices may be also restricted by social concerns or biological constraints, or they may have unacceptably high associated risks. There may be a well-justified reluctance to experiment with high-value, unique ecosystems. Risks can be especially great when management decisions have irreversible consequences, for example, the introduction of an exotic species as a biological control agent or for rearing purposes, which may prove harmful to the environment. The design of adaptive management programs involves anticipating possible consequences and evaluating trade-offs between expected costs and gains associated with the different candidate policies.

Once a plan for action has been chosen, the next stage is to implement the program in the field. This is one of the most difficult steps, because it involves a concerted and sustained effort from all sectors involved in the use, assessment, and management of the natural resources. Beyond the implementation of specific initial actions, putting in place an adaptive management program involves a long-term commitment to monitoring the compliance of the plan, evaluating the effects of management interventions, and adjusting management accordingly.

No matter how thorough and complete the initial assessment and design may have been, systems may always respond in manners that could not be

foreseen at the planning stage. Ecosystems exhibit long-term, persistent changes at the scale of decades and centuries; thus, recent experience is not necessarily a good basis for predicting future behavior. The effects of global climatic change on the dynamics of ecosystems, which are to a large extent unpredictable, will pose many such management challenges. Adaptive management programs have to include a stage of evaluation and adjustment. Outcomes of past management decisions must be compared with initial forecasts, models have to be refined to reflect new understanding, and management programs have to be revised accordingly. New information may suggest new uncertainties and innovative management approaches, leading to another cycle of assessment, design, implementation, and evaluation.

Applications. Adaptive management, although it has always been practiced in a more or less ad hoc manner, was formally developed in the 1970s by C. S. Holling and coworkers at the University of British Columbia, Canada. Since then, it has been applied to a variety of ecological problems involving the management of fisheries and forests, the control of introduced species in conservation problems, the recovery of essential habitat for threatened populations, the regulation of river flows, and the restoration of large-scale ecosystems, among others. Management of natural resources requires a much more extensive application of adaptive management approaches than is typical of current practices. Unless a wider application of adaptive management is begun, inadequate conservation of resources will continue. *See* CONSERVATION OF RESOURCES; ECOSYSTEM; FISHERIES ECOLOGY. Ana M. Parma

Bibliography. A. M. Parma et al., What can adaptive management do for our fish, forests, food, and biodiversity, *J. Integrative Biol.*, 1:16–26, 1998; C. J. Walters, *Adaptive Management of Renewable Resources*, Macmillan, New York, 1986; C. J. Walters, Challenges in adaptive management of riparian and coastal ecosystems, *Conserv. Ecol.* [online], 1:1–9, 1997; C. J. Walters and R. Green, Valuation of experimental management options for ecological systems, *J. Wildlife Manag.*, 1999.

Adaptive optics

The science of optical systems in which a controllable optical element, usually a deformable mirror, is used to optimize the performance of the system, for example, to maintain a sharply focused image in the presence of wavefront aberrations. A distinction is made between active optics, in which optical components are modified or adjusted by external control to compensate slowly changing disturbances, and adaptive optics, which applies to closed-loop feedback systems employing sensors and data processors, operating at much higher frequencies. *See* CONTROL SYSTEMS.

The practical development of adaptive optics started in the late 1960s. Its main applications have been to compensate for the effects of atmospheric turbulence in ground-based astronomical telescopes and to improve the beam quality of high-power lasers. Adaptive optics is now used routinely at several astronomical observatories.

In a typical adaptive optics system (**Fig. 1**), the distorted light beam to be compensated is reflected from the deformable mirror and is sampled by a beam splitter. The light sample is analyzed in a wavefront sensor that determines the error in each part of the beam. The required corrections are computed and applied to the deformable mirror whose surface forms the shape necessary to flatten the reflected wavefront. The result is to remove the optical error at the sampling point so that the light passing through the beam splitter may be focused to a sharp image. Nonlinear optical devices are also capable of performing some adaptive optics functions; these devices operate at high optical power levels. *See* ABERRATION (OPTICS); GEOMETRICAL OPTICS; NONLINEAR OPTICAL DEVICES.

Astronomical applications. There are two main limitations to the performance of large ground-based astronomical telescopes: (1) aberrations in the optical components, particularly the primary mirror, due to temperature variations and mechanical sagging at different pointing angles; and (2) atmospheric turbulence, which distorts the light waves received from distant objects. The angular resolution of a telescope is theoretically λ/D, where λ is the wavelength of the radiation and D is the diameter of the aperture. A 4-m (157-in.) telescope operating at visible wavelengths (0.5 μm) should ideally be capable of separating objects less than 0.03 arc-second apart. In practice, atmospheric turbulence limits the angular resolution of uncompensated telescopes to about 1 second of arc, regardless of their size. Modern telescopes make use of both active and adaptive optics to overcome these limitations.

Active optics. Active optics is employed to maintain the accuracy of the primary mirror surface to a fraction of the wavelength of the radiation being observed. Such active mirrors use either thin monolithic faceplates supported on an array of actuators, or segmented faceplates in which a number of individual, stiff panels are precisely positioned by control of their supporting points. Examples of the monolithic approach are the Very Large Telescope of the European Southern Observatory, which uses four 8.2-m (323-in.) faceplates, each supported by 150 actuators, and the Subaru telescope operated by the National Astronomical Observatory of Japan, which has a single 8.3-m (327-in.) faceplate controlled by 261 actuators. Segmented primary mirrors are used in the two 10-m (400-in.) Keck telescopes, each of which consists of 36 hexagonal elements adjustable in tip, tilt, and piston.

Another important application of active optics is in long-baseline interferometers, in which the optical path lengths of two or more arms must be precisely controlled. This is achieved by a fringe-tracking detector coupled to an actively controlled optical delay line.

Turbulence compensation. Adaptive optics is used in astronomical telescopes to compensate for the optical effects of atmospheric turbulence. Because of the large mass of primary mirrors, it is not feasible to drive them at the high speed necessary for turbulence compensation, which requires measuring and correcting the wavefront at rates between 100 and 1000 times per second. The adaptive optics systems used for this purpose employ small high-speed deformable mirrors, together with real-time wavefront sensors and data processors.

The data to drive the deformable mirror are provided by a very sensitive high-speed wavefront sensor that operates by using the radiation from a star or similar object. There are two fundamental limitations to the use of adaptive optics for turbulence compensation: the photon limit and the size of the isoplanatic patch. To make useful wavefront measurements, it is necessary to collect at least 100 photons from the reference source within the atmospheric change time. At visible wavelengths the limiting magnitude of the reference star is about the 10th visual magnitude while for infrared observations it may be 15th magnitude. The isoplanatic patch is the field of view around the reference star over which the wavefront measurement is valid. Because of turbulent layers high in the atmosphere, this angle is extremely small, often only a few seconds of arc at visible wavelengths. The result of these limitations is that turbulence compensation is possible only in the immediate vicinity of a few bright stars or similar objects.

The spatial characteristics of turbulence are a critical factor in the design of adaptive optics systems. Turbulence strength is specified in terms of the coherence length, r_0, which is defined statistically as the aperture diameter over which the wavefront has a mean square error of 1 radian squared; as the turbulence increases, the value of r_0 gets smaller. The significance of r_0, which is typically 10–20 cm (4–8 in.) at visible wavelengths, is that it determines the size of the compensation elements and consequently the number of actuators required for a given telescope aperture. At visible wavelengths, the number of actuators required to compensate a 4-m (157-in.) telescope is between 500 and 1000. The value of r_0 varies as the 6/5 power of wavelength, so that adaptive optics is much easier to implement at infrared wavelengths. Not only are fewer actuators required, but the turbulence changes at a lower rate and the isoplanatic angle is larger. At wavelengths above 10 μm, adaptive optics is no longer necessary. *See* COHERENCE; INFRARED ASTRONOMY; TURBULENT FLOW.

Laser beacons. The only method of obtaining good sky coverage at shorter wavelengths is the use of laser beacons, which may be created in the upper atmosphere either by backscatter from air molecules (Rayleigh scattering) at 10–40 km (6–25 mi) altitude or by stimulating fluorescence in the sodium layer at around 90 km (55 mi). *See* FLUORESCENCE; SCATTERING OF ELECTROMAGNETIC RADIATION.

The operating sequence (**Fig. 2**) is generally as follows. Laser pulses are projected through the telescope to produce a focused spot of length about 1.5 km (0.9 mi) and diameter about 1 m (3 ft). The backscattered light passes through the turbulent atmosphere, is collected by the telescope, and is

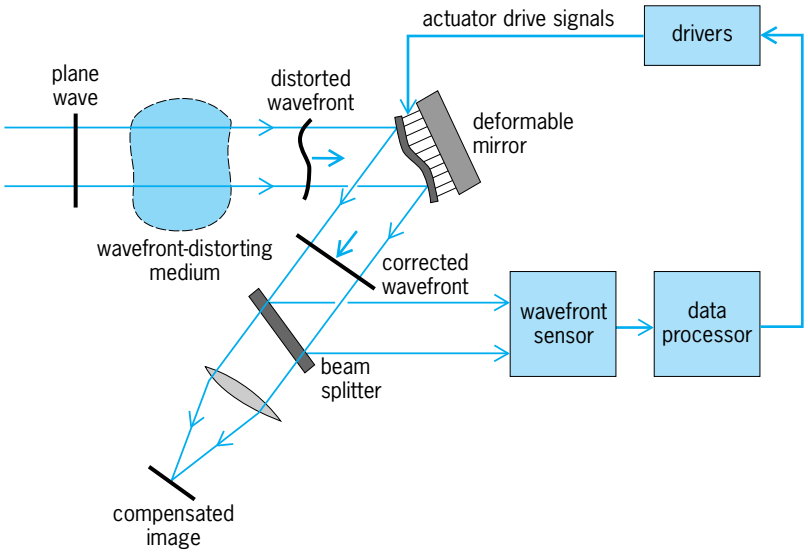


Fig. 1. Typical adaptive optics system using discrete components.

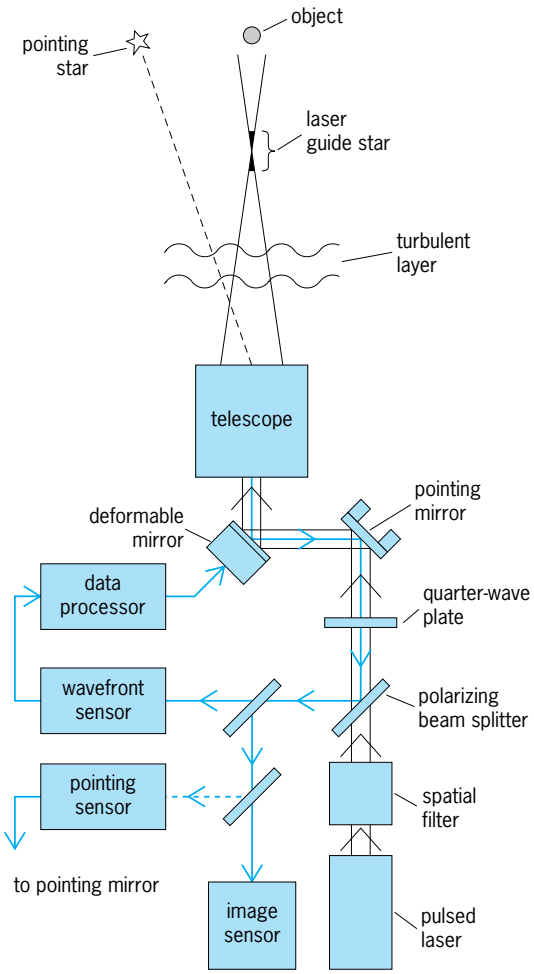


Fig. 2. Adaptive optical system using a laser beacon (guide star).

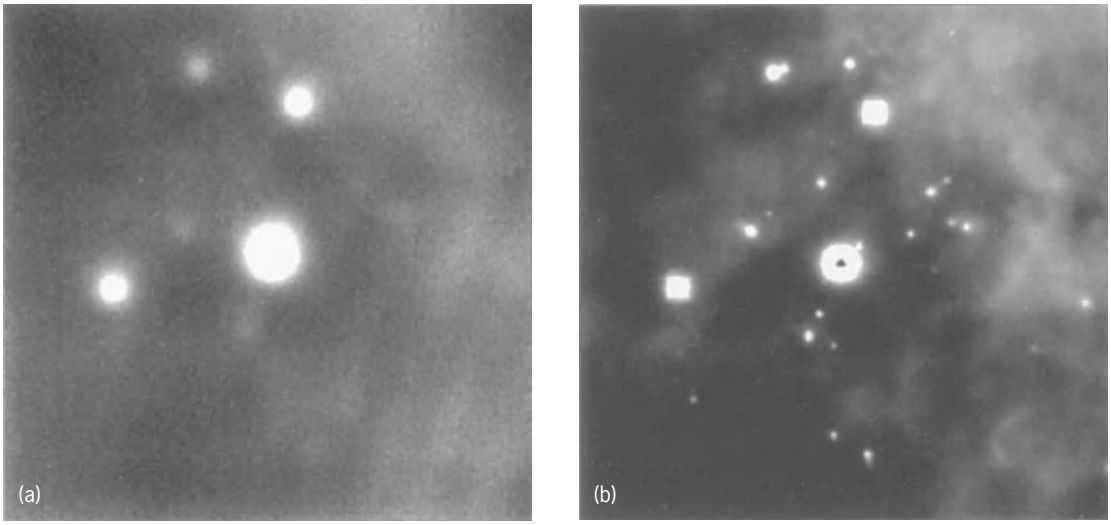


Fig. 3. Images of the Trapezium region in the Orion Nebula, obtained by using adaptive optics with a laser beacon reference source and a 241-actuator deformable mirror on a 1.5-m (59-in.) telescope. Fields of view are 40 arc-seconds square. (*a*) Uncorrected image. (*b*) Real-time image correction with laser beacon adaptive optics. (*Robert Q. Fugate, USAF Phillips Laboratory, Starfire Optical Range*)

directed to the wavefront sensor, which is range-gated to select light scattered from the required altitude. The static focus error of the received wavefront (due to the finite altitude of the laser beacon) is removed, and the residual wavefront errors are measured. The drive signals required to null the wavefront errors are then computed and applied to the deformable mirror. The imaging sensor then records a compensated image of the object (**Fig. 3**).

A laser beacon projected through the telescope cannot provide absolute pointing information, so it is necessary to use a natural star within the field of view as a direction reference; stars as faint as 20th magnitude may be used.

A single laser beacon samples a partial (conical) volume of the telescope beam, resulting in an additional error known as focal anisoplanatism. The magnitude of this error depends on the altitude of the spot and is most serious with lower-altitude Rayleigh scattering.

Many telescopes have been fitted with adaptive optics, including the Mount Wilson 2.5-m (100-in.), the Palomar 5-m (200-in.), the Canada-France-Hawaii 3.6-m (144-in.), the European Southern Observatory 3.6-m (142-in.), and the Lick Observatory 3-m (120-in.), instruments. In addition, adaptive optics will be installed in almost all of the current generation of 6–10-m (240–400-in.) giant telescopes, including the European Southern Observatory Very Large Telescope, Keck 2, Single-Mirror MMT, Gemini, and Subaru instruments. *See* TELESCOPE.

Laser applications. Adaptive optics is used to improve the quality of beams generated in laser cavities, and also to compensate for thermal effects and turbulence in the propagation path. The aberrations can be cleaned up by using an adaptive optics system similar to that discussed above. For this application adequate optical power is usually available to make the wavefront measurement. *See* LASER.

Compensation of laser beams transmitted through a turbulent medium such as the Earth's atmosphere relies on the principle of optical reciprocity. A low-power reference source, at or adjacent to the target, transmits its light through the turbulent medium to the laser aperture, where the wavefront aberrations are measured. A correction for these aberrations is applied to a deformable mirror before the laser is fired. The outgoing beam is predistorted by the deformable mirror and retraces the path taken by the light from the reference source. When it arrives at the target, its predistortion will have been exactly canceled by the atmosphere and so it will be precisely focused.

The isoplanatic limitation also applies to such a system: it is necessary for the reference beam to be aligned with the intended laser path; otherwise the measured wavefront will not be correct. If the target is moving, it is necessary to correct for the finite velocity of light by making the wavefront measurement at a point ahead of the target.

Technology. Adaptive optical systems may be broadly classified into conventional systems employing discrete optical and electronic components, and unconventional systems using integrated devices in which the basic functions are performed by physical processes within a nonlinear medium.

Conventional systems employ three main components; a wavefront sensor, a data processor, and a wavefront compensator.

Wavefront sensors. Most practical wavefront sensors measure the gradient of the wavefront over an array of subapertures in the incoming beam. These local gradient measurements are then reconstructed into a map of the wavefront error over the whole aperture. The two main techniques for gradient sensing use the imaging Hartmann sensor and the shearing interferometer. The imaging Hartmann sensor divides the optical beam into an array of contiguous subapertures, each containing a lens that brings the light from the reference source to a separate focus.

Displacement of the focused spots reveals the direction and magnitude of the two-dimensional wavefront gradient.

The principle of shearing interferometry is to split the wavefront to be measured into two replicas which are mutually displaced by a small distance and then recombined. Lateral displacement or shear is most often used, in which case the intensity of the interference pattern is proportional to the wavefront gradient in the direction of shear. *See* INTERFERENCE OF WAVES; INTERFEROMETRY.

Curvature sensing is an alternative technique that measures the local wavefront curvature (a scalar quantity) rather than the wavefront slope (a vector quantity). The curvature measurements can be applied directly to a deformable mirror, employing bimorph actuators that produce local curvature of the faceplate.

Data processors. Reconstruction of the gradient measurements of wavefront sensors into an error map involves the solution of a set of simultaneous equations, or equivalently the inversion of a matrix. The computation is usually implemented by a digital processor using an array of multiplier-accumulator devices. For large apertures the computational load is considerable. *See* DIGITAL COMPUTER; MATRIX THEORY.

Wavefront compensators. In principle, wavefront compensators may either shift the optical phase of a light beam or change the optical path length. Phase-shift devices such as electrooptical crystals have a limited range of correction and are often spectrally limited. Deformable mirrors employ mechanical displacement of the reflecting surface to control the optical path length. Advantages of deformable mirrors include high reflection efficiency, wide spectral bandwidth, and large correction range. *See* ELECTROOPTICS.

Three main types of deformable mirror have been developed: the monolithic deformable mirror, the thin-plate mirror using discrete actuators, and the segmented mirror. The monolithic mirror employs a solid block of piezoelectric ceramic such as lead zirconium titanate, with an array of electrodes at the top surface, to which is bonded a thin sheet of glass forming the reflecting surface. The thin-plate mirror employs an array of discrete multilayer piezoelectric actuators mounted on a rigid baseplate. Segmented mirrors are composed of many separate panels, each of which is normally supported on three actuators that provide tip, tilt, and piston adjustments. The panels are relatively small and easier to fabricate than one large mirror. However, maintaining the alignment of the segments to maintain a properly phased surface is quite difficult, and a separate wavefront sensor is usually employed for this purpose. To reduce the number of components in the optical train, some compensated telescopes are designed with adaptive secondaries, which employ thin deformable faceplates controlled by several hundred electromagnetic actuators. *See* PIEZOELECTRICITY.

Nonlinear optics. Nonlinear optical devices function as conjugate mirrors in which the output beam is a replica of the input except that it travels in exactly the reverse direction. If a plane wave passes through an aberrating medium, is reflected by a conjugate mirror, and then retraces its path, it will emerge in its original state, once again as a plane wave. The basic mechanism of nonlinear phase conjugation is the formation within a cell of zones of different optical refractivity that form a partially reflecting three-dimensional grating structure. There are two main types of nonlinear devices: (1) those based on stimulated Brillouin scattering and the related stimulated Raman scattering, which are produced in suitable transparent media by the self-pumping effect of high-intensity, high-power laser beams; and (2) three-wave and four-wave mixing devices, in which auxiliary high-power pump beams set up the grating structure, allowing considerable gain to be obtained. *See* NONLINEAR OPTICS; OPTICAL PHASE CONJUGATION; RAMAN EFFECT.

John W. Hardy

Bibliography. D. G. Crowe (ed.), *Selected Papers on Adaptive Optics and Speckle Imaging*, SPIE, vol. MS93, 1994; M. A. Ealey and F. Merkle (eds.), Adaptive Optics in Astronomy, *Proc. SPIE*, vol. 2201, 1994; J. W. Hardy, *Adaptive Optics for Astronomical Telescopes*, Oxford University Press, New York, 1998; J. E. Pearson (ed.), *Selected Papers on Adaptive Optics for Atmospheric Compensation*, SPIE, vol. MS92, 1994.

Adaptive signal processing

Signal processing is a discipline that deals with the extraction of information from signals. The devices that perform this task can be physical hardware devices, specialized software codes, or combinations of both. In recent years the complexity of these devices and the scope of their applications have increased dramatically with the rapidly falling costs of hardware and software and the advancement of sensor technologies. This trend has made it possible to pursue sophisticated signal-processing designs at relatively low cost. Some notable applications, in areas ranging from biomedical engineering to wireless communications, include the suppression of interference arising from noisy measurement sensors, the elimination of distortions introduced when signals travel through transmission channels, and the recovery of signals embedded in a multitude of echoes created by multipath effects in mobile communications.

Statistically based systems. Regardless of the application, any functional system is expected to meet certain performance specifications. The requirements, as well as the design methodology, vary according to the nature of the end application. One distinctive design methodology, which dominated much of the earlier work in the information sciences, especially in the 1950s and 1960s, is based on statistical considerations. This framework assumes the availability of information in advance about the statistical nature of the signals involved, and then proceeds to design systems that optimize some statistical

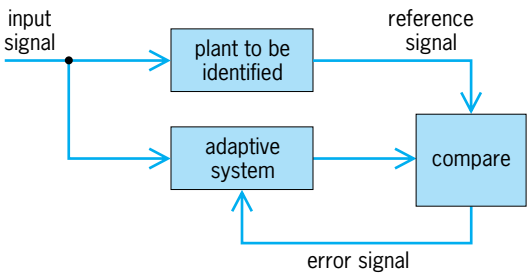
criterion. The resulting optimal designs are, in general, complex to implement. Only in special, yet important cases have they led to successful breakthroughs culminating with the Wiener and Kalman filters. *See* ESTIMATION THEORY; OPTIMIZATION.

Moreover, in many situations a design that is motivated by statistical considerations may not be immediately feasible, because complete knowledge of the necessary statistical information may not be available. It may even happen that the statistical conditions vary with time. It therefore may be expected, in these scenarios, that the performance of any statistically based optimal design will degrade, the more so as the real physical application deviates from the modeling assumptions.

Adaptive systems. Adaptive systems provide an attractive solution to the problem. They are devices that adjust themselves to an ever-changing environment; the structure of an adaptive system changes in such a way that its performance improves through a continuing interaction with its surroundings. Its superior performance in nonstationary environments results from its ability to track slow variations in the statistics of the signals and to continually seek optimal designs.

Adaptive signal processing deals with the design of adaptive systems for signal-processing applications. Related issues arise in control design, where the objective is to alter the behavior of a system, and lead to the study of adaptive control strategies; the main issue is the stability of the system under feedback. *See* ADAPTIVE CONTROL.

Operation. The operation of an adaptive system can be illustrated with a classical example in system identification. The **illustration** shows a plant (or system) whose input-output behavior is unknown and may even be time variant. The objective is to design an adaptive system that provides a good approximation to the input-output map of the plant. For this purpose, the plant is excited by a known input signal, and the response is taken as a reference signal. Moreover, a structure is chosen for the adaptive system, say a finite-impulse response structure of adequate length, and it is excited by the same input signal as the plant. At each time instant the output of the finite-impulse response system is compared with the reference signal, and the resulting error signal is used to change the coefficients of the finite-impulse response configuration. This learning process is continued over time, and the output of the adaptive system is expected to provide better tracking of the plant output as time progresses, especially when the structure of the plant is time invariant or varies only slowly with time.



Adaptive system identification.

Characteristics. Apart from emphasizing one particular application for adaptive systems, the above example also highlights several characteristics that are common to most adaptive designs:

1. An adaptive system adjusts itself automatically to a changing environment. This is achieved by changing the parameters of its internal structure in response to an error measure. Several adaptive structures have been used in practice, but the most frequent ones are linear combiners, finite-impulse response filters, infinite-impulse response filters, and linear combiners that are followed by nonlinearities such as sigmoid functions or nonlinear decision devices. *See* ELECTRIC FILTER.

2. The interaction of an adaptive system with its environment takes place through an input signal and a reference signal. The reference signal is used to evaluate the performance of the adaptation process through the computation of an error signal. The adaptive system responds to the error signal in such a way as to minimize a predetermined cost function that is computed either from the error signal directly or from a filtered version of it.

3. An adaptive system is inherently nonlinear and time variant. For this reason, it is in general considerably more difficult to analyze its performance than that of a linear time-invariant system. Nevertheless, adaptive systems offer more possibilities than conventional nonadaptive designs, and many analysis methods have been developed that offer reasonable approximate methods for performance evaluation.

4. An adaptive system can be trained to perform certain tasks. This usually involves a learning phase where the adaptive system is exposed to typical input and reference data and is left to adjust itself to them. At the end of the learning procedure, the system can be exposed to new data and will be expected to provide a reasonable response. Typical examples of this scenario abound in neural-network applications and in the equalization of communication channels. In the latter application, the outputs of communication channels (at the receiver) are equalized to compensate for the distortion of signals sent over long distances, and to thereby estimate the input (or transmitted) signals. *See* NEURAL NETWORK.

Adaptive algorithms. The performance of an adaptive system is critically dependent not only on its internal structure but also on the algorithm used to automatically adjust the parameters that define the structure. In general, the more complex the internal structure of the adaptive system, the more complex the algorithm that is needed for its adaptation. Moreover, several algorithms may exist for the same adaptive structure, and the choice of which algorithm to use is dictated by several factors that include the following:

1. The speed with which the algorithm learns in a stationary environment; that is, how fast the algorithm converges to the optimal statistical solution under stationarity assumptions. This factor determines the convergence rate of the algorithm.

2. The speed with which the algorithm tracks statistical variations. This factor determines the tracking capability of an algorithm and is a major motivation for the use of adaptive schemes, especially in recent mobile communications applications where adaptive equalizers are used to compensate for (that is, adapt to) changes in time-variant transmission channels, and thereby give good estimates of the transmitted signal from the received signal despite these changes.

3. The manner in which the performance of an adaptive scheme, operating in steady-state conditions, deviates from the performance of a statistically optimal design. This factor measures the so-called misadjustment of an adaptive scheme and serves to compare its performance with that of an optimal design in a statistical framework.

4. The amount of computational effort required to carry out each adjustment of the parameters of the adaptive system structure. Applications that require a large number of adaptive parameters tend to dictate a preference for computationally fast algorithms at the expense of other performance factors.

5. The reliability of an algorithm in finite-precision implementations. This factor is concerned with the numerical behavior of an algorithm when implemented in finite-precision arithmetic, and whether numerical effects might lead to erroneous behavior.

6. The robustness of an adaptive system to disturbances and unmodeled dynamics. This factor determines whether small disturbances can result in large errors and therefore compromise the performance of an adaptive scheme.

The above factors are often competing requirements, so that it is usually necessary to seek a compromise solution that best suits a particular application.

Derivation procedures. There have been many procedures for the derivation of adaptive algorithms, but the most frequent, at least in having had the most applications, are procedures that are based either on the method of stochastic approximation or on the least-squares criterion. In both cases, and especially for finite-impulse response adaptive structures, each criterion has led to several different variants that address in one way or another the above performance factors.

In stochastic-approximation algorithms, a recursive procedure is devised for the minimization of a predetermined cost function by iteratively approximating its minimum through a gradient descent procedure (a standard procedure in optimization theory). While this class of algorithms generally suffers from slow convergence rates, it still enjoys widespread use due to its simplicity, low computational requirements, and often observed robustness under different operating conditions. The most prominent member of the stochastic-approximation algorithms is the least-mean-squares algorithm, which is undoubtedly the most widely used adaptive filtering algorithm. Other members include filtered-error variants that are useful in active noise control and infinite-impulse response system identification problems, as well as several frequency-domain adaptive schemes. *See* ACTIVE SOUND CONTROL; STOCHASTIC CONTROL THEORY.

In least-squares algorithms, a recursive procedure is devised for the minimization of a quadratic cost function. This family of algorithms is based on the least-squares criterion, which was developed in the late eighteenth century in work on celestial mechanics. Since then, the least-squares criterion has enjoyed widespread popularity in many diverse areas as a result of its attractive computational and statistical properties. Notably, for linear data models, least-squares solutions can be explicitly evaluated in closed form, can be recursively updated as more input data are made available, and are optimal maximum likelihood estimators in the presence of gaussian measurement noise. *See* LEAST-SQUARES METHOD.

Many recursive least-squares algorithms have been developed. Several of them are computationally more demanding than least-mean-squares-type algorithms, but variants exist that are computationally competitive, although more complex. They have better convergence properties but are less robust to disturbances.

Array algorithms. Following a trend initiated in the late 1960s in Kalman filtering methods, least-squares adaptive schemes are currently more often implemented in convenient array algorithms. These algorithms are closely related to the QR method, a numerically stable algorithm for solving systems of linear equations, and have the properties of better conditioning (that is, lower sensitivity to errors in the initial data), reduced dynamical range (which favors better conditioning), and orthogonal transformations, which typically lead to better numerical performance in finite-precision arithmetic. In the array form, an algorithm is described as a sequence of elementary operations on arrays of numbers. Usually, a prearray of numbers has to be triangularized by a rotation, or a sequence of elementary rotations, in order to yield a postarray of numbers. The quantities needed to form the next prearray can then be read off from the entries of the postarray, and the procedure can be repeated. The explicit forms of the rotation matrices are not needed in most cases. Such array descriptions are more truly algorithms in the sense that they operate on sets of numbers and provide other sets of numbers, with no explicit equations involved. *See* ALGORITHM.

Ali H. Sayed; Babak Hassibi; Thomas Kailath

Bibliography. S. Haykin, *Adaptive Filter Theory*, 3d ed. 1996; J. G. Proakis et al., *Advanced Digital Signal Processing*, 1992; B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, 1985.

Adaptive wings

A wing is the primary lift-generating surface of an aircraft. Adaptive wings—also known as smart, compliant, intelligent, morphing, controllable, and reactive wings—are lifting surfaces that can change their shape in flight to achieve optimal performance at different speeds, altitudes, and ambient conditions. There are different levels of sophistication, or intelligence, that can be imbued in a particular design.

Whether the wing is rigidly attached to the fuselage for fixed-wing airplanes or rotating for helicopters, a primary design objective of such a lifting surface is to maximize the lift-to-drag ratio, which is achieved by controlling the airflow around the wing. Other design objectives for the wing include improving maneuverability and minimizing vibrations and flow-induced noise. The wing can have a set design optimized for specific flight conditions, or it can change shape to conform to a variety of conditions. Chosen judiciously, minute dynamic changes in the wing's shape can, under the right circumstances, greatly affect the airflow and thus the aircraft's performance. *See* AERODYNAMIC FORCE; AIRFOIL.

Flying insects and birds, through millions of years of evolution, can change the shape of their flapping wings, subtly or dramatically, to adapt to various flight conditions. The resulting performance and agility are unmatched by any airplane. For example, the dragonfly can fly forward and backward, turn abruptly and perform other supermaneuvers, hover, feed, and even mate while aloft (**Fig. 1**). Undoubtedly, its prodigious wings contributed to the species' survival for over 250 million years. *See* FLIGHT.

Among human-produced flyers, the Wright brothers changed the camber of the outboard tip of their aircraft's wings to generate lateral or roll control (combined with simultaneous rudder action for banked turn), thus achieving in 1903 the first heavier-than-air, controlled flight. The R. B. Racer built by the Dayton Wright Airplane Company in 1920 allowed the pilot to change the wing camber in flight using a hand crank. The wings of today's commercial aircraft contain trailing-edge flaps and leading-edge slats to enhance the lift during the relatively low speeds of takeoff and landing, and ailerons for roll control, all engaged by the pilot via clumsy, heavy, and slow servomechanisms. To equalize the lift and prevent rolling in forward flight, the rotary wings of most helicopters are cyclically feathered to increase the pitch on the advancing blade and decrease it on the retreating blade. *See* AILERON; ELEVATOR (AIRCRAFT); FLIGHT CONTROLS; HELICOPTER.

Fig. 1. Male and female Cardinal Meadowhawk dragonflies following airborne mating. The male has towed the just-inseminated female to a pond and is dipping her tail in the water so she can deposit her eggs. (*Reprinted with permission, The Press Democrat, Santa Rosa, CA*)

While bird wings are quite smart, human-designed ones are not very intelligent. With few exceptions, the level of autonomous adaptability sought in research laboratories is some years away from routine field employment. Intelligent control of the wing shape involves embedded sensors and actuators with integrated control logic; in other words, the wing's skin is made out of smart materials. The sensors detect the state of the controlled variable, for example the wall shear stress, and the actuators respond to the sensors' output based on a control law to effect the desired in-flight metamorphosing of the wing. For certain control goals, for example skin-friction drag reduction, required changes in the wing shape can be microscopic. For others, for example morphing the wing for takeoff and landing, dramatic increases in camber may be needed.

Smart materials. Adaptive wing design involves adding smart materials to the wing structure and using these materials to effect flow changes. Smart materials are those that undergo transformations through physical interactions. Such materials sense changes in their environment and adapt according to a feedforward or feedback control law. Smart materials include piezoelectrics, electrostrictors, magnetostrictors, shape-memory alloys, electrorheological and magnetorheological fluids, and optical fibers. For no rational reason, several other types of sensors and actuators that fall outside those categories are not usually classified as constituting elements of smart structures.

The piezoelectric effect is displayed by many noncentrosymmetric ceramics, polymers, and biological systems. The direct effect denotes the generation of electrical polarization in the material in response to mechanical stress; the poled material is then acting as a stress or strain sensor. The converse effect denotes the generation of mechanical deformation upon the application of an electrical charge; in this case the poled material is acting as an actuator. The most widely used piezoceramic and piezopolymer are, respectively, lead zirconate titanate (PZT) and polyvinylidene fluoride (PVDF). Piezoelectrics are the most commonly used type of smart materials and the only ones that can be used readily as both sensors and actuators. *See* PIEZOELECTRICITY.

Electrostrictive materials are dielectrics that act similarly to piezoelectric actuators, but the relation between the electric charge and the resulting deformation in this case is nonlinear. Examples of such materials are lead magnesium niobate (PMN) compounds, which are relaxor ferroelectrics. Magnetostrictive materials, such as Terfenol-D, are actuators that respond to a magnetic field instead of an electric field. *See* ELECTROSTRICTION; MAGNETOSTRICTION.

Shape memory alloys, such as a nickel–titanium alloy known as Nitinol, are metal actuators that can sustain large deformation and then return to their original shape by heating without undergoing plastic deformation. Electrorheological and magnetorheological fluids rapidly increase in viscosity—by several orders of magnitude—when placed in, respectively, electric or magnetic fields. Both kinds of fluids can provide significant actuation power and are therefore considered for heavy-duty tasks such as shock absorbing for large-scale structures. Finally, optical fibers are sensors that exploit the refractive properties of light to sense acoustical, thermal, and mechanical-strain perturbations. *See* FIBER-OPTIC SENSOR; SHAPE MEMORY ALLOYS.

Outstanding issues to be resolved before smart materials for aircraft wings become routine include cost; complexity; computer memory, speed, and software; weight penalty; maintenance; reliability; robustness; and the integrity of the structure on which the sensors and actuators are mounted. Sensors and actuators that have length scales between 1 and 1000 micrometers constitute a special domain of smart materials that in turn is a cornerstone of micro-electro-mechanical systems (MEMS). *See* MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS).

Flow control. An external wall-bounded flow, such as that developing on the exterior surfaces of a wing, can be manipulated to achieve transition delay, separation postponement, lift increase, skin-friction and pressure drag reduction, turbulence augmentation, mixing enhancement, and noise suppression. These objectives are interrelated (**Fig. 2**). If the boundary layer around the wing becomes turbulent, its resistance to separation is enhanced and more lift can be obtained at increased incidence. On the other hand, the skin-friction drag for a laminar boundary layer can be as much as an order of magnitude less than that for a turbulent one. If transition is delayed, lower skin friction and lower flow-induced noise are achieved. However, a laminar boundary layer can support only very small adverse pressure gradients without separation and, at the slightest increase in angle of attack or some other provocation, the boundary layer detaches from the wing's surface and subsequent loss of lift and increase in form drag occur. Once the laminar boundary layer separates, a free-shear layer forms, and for moderate Reynolds numbers transition to turbulence takes place. Increased entrainment of high-speed fluid due to the turbulent mixing may result in reattachment of the separated region and formation of a laminar separation bubble. At higher incidence, the bubble breaks down, either separating completely or forming a longer bubble.

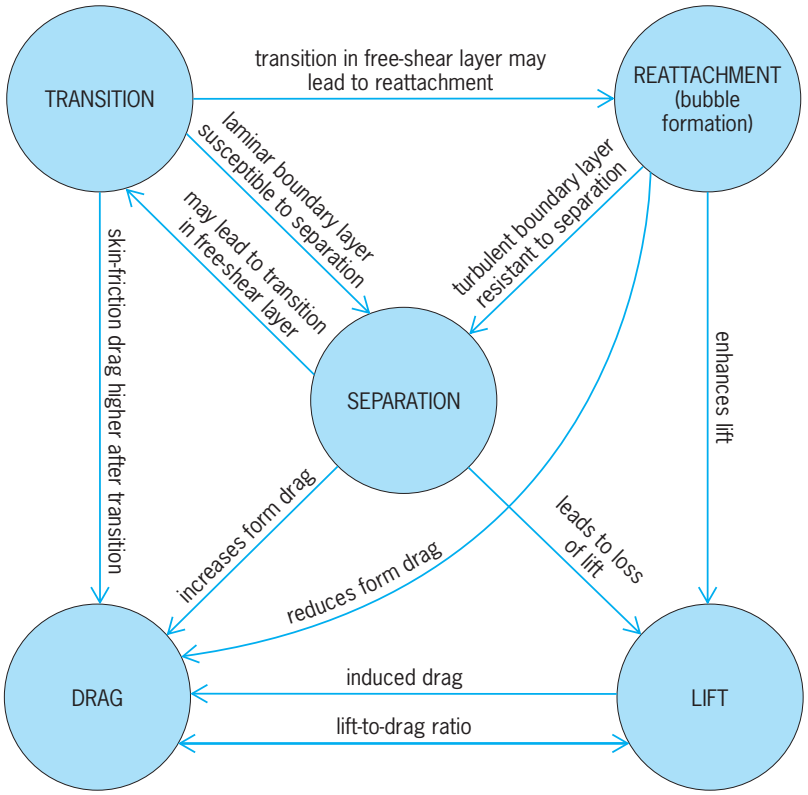


Fig. 2. Partial representation of the interrelation among flow control goals. (*After M. Gad-el-Hak, Flow Control: Passive, Active and Reactive Flow Management, Cambridge University Press, London, 2000*)

In either case, the form drag increases and the lift curve's slope decreases. The ultimate goal of all this is to improve the airfoil's performance by increasing the lift-to-drag ratio. However, induced drag is caused by the lift generated on a wing with a finite span. Moreover, more lift is generated at higher incidence, but form drag also increases at these angles. *See* BOUNDARY-LAYER FLOW; FLUID MECHANICS; LAMINAR FLOW; REYNOLDS NUMBER; TURBULENT FLOW.

Flow control is most effective when applied near the transition or separation points, where conditions are near those of the critical flow regimes where flow instabilities magnify quickly. Therefore, delaying or advancing the laminar-to-turbulence transition and preventing or provoking separation are easier tasks to accomplish. Reducing the skin-friction drag in a nonseparating turbulent boundary layer, where the mean flow is quite stable, is a more challenging problem. Yet, even a modest reduction in the fluid resistance to the motion of, for example, the worldwide commercial airplane fleet is translated into annual fuel savings estimated to be in billions of dollars. Newer ideas for turbulent flow control focus on targeting coherent structures, which are quasiperiodic, organized, large-scale vortex motions embedded in a random, or incoherent, flow field (**Fig. 3**).

Future systems. Future systems for control of turbulent flows in general and turbulent boundary layers in particular could greatly benefit from the merging of the science of chaos control, the technology of microfabrication, and the newest

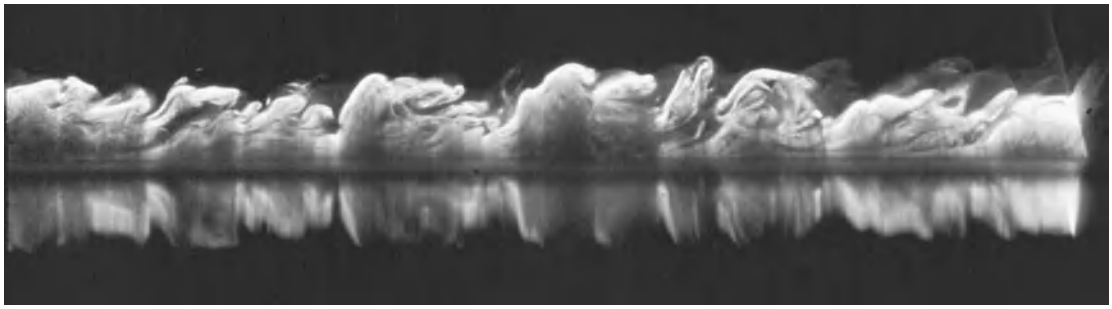


Fig. 3. Large-eddy structures (one type of coherent structure) in a turbulent boundary layer. The side view is visualized using a sheet of laser light and a fluorescent dye. (*From M. Gad-el-Hak, R. F. Blackwelder, and J. J. Riley, On the interaction of compliant coatings with boundary layer flows, J. Fluid Mech., 140:257–280, 1984*)

computational tools collectively termed soft computing. Control of chaotic, nonlinear dynamical systems has been demonstrated theoretically as well as experimentally, even for multi-degree-of-freedom systems. Microfabrication is an emerging technology which has the potential for mass-producing inexpensive, programmable sensor–actuator chips that have dimensions of the order of a few micrometers. Soft computing tools include neural networks, fuzzy logic, and genetic algorithms. They have advanced and become more widely used in the last few years, and could be very useful in constructing effective adaptive controllers. Such futuristic systems are envisaged as consisting of a colossal number of intelligent, interactive, microfabricated wall sensors and actuators arranged in a checkerboard pattern and targeted toward specific organized structures that occur quasirandomly (or quasiperiodically) within a turbulent flow. Sensors would detect oncoming coherent structures, and adaptive controllers would process the sensors' information and provide control signals to the actuators that in turn would attempt to favorably modulate the quasiperiodic events. A finite number of wall sensors perceives only partial information about the flow field. However, a low-dimensional dynamical model of the near-wall region used in a Kálmán filter can make the most of this partial information. Conceptually all of that is not too difficult, but in practice the complexity of such control systems is daunting and much research and development work remain. *See* CHAOS; COMPUTATIONAL INTELLIGENCE; ESTIMATION THEORY; FUZZY SETS AND SYSTEMS; GENETIC ALGORITHMS; MICROSENSOR; NEURAL NETWORK.

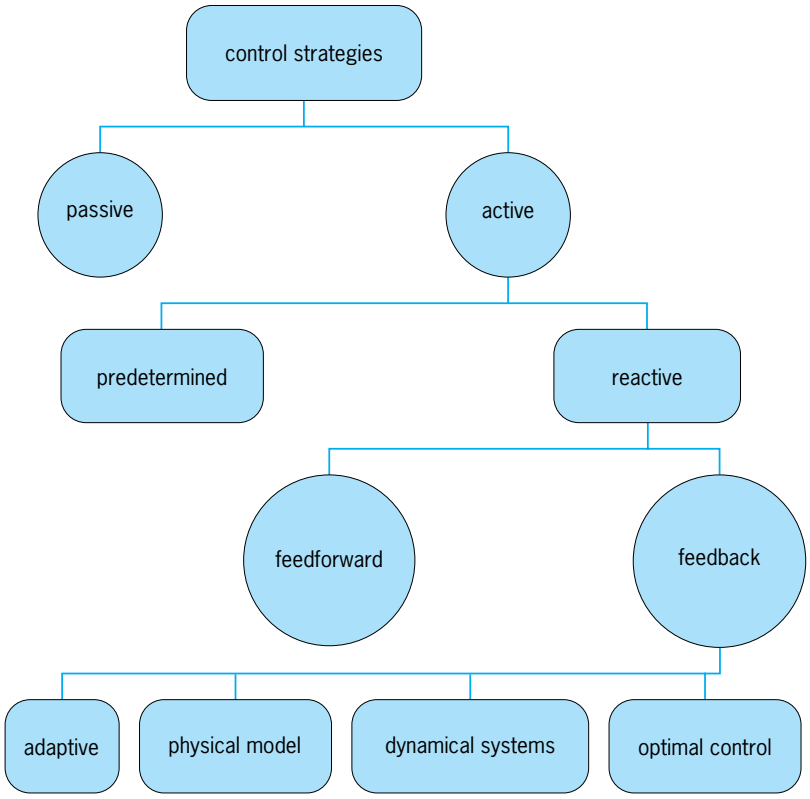


Fig. 4. Classification of control strategies. (*After M. Gad-el-Hak, Flow Control: Passive, Active and Reactive Flow Management, Cambridge University Press, London, 2000*)

Control strategies. Different levels of intelligence can be imbued in a particular control system (**Fig. 4**). The control can be passive, requiring no auxiliary power and no control loop, or active, requiring energy expenditure. Manufacturing a wing with a fixed streamlined shape is an example of passive control. Active control requires a control loop and is further divided into predetermined or reactive. Predetermined control includes the application of steady or unsteady energy input without regard to the particular state of the system—for example, a pilot engaging the wing's flaps for takeoff. The control loop in this case is open, and no sensors are required. Because no sensed information is being fed forward, this open control loop is not a feedforward one. This subtle point is often confused, blurring predetermined control with reactive, feedforward control.

Reactive, or smart, control is a special class of active control where the control input is continuously adjusted based on measurements of some kind. The control loop in this case can be either an open, feedforward one or a closed, feedback loop. Achieving that level of autonomous control (that is, without human interference) is the ultimate goal of smart-wing designers. In feedforward control, the measured variable and the controlled variable are not necessarily the same. For example, the pressure can be sensed at an upstream location, and the resulting signal is used together with an appropriate control law to actuate a shape change that in turn influences the shear stress (that is, skin friction) at

a downstream position. Feedback control, on the other hand, necessitates that the controlled variable be measured, fed back, and compared with a reference input. Reactive feedback control is further classified into four categories: adaptive, physical model–based, dynamical systems–based, and optimal control. An example of reactive control is the use of distributed sensors and actuators on the wing surface to detect certain coherent flow structures and, based on a sophisticated control law, subtly morph the wing to suppress those structures in order to dramatically reduce the skin-friction drag.

Prospects. It is anticipated that the first field applications of adaptive wings will probably take place in micro-air-vehicles (MAV) and unmanned aerial vehicles (UAV). The next generations of cruise missiles and supermaneuverable fighter aircraft will probably have adaptive wings. In the 2010s or 2020s, commercial aircraft may have smart wings, fuselage, and vertical and horizontal stabilizers. *See* DRONE; WING.

Mohamed Gad-el-Hak

Bibliography. H. T. Banks, R. C. Smith, and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Wiley, New York, 1996; M. Gad-el-Hak, *Flow Control: Passive, Active and Reactive Flow Management*, Cambridge University Press, London, 2000; M. Gad-el-Hak (ed.), *The MEMS Handbook*, vols. 1–3, 2d ed., Taylor & Francis/CRC Press, Boca Raton, FL, 2006; M. Gad-el-Hak, R. F. Blackwelder, and J. J. Riley, On the interaction of compliant coatings with boundary layer flows, *J. Fluid Mech.*, 140:257–280, 1984; M. Schwartz (ed.), *Encyclopedia of Smart Materials*, vols. 1 and 2, Wiley-Interscience, New York, 2002.

Addictive disorders

Addictive disease disorders are characterized by the chronic use of a drug (such as heroin, cocaine, or amphetamines), alcohol, or similar substance. These disorders usually result in (1) the development of tolerance for the substance, with the need for increasing amounts to achieve the desired effect; (2) physical dependence, characterized by a sequence of well-defined signs and physiological symptoms, such as the withdrawal or abstinence syndrome on cessation of use of the substance; and (3) compulsive drug-seeking behavior, with chronic, regular, or intermittent use, despite possible harm to self or others. Since the early 1960s, research has been increasing in the biology of addictive diseases, and emphasis has shifted from only psychological, sociological, and epidemiological studies to investigations of the metabolic, neurobiological, and molecular bases of addiction.

The four major addictive diseases are alcoholism, narcotic (or opiate) addiction, cocaine and other stimulant addiction, and nicotine addiction. Drug addiction may also occur after chronic use of other types of agents such as barbiturates, benzodiazepines, and marijuana. Abusive use of other, licit and illicit drugs and other chemical agents is common. However, many types of drug abuse, such as hallucinogen abuse, do not necessarily result in drug addiction but only in drug-using behavior. *See* ALCOHOLISM; BARBITURATES; COCAINE; NARCOTIC; OPIATES.

Opiate receptors and endogenous opioids. The conclusive identification of opiate receptors (cell structures that function as an intermediary between the opioid and the physiological response) in mammals (including humans) in 1973 has been extremely important for many aspects of physiology and pathology. Subsequently, it has been determined by use of selective chemical ligands that there are at least three separate types of opioid receptors—mu receptors, delta receptors, and kappa receptors—all of which meet the accepted criteria of being specific opiate receptors. Opioid (opiatelike) drugs of different types bind to these receptors. This binding may be displaced by a specific opioid (or narcotic) antagonist such as naloxone, and this binding leads to some demonstrable effect. The genes encoding each of these receptor types were cloned for the first time in 1992, beginning with the delta opioid receptor.

Subsequent to the discovery of specific opioid receptors, endogenous ligands which bind to these receptors, the so-called endogenous (originating from within) opioids were discovered. Opioids include substances that are produced endogenously (such as the enkephalins, endorphins, and dynorphins) and may be produced synthetically. Exogenous synthetic opioids are used extensively in the treatment of pain.

The first of these endogenous opioids to be found were met-enkephalin and leu-enkephalin, which are now known to be derived from a single gene product, preproenkephalin. This propeptide is further processed to yield at least five active endogenous opioids. It is produced in the adrenal medulla as well as in the brain and in possibly many other peripheral sites.

The second class of endogenous opioids to be delineated was beta endorphin. This has been shown to be biotransformed from beta lipotropin in equimolar amounts with the release of adrenocorticotropic hormone (ACTH; stimulates the adrenal cortex). Each of these peptides is derived from a single precursor peptide, proopiomelanocortin (POMC). These peptides are found also predominantly in the anterior pituitary. However, the POMC gene is also expressed in the hypothalamus and in other regions of the brain, as well as at peripheral sites.

The third class of endogenous opioids to be delineated were the dynorphins. They consist of multiple active compounds coming from a precursor peptide, preprodynorphin, found in the brain, the spinal cord, and other sites. *See* ENDORPHINS.

Role of endogenous opioids. It is not known to what extent the endogenous opioids play a role in addictive diseases. This is especially true for narcotic (primarily heroin) addiction and alcoholism, and more recently for cocaine addiction. It has been suggested that narcotic addiction may be a disorder characterized by (1) a relative or absolute deficiency of

endogenous opioids; (2) an end organ or receptor failure to respond to normal or possibly increased levels of endogenous opioids; (3) genetic or acquired (for example, short-acting opiate or stimulant drug–induced) abnormalities in the feedback control of the synthesis, release, and processing, or degradation of one or more types of the endogenous opioids; or (4) genetic variations in the opioid receptors. It has also been documented that acute or chronic use of short-acting narcotics, such as heroin or morphine, results in a suppression of levels of beta endorphin as well as levels of ACTH. These two peptides are released together from their precursor peptide. Cortisol (a steroid hormone produced by the adrenal cortex) levels are also lowered and their diurnal rhythm is blunted during use of short-acting narcotics. Conversely, during chronic long-term administration of the long-acting narcotic methadone, in the maintenance treatment of addiction in humans, normalization of these abnormalities occurs, with restoration of normal levels of beta endorphin, ACTH, and cortisol. Also, the normal circadian rhythm of these hormones is restored. Such normalization occurs at a time when there is no longer any drug-seeking behavior, although a high degree of tolerance to mu opioid receptor agonists and of physical dependence exists. When the beta endorphin–ACTH–cortisol levels become normalized, other physiological alterations caused by chronic short-acting narcotic use, such as suppression of luteinizing hormone (secreted by the anterior pituitary, and participating in the menstrual cycle) and peripheral reproductive steroid hormones, also become normalized. Menstrual function and fertility return to normal in women under chronic methadone treatment. Studies have shown that immune function disrupted during heroin addiction may normalize during methadone maintenance treatment. *See* ENDOCRINE MECHANISMS.

It has been found that full tolerance does not develop to one well-documented physiological effect of the acute and subacute administration of any narcotic—that is, the effect of exogenous opioid on release of prolactin (a hormone of the anterior pituitary that stimulates lactation in females). Even during chronic long-term steady-state methadone maintenance treatment in which sustained plasma levels of methadone are achieved, full tolerance does not develop to the prolactin-releasing effect of peak levels of opioid. The peak plasma levels of prolactin occur around the time of peak plasma levels of methadone, that is, around 2–4 h after an oral dose of methadone, rather than in the early morning hours, the normal time of peak levels of prolactin. However, the actual elevations in levels of prolactin caused by the opioids do not, in most cases, exceed the upper limits of prolactin normally observed during spikes in prolactin levels. Since prolactin release is normally under the tonic inhibitory control by dopamine or dopaminergic mechanisms, these findings suggest that even during chronic steady-state treatment methadone suppresses dopaminergic activity, and that tolerance does not develop to this effect during chronic use. It is also of interest that several psychotropic (affecting the mind) drugs, especially the neuroleptic agents (substances that act on the nervous stystem such as antipsychotics), have similar effects on prolactin release, and tolerance apparently does not develop to this effect of neuroleptics. The mechanism underlying the prolactin release by neuroleptics may be similar to the mechanism of action of the exogenous opioids. Both exogenous and endogenous kappa opioid receptor–preferring ligands (such as dynorphins) may modulate prolactin release, as well as exogenous and endogenous mu opioid receptor–directed ligands (such as beta endorphin).

The possible role of endogenous opioids in alcoholism is not clear. It has been shown that a specific opioid antagonist (inhibits the effect of the drug), naloxone, may reverse or ameliorate some of the signs and symptoms and physical abnormalities of the acute alcohol intoxication syndrome. However, this apparent beneficial effect of an opioid antagonist may not be related to the addictive disease per se, but may be due to the counteracting of the possible acute release of large amounts of endogenous opioids in response to excessive acute alcohol ingestion. Theories with some experimental data have further linked alcoholism and narcotic addiction. It has been shown that acetaldehyde, the first major metabolite of ethyl alcohol, can condense in the body with endogenous amines, particularly some specific neurotransmitters, to form opiatelike compounds. Laboratory studies have also shown that a very large amount of ethanol will displace some endogenous opioids from binding to the opioid receptor. Several recent studies of these specific opioid antagonists, naltrexone and nalmefene, may be beneficial in treating chronic alcoholism. Laboratory-based clinical studies have suggested that part of these beneficial effects of opioid antagonists in the management of alcoholism may be mediated through the effects of the endogenous opioid system in modulating the stress-responsive hypothalamic-pituitary-adrenal axis. Whether there are some common mechanisms with a genetic, metabolic, or purely behavioral basis underlying these two addictive diseases has not yet been defined.

Treatment. There are three approaches to the management of opiate (primarily heroin) addiction: pharmacological treatment, drug-free treatment following detoxification, and incarceration. Statistical data over the last century reveal that after release from prison, completion of drug-free treatment, or discontinuation of chronic pharmacological treatment with methadone or other mu agonists, partial agonists, or antagonists, fewer than 30% of former long-term heroin addicts are able to stay drug-free. (Long-term addiction is defined as more than 1 year of daily use of several doses of heroin, with tolerance, physical dependence, and drug-seeking behavior.)

Methadone. Since the 1960s, methadone maintenance treatment has been documented to be medically safe and effective, and the most effective available treatment for heroin addiction. Unlike heroin, which has a 3-min half-life in humans, and its major

metabolite, morphine, which has a half-life of 4–6 h, methadone has a 24-h half-life. Therefore, when given orally on a daily basis, methadone prevents the signs and symptoms of narcotic abstinence and also prevents drug hunger, without causing any narcotic-induced euphoria or high during the 24 h between doses. Steady moderate to high doses of methadone can be used in the treatment of addiction over long periods of time without tolerance developing to these desired effects. Because of the high degree of cross-tolerance developed for other narcotics, a patient receiving methadone maintenance treatment does not experience any narcotic high or other effects after self-administration of a dose of short-acting narcotic such as heroin. Through this cross-tolerance mechanism, methadone blocks any illicit narcotic effect.

Over 60–90% of heroin addicts who have entered into methadone maintenance treatment will stay in treatment voluntarily for 1 year or more. Illicit opiate abuse drops to less than 15% during such treatment when adequate doses of methadone (usually 60–150 mg per day) are administered. Cocaine or polydrug abuse or alcohol abuse may persist in 20–30% of patients even in excellent methadone programs, since methadone maintenance is specific for treatment of narcotic dependency.

Chronic methadone treatment also brings about the normalization of the multiple physiological alterations caused by chronic heroin abuse. Methadone-maintained patients, who are not chronic abusers of other drugs or alcohol, are able to work, to attend school, and to take part in normal socialization, including restoration of family life. Long-term follow-up studies have shown that methadone is medically safe. All these effects are due to the long-acting pharmacokinetic properties of methadone, which yields no euphoria or high, yet does not allow any withdrawal or sick period during a 24-h dosing interval. However, when maintenance treatment is discontinued following slow or rapid dose reduction and elimination, using methadone alone or combined with other agents such as clonidine, over 70% of all patients will return to opiate abuse within 2 years. Few studies focusing on physiological and medical effects have been conducted, but the longer-acting congener (related substance) of methadone, l-alpha-acetylmethadol (LAAM), has been found to be similarly effective and medically safe in the treatment of heroin addiction.

Naltrexone. A second type of pharmacological treatment of narcotic addiction is chronic treatment with a specific narcotic antagonist, naltrexone. This drug also prevents any narcotic effect from any illicitly administered heroin, but does so by a different mechanism than methadone: by way of direct opioid receptor blocking, with displacement of endogenous or exogenous opioids from receptor binding. Naltrexone will precipitate narcotic withdrawal symptoms in narcotic-dependent individuals, and thus may be given only after detoxification from heroin has been completed. Naltrexone will not prevent drug hunger. Naltrexone treatment has limited acceptance by unselected opiate addicts; only 15–30% of former narcotic addicts entering treatment with this agent remain in treatment for 6 months or more. Therefore, although it may be a worthwhile treatment for some small defined special populations, it does not seem to be an effective treatment for the majority of unselected long-term heroin addicts. Naltrexone does not allow the normalization of all physiological functions disrupted by chronic heroin use. However, it may be useful in the treatment of early, sporadic or intermittent narcotic abusers.

Drug-free treatment. Drug-free treatment involving either long-term institutionalization in a drug-free residence or, less frequently, attendance at an outpatient resource, has resulted in long-term success in approximately 10–30% of all unselected heroin addicts who enter treatment. Many addicts remain in treatment only a very short time, and unfortunately sometimes are not included in calculations of treatment outcome. Most drug-free treatments include a 1–3-year residential period followed usually by a 1–3-year nonresidential follow-up period. Only a very small percentage of those who initially enter treatment (though a larger percentage complete this 2–5 years of treatment) will remain drug-free 1 year or more after discharge from treatment.

Alcoholism. Alcoholism has been difficult to treat over a long-term period. There is no specific pharmacological replacement treatment for alcoholism. Disulfiram (Antabuse) is an agent which blocks the metabolism of acetaldehyde, the major metabolite of ethanol. When a person treated with Antabuse drinks alcohol, there is a rapid buildup of acetaldehyde and a severe physiological syndrome ensues, which frequently prevents or modifies further immediate drinking behavior. Most former alcoholics stop the use of disulfiram prior to returning to alcohol use.

The most widely used treatment of alcoholism is self-help groups, such as Alcoholics Anonymous (AA), where mutual support is available on a 24-h basis, along with group recognition of chronic problems with alcohol. However, only around 20–40% of severe chronic alcoholics are able to stay alcohol-free for more than 2 years even under such management. Early studies of the use of an opioid antagonist naltrexone or nalmefene for the treatment of chronic alcoholism have shown that about 50% of the subjects had reduced numbers and magnitudes of relapse events. Ethanol, unlike heroin and other narcotics, has many well-defined severe specific toxic effects on several organs, including the liver, brain, and reproductive system, so even relatively short drug-free intervals will decrease exposure to this toxin. Therefore, it has been suggested that success in treatment of alcoholism be measured in part by increasing lengths of alcohol-free intervals, and not just by permanent restoration of the alcohol-free state.

Cocaine addiction. Cocaine addiction may also involve disruption of the endogenous opioid system in addition to the well-known primary effect of cocaine in blocking reuptake of dopamine by the synaptic

dopamine transporter protein. This effect results in the accumulation of dopamine in the synapse and similar actions at the serotonin and norepinephrine reuptake transporter. Demonstrated changes in the mu and kappa endogenous opioid system increase the complexity of the effect of cocaine and may contribute to its resultant reinforcing properties leading to addiction, as well as to the drug craving and relapse. This may explain in part the resultant difficulties in developing a pharmacotherapeutic approach for the treatment of cocaine addiction. Mary Jeanne Kreek

Bibliography. C. Bond et al., Single nucleotide polymorphism in the human mu opioid receptor gene alters beta-endorphin binding and activity: Possible implications for opiate addiction, *Proc. Nat. Acad. Sci.*, 95:9608–9613, 1998; M. J. Kreek, Clinical update of opioid agonist and partial agonist medications for the maintenance treatment of opioid addiction, *Sem. Neurosci.*, 9:140–157, 1997; M. J. Kreek, Opiate and cocaine addictions: Challenge for pharmacotherapies, *Pharm. Biochem. Behav.*, 57:551–569, 1997; M. J. Kreek, Opiates, opioids and addiction, *Mol. Psych.*, 1:232–254, 1996; M. J. Kreek, Opioid receptors: Some perspectives from early studies of their role in normal phsyiology, stress responsivity and in specific addictive diseases, *J. Neurochem. Res.*, 21:1469–1488, 1996; J. H. Lowinson et al. (eds.), *Substance Abuse: A Comprehensive Textbook*, 3d ed., Williams & Wilkins, 1997; C. P. O'Brien and J. H. Jaffe (eds.), *Addictive States*, Raven Press, 1992; R. A. Rettig and A. Yarmolinsky (eds.), *Federal Regulation of Methadone Treatment*, National Academy Press, 1994.

Addition

One of the four fundamental operations of arithmetic and algebra. The symbol + of addition is thought to be a ligature for "et," used in a German manuscript of 1456 to denote addition. Its first printed appearance is in Johann Widman's *Behennede und hüpsche Rechnung*, Leipzig, 1489. As a symbol of operation, the plus sign appeared in algebra before arithmetic, and now the term addition, together with its symbol, is applied to many kinds of objects other than numbers. For example, two vectors $\mathbf{x}$, $\mathbf{y}$ are added to produce a third vector $\mathbf{z}$ obtained from them by the "parallelogram" law, and two sets A, B are added to form a third set C consisting of all the elements of A and of B. *See* CALCULUS OF VECTORS.

As an operation on pairs of real or complex numbers, addition is associative, Eq. (1), and commutative, Eq. (2); and multiplication is distributive over addition, Eq. (3). There are important mathematical structures in which an addition is defined that lacks one or more of these properties. Although addition is frequently a primitive concept (defined only by properties assumed for it), it is explicitly defined in Peano's postulates for the natural numbers, in terms of the primitive operation "successor of." When this is denoted by $'$, for any two natural numbers a and b, $a + 1 = a'$ and $a + b' = (a + b)'$. *See* ALGEBRA; DIVISION; MULTIPLICATION; NUMBER THEORY; SUBTRACTION. Leonard M. Blumenthal

$$a + (b + c) = (a + b) + c \tag{1}$$

$$a + b = b + a \tag{2}$$

$$a(b + c) = a \cdot b + a \cdot c \tag{3}$$

Adenohypophysis hormone

Of the endocrine glands, the anterior pituitary, or adenohypophysis, occupies the prime place because, through the secretion of various hormones, it controls the functioning of certain other endocrine glands, namely, the adrenal cortex, the thyroid, and the gonads. In addition, hormones from the anterior pituitary influence the growth and metabolism of the organism through direct action on skeletal, muscular and other tissues. The pituitary maintains control over the various target organs by a type of feedback mechanism which is sensitive to circulating levels of hormones from the target organs. When, for example, the level of thyroxine falls below that required for the maintenance of metabolic homeostasis, the pituitary secretes thyroid-stimulating hormone which induces the thyroid gland to produce additional thyroxine. Pituitary hormones are also released in response to other metabolic conditions which they help to control. One factor influencing the secretion of growth hormone, for example, is the level of blood glucose. *See* HORMONE.

Composition. There are eight anterior pituitary hormones whose existence has been firmly established for some time. They include the two gonadotropic hormones, interstitial-cell stimulating hormone (ICSH, or luteinizing hormone, LH) and follicle-stimulating hormone (FSH); thyrotropic hormone (thryoid-stimulating hormone, TSH); lactogenic hormone (prolactin); growth hormone (GH, or somatotropin, STH); adrenocorticotropic hormone (ACTH, or adrenocorticotropin, or corticotropin); and the two melanocyte-stimulating hormones (α-MSH and β-MSH). In addition, two peptides have been isolated which are structurally related to ACTH and the MSHs. These peptides have been designated β-lipotropic hormone (β-LPH) and γ-lipotropic hormone (γ-LPH).

The hormones of the adenohypophysis are composed of amino acids in peptide linkage and are therefore classed as either polypeptides or proteins, depending on their size. In addition, three of the hormones, TSH, ICSH, and FSH, contain carbohydrate and are therefore categorized as glycoproteins. *See* AMINO ACIDS; PROTEIN.

Gonadotropic hormones. In the female, FSH initiates the development of ovarian follicles. ICSH, acting synergistically with FSH, is necessary for the final stages of follicular maturation and the production of estrogen. ICSH also stimulates the development of the corpus luteum. In the male, FSH stimulates spermatogenesis through its action on the germinal epithelium of the testis, and ICSH primarily activates

the Leydig cells which produce androgen.

ICSH has been isolated from sheep, beef, human, and pork pituitaries. The amino acid composition of ICSH from each of these sources is known, but a complete amino acid sequence has not yet been determined. As is generally true regarding species variation of a given protein hormone, the ICSHs isolated from two closely related species have greater chemical similarities than do ICSHs from distantly related species. Ovine and bovine ICSHs, for example, are more alike in amino acid composition and carbohydrate content than are ovine and human ICSHs. Ovine ICSH consists of 202 amino acids and contains carbohydrate in the form of hexose and hexosamine. The molecule has a weight of approximately 30,000 and under acidic conditions can be dissociated into two subunits having molecular weights of approximately 15,000. The subunits, however, are not identical, having different amino acid and carbohydrate compositions. Each subunit separately has no ICSH activity. Under suitable chemical conditions, however, the two subunits can be recombined to form the active hormone. Thus far ICSH is the only pituitary hormone known to have a subunit structure. *See* CARBOHYDRATE.

FSH has also been isolated from a variety of sources, but highly purified quantities of FSH have proven more difficult to obtain than ICSH, since FSH seems to be present in smaller quantities in the pituitary and is more subject to inactivation during isolation. Ovine FSH has an approximate molecular weight of 25,000.

Metabolic hormones. TSH stimulates the growth of the thyroid gland and the secretion of thyroid hormones. Although TSH has been isolated from beef and human pituitary glands, difficulties in purifying the hormone have hindered studies on its chemical properties. At one time it was believed bovine TSH had a rather low molecular weight of approximately 10,000, but studies now indicate the TSHs from the above-mentioned sources have molecular weights in the range of 25,000. *See* THYROID GLAND.

Lactogenic hormone. Through a process which requires other hormones as well, lactogenic hormone stimulates the mammary gland to secrete milk. There is evidence that in some species of mammals lactogenic hormone also plays a role in maintaining the corpus luteum in the ovary. Lactogenic hormone has been isolated from beef and sheep pituitaries. The chemistry of ovine lactogenic hormone has been substantially investigated, and the entire amino acid sequence of the molecule has been elucidated. The ovine molecule is a chain of 198 amino acids, beginning at the NH_2 terminus with threonine and ending at the COOH terminus with half-cystine. Three disulfide linkages cause the formation of three loops in the chain. In contrast to all nonprimate species investigated, a distinct lactogenic hormone has never been isolated from the pituitaries of the monkey or the human. In these two species a hormone containing lactogenic activity can indeed be isolated from the pituitary, but it also has growth-promoting activity and in major respects corresponds to growth hormones isolated from nonprimate species. *See* LACTATION.

Growth hormones. Thus, in nonprimates, a distinct growth hormone, having no lactogenic activity, can be isolated from the pituitary gland. In the primates, however, growth-promoting and lactogenic activities are present in the same molecule. In spite of its dual activities, the primate hormone is usually referred to simply as growth hormone, which promotes an increase in body size. It stimulates the growth of bones, muscles, and other tissues and enhances the effects of other pituitary hormones on their target organs. Although certain cellular effects of growth hormone are known, such as increasing incorporation of amino acids into muscle protein, the biochemical mechanisms whereby this hormone exerts its effects at the cellular level remain a mystery. The first highly purified growth hormone preparation was obtained from beef pituitaries. Although bovine growth hormone is active in the rat and mouse, it was found to have no effect on human growth when attempts were made to use it for the treatment of human dwarfism. When growth hormones were isolated from human and monkey pituitaries, it was found that only these primate growth hormones were active in humans. Human growth hormone has a molecular weight of 21,500, and the complete sequence of its 191 amino acids has been determined. Also, since the protein is not glycosylated, human growth hormone has been successfully produced via recombinant DNA technology. *See* ANIMAL GROWTH.

Adrenocorticotropin. The hormone ACTH stimulates the growth of the adrenal cortex and the secretion of cortisol and other cortical hormones. ACTHs have been isolated in pure form from sheep, beef, and hog pituitaries, and the amino acid sequences of these molecules have been determined (see **illus.**). Slight variations exist in the structures, but all of the molecules consist of a peptide chain containing 39 amino acids, having an NH_2-terminal serine and COOH-terminal phenylalanine and a molecular weight of about 4500. In 1960–1961 three groups of investigators synthesized peptides having ACTH activity. These peptides represented partial sequences of the ACTH molecule; one, for example, being a nonadecapeptide corresponding to the first 19 amino acids from the NH_2 terminus of ACTH. The entire ACTH molecule was subsequently synthesized, but the synthesis of partial sequences showed that the entire molecule was not necessary for adrenocorticotropic activity. An interesting aspect of the ACTH molecule is that it contains in part a sequence found in the melanocyte-stimulating hormones and lipotropic hormones (see below). In accordance with their chemical similarities, it is not surprising that all of these hormones exhibit both melanocyte-stimulating and lipolytic activities. The complete structural requirements for adrenocorticotropic activity, however, are not present in the MSHs and LPHs, since these hormones do not significantly increase the output of corticoid hormones when assayed in various adrenal-stimulating assays.

Intermedins. The melanocyte-stimulating hormones are also called intermedins, since they can be isolated from the intermediate lobe of the pituitary in those animals which have a distinct intermediate lobe.

Structure of bovine ACTH

```
Ser-Tyr-Ser-Met-Glu-His-Phe-Arg-Try-Gly-Lys-
 1   2   3   4   5  6   7   8   9  10  11
Pro-Val-Gly-Lys-Lys-Arg-Arg-Pro-Val-Lys-Val-
12  13  14  15  16  17  18  19  20  21  22
                                        NH2
                                         |
Tyr-Pro-Asp-Gly-Glu-Ala-Glu-Asp-Ser-Ala-Glu-
23  24  25  26  27  28  29  30  31  32  33
                    Ala-Phe-Pro-Leu-Glu-Phe
                    34  35  36  37  38  39
```

Species	Amino acid residue in position
	25 26 27 28 29 30 31 32 33
Pig	Asp-Gly-Ala-Glu-Asp-Glu-Leu-Ala-Glu (NH2 at 30)
Sheep	Ala-Gly-Glu-Asp-Asp-Glu-Ala-Ser-Glu (NH2 at 33)
Beef	Asp-Gly-Glu-Ala-Glu-Asp-Ser-Ala-Glu (NH2 at 33)

Amino acid composition and chemical structure of bovine ACTH compared with ACTH of pig, sheep, and beef. Abbreviations are: Ala, alanine; Arg, arginine; Asp, aspartic acid; Gly, glycine; Glu, glutamic acid; His, histidine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Try, typtophan; Tyr, tyrosine; and Val, valine.

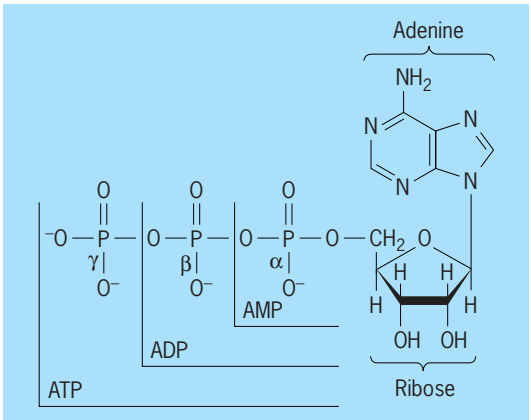
Two types of MSHs have been isolated. α-MSH isolated from pig, beef, and horse consists of 13 amino acids and has an acetylated NH_2-terminal serine and a COOH-terminal valine amide. With the exception of the NH_2-acetyl group, α-MSH is identical to the first 13 residues in ACTH. β-MSH from pig, beef, and horse consists of 18 amino acids, and the sequence from residues 7–13 is identical to the sequence from residues 4–10 in α-MSH and ACTH. Human β-MSH differs from the other known β-MSHs in having 22 amino acids, the 4 additional ones being attached to the NH_2 terminus. Melanocyte-stimulating activity refers to the ability of these hormones to cause dispersion of pigment granules in melanocytes, producing a darkening of the skin, an effect that is usually measured on frog skins. α-MSH is the most potent melanocyte-stimulating hormone as determined by frog skin assay. β-MSH has about 50% of the activity of α-MSH, while ACTH and the LPHs have about 1%.

Lipotropic hormones. In 1965–1966 two new substances, β-lipotropic hormone and γ-lipotropic hormone, were isolated from the pituitaries of sheep and were chemically characterized. The name lipotropic refers to the lipolytic activity of these substances on adipose tissue, which was the initially studied biological action of these hormones. Subsequently, it was learned that they also possessed melanocyte-stimulating activity. Lipolytic activity refers to the ability of certain hormones to stimulate the breakdown of lipid in adipose tissue to free fatty acids and glycerol. As has been mentioned, ACTH and the MSHs are also lipolytic hormones, and are, in fact, of greater potency in this regard than the LPHs. β-LPH has been isolated from beef, pork, and human pituitaries, as well as sheep, and the molecular weight of β-LPH from all of these sources appears to be about 10,000. The amino acid sequence of the ovine β-LPH molecule, which consists of 90 amino acids, has also been determined. In the middle of the chain the entire sequence of β-MSH is present, and therefore it is not surprising that the two molecules exhibit the same activities. γ-LPH, isolated thus far only from sheep pituitaries, consists of the first 58 amino acid residues of β-LPH, ending at the COOH terminus with the complete sequence of β-MSH. The discovery of the LPHs has raised speculation concerning the manner in which the pituitary gland synthesizes a number of substances having common sequences and suggests the possibility that even more peptides related to ACTH and MSH may be present in the pituitary gland. *See* PITUITARY GLAND. Choh Hao Li

Adenosine triphosphate (ATP)

A nucleotide composed of three subunits: ribose (5-carbon sugar), adenine (a nitrogenous base composed of two carbon-nitrogen rings), and a triphosphate group (a chain of three phosphate groups, designated α, β, and γ) [see **illus.**]. ATP is a vital energy-rich chemical found in all living cells. The oxidation of carbohydrates, fats, and proteins provides chemical energy that is used to synthesize ATP, which then serves as the source of the chemical energy used for biosynthesis, ion transport, and muscle contraction. The idea that ATP serves as the common currency of energy exchange in all cells is a cornerstone of biology that was first described by Fritz Lipmann in 1941.

Turnover. ATP can be broken down to adenosine diphosphate (ADP), which contains ribose and adenine but just two phosphate groups, α and β, and inorganic phosphate (P_i) with the liberation of chemical energy. Alternatively, ATP can be broken down to adenosine monophosphate (AMP) and inorganic



Structure of ATP, ADP, and AMP.

pyrophosphate (PP_i) with the liberation of energy. The magnitude of the synthesis and breakdown of ATP in humans is remarkable. A 110-lb (50-kg) adult female on a 2000-calorie (8360-kilojoule) diet breaks down and resynthesizes about 80 moles of ATP daily. This corresponds to 90 lb (41 kg) of ATP, and each molecule of ATP is broken down and resynthesized about every 1.5 min. *See* BIOLOGICAL OXIDATION; ENERGY METABOLISM; METABOLISM.

Synthesis. ATP is synthesized in cells via three general mechanisms: substrate-level phosphorylation, oxidative phosphorylation, and photophosphorylation.

Substrate-level phoshorylation. In substrate-level phosphorylation, ATP is generated from ADP and P_i as a consequence of energy-yielding oxidation-reduction reactions mediated by nonmembrane-bound enzymes. For example, during the breakdown of glucose, the sequence of reactions involved in the oxidation of glyceraldehyde 3-phosphate to 3-phosphoglyceric acid is coupled to the uptake of P_i and the formation of ATP. The breakdown of fat (or fatty acids) also generates large amounts of ATP. *See* CARBOHYDRATE METABOLISM.

Oxidative phosphorylation. During oxidative phosphorylation, ATP is generated while electrons are transported from organic compounds to oxygen by an electron transport chain. Electron transport produces an increase in the proton concentration outside the inner mitochondrial membrane in animals and plants or outside the cell membrane of oxidative bacteria. Protons, also called hydrogen ions (H^+), bear a positive charge, and the generation of a concentration gradient across a membrane leads to an electrochemical potential difference between the two sides. The protons return to the mitochondria or bacterial cell through a membrane-bound ATP synthase, and this energetically favorable process drives the energetically unfavorable conversion of ADP and phosphate to ATP. *See* ELECTRON-TRANSFER REACTION; MITOCHONDRIA.

Photophosphorylation. In photosynthetic organisms, ATP is generated as a result of photochemical reactions during photophosphorylation. Light energy generates reductants that donate electrons to a photosynthetic electron transport chain. As a result of electron transport, the proton concentration on one side of the inner mitochondrial membrane is increased, and protons pass down their concentration gradient through a membrane-bound ATP synthase resulting in ATP formation. *See* CYTOCHROME; PHOTOSYNTHESIS.

Functions. ATP is a powerful donor of phosphate groups to suitable acceptors because of the energy-rich nature of the bonds between the α- and β-phosphates and between the β- and γ-phosphates (see illus.). For instance, in the phosphorylation of glucose, which is an essential reaction in carbohydrate metabolism, the enzyme hexokinase catalyzes the transfer of the terminal phosphoryl group from ATP to glucose (a carbohydrate) as shown in reaction (1).

$$\text{ATP} + \text{glucose} \rightarrow \text{glucose 6-phosphate} + \text{ADP} \quad (1)$$

Biosynthesis reactions. ATP is involved directly in many biosynthetic reactions. It is required for reactions that lead to the formation of the bases that occur in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) and the activation of amino acids, a necessary step in protein synthesis. ATP is also used for the synthesis of amino acids, the building blocks of proteins, and nitrogen fixation in plants. ATP generated during photosynthetic phosphorylation is used to convert carbon dioxide to carbohydrate. *See* NITROGEN FIXATION.

ATP supports many other biosynthetic pathways indirectly. It converts guanosine diphosphate (GDP) to guanosine triphosphate (GTP), which is used in protein synthesis and in many other cellular processes. ATP also converts uridine diphosphate (UDP) to uridine triphosphate (UTP), which is used in carbohydrate biosynthesis, and cytidine diphosphate (CDP) to cytidine triphosphate (CTP), which is used in lipid biosynthesis. The enzyme nucleoside diphosphate catalyzes each of these transphosphorylation reactions, in which a phosphoryl group is transferred from ATP to a nucleoside diphosphate (GDP, UDP, or CDP) to produce a nucleoside triphosphate (GTP, UTP, or CTP) and ADP. These nucleoside triphosphates, along with ATP, are required for RNA synthesis. *See* LIPID METABOLISM; NUCLEIC ACID; PROTEIN; RIBONUCLEIC ACID (RNA).

Enzyme pump. In animals, a considerable amount of ATP is expended to maintain ion gradients using the sodium/potassium ATPase. This enzyme pumps three sodium ions out of cells into the extracellular compartment, and it pumps two potassium ions into cells for each ATP that is hydrolyzed to ADP and P_i.

Muscle metabolism. ATP serves as the source of energy for the mechanical work performed by muscle. Myosin, the chief muscle protein, is the site of energy transduction, by which the chemical energy of ATP is converted into the mechanical energy of muscle contraction. Myosin possesses ATPase activity; it catalyzes the reaction of ATP and water to produce ADP, P_i, and mechanical energy. In addition to oxidative phosphorylation and the breakdown of glucose (substrate-level phosphorylation), ATP synthesis takes place in muscle and the brain during brief periods of energy need through the use of creatine phosphate. Creatine phosphate possesses an energy-rich linkage and is able to transfer the terminal phosphoryl group to ADP to form ATP, which can be used to energize muscle contraction when ATP would otherwise be depleted. Creatine kinase is the enzyme that mediates the reaction (2). When

$$\text{Creatine phosphate} + \text{ADP} \rightleftharpoons \text{ATP} + \text{creatine} \quad (2)$$

the energy stores of muscle return, ATP mediates the rephosphorylation of creatine to form creatine phosphate.

The adenylate kinase, or myokinase, reaction mediates the transfer of a phosphoryl group from one ADP to another, generating a molecule of ATP that can be used for myosin ATPase and a molecule of

adenylic acid (AMP) as shown in reaction (3). This

$$\text{ADP} + \text{ADP} \rightleftharpoons \text{ATP} + \text{adenylic acid} \qquad (3)$$

reaction releases energy from ADP to produce ATP in times of energy need. When the energy stores of muscle return, the reaction runs in reverse—ATP donates its terminal phosphoryl group to adenylic acid to form two molecules of ADP. *See* MUSCLE; MUSCLE PROTEINS.

Robert Roskoski, Jr.

Bibliography. J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 5th ed., 2001; C. Mathews, K. Van Holde, and K. Ahern, *Biochemistry*, 3d ed., 2000; D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 3d ed., 2000; D. G. Nicholls and S. J. Ferguson, *Bioenergetics 3*, 2002; R. Roskoski, Jr., *Biochemistry*, 1996.

Adeno-SV40 hybrid virus

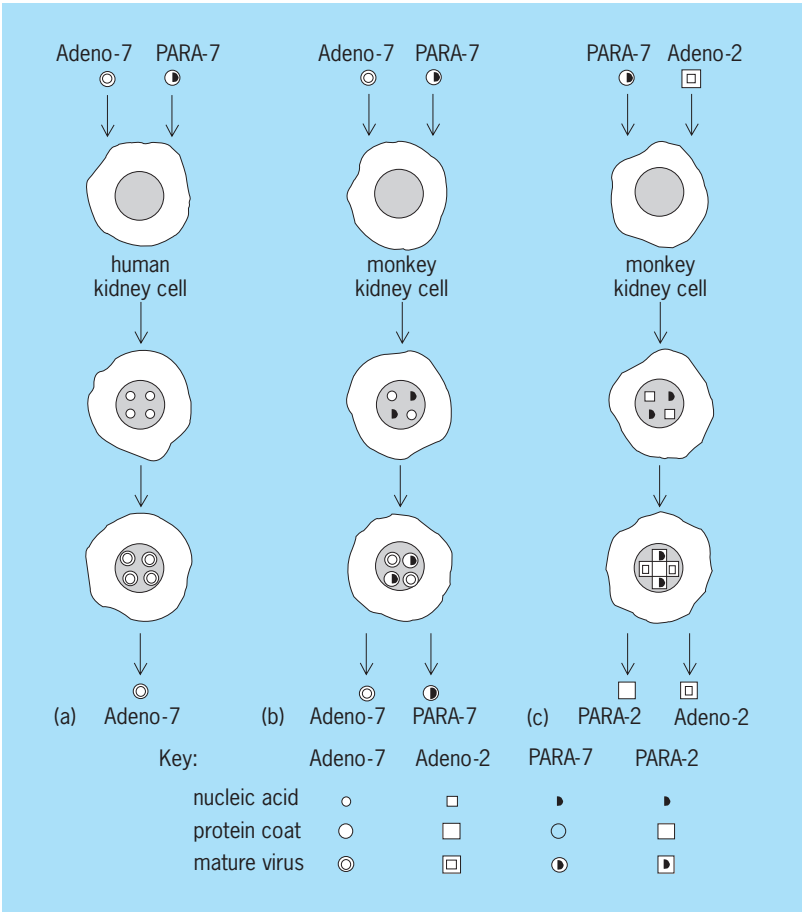
A type of defective virus particle in which part of the genetic material of papovavirus SV40 is encased within an adenovirus protein coat (capsid). Human adenoviruses require human cells for their propagation; however, papovavirus SV40 can serve as a "helper," enabling human adenoviruses to replicate in monkey cells.

Characteristics. After numerous continuous passages of adenovirus and SV40 together in monkey cells, adenovirus progeny may be obtained that possesses properties different from those of the original parent adenovirus. It then produces tumors in newborn hamsters; it replicates in monkey cell cultures; and although it does not produce infectious SV40, it causes monkey cells in culture to produce a new cellular antigen known to be specifically induced by SV40—the SV40 tumor, or T, antigen. The new virus stock therefore behaves like a hybrid. It has proved to be a population of two distinct kinds of virus particles. One kind of particle is a true adenovirus. The other particle is the adeno-SV4 hybrid, which has an adenovirus coat, but whose genetic material appears to consist of defective adenovirus type 7 DNA, representing about 85% of the adenovirus genome, covalently linked to a portion (about 50%) of an SV40 genome.

In this virus population the particle carrying the SV40 genetic material has been termed PARA (particle aiding replication of adenovirus). The PARA is considered an unconditionally defective virus, since under no known conditions can it reproduce itself except in a cell coinfected with adenovirus. The adenovirus is considered conditionally defective, since it can reproduce independently in human cells but not in monkey cells (**illus.** *a* and *b*).

When this hybrid virus stock is cultivated serially in human cell cultures, only the true adenovirus particles reproduce. The SV40 genetic determinants are no longer present in the progeny, and the new generations of virus, if inoculated into monkey cells, cannot reproduce themselves. On the other hand, if the hybrid stock population is cultivated in monkey cells, the two kinds of particles complement each other, and new generations contain both kinds of particles. Thus there is a partnership in which the adeno-SV40 hybrid particle supplies genetic information that allows the human adenovirus to reproduce in these foreign cells, and the adenovirus supplies genetic information for synthesis of the protein coat that endows the hybrid particle with the power to attach to cells and to infect them.

Transcapsidation. Type 7 and a number of other adenovirus types can act as helpers for the replication of PARA. In each case the PARA progeny takes on the type-specific protein coat, or capsid, of the helper adenovirus. This change in the capsid of PARA from one type of adenovirus to another is termed transcapsidation (illus. *c*). Some human adenoviruses are able by themselves to induce formation of tumors in hamsters, but other adenoviruses are nononcogenic. After transcapsidation of the PARA particle to some of these previously nononcogenic adenoviruses, they acquire the ability to induce tumors in newborn hamsters. When the tumors are analyzed, both SV40 tumor antigen and adenovirus tumor antigen can be detected in the tumor cells,



Interactions between PARA-7 particles and adenoviruses. (*a, b*) Mutual dependence between adeno particles and PARA in hybrid populations. The synthesis of SV40 tumor antigen in *b* by a hybrid requires the multiplication of both adeno particles and PARA. (*c*) Transcapsidation. PARA-7 particles (from an SV40-adenovirus type 7 hybrid) grown in the presence of adenovirus type 2 acquire the protein coat of the helper adenovirus. This results in an SV40-adenovirus type 2 hybrid population containing both pure adeno-2 and PARA-2 particles.

and the tumor-bearing animals develop antibodies to both of these antigens. The tumors are of three main types: those with histopathology characteristic for adenovirus tumors, those typical of SV40 tumors, and those with a mixture of both types of pathology.

A second type of hybrid, the $Ad2^{+}ND$ viruses, consists of a series of nondefective adenovirus type 2 isolates carrying different amounts (5–44%) of the SV40 genome. Hybrids of still another type, designated $Ad2^{++}$, contain the entire SV40 genome and are still defective, but they can be propagated in cells coinfected with wild adenovirus type 2.

MAC adenovirus hybrid. Other stocks of human adenovirus have been found that are able to replicate in monkey cells when neither SV40 nor PARA is present. Such stocks contain two types of particles similar to the adenovirus-plus-PARA populations described in the foregoing. The "hybrid" particle containing the nonadenovirus genetic material which makes the adenovirus virulent for monkey cells is known as MAC (monkey adapting component); like PARA, it can be transcapsidated to other types of adenovirus. The source from which this foreign genetic material originally was picked up by the adenovirus is unknown.

Significance. The discovery and exploration of the properties of these hybrids have shown that one virus can usurp the protein coat of another and, while concealed from the usual viral identity tests, can endow a previously mild virus with new powers—which can include the power to produce virulent infection in some cases or cancer in other instances.

These hybrid viruses have been useful in the genetic analysis of SV40. Through the use of physical mapping techniques, the precise amount of SV40 information has been determined for each hybrid. Through comparison of the SV40 antigens (T, U, TSTA) expressed in cells infected by each of the hybrids, a map has been constructed for SV40 genes that code for the different antigens. *See* ADENOVIRIDAE; ANIMAL VIRUS; VIRUS, DEFECTIVE.

Joseph L. Melnick

Bibliography. T. J. Kelly, Jr., et al., *Cold Spring Harbor Symp. Quant. Biol.*, 39:409–417, 1975; F. Rapp and J. L. Melnick, The footprints of tumor viruses, *Sci. Amer.*, 214:34–41, 1966.

Adenoviridae

A family of viral agents. They have been associated with pharyngoconjunctival fever (types 3, 4, and 7, in particular), acute respiratory disease (ARD), epidemic keratoconjunctivitis (type 8), and febrile pharyngitis in children (types 1, 2, and 5). A number of types have been isolated from tonsils and adenoids removed from surgical patients. Although most of the illnesses caused by adenoviruses are respiratory, adenoviruses are frequently excreted in stools, and certain adenoviruses have been isolated from sewage. Distinct serotypes of mammalian (genus *Mastadenovirus*) and avian (genus *Aviadenovirus*) are known. These genera contain 87 and 14 species, respectively. *See* ANIMAL VIRUS.

Infective virus particles, 70 nanometers in diameter, are icosahedrons with shells (capsids) composed of 252 subunits (capsomeres). No outer envelope is known. The genome is double-stranded deoxyribonucleic acid (DNA), with a molecular weight of $20-25 \times 10^6$. Three major soluble antigens are separable from the infectious particle by differential centrifugation. These antigens—a group-specifc antigen common to all adenovirus types, a type-specific antigen unique for each type, and a toxinlike material which also possesses group specificity—represent virus structural protein subunits that are produced in large excess of the amount utilized for synthesis of infectious virus.

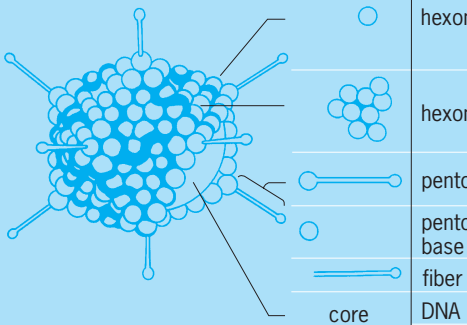
The known types of adenoviruses of humans total at least 33, and previously unrecognized types continue to be isolated. The serotypes are antigenically distinct in neutralization tests, but they share a complement-fixing antigen, which is probably a smaller soluble portion of the virus.

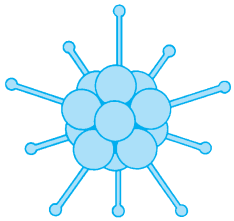
Morphologic and antigenic subunits. Adenoviruses contain four complement-fixing antigens—A, B, C, and P—whose characteristics are described in the **table**. The hexons, each surrounded by six neighboring capsomeres, constitute the majority of the capsomeres of the adenocapsid (240); they are tubelike and are 8 nm in diameter. The pentons, found at the 12 vertexes of the capsid, are also 8 nm in diameter. Associated with the penton base antigen is a toxinlike activity which causes tissue culture cells to detach from the surface on which they are growing. Attached to each penton base is a fiber antigen. For adenovirus type 5, the fiber antigen consists of a rod (2×20 nm) attached to the penton, with a 4-nm knob at the end. However, the dimensions of the fiber antigen vary according to the adenovirus type. The P antigen of adenoviruses is an internal antigen which is released upon virus disruption and is very unstable. The hemagglutinating activity of adenovirus is associated with the penton and fiber antigens. Excess pentons produced in cells infected with some types of adenovirus (3, 4, 7, 9, 11, and 15) form dodecahedral aggregates of 12 pentons (dodecons) which have hemagglutinating activity.

Protein components. Polyacrylamide gel electrophoresis of purified adenovirus and adenovirus subunits after disruption by detergent (sodium dodecyl sulfate) and urea has revealed nine virus-specific polypeptides, three of which are associated with the DNA-containing viral core. The hexon capsomere is composed of three to six molecules of a single type (component II) which makes up about 50% of the total virion. Groups of these hexon capsomeres ("groups of nine") are associated with two additional minor polypeptides (components VIII and IX). The penton base is composed of a single type of peptide (component III), as is the fiber antigen (component IV). The internal core released by treatment of the virion with 5 *M* urea contains the viral DNA in association with three arginine-rich peptides (components V, VI, and VII) which constitute about

Table showing comparative data on adenovirus type 2 morphologic and antigenic subunits and protein components

	Morphologic subunits			Antigenic subunits		Protein components
Appearance	Name	Number per virion	Molecular weight	Antigen	Specificity	
virion	DNA		23,000,000			
	protein		150,000,000			
	hexon	240	210,000 400,000 320,000 360,000	A	group	II
	hexons	20	3,600,000			II, VIII, IX
	penton	12	280,000 1,100,000			III, IV
	penton base	12	210,000	B	subgroup	III
	fiber	12	70,000	C	type	IV
core	DNA	1	23,000,000	P		
	protein		29,000,000			V, VI, VII
	protein		13,000			VIII, IX
	protein		7,500			X





dodecon: hemagglutinin made up of 12 pentons with their fibers

20% of the total virion protein. In addition, there is a very small polypeptide (component X) whose role and location in the virion remain unknown.

Subgroups. A hemagglutinin has been used to separate the human adenoviruses into subgroups. Group A (types 3, 7, 11, 14, 16, 20, 21, 25, and 28) agglutinates rhesus but not rat erythrocytes; group B (8, 9, 10, 13, 15, 17, 19, 22, 23, 24, 26, 27, 29, and 30) agglutinates rat cells but not (or hardly) rhesus cells; group C (1, 2, 4, 5, and 6) fails to agglutinate rhesus cells and only partially agglutinates rat cells. Types 12, 18, and 31 do not usually agglutinate, but some strains partially agglutinate rat cells. The hemagglutination inhibition test is being used for type-specific identification. Some cross-reactions have been noted, however.

Pathogenicity and tumors. The virus does not commonly produce acute disease in laboratory animals but is cytopathogenic, that is, destroys cells, in cultures of human tissue. Certain human adenovirus serotypes produce cancer when injected into newborn hamsters.

In cells derived from other species the human adenoviruses undergo an abortive replicative cycle. Adenovirus tumor antigen, messenger RNA (mRNA), and DNA are all synthesized, but no capsid proteins or infectious progeny are produced. Although complete, infective virus cannot be recovered from adenovirus-induced hamster tumors, a new antigen induced by the virus can be detected by complement fixation or immunofluorescence techniques. This antigen, called tumor or T antigen because of its association with tumor or transformed cells, can also be detected in the cytolytic cycle of the virus. Hamster cells can be transformed in tissue cultures by the oncogenic human adenoviruses. These cells contain the adenovirus tumor antigen, do not contain infectious virus, and produce tumors when inoculated into adult hamsters. Adenovirus mRNA can be detected in the cytoplasm and nucleus of both transformed and tumor cells.

Base ratio determinations have revealed three distinct groups of adenoviruses: those with a low guanine plus cytosine (G + C) content (48–49%); those with an intermediate G + C content (50–53%); and those with a high G + C content (56–60%). The strongly oncogenic adenovirus types 12, 18, and 31 are the only members of the group with low G + C, and certain adenoviruses in the intermediate group (types 3, 7, 14, 16, and 21) are mildly oncogenic. The adenovirus mRNA observed in transformed and tumor cells has a G + C content of 50–52% in the DNA. This suggests that viral DNA regions containing 47–48% G + C are integrated into the tumor cells or that such regions are preferentially transcribed. However, the mRNA from tumor cells induced by one

subgroup such as the highly oncogenic adenoviruses (types 12 and 18) do not hybridize with DNA from the other two subgroups. Apparently, different viral-coded information is involved in carcinogenesis by the three different groups of adenoviruses.

With simian adenovirus 7 (SA7), the intact genome, as well as the heavy and light halves of the viral DNA, is capable of inducing tumors when injected into newborn hamsters. Extensive studies have failed to demonstrate adenovirus DNA or viral-specific mRNA in human tumors.

Adenoviruses, especially types 3, 4, 7, 17, and 21, cause by far the largest number of disabling respiratory diseases among military recruits. In seasoned troops, adenovirus disease is not a serious problem; but the rate in recruits is 33 times higher than in the older group. Presumably, seasoned troops have acquired immunity to the common adenoviruses either before or during their basic training.

Live virus vaccines against type 4 and type 7 have been developed and used extensively in military populations. When both are administered simultaneously, vaccine recipients respond with neutralizing antibodies against both virus types. *See* ADENO-SV40 HYBRID VIRUS; ANTIGEN; COMPLEMENT-FIXATION TEST; NEUTRALIZATION REACTION (IMMUNOLOGY); VIRUS CLASSIFICATION.

Joseph L. Melnick; M. E. Reichmann

Bibliography. E. Everitt et al., Structural proteins of adenoviruses, X: Isolation and topography of low molecular weight antigens from the virion of adenovirus type 2, *Virology,* 52:130–147, 1973; H. Fraenkel-Conrat and R. R. Wagner (eds.), *Comprehensive Virology,* vol. 3: *Reproduction, DNA Animal Viruses,* pp. 143–228, 1974; G. G. Jackson and R. L. Muldoon, Viruses causing common respiratory infections in man, IV: Reoviruses and adenoviruses, *J. Infect. Dis.,* 128:811–866, 1973; B. A. Rubin and H. Tint, The development and use of vaccines based on studies of virus substructures, *Progr. Med. Virol.,* 21:144–157, 1975.

Adhesive

A material capable of fastening together two other materials by means of surface attachment. The terms glue, mucilage, mastic, and cement are synonymous with adhesive. In a generic sense, an adhesive is any material capable of fastening by means of surface attachment, and thus includes inorganic materials such as portland cement and solder. Practically, however, adhesives are a broad set of materials composed of organic, primarily polymeric materials that can be used to fasten two other materials together. The materials being fastened are often called adherends, and the resulting assembly is called an adhesive joint or adhesive bond. *See* ADHESIVE BONDING.

Theories of adhesion. Adhesion is the physical attraction between two surfaces. This attraction is the result of intermolecular interactions that provide the attraction between atoms and molecules, such as electrostatic, van der Waals forces (dipole-dipole interactions, dipole–induced dipole interactions, and dispersion forces), hydrogen bonding, and covalent bonding. With the exception of electrostatic, all of these interactions are strong only at short (nanometer-range) distances. This leads directly to the primary theory of adhesion, the wettability-adsorption theory, in which an adhesive must come into intimate contact with or wet the adherend in order for it to adhere. In this theory, surface energetics and the measurement and interpretation of contact angles play an important role. The diffusion theory of adhesion is used to describe the situation in which an adhesive and adherend are soluble in one another. Here the solubility parameter or the χ-parameter plays an important role. Covalent bonding at interfaces is usually not necessary for a strong adhesive joint. However, it has been found that interfacial covalent bonding is important when the adhesive bond is exposed to adverse environmental conditions such as temperature and humidity. *See* ADSORPTION; CHEMICAL BONDING; INTERFACE OF PHASES; INTERMOLECULAR FORCES; SOLUTION; SURFACE TENSION.

Even though the strength of an adhesive joint depends fundamentally on adhesion, the strength of an adhesive joint is not equal to the force of adhesion. Rather, the strength of an adhesive joint depends in a complex way on many factors, including adhesion, the design of the adhesive joint, and the physical properties of the adhesive and the adherend. *See* JOINT (STRUCTURES).

Types of adhesives. Most polymeric materials have been evaluated as adhesives. Materials that are recognized as adhesives come into intimate contact with the adherends as liquids and then solidify to form the adhesive bond.

Adhesives have been used since ancient times, with the earliest adhesives coming from flora (tree resins and vegetable oils), fauna (casein, blood, and collagen), or a combination of these from prehistoric times, such as bitumen and tar. Generically, these adhesives are called bioadhesives or adhesives of natural origin. These materials are still used. Protein-based adhesives are used to make interior-grade plywood. Some protein-based adhesives are used as tissue sealants during surgery. Cellulosic and starch-based adhesives are used in paper binding. *See* BIOPOLYMER.

The most significant growth in the development and use of adhesives came with the development of synthetic (organic) polymers. The first of these, based on the reaction of phenol and formaldehyde, were phenolic resins. Broadly, organic polymers are classified as thermosets and thermoplastics. Thermoplastics become soft or liquid upon heating and are soluble. Upon heating or curing, thermosets become insoluble and infusible. Both polymer types are used as adhesives. Depending on the ambient temperature, thermoplastics and cured thermosets can be glassy or elastomeric. *See* PHENOLIC RESIN; POLYMER.

Pressure-sensitive adhesives. Pressure-sensitive adhesives are primarily thermoplastic elastomers but may

have some thermoset character. These materials exhibit tack, the property of an adhesive that enables it to form a bond of measurable strength immediately after the adherend and adhesive are brought into contact under low pressure. Thus, when an adhesive is on a backing (as in the case of a pressure-sensitive adhesive tape), just touching the adhesive to a surface is enough to provide measurable adhesive bond strength.

The performance of a pressure-sensitive adhesive results from its viscoelasticity; that is, the adhesive displays both liquidlike and solidlike properties that depend upon temperature and rate. The silicone-based polymer Silly Putty® provides a good example of viscoelasticity. If one places Silly Putty on a surface, after sufficient time (low rate of application of force) it will flow like a viscous liquid. However, if one rolls the Silly Putty into a ball and bounces it against a surface (high rate of application of force), it responds like a solid or elastically.

Pressure-sensitive adhesives are made from a wide range of elastomeric materials, including natural rubber, acrylic resins, poly(vinyl ethers), styrene-butadiene copolymers, and silicone, to name some. Pressure-sensitive adhesives are primarily used for adhesive tapes, which come in various types depending upon the backing and the adhesive. Uses range from paper mending tape, to packaging tape, to removable notes, to adhesive bandages. *See* COPOLYMER; POLYACRYLATE RESIN; POLYVINYL RESINS; RUBBER; SILICONE RESINS.

Semistructural adhesives. Semistructural adhesives are able to support a small load (50–500 lb/in.^{2}; 0.3–3 MPa) for a long time. Semistructural adhesives are either elastomer-based or hot-melt. The elastomer-based adhesives are usually partially thermoset materials formulated with various additives in solvent. These adhesives are used for applying ceramic tile on walls, adhering carpeting to floors, laminating furniture, and many other applications. Elastomers often used in this type of adhesive are chloroprene and nitrile rubber.

Hot-melt adhesives are primarily thermoplastic in character. They are solid materials formulated or synthesized to provide desired physical properties. The melted (liquid) adhesive wets the surface. Upon cooling, the adhesive solidifies and provides adhesive strength. In recent years, the development and use of hot-melt adhesives has become popular because these adhesives do not require the use of solvents for their formulation or application. Hot-melt adhesives, therefore, are environmentally friendly. The materials most often used in hot-melt adhesives are the semicrystalline thermoplastics such as ethylene–vinyl acetate co-polymers and other polyolefins, but polyamides and polyesters are also used. The uses of hot-melt adhesives range from paper and book binding, to laminating, to craft and product assembly. *See* POLYAMIDE RESINS; POLYESTER RESINS; POLYOLEFIN RESINS.

Structural adhesives. Structural adhesives are primarily thermosets and, once cured, have the ability to bond high-strength solids such as wood and metal to create adhesive joints that can bear a significant load (>1000 psi; 7 MPa) for a long time. These adhesives are liquids or become liquids during application. The adhesive solidifies (cures) by means of a chemical reaction that can be initiated by heat or some other means.

Structural adhesives are chemically quite varied, ranging from phenolics and epoxies to polyurethanes and acrylics and encompassing cyanates and polyimides. The chemistry of the adhesive governs the application and use conditions. Epoxy adhesives can be cured at temperatures ranging from room temperature to 170°C (338°F) and can provide strength to temperatures as high as 200°C (392°F). Polyimides cure at high temperatures (in excess of 170°C or 338°F) and provide strength to 350°C (662°F). Acrylic adhesives provide high strength and rapid cure at room temperature. *See* EPOXIDE; HETEROCYCLIC POLYMER; POLYETHER RESINS; POLYURETHANE RESINS.

A well-known structural adhesive is the cyanoacrylate Super-Glue™. Recently, a cyanoacrylate-based adhesive has been approved for medical applications such as wound closure. The primary use of structural adhesives is in load-bearing structures such as aircraft, automobiles, and building materials (such as plywood, pressboard, and strand board). *See* COMPOSITE MATERIAL; WOOD ENGINEERING DESIGN.

Alphonsus V. Pocius

Bibliography. A. J. Kinloch, *Adhesion and Adhesives*, 1987; J. D. Minford, *Treatise on Adhesion and Adhesives*, vol. 7, 1991; R. L. Patrick (ed.), *Treatise on Adhesion and Adhesives*, vols. 1–6, 1967–1990; A. V. Pocius, *Adhesion and Adhesives Technology: An Introduction*, 2d ed., 2002; A. V. Pocius, D. A, Dillard, and M. K. Chaudhury (eds.), *Adhesion Science and Engineering*, vols. 1 and 2, 2002; D. Satas (ed.), *Handbook of Pressure Sensitive Adhesives*, 3d ed., 1999; I. Skeist (ed.), *Handbook of Adhesives*, 3d ed., 1990.

Adhesive bonding

The process of using an adhesive to manufacture an assembly. In this article, the adhesive-bonded assembly will be called an adhesive joint, and the materials to which the adhesive adheres will be called the adherends. The adhesive bonding process consists of adhesive joint design, adherend surface preparation, adhesive application, adhesive joint assembly, solidification of the adhesive (if necessary), adhesive joint evaluation, and use of the assembly. *See* ADHESIVE.

Adhesive joint design. As shown in **Fig. 1**, there are three primary ways in which a load can be applied to an adhesive joint: tension, shear, and cleavage or peel. In general, adhesive joints are designed to perform under a shear load, although adhesive joints can be loaded in all three modes. As with all materials, adhesives are weaker in cleavage than they are in shear or tension. Adhesive joints are designed to minimize the possibility of cleavage loading. *See* SHEAR.

Adhesive joints are designed by first determining the loads that are to be supported. Adherends and adhesives are chosen according to the needs of the application. Properties such as stiffness, toughness (fracture resistance), and elongation are considered. Mechanical engineering principles are applied to ensure that the joint can support the necessary load. A properly designed adhesive joint will provide for adherend failure rather than adhesive failure unless the joint is designed to be reworked or reused. Usually, the design is subjected to a test protocol before production.

Adherend surface preparation. Adhesive joints are made by means of surface attachment; thus the condition of the adherend surface must be taken into account. This is particularly important when the adhesive joint is to be exposed to adverse environmental conditions such as temperature and humidity. In general, surface preparation is done to remove weak boundary layers (such as oils and greases), increase the adherend surface energy, and provide a surface with enough mechanical roughness to "key" the adhesive into the surface of the adherend.

Metal adherend surface preparation methods include abrasion (for example, sanding or grit blasting) and electrochemical methods such as etching for most metals or anodization for metals such as aluminum, titanium, and magnesium. Other inorganic adherends, such as glass and ceramics, can be mechanically abraded or etched. Polymer surface preparation methods include abrasion, chemical etching, corona and flame treatment, and plasma treatment. Bicyclists are familiar with the use of sandpaper or other abrasive to "roughen up" the surface of an inner tube in order to get better adhesion of a patch. In general, a wood surface is not prepared except for sanding the surfaces to be joined in order to ensure proper mating.

In some cases, a primer is applied to the adherend before applying the adhesive. For plastics, a primer can sometimes take the place of a surface preparation. For inorganic materials, a primer can assist formation of covalent chemical bonds between the adherend and the adhesive. Silanes are commonly used to achieve interfacial covalent bonding between an inorganic surface and an organic adhesive. For metals, primers often have other functions such a corrosion protection. *See* CHEMICAL BONDING; ORGANOSILICON COMPOUND; SILICON.

Adhesive application and joint assembly. For a proper adhesive joint, the adhesive must come into intimate contact with or wet the adherend. As a guideline, the adhesive must have a liquid surface energy less than the critical wetting tension of the adherend. If the adherend's surface has been properly prepared or primed, this is usually achievable. Alternatively, the adhesive can be chosen such that intimate contact is achievable. *See* INTERFACE OF PHASES; SURFACE TENSION.

The adhesive can be applied to the adherend by a number of means. The adhesive can be troweled into place as is often done in the application of ceramic tile or floor tile. It can be roller-coated or sprayed as is often done in lamination of surfacing materials onto furniture. The adhesive can be caulked into place. Adhesives in film form are applied as a solid material that is melted and then cooled (solidified). An adhesive can be coated on a backing, such as a pressure-sensitive adhesive tape, and applied by pressure with no further processing steps required.

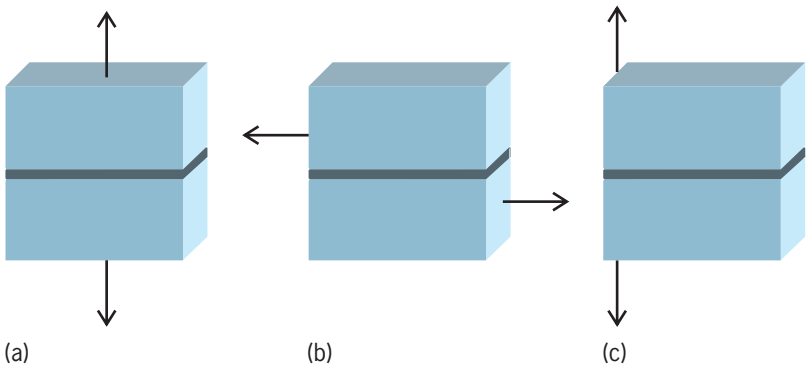


Fig. 1. Modes in which adhesive bonds can be loaded: (*a*) tension, (*b*) shear, and (*c*) cleavage. The layer in the middle of the specimens is the adhesive. In general, adhesive bonds are strongest in shear and weakest in cleavage.

Joint assembly is an important consideration in adhesive bonding. In many cases, the adhesive has a set time, a period in which the adhesive has little or no strength until some solidification takes place. During the solidification process, the adherends must be kept in place by some means. This can be done with materials as crude as simple weights, but often the adherends are kept in place by means of clamps. In some applications, such as aerospace adhesive bonding, the adhesive joint is constructed in a specially made tool that holds the parts in place during solidification. Care must be taken that the force used to hold the adherends in place does not squeeze out significant quantities of the adhesive or the strength of the joint could be adversely affected.

Adhesive solidification. Pressure-sensitive adhesives usually require no processing to solidify, as they are already viscoelastic solids; that is, they have both liquidlike and solidlike character. Adhesives such as rubber-based adhesives and contact bond adhesives require the evaporation of solvent or water to solidify. Thus, some provision must be made to remove the solvent. Adhesives may undergo a chemical reaction to solidify. For example, two-part epoxy adhesives must be properly mixed in order to effect the solidification (cure) of the adhesive. Some adhesives require the application of heat to cure the adhesive, while other adhesives are cured by ultraviolet or visible light. Hot-melt adhesives are applied in the liquid state and solidify upon cooling. *See* POLYETHER RESINS; POLYMER; RUBBER.

Adhesive evaluation. In the majority of cases, adhesive joints cannot be tested for strength without actually breaking the joint. Evaluation of adhesive bonding processes is usually done by coupon tests. That is, model adhesive joints are generated on the same processing line with the actual joints and then tested to failure. If the model joints have strength that is in line with specifications, it can be assumed that the

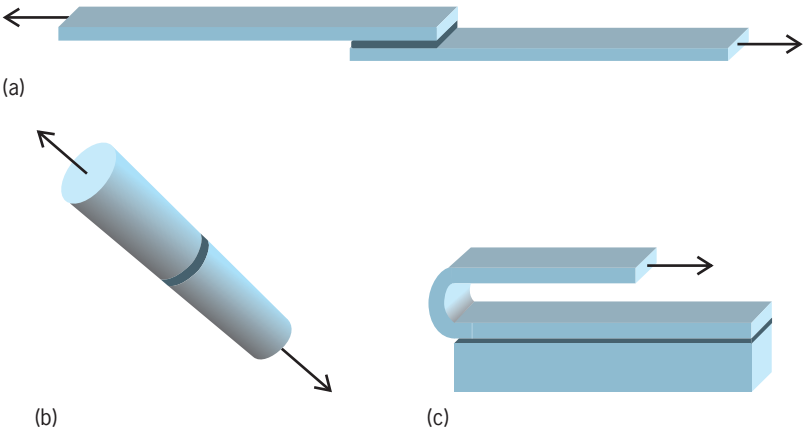


Fig. 2. Examples of adhesive bond test specimens. (*a*) Lap shear test specimen (ASTM D1002). (*b*) Butt tensile specimen (ASTM D2095). (*c*) 180° peel specimen (ASTM D903).

actual joints will perform properly. Model adhesive test joints, as considered in the American Society for Testing and Materials (ASTM), include the lap shear specimen, tensile specimens, and various peel specimens. **Figure 2** shows several of the model adhesive joints that are used to evaluate adhesives and adhesive bonding processes. The lap shear joint is used to test the shear properties of the adhesive, while peel tests are used to evaluate the fracture resistance (toughness) of the adhesive.

Adhesive uses. Adhesives have a wide range of uses. Many types of electronic assemblies use adhesives. For example, silicon chips (dies) are attached to lead frames by means of a die-attach adhesive. This type of adhesive is usually heat-cured, and the cured adhesive is thermally conductive. Adhesives are also used to rigidize parts on a circuit board.

In the automotive industry, adhesives are used to attach decorative features such as body side moldings. They also are used to attach brake linings to brake shoes and friction surfaces in transmissions. Many parts of the automobile are bonded in a hem flange arrangement, in which an inner panel is bonded to an outer skin using an adhesive. The outer skin is then crimped over the inner panel.

Adhesives are used extensively in furniture construction, with rubber-based adhesives used to adhere fabric over foam cushions. Rubber-based adhesives also are used to laminate Formica™ to base wood to produce counter tops. Floor tile and wall tile are attached by means of rubber-based adhesives. Plywood manufacturers are one of the biggest users of adhesive. *See* PLYWOOD.

Aircraft manufacturers extensively use adhesive bonding because of several favorable attributes of adhesives. These include their ability to bond and seal simultaneously, the low specific gravity of polymeric adhesives, and the ability of the adhesive to dampen vibrations so that the adhesive joint is more fatigue-resistant than a welded or mechanically fastened joint.

Adhesives are also used in many noncritical applications such as paper binding, carton sealing (hot-melt adhesives), and envelope sealing. Pages in books have been mended with pressure-sensitive adhesive-backed tapes. Information is transmitted using Post-It™ notes.

Adhesives are being used more often in medicine. In addition to the adhesive bandage that uses a pressure-sensitive adhesive, adhesives are used as tissue sealants during surgeries, and have recently received permission to be used as wound closure materials. Adhesive bonds are also used in transdermal drug delivery systems (such as the nicotine patch). *See* DRUG DELIVERY SYSTEMS.

Alphonsus V. Pocius

Bibliography. A. J. Kinloch, *Adhesion and Adhesives*, Kluwer Academic, 1987; J. D. Minford (ed.), *Treatise on Adhesion and Adhesives*, vol. 7, Marcel Dekker, 1991; R. L. Patrick (ed.), *Treatise on Adhesion and Adhesives*, vol. 6, Marcel Dekker, 1990; A. V. Pocius, *Adhesion and Adhesives Technology: An Introduction*, 2d ed., Hanser Gardner, 2002; A. V. Pocius, D. A. Dillard, and M. K. Chaudhury (eds.), *Adhesion Science and Engineering*, vols. 1 and 2, Elsevier Health Sciences, 2002; D. Satas (ed.), *Handbook of Pressure Sensitive Adhesives*, 3d ed., Satas & Associates, 1999; I. Skeist (ed.), *Handbook of Adhesives*, 3d ed., Kluwer Academic, 1990.

Adiabatic demagnetization

The removal or diminution of a magnetic field applied to a magnetic substance when the latter has been thermally isolated from its surroundings. The process concerns paramagnetic substances almost exclusively, in which case a drop in temperature of the working substance is produced (magnetic cooling). *See* PARAMAGNETISM.

Thermodynamic principles. When a specimen is magnetized in a magnetic field $\mathbf{H}$ and a magnetic moment $\mathbf{M}$ is produced, there results a contribution to the energy of the specimen of magnitude $-\mathbf{M} \cdot \mathbf{H}$. By restricting the discussion to isotropic media for simplicity, and taking the ambient pressure to be modest and the compressibility of the substance to be very small, the thermodynamics of the situation may be taken over from the conventional results entirely by substituting $-MH$ for $+PV$ (where P is the pressure and V is the volume) everywhere. One particular result is the relation below, where T is the

$$(\partial T/\partial H)_S = (\partial M/\partial S)_H = -(T/C_H)(\partial M/\partial T)_H$$

temperature, S is the entropy, and C_H is the heat capacity at constant magnetic field. As C_H is always positive, the sign of the magnetocaloric effect is determined by (and is opposite to the sign of) the quantity $(\partial M/\partial T)_H$. Where the latter is negative, as is the case for paramagnetics, $(\partial T/\partial H)_S$ is positive and magnetization produces heating, and vice versa. The maximum theoretical effect will not be achieved in practice unless the process is truly isentropic. In addition, it is obviously desirable that nonmagnetic contributions to C_H be kept small. In nearly all cases, this requires restricting the process to a region far

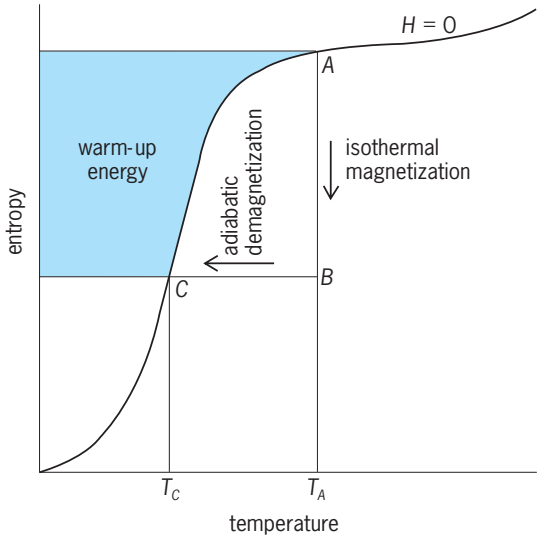


Fig. 1. Entropy–temperature diagram of a paramagnetic substance, showing process of adiabatic demagnetization.

below room temperature, where the lattice specific heat is very small in comparison with that of the assembly of magnetic ions (the spin system). Adiabatic demagnetization was employed from 1933 on to produce temperatures below those readily obtainable by using only liquid helium (that is, below 1 K). *See* LIQUID HELIUM; MAGNETOCALORIC EFFECT.

The process is most perspicuously discussed in terms of entropy rather than energy or heat capacity (**Fig. 1**). The atomic magnets (electronic or nuclear), of moment μ, interact weakly with each other, but at sufficiently low temperatures ($T \sim U/k$, where k is Boltzmann's constant) the interaction energy U will restrict the motions of the atomic magnets and cause a drop in entropy S (Fig. 1). Well above this temperature region, the orientation of μ is practically random over the possible $(2J + 1)$ orientations (where J is the quantum number appropriate to the situation under discussion), and the "infinite temperature" entropy of $R \cdot \ln (2J + 1)$ per mole is approached, where R is the gas constant. At still higher temperatures, the lattice entropy begins to rise to nonnegligible values. Consider a starting condition represented by a state A, the temperature being T_A. Application of a magnetic field isothermally will reduce the entropy to a value corresponding to a state B, say, governed by the thermodynamic relation $(\partial S/\partial H)_T = (\partial M/\partial T)_H$. (The energy which is released by the magnetization process is carried away by exchange gas, helium at low pressure, or a heat switch.) After reaching B, the substance is isolated from its surroundings and H is reduced to zero. Now, no energy exchange with the surroundings can occur, the system returns to the zero-field entropy curve along an isentrope (horizontal line BC), and it arrives at a state C, where the temperature T_C is much lower than temperature T_A. *See* ENTROPY; THERMODYNAMIC PRINCIPLES.

Limiting temperatures. The thermodynamic properties are found to be functions of the quantity $(\mu H/kT)$. The situation may be thought of as a competition between the magnetic (restraining) energy μH and the thermal (disruptive) energy kT. If the interaction U is thought of as arising from an internal magnetic field b, then b at a temperature T_C and H at T_A are equally effective in holding the system at the diminished entropy represented by the isentrope CB. Hence, given that S and, therefore, $(\mu H/kT)$ are constant, the equations below provide a rough approximation. In order to produce a very low temperature, T_C, it is necessary that b and T_A be small and H large. Traditionally, low-temperature physicists have employed electronic paramagnets (paramagnetic salts) to reach limiting temperatures of 2 to 30 millikelvins with $T_A \sim 1$ K and H lying in the range 10–60 kilooersteds or 0.8–4.8 MA/m (magnetic inductions of 10–60 kilogauss or 1–6 teslas). The lower the final temperature T_C, the more difficult it is to maintain the sample in that very low-temperature region, given a fixed background heat influx. The "warm-up energy" (Fig. 1) will be obviously very small if the steep fall in S occurs at a very low temperature. Compromises may be effected by working with different salts or, if feasible, carrying out the adiabatic demagnetization to a nonzero final value of H, which results in a higher final temperature but an enhanced heat capacity.

$$\mu H/T_A = U/T_C = \mu b/T_C$$

$$T_C/T_A = U/\mu H = b/H$$

Nuclear adiabatic demagnetization. Nuclear magnetic moments are one or two thousand times smaller than their ionic (that is, electronic) counterparts, and the characteristic temperature U/k of their mutual interaction lies in the microkelvin rather than millikelvin region. Successful experiments in nuclear adiabatic demagnetization date from the mid-1950s. First, as will be readily deduced from the above discussion, the total thermal energy associated with a loss (or restoration) of a fixed amount of entropy at microkelvin temperatures is very small; hence special precautions for thermal isolation are required. It has, in fact,proved possible to routinely reduce heat leaks from the 0.1 erg s^{-1} (10^{-8} J · s^{-1}) level to 0.1 erg min^{-1} (1.7×10^{-10} J · s^{-1}) in a typical ensemble. Second, as the entropy leverage in the basic process is a function of $\mu H/T_A$, very intense magnetic fields or very low starting temperatures (or a combination of both) must be utilized to compensate for the thousandfold diminution in μ. As there are severe practical limits to substantially increasing H—a maximum induction of 15 teslas, say, using modern superconducting-solenoid technology—salvation is mainly sought in reducing T_A. A factor of 100 may be gained here, that is, from 1 K down to 0.01 K; the latter may be reached by a conventional electronic stage in a two-stage process (**Fig. 2**) or by other modern techniques, such as Pomeranchuk cooling or ^{3}He-^{4}He dilution refrigeration.

In moving from electronic to nuclear magnetic cooling, many difficulties and subtle features of design must be dealt with. Thermal conductivities, heat transfer coefficients, intersystem relaxation times,

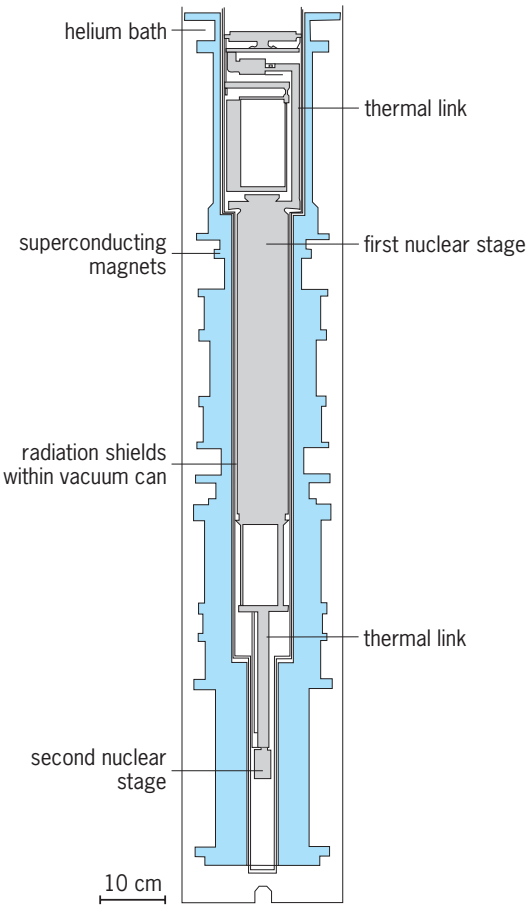


Fig. 2. Low-temperature part of two-stage nuclear refrigerator with superconducting magnets. Demagnetization of the second stage from an initial magnetic field of 9 teslas gave measured electronic temperatures of 10–12 μK. *(After F. Pobell, Matter and Methods at Low Temperatures, Springer, 1992)*

and irreversibilities must be understood and taken into account. (It may, for example, require several hours, rather than minutes, to remove the initial heat of magnetization.) Refinements have also entered through the utilization of the very intense, naturally occurring magnetic fields which make it possible to substitute more subtle and more powerful techniques for the straightforward "brute force" technique discussed so far. Through the nuclear hyperfine interaction, the atomic nuclear moments may be subject to enormous magnetic fields arising from the same (or neighboring) atoms' electronic moments. In general, these hyperfine fields are constant and thus cannot be used for magnetic cooling. In certain systems, however, the electronic magnetic moment is suppressed, or quenched, by crystal electric field effects, and may be partially restored by an external magnetic field, as a result of quantum-mechanical "mixing" of excited states and the ground state of the atom. This induced hyperfine field may correspond to an enhancement of the external magnetic field by a factor of about 10 in praseodymium compounds and about 100 in thulium compounds, for example. The fact that these substances are metals also has practical advantages in energy transmission at very low temperatures. *See* ABSOLUTE ZERO; CRYOGENICS; HYPERFINE STRUCTURE; LOW-TEMPERATURE PHYSICS.

Ralph P. Hudson

Bibliography. R. P. Hudson, *Principles and Application of Magnetic Cooling,* 1972; O. V. Lounasmaa, *Experimental Principles and Methods below 1K,* 1974; F. Pobell, *Matter and Methods at Low Temperatures*, 2d ed., 1996.

Adiabatic process

A thermodynamic process in which the system undergoing the change exchanges no heat with its surroundings. An increase in entropy or degree of disorder occurs during an irreversible adiabatic process. However, reversible adiabatic processes are isentropic; that is, they take place with no change in entropy. In an adiabatic process, compression always results in warming, and expansion always results in cooling. *See* ENTROPY; ISENTROPIC PROCESS.

By the first law of thermodynamics, the change of the system's internal energy U in any process is equal to the sum of the heat Q gained and the work W done *on* the system, as expressed in Eq. (1).

$$\Delta U = Q + W \tag{1}$$

For an adiabatic process, which involves no heat flow to or from the system, the change in internal energy when the system goes from state 1 to state 2 is equal to the external work performed on the system (which brings about the change). In Eq. (2), if U_2 is less than U_1, then W is negative and $-W$ is the work done *by* the system.

$$U_2 - U_1 = W \tag{2}$$

From the ideal gas equation of state, it can be shown that Eq. (3) is followed during an adiabatic

$$P_1 V_1^{\gamma} = P_2 V_2^{\gamma} = \text{constant} \tag{3}$$

isentropic process. Here, γ is defined in terms of the isobaric and isochoric heat capacities, C_P and C_V, respectively, by Eq. (4), where R is the gas constant

$$\gamma = \frac{C_P}{C_V} = \frac{C_V + nR}{C_V} = 1 + \frac{nR}{C_V} \tag{4}$$

and n is the number of moles in the system. *See* GAS; GAS CONSTANT; HEAT CAPACITY; ISOBARIC PROCESS.

During an adiabatic process, temperature changes are due to internal system fluctuations. For example, the events inside an engine cylinder are nearly adiabatic because the wide fluctuations in temperature take place rapidly, compared to the speed with which the cylinder surfaces can conduct heat. Similarly, fluid flow through a nozzle may be so rapid that negligible exchange of heat between fluid and nozzle takes place. The compressions and rarefactions of a sound wave are rapid enough to be considered adiabatic. *See* NOZZLE; SOUND; THERMODYNAMIC PROCESSES.

Philip E. Bloomfield

Bibliography. D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 5th ed., 1996; J. R. Howell and R. O. Buckius, *Fundamentals of Engineering Thermodynamics*, 2d ed., 1992; G. J. Van Wylen, R. E. Sonntag, and C. Borgnakke, *Fundamentals of Classical Thermodynamics*, 4th ed., 1994; M. W. Zemansky, M. M. Abbott, and H. C. Van Ness, *Basic Engineering Thermodynamics*, 2d ed., 1975.

Adipose tissue

A type of connective tissue that is specialized for the storage of neutral fats (lipids). Adipose cells originate mainly from fibroblasts. In the mammalian embryo, development of adipose cells starts at an early stage. The two types of adipose cells have names reflecting their gross physical appearance: white fat, which can be yellowish if the animal's diet is rich in carotenoids (as found in tomatoes), and brown fat, containing vascularization and respiratory pigments.

White fat. White fat is the more common type of adipose cell. These cells are found in a wide variety of locations in the mammalian body, and their function varies from location to location. For example, they may act to store food reserves and to provide thermal and physical insulation. Locations of stored food reserves include abdominal areas (such as the greater omentum) and around the kidneys and other internal organs. Because lipids are poor conductors of heat, layers of fat beneath the skin serve as heat barriers. This is especially significant in aquatic mammals, such as seals and whales, residing in polar waters. Adipose tissue in the palms of the hands, soles of the feet, and around the eyes acts as a physical cushion.

The number of white adipose cells and the amount of fat in a cell are regulated by various factors, including genetics, hormones, diet, innervation, and physical activity. Studies of identical twins revealed similar fat location but sometimes different amounts of fat tissue. The sympathetic nervous system and norepinephrine are involved in fat mobilization, especially during times of reduced caloric intake or in cold environments. Some hypothalamic neurosecretions play a role in increased body fat. Many animals, especially migratory and hibernating mammals, greatly increase their fat reserves in preparation for travel or for hibernation.

A single white fat cell (an adipocyte; **Fig. 1**) contains a large, single lipid droplet, not bound by membranes, occupying 90% of the cell's volume. The nucleus and other organelles are peripheral at one side of the droplet, with only a thin ring around the remainder of the droplet. While fairly constant in size, the droplet is constantly undergoing a turnover in lipid content. Normal histologic tissue preparation removes the lipid, leaving a clear, open area with the nucleus on one side. Special techniques, such as osmium tetroxide fixation (making fat droplets appear black) or cryostatic sections (frozen sections), are needed to observe lipids intact.

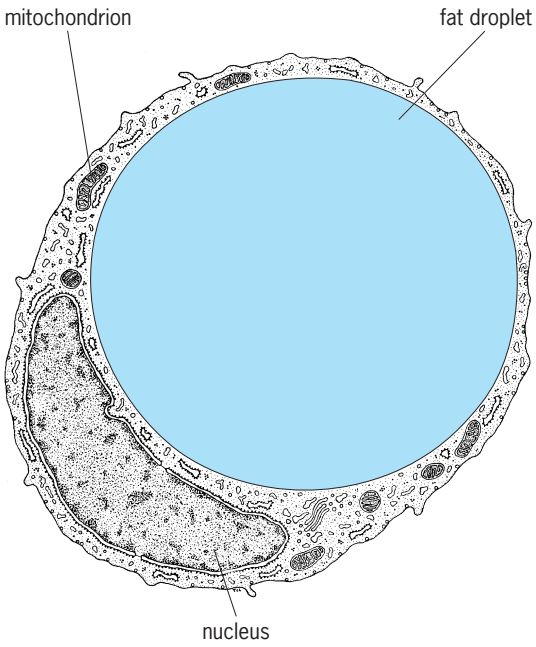


Fig. 1. White adipose cell with a large single fat droplet. (*After T. L. Lentz, Cell Fine Structure: An Atlas of Drawings of Whole-Cell Structure, W. B. Saunders, 1971*)

Brown fat. Brown adipose tissue is more limited in anatomical distribution. It is mainly found in subscapular, interscapular, and mediastinal areas. Brown adipose cells differ from white cells in that they have a centrally located nucleus, many small fat droplets, and abundant mitochondria (**Fig. 2**). These

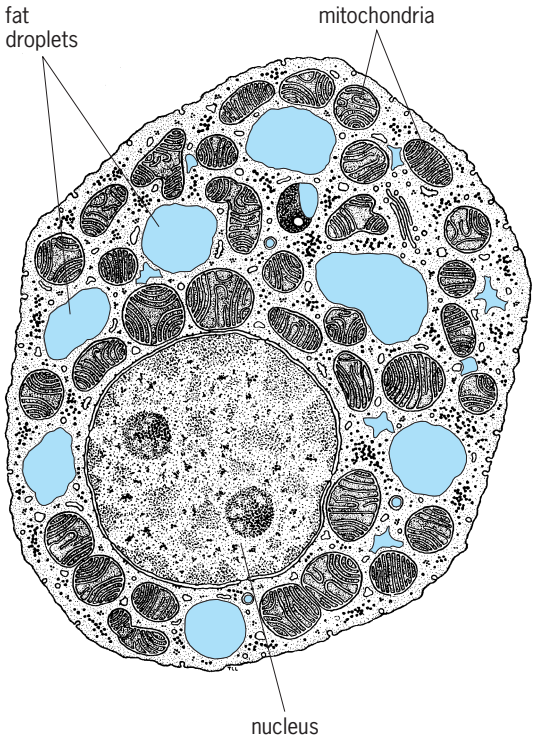


Fig. 2. Brown adipose cell with many small lipid droplets. (*After T. L. Lentz, Cell Fine Structure: An Atlas of Drawings of Whole-Cell Structure, W. B. Saunders, 1971*)

structural differences reflect the thermogenic (heat production) function of these cells, mainly associated with hibernating and newborn mammals. Brown adipose tissue is also highly vascular, aiding in distribution of heat and innervated by the sympathetic nervous system. Brown fat in nonhibernating animals gradually disappears with age and increased size. *See* CONNECTIVE TISSUE; LIPID. Todd Georgi

Bibliography. G. A. Bray, Progress in understanding the genetics of obesity, *J. Nutr.*, 127(5):940S–942S, 1997; R. E. Frisch (ed.), *Adipose Tissue and Reproduction*, 1990; S. Klaus, Functional differentiation of white and brown adipocytes, *Bioessays*, 19(3):215–219, 1997; M. H. Ross, L. J. Romrell, and G. Kaye, *A Text and Atlas*, 1995.

Admittance

The ratio of the current to the voltage in an alternating-current circuit.

In terms of complex current I and voltage V, the admittance of a circuit is given by Eq. (1), and is related to the impedance of the circuit Z by Eq. (2). Y is a complex number given by Eq. (3). G, the real part

$$Y = \frac{I}{V} \tag{1}$$

$$Y = \frac{1}{Z} \tag{2}$$

$$Y = G + jB \tag{3}$$

of the admittance, is the conductance of the circuit, and B, the imaginary part of the admittance, is the susceptance of the circuit. The units of admittance are called siemens or mhos (reciprocal ohms). *See* CONDUCTANCE; SUSCEPTANCE.

The modulus of the admittance is given by Eq. (4),

$$|Y| = \sqrt{G^2 + B^2} \tag{4}$$

which is the ratio of the maximum current to the maximum voltage. The phase angle ψ of the admittance is given by Eq. (5) and is the angle by which

$$\psi = \tan^{-1} \frac{B}{G} \tag{5}$$

the current leads the voltage. The power factor of the circuit is $\cos \psi$, given by Eq. (6).

$$\cos \psi = \frac{G}{\sqrt{G^2 + B^2}} \tag{6}$$

From Eq. (2) it follows that Eqs. (7), (8), and (9) are valid, where R, the resistance, and X, the reactance, are the real and imaginary parts of the circuit impedance. *See* ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRICAL RESISTANCE; REACTANCE.

$$|Y| = \frac{1}{|Z|} \tag{7}$$

$$G = \frac{R}{R^2 + X^2} \tag{8}$$

$$B = \frac{-X}{R^2 + X^2} \tag{9}$$

J. O. Scanlan

Adrenal gland

A complex endocrine organ in proximity to the kidney. Adrenal gland tissue is present in all vertebrates from cyclostomes to placental mammals. The adrenal consists of two functionally distinct tissues: steroidogenic cells of mesodermal origin and neural crest–derived catecholamine-secreting cells. While "adrenal" refers to the gland's proximity to the kidney, significant variation exists among vertebrates in its anatomic location as well as the relationship of the two endocrine tissues which make up the gland. In mammals, steroidogenic cells are separated into distinct zones that together form a cortex. This cortical tissue surrounds the catecholamine-secreting cells, constituting the medulla. In most other vertebrates, this unique anatomic cortical-medullary relationship is not present. In species of amphibians and fish, adrenal cells are found intermingling with kidney tissue, and the steroidogenic cells are often termed interrenal tissue.

Cortex and medulla. In mammals, the steroid secretory products of the cortex include glucocorticoids, such as cortisol and corticosterone, and mineralocorticoids, primarily aldosterone. Secretion of glucocorticoids is principally regulated by a protein hormone from the pituitary gland, adrenocorticotropin (ACTH). This hormone has been found to regulate steroidogenic secretion in virtually all vertebrates. In contrast, mineralocorticoid secretion is regulated by several factors, including the renin–angiotensin system and the plasma concentration of potassium. Glucocorticoids and mineralocorticoids act on target cells by binding to steroid-specific, intracellular receptors. In some mineralocorticoid-responsive tissues (such as the kidney), target cells contain an enzyme, 11β-hydroxysteroid dehydrogenase, that metabolizes any glucocorticoid entering these cells to inactive by-products. The adrenal cortex is essential for life; glucocorticoids have diverse functions and are especially important in carbohydrate, protein, and water metabolism, while mineralocorticoids are of fundamental importance in controlling sodium balance and extracellular fluid volume. *See* STEROID.

The adrenal medulla contains specialized neurons lacking axons that release catecholamine hormones into circulation. Treatment of medullary cells with potassium dichromate or chromic acid results in a yellow-brown staining pattern termed the chromaffin reaction. Consequently, medullary cells are termed chromaffin cells, although other cells in the body show similar staining. The adrenal medulla is really a ganglion (junction between pre- and postganglionic nerves) in the sympathetic nervous system. The chromaffin cells are unique in that they have lost their axons and instead are specialized for hormone secretion. Two major endocrine cells make up the

medulla, and contain and secrete primarily either epinephrine (adrenaline) or norepinephrine (noradrenaline). Small amounts of dopamine are also secreted. The medulla also includes the nerve terminals of preganglionic neurons that form synapses with the chromaffin cells. Catecholamines are released in response to a variety of stressful events and emergencies, and their hormones help the organism respond to such challenges by mobilizing necessary nutrients, increasing the metabolic rate, increasing alertness, and enhancing activity of the cardiovascular system.

Development. The adrenal gland forms from two primordia: cells of mesodermal origin which give rise to the steroid-secreting cells, and neural cells of ectodermal origin which develop into the catecholamine-secreting tissue. In higher vertebrates, mesenchymal cells originating from the coelomic cavity near the genital ridge proliferate to form a cluster of cells destined to be the adrenal cortex. Another proliferation of cells from the coelomic mesothelium occurs later in fetal life and is added to the original group; this latter group of cells eventually develops into the definitive or adult cortex, while the former proliferation forms the fetal or primary cortex, a structure which disappears in human infants within the first year of life. During the second month of human development, cells of the neural crest migrate to the region of the developing adrenal and begin to proliferate on its surface. The expanding cortical tissue encapsulates the neural cells forming the cortex and medulla. In mammals, three distinct zones form within the cortex: the outermost zona glomerulosa, the middle zona fasiculata, and the inner zona reticularis. The glomerulosa cells contain an enzyme, aldosterone synthase, which converts corticosterone to aldosterone, the principal steroid (mineralocorticoid) secreted from this zone. In addition, new cortical cells originate in the zona glomerulosa, supplying the inner two zones. These inner zones (fasiculata and reticularis) primarily secrete glucocorticoids and large amounts of sex steroid precursors. In many lower vertebrates, the two tissues form from similar primordia but migrate and associate in different ways to the extent that in some cases the two tissues develop in isolation from each other.

Comparative anatomy. While the paired adrenals in mammals have a characteristic cortical-medullary arrangement with distinct zonation present in the cortex, such distinctions are lacking in nonmammalian species. In more primitive fishes, chromaffin cells form in isolation from steroidogenic tissue. A general trend is present, however, throughout vertebrates for a closer association of chromaffin and steroidogenic tissues. Zonation in steroidogenic tissue is largely confined to mammals, although suggestions of separate cell types have been postulated in birds and in some other species.

Cyclostomes (hagfish, lamprey). Cells of adrenal nature that are presumably steroidogenic are found scattered along the walls of the cardinal veins and near the kidneys. Chromaffin cells are similarly located in clumps in these sites.

Elasmobranchs (sharks, rays). Steroidogenic cells are located in a single gland between the kidneys that is termed the interrenal gland. The gland assumes a variety of shapes in different species, ranging from rod-like to horseshoe-shaped. Chromaffin cells are found in paired clusters along the length of the kidney, not associated with steroidogenic cells.

Teleosts (bony fish). Steroidogenic cells are found concentrated in cords or clusters in the region near the posterior cardinal vein and the anterior portion of the kidneys. This structure is termed the head kidney, and additionally contains lymphoid cells. In most teleosts, chromaffin cells are found scattered in the region of the head kidney.

Amphibians. The steroidogenic cells of the adrenal are distributed in a variety of ways in this class, but in many species they are located in islets or nodules near the ventral surface of the kidneys. In some cases, this tissue forms interrenal glands on the ventrum of the kidney. This interrenal tissue is often found intermingling with renal tubules and interstitium. Chromaffin cells are found diffusely intermingling with steroidogenic cells. Another cell, called the Stilling or summer cell, is found in interrenal tissue of certain frogs, and appearance of this cell varies with season, being evident in summer and regressing in winter.

Reptiles. In most reptiles, the adrenals are present as paired, elongated glands anterior to the kidneys (suprarenal). The adrenals, which are covered by a distinct capsule, comprise a mixture of chromaffin and steroidogenic cells. In some reptilian species, chromaffin cells tend to concentrate in masses or clumps, often in the periphery of the gland. In lizards and some snakes, this chromaffin zone almost surrounds the steroidogenic cells, resulting in a gland with an anatomic arrangement opposite to that in mammals.

Birds. In the majority of birds, the adrenals are paired structures near the anterior end of the kidney. The glands consist of a mixture of chromaffin cells and steroidogenic cells, the former present in islets, strands, or larger masses within the gland; the latter as tortuous cords of cells radiating from the center of the gland outward. Evidence for rudimentary zonation of the steroidogenic cells is apparent in some species.

Mammals. The adrenals or suprarenal glands are located as paired structures near the anterior poles of the kidneys. The gland is distinctly separated into two segments, the cortex containing steroidogenic cells concentrically arranged in three zones and the inner medulla containing chromaffin tissue.

Comparative endocrinology. Hormones are secreted from the cells of both the medulla and the cortex.

Chromaffin cells. In all vertebrates, chromaffin cells secrete catecholamines into circulation. These substances are aminogenic derivatives of the amino acid tyrosine. In most species, the major catecholamine secreted is epinephrine, although significant amounts of norepinephrine are released by many animals. Some dopamine is also secreted. No phylogenetic trend is obvious to explain or predict

the ratio of epinephrine to norepinephrine secreted in a given species. Lampreys, for example, release more norepinephrine than epinephrine; the reverse is true in several teleosts and amphibians. Even in mammals the ratios vary: Cats secrete mainly norepinephrine while dogs and humans predominantly secrete epinephrine. A given species may release the two catecholamines in different ratios, depending on the nature of the stimulus. The great majority of the norepinephrine in circulation actually originates from that which is released from non-adrenal sympathetic nerve endings and leaks into the bloodstream. Epinephrine is formed from norepinephrine by the action of the enzyme phenylethanolamine-*N*-methyltransferase. This enzyme is induced in mammalian chromaffin cells by the high levels of glucocorticoids entering the tissue from drainage of blood from the cortex. Animals showing low activity of this enzyme tend to secrete more norepinephrine than epinephrine; it is possible that the physical separation of chromaffin from steroidogenic cells accounts in part for this finding. In addition to catecholamines, chromaffin cells secrete an array of other substances, including proteins such as chromogranin A and opioid peptides. *See* EPINEPHRINE.

Biologic effects of catecholamines are mediated through their binding to two receptor classes, α- and β-adrenergic receptors. Further examination of these receptors has revealed that subclasses of each type exist and likely account for the responses on different target tissues. In general, biologic responses to catecholamines include mobilization of glucose from liver and muscle, increased alertness, increased heart rate, and stimulation of metabolic rate.

Steroid hormones. In broad terms, most steroids secreted by adrenal steroidogenic cells are glucocorticoids, mineralocorticoids, or sex hormone precursors. However, these classes have been established largely on the basis of differential actions in mammals. The principal glucocorticoids are cortisol and corticosterone, while the main mineralocorticoid is aldosterone. This division of action holds for mammalian species and likely for reptiles and birds. In other vertebrates, such as fish and amphibians, steroids from the interrenal tissue do not show such specialized actions; instead, most show activities of both glucocorticoid and mineralocorticoid type. Mammals, birds, reptiles, and amphibians secrete cortisol, corticosterone, and aldosterone. The ratios of the two glucocorticoids vary across species; in general, corticosterone is the more important product in nonmammalian species. Even within mammals, a large variation exists across species, due to the relative ratio of cortisol to corticosterone from the adrenal cortex. Interestingly, the dominant steroid found in teleosts is cortisol. In cyclostomes and elasmobranchs, corticosterone, cortisol, and 11-desoxy-steroids appear to be the major steroid products. In elasmobranchs, a unique steroid, 1-α-hydroxycorticosterone, has been reported. In cyclostomes, elasmobranchs, and teleosts, 18-hydroxylation of corticosterone may not occur (no aldosterone synthase), thus accounting for the lack of aldosterone in these species. Consequently, in these species corticosterone and other glucocorticoids have dual roles in activating both glucocorticoid-type and mineralocorticoid-type responses.

Effects of adrenal-derived steroids in lower vertebrates involve a diverse array of actions, including control of distribution and availability of metabolic fuels such as glucose, and regulation of sodium and extracellular fluid volume. Most studies concerning the role of adrenal steroids in nonmammalian species have centered on their effects on salt and water transport, actions mediated by aldosterone and other mineralocorticoids in mammals. In nonmammalian vertebrates, corticosterone, cortisol, and aldosterone possess mineralocorticoid effects. Some fish can survive in both salt- and fresh-water environments, and other species of vertebrates adapt well to arid climates. Adrenal steroids play important roles in such animals to help regulate salt and water movement inside versus outside the body and between tissue compartments. Targets for these electrolyte- and water-regulating actions of steroids in nonmammalian species include the kidney, gills, muscle, skin, urinary bladder, and alimentary tract. In reptiles and birds, steroids act on the salt (nasal) glands to help regulate excretion of electrolytes. In birds, these glands are located in the orbit above the eyes. Birds that ingest salt water eliminate the excess salt by excretion through these glands; adrenal steroids are necessary for this process. Other hormones, such as prolactin, vasopressin (antidiuretic hormone), and thyroid hormones, act in concert with steroids to accomplish salt and water regulation. Other areas where adrenal steroids likely contribute to biologic processes include control of protein, fat, and carbohydrate balance; reproduction; and growth and development.

Robert J. Kemppainen

Bibliography. I. Chester-Jones, P. M. Ingleton, and J. G. Phillips, *Fundamentals of Comparative Vertebrate Endocrinology*, Plenum Press, New York, 1987; W. Hanke and W. Kloas, Comparative aspects of regulation and function of the adrenal complex in different groups of vertebrates, *Hormone Metab. Res.*, 27:389–397, 1995; D. O. Norris, *Vertebrate Endocrinology*, 2d ed., Lea & Febiger, Philadelphia, 1985; G. P. Vinson, B. Whitehouse, and J. Hinson, *The Adrenal Cortex*, Prentice Hall, Englewood Cliffs, NJ, 1992.

Adrenal gland disorders

Malfunctions of the paired adrenal glands. The adrenal glands can have congenital defects and can be damaged by infections and destructive tumors. Anencephaly, that is, gross underdevelopment of the brain in fetal life, leads to hypoplasia of the adrenal cortex. The cortex is also, although rarely, invaded by acute bacterial diseases, for example, septicemia due to the colon bacillus or meningococcus, which can result in failure of the adrenal cortex followed by shock and death (Waterhouse-Friderichsen

syndrome). Chronic infection of the cortex by the tubercle bacillus or a fungal infection such as histoplasmosis can cause primary adrenal deficiency. Cancer may spread to the adrenal glands, and there are uncommon primary tumors of the adrenal that do not secrete hormones. The most important disorders of the adrenal cortex and medulla are those characterized by abnormal hormone secretion. *See* ADRENAL GLAND.

Despite their anatomic proximity, the two elements of the mammalian adrenal gland are embryologically distinct: the cortex is mesodermal, while the medulla is ectodermal in origin. They also secrete different types of hormones: the cortex secretes steroids and the medulla secretes amines. The cortex is controlled by hormones from the anterior pituitary and the kidney; the medulla is directly controlled by the nervous system. The cortex and medulla are also functionally separate, except that both contribute to the mammalian body's response to outside stimuli, especially stress. *See* PITUITARY GLAND; STEROID.

Adrenal cortex. There are a number of disorders of the adrenal cortex, including primary adrenal cortical insufficiency, secondary cortical insufficiency, adrenal cortical hyperfunction, hyperaldosteronism, secondary hyperaldosteronism, Cushing's syndrome, and virilizing syndromes.

Primary adrenal cortical insufficiency. Primary adrenal cortical insufficiency, or Addison's disease, is rare, but if it is not treated it always results in death. Onset of the disease is subtle and insidious and can mimic malingering or psychoneurosis. The disease is usually caused by autoimmune destruction of the adrenals (cause unknown), by tuberculosis, or less often by chronic fungal diseases. Symptoms include lassitude, muscular weakness, weight loss, prostration during minor illnesses, low blood pressure, and brown pigmentation of the skin. In individuals with this disorder, acute intermittent diseases, surgery, or trauma precipitate Addisonian crisis; shock is reversible only by intravenous administration of water, salt, and large doses of adrenal cortical steroids. The water and electrolyte loss associated with this condition results from deficiency of the adrenal hormone, aldosterone; lack of steroids (for example, cortisol) accounts for the skin color and the extreme susceptibility to trivial stress. Maintenance therapy with small daily doses of hydrocortisone and synthetic, salt-retaining steroids permits a normal life, including the ability to do vigorous physical exercise, to become pregnant, and to have normal longevity. *See* ALDOSTERONE; TUBERCULOSIS.

Secondary adrenal cortical insufficiency. Secondary adrenal cortical insufficiency sometimes accompanies pituitary failure due to tumors, vascular accidents, chronic infectious diseases, or granulomas. The clinical picture differs from Addison's disease: weak, listless, and intolerant of minor illness, individuals also are pale and usually show neurologic signs related to the pituitary lesion. Since the thyroid gland and gonads, like the adrenal cortex, depend on the pituitary gland, hypothyroidism and testicular or ovarian deficiency also are often present. Treatment consists of cortisol, thyroid hormone, and sex steroids. *See* THYROID GLAND.

A more common form of secondary adrenal failure, which involves a feedback mechansim, follows long-term administration of high doses of semisynthetic steroids for treatment of such diseases as rheumatoid arthritis, asthma, and acute lymphoblastic leukemia. Individuals who are treated with these steroids have the general symptoms of spontaneously occurring Cushing's syndrome; however, hormonal tests show that the adrenal gland secretes very little hydrocortisone, not too much as in the natural disease. This deficiency requires the administration of large amounts of cortisol during surgical procedures or severe acute illness.

Adrenal cortical hyperfunction. Manifestations of adrenal cortical hyperfunction, or hyperadrenocorticism, can result from excessive secretion of steroid hormones normally secreted by the adrenal cortex. The four classes of adrenal cortical hormones include mineralocorticoids, such as aldosterone, which regulate salt retention; glucocorticoids, such as cortisol, that affect carbohydrate and protein metabolism, muscle function, and blood pressure; androgens, such as dehydroepiandrosterone; and estrogens, such as estradiol. Estrogens are secreted by very rare adrenocortical tumors (usually malignant) that are most easily recognized in men by signs of feminization (breast enlargement, loss of male body hair, and impotence). *See* ANDROGENS; ESTROGEN.

Hyperaldosteronism. Excess secretion of aldosterone produces hyperaldosteronism. The classic primary form arises from a small unilateral tumor known as an adenoma of the adrenal cortex, from bilateral nodular overgrowth of the cortex, or sometimes from overfunction of the kidney's renin-angiotensin system which stimulates the cortex to secrete too much aldosterone (Bartter's syndrome). The affected individuals have high blood pressure (they make up 1% or less of the hypertensive population), muscle weakness, and decreased levels of potassium in the blood. Diagnosis is facilitated by modern methods of radiologic examination, including computerized tomography and magnetic resonance imaging, together with hormone measurements. The tumors can be removed surgically. In cases where there is no tumor, the individual can be treated with diuretics that act by physiologically opposing the action of aldosterone on the kidney.

Secondary hyperaldosteronism. Secondary hyperaldosteronism occurs in very severe hypertension and in advanced liver disease (cirrhosis), kidney disease (nephrotic syndrome), and heart failure when the body retains excess salt and water with swelling.

Cushing's syndrome. Excessive secretion of cortisol produces hypercortisolism, or Cushing's syndrome. Clinical manifestations include obesity of the trunk with thin arms and legs, round red face, thin skin, brittle bones, high blood pressure, diabetes mellitus, and, in women, virilism (masculinity). The disease must be distinguished from ordinary obesity and other virilizing disorders of the adrenal gland

or ovary. If not treated, Cushing's syndrome is fatal within five years because of complications of hypertension or infections. Hormonal testing and radiographic imaging techniques are needed to determine cause. Small unilateral adenomas of the cortex are easily removed by surgery, and the individual will be completely cured. Carcinomas of the cortex (often virilizing) are malignant; survival after surgery averages 2 years even with chemotherapy. More and more cases are being found where the cause is secretion of adrenocorticotropic hormone by a cancer (such as of the lung, pancreas, and thymus), which stimulates oversecretion of cortisol by the adrenals. In these malignant disorders, treatment of the hormonal disorder is attempted by chemotherapy with drugs that oppose the secretion or action of cortisol. *See* CANCER (MEDICINE).

The most common cause of Cushing's syndrome is overactivity of the pituitary gland, often due to a small tumor that secretes excess adrenocorticotropic hormone. The disease is difficult to treat; methods include surgical removal of the pituitary tumor, x-ray or proton-beam treatment of the pituitary region, chemotherapy, or surgical removal of both adrenals. If the excess cortisol secretion is controlled, the individual can lead a normal life, although sometimes requiring hormone replacement.

Virilizing syndromes. Virilizing syndromes may accompany benign or malignant androgen-secreting tumors of the adrenal cortex. The most common and most important virilizing conditions result from congenital adrenocortical hyperplasia, which is often recognizable at birth or soon after. The cause is deficiency of one of the several adrenocortical enzymes involved in steroid biosynthesis; normal adrenal hormones are not produced in sufficient quantities, and androgenic hormones are synthesized instead. The most common defect, 21-hydroxylase deficiency, is not rare, but it is compatible with a normal life-span and good health; however, the great oversecretion of androgens impairs fertility in both sexes, and females show obvious masculinization. Other forms are accompanied by genital malformation, high blood pressure, or severe salt loss by way of the kidney; the exact nature of the clinical syndrome depends on which adrenal enzyme is lacking. Treatment involves suppressing the pituitary-adrenal axis with cortisol and its derivatives. In special cases, salt-active hormones must also be given.

Adrenal medulla. There is no recognized hormone deficiency of the adrenal medulla in humans. Excess secretion of norepinephrine and epinephrine accompanies pheochromocytoma, a rare adrenal medullary tumor that may be single or multiple, benign or malignant. Sometimes the tumor occurs in multiple endocrine adenomatosis, type II (Sipple's syndrome), characterized clinically by pheochromocytoma, tumors of the parathyroid glands with high blood calcium, and carcinoma of the thyroid. Tumors of the medulla account for fewer than 0.1% of cases of hypertension and are difficult to detect unless the blood pressure rises intermittently, in sudden attacks accompanied by obvious symptoms of excess epinephrine. Diagnosis can be made with imaging methods and by measuring urinary catecholamines, that is, metabolic products of the medullary amines. Treatment is difficult, and consists of careful surgical removal of the tumor. *See* EPINEPHRINE.

Neuroblastoma (in children), ganglioneuroblastoma, and ganglioneuroma are malignant tumors of the adrenal medulla or sympathetic ganglia; they may secrete amines that have only a minor effect on blood pressure. The clinical importance of these tumors arises from their extreme malignancy, but some are known to regress spontaneously. *See* ENDOCRINE MECHANISMS.

Nicholas P. Christy

Bibliography. N. P. Christy (ed.), *The Human Adrenal Cortex*, 1971; H. Cushing, The basophil adenomas of the pituitary body and their clinical manifestations (pituitary basophilism), *Bull. Johns Hopkins Hosp.*, 50:137–195, 1932; L. J. DeGroot (ed.), *Endocrinology*, 3rd ed., 1994; E. M. Gold, The Cushing syndromes: Changing views of diagnosis and treatment, *Ann. Intern. Med.*, 90:829–844, 1979; J. D. Wilson and D. W. Foster (eds.), *Williams Textbook of Endocrinology*, 9th ed., 1998.

Adsorption

A process in which atoms or molecules move from a bulk phase (that is, solid, liquid, or gas) onto a solid or liquid surface. An example is purification by adsorption where impurities are filtered from liquids or gases by their adsorption onto the surface of a high-surface-area solid such as activated charcoal. Other examples include the segregation of surfactant molecules to the surface of a liquid, the bonding of reactant molecules to the solid surface of a heterogeneous catalyst, and the migration of ions to the surface of a charged electrode.

Adsorption is to be distinguished from absorption, a process in which atoms or molecules move into the bulk of a porous material, such as the absorption of water by a sponge. Sorption is a more general term that includes both adsorption and absorption. Desorption refers to the reverse of adsorption, and is a process in which molecules adsorbed on a surface are transferred back into a bulk phase. The term adsorption is most often used in the context of solid surfaces in contact with liquids and gases. Molecules that have been adsorbed onto solid surfaces are referred to generically as adsorbates, and the surface to which they are adsorbed as the substrate or adsorbent. *See* ABSORPTION.

Process. At the molecular level, adsorption is due to attractive interactions between a surface and the species being adsorbed. The magnitude of these interactions covers approximately two orders of magnitude (8–800 kilojoules/mole), similar to the range of interactions found between atoms and molecules in bulk phases. Traditionally, adsorption is classified according to the magnitude of the adsorption forces. Weak interactions ($<$40 kJ/mol) analogous to those between molecules in liquids give rise to what is called physical adsorption or physisorption.

Strong interactions (>40 kJ/mol) similar to those found between atoms within a molecule (for example, covalent bonds) give rise to chemical adsorption or chemisorption. In physisorption the adsorbed molecule remains intact, but in chemisorption the molecule can be broken into fragments on the surface, in which case the process is called dissociative chemisorption.

The extent of adsorption depends on physical parameters such as temperature, pressure, and concentration in the bulk phase, and the surface area of the adsorbent, as well as on chemical parameters such as the elemental nature of the adsorbate and the adsorbent. Low temperatures, high pressures, high surface areas, and highly reactive adsorbates or adsorbents generally favor adsorption.

A thermodynamic description of adsorption quantifies the driving force for adsorption in terms of these parameters. An important quantity that can be defined in this regard is the Gibbs free energy of the surface, that is, the surface free energy. For a one-component system, surface free energy is also called surface tension, and it corresponds to the so-called reversible work required to create a surface. Surface free energy is the two-dimensional analog of pressure, and whereas pressure is measured as force per unit area, surface free energy is given as force per unit length. Common units for surface free energy are dynes/cm or, equivalently, ergs/cm^2. Easily created surfaces such as those for liquids and layered solids (for example, graphite and mica) have low surface free energies of 0–500 ergs/cm^2. Clean surfaces of tough, refractory materials have high free energies of up to several thousand ergs per square centimeter. At constant temperature and pressure, spontaneous processes lower the free energy of a system, and adsorption generally lowers the free energy of a surface; therefore, adsorption tends to occur most readily on materials with high surface free energies.

A common way to portray the results of adsorption studies on solid surfaces is in the form of an adsorption isotherm. Such a diagram gives the amount of adsorbed material per surface area at a constant temperature as a function of pressure or concentration in the bulk phase. Adsorption often initially increases with pressure and then saturates at a value that corresponds to one layer (a monolayer) of adsorbed molecules. The adsorbed monolayer has a lower free energy than the clean surface, so the driving force for adsorption is decreased upon completion of the monolayer. At still higher pressures, however, the free energy of the molecules in the gas is high enough that there is measurable adsorption of molecules on top of the monolayer and multiple layers or multilayers can be condensed onto the surface.

An adsorption isotherm represents a chemical equilibrium in which molecules are simultaneously adsorbing to and desorbing from the surface. The isotherm shows the extent of the adsorbed layer that is the net result of these two competing processes. For many adsorption systems, particularly those involving physical adsorption, the potential energy decreases monotonically to a minimum as the adsorbate approaches the surface so that there is no energy barrier for adsorption, and equilibrium is readily achieved. Other systems involve activated adsorption, and the adsorbate encounters a potential energy barrier as it approaches the surface. This phenomenon is commonly encountered in dissociative chemisorption; the process can be depicted schematically in a one-dimensional potential diagram (**Fig. 1**). The height of the potential energy barrier in such a system is determined by the extent to which the molecular fragments bond to the surface as the bond within the molecule is being broken. *See* CHEMICAL THERMODYNAMICS.

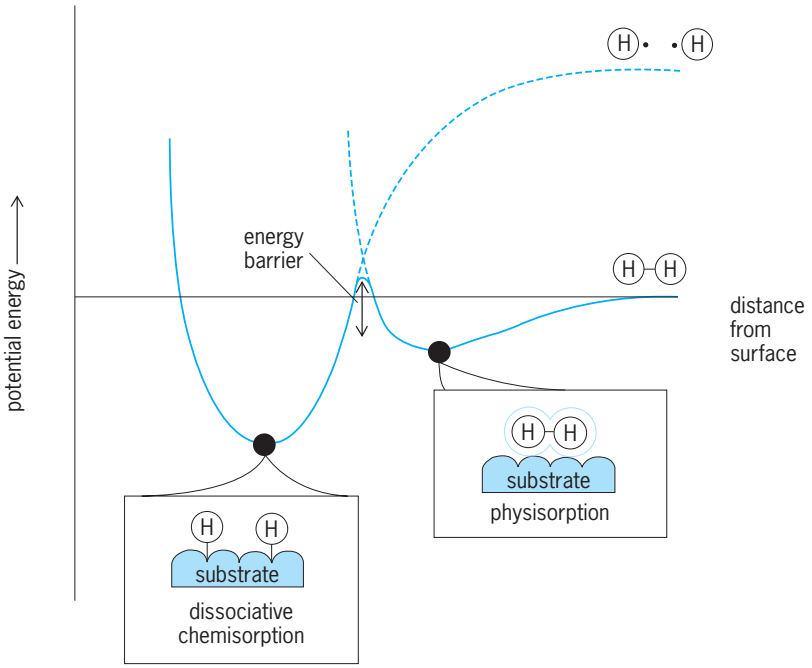


Fig. 1. One-dimensional potential energy curves typical for hydrogen molecules (H_2) and atoms (H) as they approach the surface of a metal substrate. The solid curve, which involves a crossing from the molecular curve to the atomic curve, shows the type of potential profile expected for dissociative chemisorption of H_2 on the surface.

Detection and study. Studying adsorption, particularly in the monolayer regime, is difficult because of the small numbers of molecules involved. For 1 cm^2 of surface area, the number of surface atoms or adsorbed molecules in a monolayer is of the order of 10^{15}, which is only $\sim 10^{-9}$ mole. Often, however, adsorbents have surface areas of hundreds of square centimeters per gram, and adsorption can be quantified simply by measuring the change in pressure or concentration of the bulk phase. Adsorption can also be quantified by measuring the change in mass of the adsorbent and by monitoring changes in the properties of the surface of the adsorbent such as the surface tension and the surface potential. By determining the number of molecules in an adsorbed monolayer, constructing an isotherm, and knowing the size of the adsorbed molecule, the area of the surface can be determined. Such a procedure is the basis for the Brunauer-Emmett-Teller (BET) method for determining surface areas.

Spectroscopies are commonly employed to study

directly the structure and bonding of atoms or molecules in adsorbed monolayers. Spectroscopies that measure the absorption of light have the advantage of high resolution but the disadvantage of a low absorption cross section. High-intensity light sources (lasers and synchrotrons) and high-sensitivity detectors are often required to compensate for low absorption cross sections. If, however, the adsorbent does not interact strongly with the light, optical spectra of adsorbates can be obtained by using high-surface-area adsorbents.

Higher-sensitivity detection of adsorbed monolayers can be achieved by using electrons, ions, and atoms as probes. An advantage of these techniques is that because the probing particles do not significantly penetrate the solid, the information that they provide is surface-specific. Interactions of the probing particles with gas phase and liquid molecules, however, limit these types of studies to vacuum conditions where the particles can be scattered from adsorbed monolayers without interference from the gas phase. *See* MONOMOLECULAR FILM.

The application of spectroscopies to study adsorbed monolayers has formed the basis for what is known as surface science. A wide variety of spectroscopic and diffraction techniques that utilize various combinations of electrons, photons, atoms, and ions as probes have been developed as part of surface science. These include Auger electron spectroscopy (AES), x-ray photoelectron spectroscopy (XPS), electron energy loss spectroscopy (EELS), ion scattering spectroscopy (ISS), and electron or atom diffraction. *See* AUGER EFFECT; ELECTRON DIFFRACTION; SPECTROSCOPY; X-RAY SPECTROMETRY.

Another vacuum-based technique that is frequently applied to probe the strength of surface-adsorbate bonds and to study chemical reactions in adsorbed monolayers is temperature-programmed desorption (TPD). In a temperature-programmed desorption experiment, a substrate covered with an adsorbed monolayer is heated, and the species that desorb from the surface are detected (typically with a mass spectrometer) as a function of surface temperature. The resulting temperature-programmed desorption spectrum consists of peaks that are proportional to the amount of material desorbed, and the peak temperatures are related to the strengths of the adsorbate-surface bonds—higher temperatures corresponding to stronger bonds. If a reaction occurs in the adsorbed monolayer during heating, a temperature-programmed desorption spectrum shows the products that are evolved from the surface and indicates the rate at which they are produced.

Many surface science studies are carried out under ultrahigh-vacuum conditions (pressure = 10^{-8} to 10^{-11} torr, or 10^{-3} to 10^{-6} pascal). Ultrahigh-vacuum conditions make it possible to prepare and maintain atomically clean surfaces of materials that react with air. Often in these studies, samples of single crystals are utilized, and surfaces that are atomically smooth over the range of thousands of atoms can be prepared. These well-defined surfaces can serve as model systems for understanding the molecular details of adsorption.

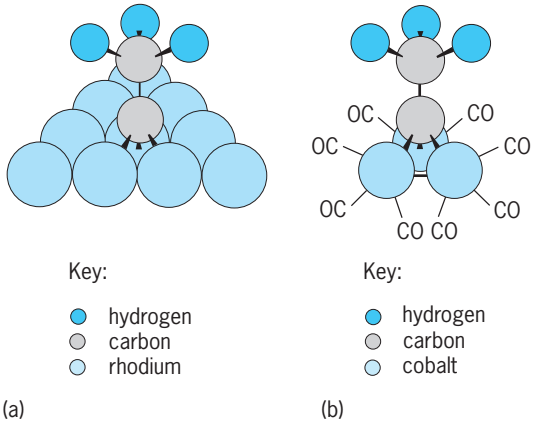


Fig. 2. Analogy between the bonding of (*a*) the ethylidyne ligand (CCH_3) to a rhodium (Rh) single crystal surface and (*b*) to a cobalt carbonyl cluster [$H_3CC(Co)_2(CO)_7$]. The bond lengths and angles, as determined by diffraction techniques, are virtually identical in the two systems.

A general finding from surface science studies of adsorption on single crystal surfaces is that, even for strongly chemisorbed layers, the adsorbate bonding to the surface is directly analogous to ligand bonding in discrete molecular compounds. An example is the ethylidyne ligand (CCH_3) bound to a rhodium surface and to a molecular cobalt complex (**Fig. 2**). *See* INTERMOLECULAR FORCES.

An exciting frontier in surface science studies of adsorption is the application of scanning tunneling microscopy (STM). This technique can provide an atomic resolution image of the electronic structure at the surface (**Fig. 3**). *See* SCANNING TUNNELING MICROSCOPE.

Applications. Direct applications of adsorption in processes such as filtration and detergent action are well known. Adsorption is also the basis for a series of vacuum pumps known as getter pumps. In these pumps, molecules are removed from the gas phase either by physisorption on high-surface-area materials at low temperatures (cryopumps and sorption pumps) or by chemisorption on highly reactive metal surfaces (ion pumps and titanium sublimation pumps).

Adsorption also plays an important role in processes such as heterogeneous catalysis, electrochemistry, adhesion, lubrication, and molecular recognition. In heterogeneous catalysis, gas or solution-phase molecules adsorb onto the catalyst surface, and reactions in the adsorbed monolayer lead to products which are desorbed from the surface. In electrochemistry, molecules adsorbed to the surface of an electrode donate or accept electrons from the electrode as part of oxidation or reduction reactions. In adhesion and lubrication, the chemical and mechanical properties of adsorbed monolayers play a role in determining how solid surfaces behave when in contact with one another. In biological systems, the adsorption of atoms and molecules onto the surface of a cell membrane is the first step in

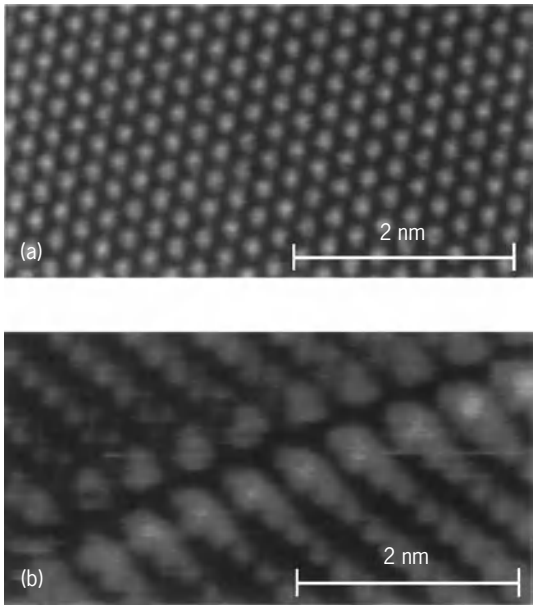


Fig. 3. Scanning tunneling micrographs of 5× 2.25 nanometer areas of (*a*) the basal plane of a graphite surface and (*b*) the same surface after being covered with a liquid drop of 1-triacontanol. The bright spots in *a* correspond to every other carbon atom on the graphite surface. The lumps along the fibrous strands in *b* can be correlated with every other CH_2 group in the 30-carbon chain of 1-triacontanol ($C_{30}H_{61}OH$), which is adsorbed as an ordered monolayer on the graphite surface. (*Courtesy of B. Venkataraman, J. Breen, and G. W. Flynn*)

molecular recognition. *See* ELECTROCHEMISTRY; HETEROGENEOUS CATALYSIS; MOLECULAR RECOGNITION.

Because of the many important technological applications of adsorption, studies of adsorption are important in disciplines ranging from solid-state physics and physical chemistry to materials science and molecular biology. *See* SURFACE AND INTERFACIAL CHEMISTRY; SURFACE PHYSICS. Brian E. Bent

Bibliography. A. W. Adamson, *Physical Chemistry of Surfaces,* 6th ed., 1997; A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications,* 1980; I. Langmuir, The adsorption of gases on plane surfaces of glass, mica, and platinum, *J. Amer. Chem. Soc.,* 40:1361, 1918; J. E. Lennard-Jones, Processes of adsorption and diffusion on solid surfaces, *Trans. Faraday Soc.,* 28:333, 1932; G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis,* 1994.

Adsorption operations

Processes for separation of gases based on the adsorption effect. When a pure gas or a gas mixture is contacted with a solid surface, some of the gas molecules are concentrated at the surface due to gas-solid attractive forces, in a phenomenon known as adsorption. The gas is called the adsorbate and the solid is called the adsorbent.

Adsorption can be either physical or chemical. Physisorption resembles the condensation of gases to liquids, and it may be mono- or multilayered on the surface. Chemisorption is characterized by the formation of a chemical bond between the adsorbate and the adsorbent.

If one component of a gas mixture is strongly adsorbed relative to the others, a surface phase rich in the strongly adsorbed species is created. This effect forms the basis of separation of gas mixtures by gas adsorption operations. Gas adsorption has become a fast-growing unit operation for the chemical and petrochemical industries, and it is being applied to solve many different kinds of gas separation and purification problems of practical importance. *See* ADSORPTION; UNIT OPERATIONS.

Adsorption equilibrium. The extent of adsorption is measured in terms of the surface excess (n^e), which is the actual amount adsorbed (n) minus the amount of gas that would be in the adsorbed phase if the solid exerted no intermolecular forces. The n^e is approximately equal to n when the gas pressure (P) is low. Both n and n^e are expressed in specific terms, that is, in number of moles of gas per unit mass of adsorbent. The relationship between n and P for a pure gas is called the adsorption isotherm. Its shape depends on the nature of the gas and the solid, as well as on the temperature.

Practical considerations require that the adsorbents have a high surface area so that a large amount of gas can be adsorbed in a small volume of the solid. This is achieved by using porous adsorbents which provide a large surface area on the internal pore walls. **Figure 1** shows an example of adsorption isotherms on such a porous solid. These gas loadings are for equilibrium conditions; the actual amounts adsorbed may be less if the contact time is too short. The equilibrium adsorption reaches a maximum with increasing pressure because the total adsorption space is limited by the pore volume of the solid. The type I adsorption isotherms shown on Fig. 1 are the most common, but other shapes are also possible.

The adsorption process is exothermic, and therefore the amount adsorbed decreases with increasing temperature at any given pressure. The heat of adsorption can be defined in various ways, the most practical of which is the isosteric heat of adsorption

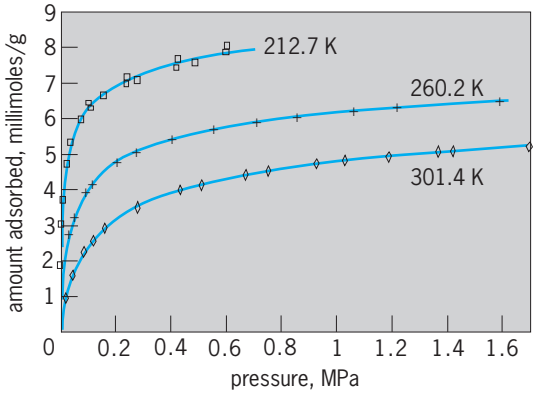


Fig. 1. Adsorption isotherms for ethylene gas on activated carbon. Points are experimental data and solid lines are Eq. (6).

(q) given by Eq. (1). The value for q can be calculated

$$q = RT^2 \left(\frac{\partial \ln P}{\partial T} \right)_n \quad (1)$$

from isotherms measured at different temperatures using Eq. (1).

The amount adsorbed of component i (n_i) from a multicomponent gas mixture of composition y_i depends on P, y_i, and T. A practical means of representing the competitive adsorption of component i with respect to component j of a mixture is in terms of the selectivity (s_{ij}) parameter [Eq. (2)].

$$s_{ij} = \frac{n_i y_j}{n_j y_i} \quad (2)$$

The selectivity, which is a function of P, y_i, and T, forms the basis of separation of a gas mixture by adsorption. Component i is selectively adsorbed over component j if s_{ij} is greater than 1. The higher the value of s_{ij}, the better the separation of the components by adsorption.

Model isotherms. The simplest model for type I isotherms of a pure gas is the Langmuir equation (3), where the value for K is derived from Eq. (4).

$$n = mKP(1 + KP)^{-1} \quad (3)$$

$$K = K_0 \exp \frac{q}{RT} \quad (4)$$

In Eq. (3), the term m is the saturation adsorption capacity and K is a temperature-dependent constant. The two-parameter (m, K) Langmuir isotherm is applicable to homogeneous adsorbents for which the heat of adsorption (q) is independent of coverage (n).

Most commercial adsorbents are, however, energetically heterogeneous, as indicated by the variation of q with n. A common approach to account for heterogeneity is to assume that the adsorbent consists of a distribution of homogeneous patches and that the overall amount adsorbed can be obtained by integration of the contributions of each patch [Eq. (5)], where n_H is the homogeneous isotherm

$$n(P) = \int n_H \lambda(q)\, dq \quad (5)$$

(such as Langmuir's) on a patch defined by the energy q, and $\lambda(q)$ is the probability density function of q on the adsorbent surface.

One such heterogeneous isotherm is that of Toth [Eq. (6)], where the constants m, b, and t are functions of temperature. The Toth equation has been successful in describing type I isotherms of pure gases as shown in Fig. 1. Other three-parameter heterogeneous isotherms have been derived by different choices of n_H and $\lambda(q)$ in Eq. (5).

$$n = mP(b + P^t)^{-1/t} \quad (6)$$

Multicomponent equilibria. Many theories are available for calculating adsorption from a multicomponent gas mixture. The simplest of these for a porous adsorbent is the Langmuir equation (3), written for mixtures [Eq. (7), where K_i is the Langmuir constant for the ith pure gas at the same temperature]. Equation (7) incorrectly predicts that the isothermal selectivity s_{ij} is constant and therefore is not very accurate for heterogeneous adsorbents. Other methods based on adsorption thermodynamics such as ideal adsorbed solution theory are used in practice.

$$n_i = mK_iPy_i(1 + \Sigma K_iPy_i)^{-1} \quad (7)$$

Adsorption kinetics. The actual physisorption process is generally very rapid. However, the measured rate of uptake of a gas by a porous adsorbent particle may be limited by its rate of transfer to the adsorption sites. Slow mass transfer may be caused by the gas-film resistance outside the particle; an anisotropic skin of binding material at the surface of the particle; and internal macro- and micropore diffusional resistances.

The rate of uptake may also be limited by the rate of dissipation of heat liberated during the sorption process. Thus, unlike the equilibrium isotherms, which can be defined without specifying the adsorbent structure or the mechanisms for mass and heat transfer, mathematical models are needed to describe the sorption kinetics. Kinetic models based upon linearized driving forces for heat and mass transfer, or Fickian diffusion of adsorbates into the pores of the solid, are well developed. *See* DIFFUSION.

Direct measurement of the intracrystalline self-diffusion of gases in zeolitic adsorbents can be achieved by the nuclear magnetic resonance pulsed-field gradient technique. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

Column dynamics. Most separations and purifications of gas mixtures are carried out in packed columns (that is, columns filled with solid adsorbent particles). The dynamics of column adsorption can be complex, and it has been the subject of numerous studies.

Adsorption. For adsorption of a single adsorbate of mole fraction y_0 from an inert carrier gas, the process consists of flowing the feed mixture through an isobaric and adiabatic column which has been previously saturated with the pure carrier gas at the temperature (T_0) and pressure (P_0) of the feed gas. Two types of behavior may be observed. (1) In type I behavior, two pairs of heat- and mass-transfer zones are formed in the column as shown in **Fig. 2**. The column ahead (section 1) of the front zones (section 2) remains saturated with carrier gas at the initial conditions. The column (section 3) between the front and rear (section 4) zones is equilibrated with a gas mixture of composition y^* ($y_0 > y^*$) at temperature T^* ($T^* > T_0$). The column (section 5) behind the rear zones is equilibrated with the feed gas at feed conditions. (2) In type II behavior, a pure heat-transfer zone is formed, followed by a pair of mass- and heat-transfer zones as shown in **Fig. 3**. The adsorbate is absent in sections 1 to 3 in this case. The column

behind the rear zones remains equilibrated with the feed gas at feed conditions.

In both cases, the zones propagate through the column after their formation as more feed gas is introduced. The front zones move faster than the rear zones. The shapes and sizes of the zones can be measured by monitoring the $y(t)$ and $T(t)$ profiles in the effluent gas from the column. The effluent profiles are known as breakthrough curves, which are mirror images of the profiles within the column. The zones may expand (proportionate pattern) or retain their shape (constant pattern) as they move through the column.

Constant-pattern front zones for type I systems and constant-pattern rear zones for type II systems are usually formed in a long column when the isotherm of the adsorbate is sufficiently concave toward the pressure axis. Approximate criteria for the formation of type I and type II systems are given in Eqs. (8), where n_0 is the adsorbate capacity at feed

$$\frac{n_0}{y_0} < \frac{c_s}{c_g} \qquad \text{(type I)} \qquad (8a)$$

$$\frac{n_0}{y_0} > \frac{c_s}{c_g} \qquad \text{(type II)} \qquad (8b)$$

conditions and c_s [$J/(g - K)$] and c_g [$J/(\text{mol} - K)$] are, respectively, the heat capacities of the adsorbent and the gas mixture. Type I behavior is common for adsorption of bulk adsorbate systems, while type II is favored when the adsorbate is dilute ($y_0 \ll 1$) and strongly adsorbed.

Multiple transfer zones and equilibrium sections are formed in multicomponent adsorption, which is also characterized by roll-over effects, where the more strongly adsorbed species displaces the weaker

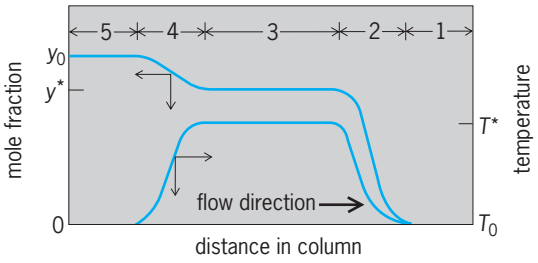


Fig. 2. Composition and temperature profiles during adsorption of type I systems in packed columns. Arrows at the curves indicate the relevant axes.

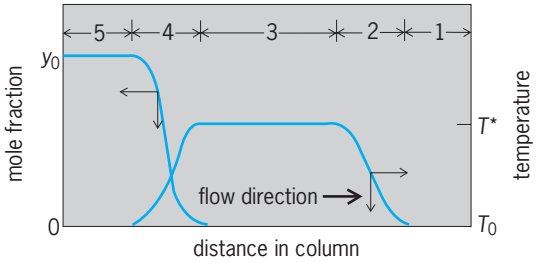


Fig. 3. Composition and temperature profiles during adsorption of type II systems in packed columns. Arrows at the curves indicate the relevant axes.

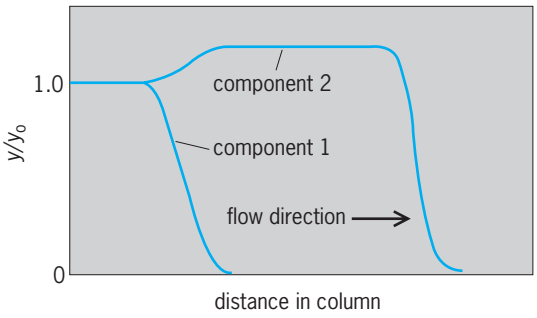


Fig. 4. Roll-over effect showing displacement of more weakly adsorbed component 2 by component 1 during isothermal adsorption from an inert carrier gas.

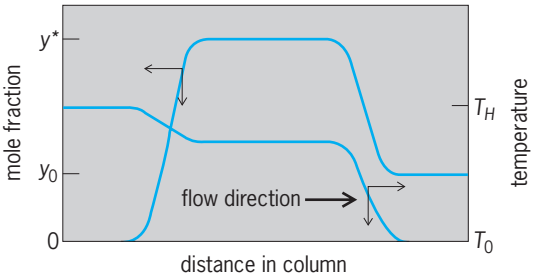


Fig. 5. Composition and temperature profiles during thermal desorption in packed columns. Arrows at the curves indicate the relevant axes.

ones as the zones propagate. **Figure 4** shows, for example, y profiles for an isothermal column adsorbing two substances from an inert gas; component 1 is more strongly adsorbed.

The general behavior of a column can be estimated by the simultaneous solution of the partial differential equations describing the mass and heat balances for each adsorbate species present in the column, in conjunction with their kinetic and equilibrium adsorption properties. Numerical solutions requiring considerable computation time are generally needed. Other factors influencing the dynamics are gas maldistribution, channeling, and nonadiabatic columns.

Desorption. Desorption of adsorbates from a column is usually accomplished by heating the column with a hot, weakly adsorbed gas; lowering the column pressure; purging the column with a weakly adsorbed gas; or combinations of these methods.

Like adsorption, well-defined transfer zones and equilibrium sections can be formed during desorption by the methods of heating and purging as shown in **Figs. 5** and **6**, respectively, for the case of a single adsorbate. Two pairs of mass- and heat-transfer zones are formed in both cases. The sections ahead of the front zones remain equilibrated at the initial column conditions. The middle equilibrium sections, however, attain temperatures (T^*) and compositions (y^*) greater or less than T_0 and y_0 depending on whether the desorption is by the heating or the purging method, respectively. The section behind the rear transfer zones is equilibrated at the conditions of the purge gas.

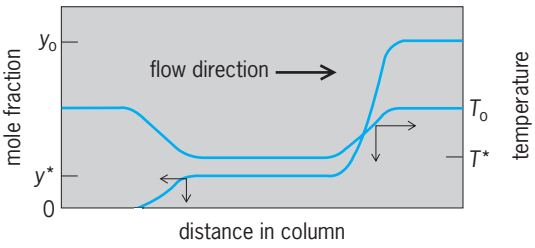


Fig. 6. Composition and temperature profiles during purge desorption in packed columns. Arrows at the curves indicate the relevant axes.

Adsorbents. Microporous adsorbents like zeolites, activated carbons, silica gels, and aluminas are commonly employed in industrial gas separations. These solids exhibit a wide spectrum of pore structures, surface polarity, and chemistry which makes them specifically selective for separation of many different gas mixtures. Separation is normally based on the equilibrium selectivity. However, zeolites and carbon molecular sieves can also separate gases based on molecular shape and size factors which influence the rate of adsorption. *See* ACTIVATED CARBON; MOLECULAR SIEVE; ZEOLITE.

Industrial applications. By far the most frequent industrial applications of gas adsorption have been the drying of gases, solvent vapor recovery, and removal of impurities or pollutants. The adsorbates in these cases are present in dilute quantities. These separations use a thermal-swing adsorption process whereby the adsorption is carried out at a near-ambient temperature followed by thermal regeneration using a portion of the cleaned gas or steam. The adsorbent is then cooled and reused. Many variations of this theme are practiced to reduce the energy consumption and the adsorbent inventory.

Pressure-swing adsorption processes are employed for separation of bulk gas mixtures. The adsorption is carried out at an elevated pressure level to give a product stream enriched in the more weakly adsorbed component. After the column is saturated with the strongly adsorbed component, it is regenerated by depressurization and purging with a portion of the product gas. The cycle is repeated after raising the pressure to the adsorption level. Numerous combinations of these cyclic steps in conjunction with internal recycling of gases have been patented. Key examples of pressure-swing adsorption processes are production of enriched oxygen and nitrogen from air; production of ultrapure hydrogen from various hydrogen-containing streams such as steam-methane reformer off-gas; and separation of normal from branched-chain paraffins. Both thermal-swing and pressure-swing adsorption processes use multiple adsorbent columns to maintain continuity, so that when one column is undergoing adsorption the others are in various stages of regeneration modes. The thermal-swing adsorption processes typically use long cycle times (hours) in contrast to the rapid cycle times (minutes) for pressure-swing adsorption processes.

A. L. Myers; S. Sircar

Bibliography. S. J. Gregg and K. S. W. Sing, *Adsorption, Surface Area and Porosity*, 1982; A. L. Myers and G. Belfort (eds.), *Fundamentals of Adsorption*, 1984; R. H. Perry and D. Green (eds.), *Perry's Chemical Engineers' Handbook*, 7th ed., 1997; D. M. Ruthven, *Principles of Adsorption and Adsorption Processes*, 1984; D. M. Young and A. D. Crowell, *Physical Adsorption of Gases*, 1962.

Aerial photography

A photograph of a portion of the Earth's surface taken from an aircraft or from a satellite. Most often, these photographs are taken sequentially and overlap each other, to give complete coverage of the area of interest. Thus they may be viewed stereoscopically to give a three-dimensional view of the Earth's surface (**Fig. 1**). Although the camera in the aircraft may be pointed obliquely to the side of the line of flight, by far the most common type of aerial photograph is taken with the camera pointed vertically downward beneath the plane. *See* STEREOSCOPY.

The entire United States has been photographed, generally more than once, and most of the rest of the world has also been covered. In the United States, the federal government is the major holder and seller of aerial photographs, although private companies also make and sell them. Federal agencies such as the Geological Survey, the Department of Agriculture, NASA, and others routinely use aerial photographs.

Aerial photographs may be made with a variety of films and filters. By far the most common is black-and-white panchromatic film used with a yellow filter. Other films are black-and-white infrared, color, and color infrared. The color and color infrared films are usually flown by companies on contract with the user and are rarely available from the government. The scales of aerial photographs are generally between 1 to 2500 and 1 to 80,000, and the majority have been taken at a scale of about 1 to 20,000. On a 1-to-20,000 photograph, 1 mi on the ground is about 3 in. on the photograph (1 km = 5 cm). As the most frequently available picture size is 9 in. by 9 in. (23 cm by 23 cm), most available federal aerial photographs cover about 9 mi^2 (23 km^2) each.

Photogrammetry. Aerial photographs have two main uses, the making of planimetric and topographic maps, and interpretation or data-gathering in a variety of specialized fields. All modern topographic (contour) maps are made from stereoscopic aerial photographs, usually black-and-white panchromatic vertical photographs. The science of making accurate measurements and maps from aerial photographs is called photogrammetry. Because vertical aerial photographs are central perspective views of the Earth (seen through a single small camera lens), and because the camera may not be precisely vertical when the picture is taken, the individual aerial photograph is not an exact map or geometric representation of the Earth's surface. The geometric discrepancy between the distribution of features seen on the photograph and their actual distribution on the

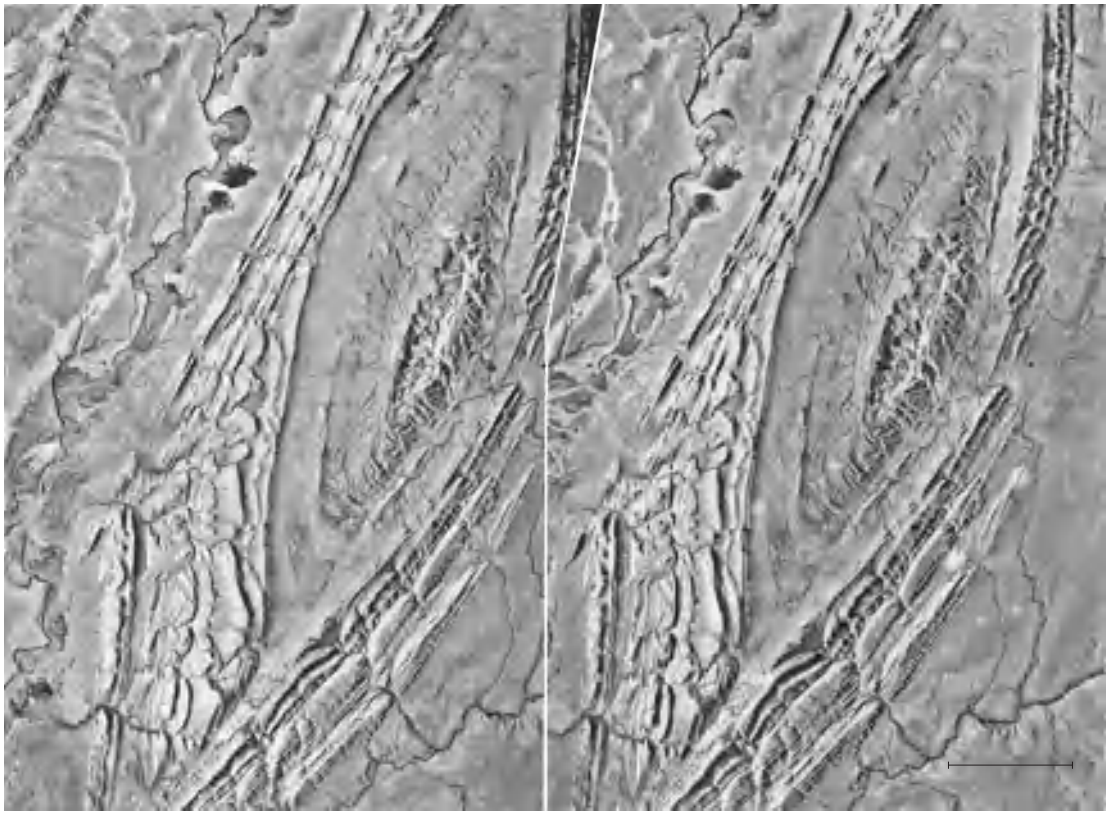


Fig. 1. Stereogram of sedimentary rocks outcropping around eroded structural dome in Wyoming. Such photographs supply much of the information needed in preparing geologic maps. Structures of this type are commonly important in localizing petroleum accumulations. To obtain a three-dimensional image, a standard lens-type stereoscope may be used. *(USGS)*

Earth's surface depends on several variables, the most important of which is the relief of the surface. The error becomes greater as relief increases. Photogrammetry deals with these errors and allows highly accurate and complete map production. Topographic maps are made from one or more stereoscopic pairs of aerial photographs by means of optical-mechanical plotting machines or mathematical methods (analytical photogrammetry). In addition to topography, many other features such as vegetation and structures made by people can be transferred from the photographs to the maps. There is not now available any method of map production which can compete in versatility, speed, and completeness with aerial photographs. *See* CARTOGRAPHY.

Photo interpretation. The other major use of aerial photographs is photo interpretation, utilizing all types of film. Photo interpretation is the attempt to extract information about the Earth's surface (and sometimes, subsurface) from aerial photographs. It is usually carried out by specialists in a subject area. Some users of photo interpretation methods are geologists, foresters, hydrologists, agricultural specialists, soil scientists, land-use planners, environmentalists, military intelligence personnel, engineers, and archeologists. The systematic study of aerial photographs, particularly when they are viewed stereoscopically, gives these specialists the ability to gather rapidly, record, map, and interpret a great deal of information. In addition, an aerial photograph is a record at a moment in time; photographs taken at intervals give an excellent record of the change with time in surface features of the Earth. In some specialties this is of vital importance. The use of different types of film, such as black-and-white infrared or color infrared, allows the interpreter to see the Earth's surface by reflected light energy beyond the range of the human eye. Thus features normally difficult to see may be made apparent.

Applicability. Some examples of the use of aerial photographs may make clear their value. Geologists use the photographs primarily for geologic mapping. It is often possible to see different types of rocks and trace their distribution. This, combined with fieldwork, enormously increases the efficiency and completeness of geologic mapping. But, in addition, the synoptic view of the Earth's surface given by the photograph enables the geologist to map fractures in the Earth's surface, subtle tonal variation in the soil, and other features difficult or impossible to see on the ground. The land-use planner obtains a synoptic view of the surface and of human activities. This, coupled with aerial photographs taken at various times, gives a spatial and temporal database vital to planning. The forester has learned to recognize various tree types and calculate tree sizes, thus enabling rapid and accurate forest inventories to be done. Importantly, various types of film allow rapid recognition of diseased trees and mapping of the

Fig. 2. Aerial view over Sinai Peninsula from *Gemini II*, looking northeast. Eastern desert of Egypt is in foreground, with Israel and Jordan in background. Note detail in which topographic and drainage features are shown. (*NASA*)

extent of the disease. Agricultural specialists can recognize and map various crops, thus making crop inventories in much less time and with much less expense than could be done purely by ground methods. Disease, drought, and other crop stresses can be recognized and mapped. Archeologists can see ancient occupation sites, travel routes, and other features on aerial photographs which are very difficult to see on the ground. Environmental studies are greatly aided by the enlargement of the aerial photograph, and environmental monitoring gains much from the study of changes seen on aerial photographs taken at different times. The use of aerial photographs for military intelligence gathering, bomb damage assessment, and other uses is well known. This short, incomplete list of the wide applicability of photo interpretation to many fields of study indicates the great value of aerial photographs. *See* AGRICULTURE; LAND-USE PLANNING; TOPOGRAPHIC SURVEYING AND MAPPING; VEGETATION AND ECOSYSTEM MAPPING.

Remote sensing. Photo interpretation depends on two things: the appearance and distribution (geometry) of familiar objects, and their spectral reflectance characteristics. The latter category is of increasing importance and is based on the fact that objects on the Earth's surface reflect different amounts of energy (light) in different parts of the electromagnetic spectrum. Thus by careful selection of films and filters with different light sensitivity in different wavelengths of the spectrum, individual "target" objects may be emphasized. This technique is of considerable importance in the interpretation of satellite imagery, such as that transmitted from *Landsat*. In this case the original output is not a photograph but is digital and may be treated mathematically before being made into a photographic image. The field of remote sensing is much concerned with these techniques. Aerial photographs, as well as satellite imagery and airborne radar imagery, are remote-sensed images which, taken as a group, provide a view of the Earth's surface that was never before available (**Fig. 2**). *See* REMOTE SENSING. L. H. Lattman

Bibliography. J. Ciciarelli, *A Practical Guide to Aerial Photography*, 1991; R. Graham, *Digital Aerial Survey: Theory and Practice*, 2002; H. Lloyd, *Aerial Photography: Professional Techniques and Applications*, 1990; E. M. Mikhail et al., *Introduction to Modern Photogrammetry*, 2001; D. P. Paine and J. D. Kiser, *Aerial Photography and Image*, 2003.

Aerodynamic force

The force exerted on a body whenever there is a relative velocity between the body and the air. There are only two basic sources of aerodynamic force: the pressure distribution and the frictional shear stress distribution exerted by the airflow on the body surface. The pressure exerted by the air at a point on the surface acts perpendicular to the surface at that point; and the shear stress, which is due to the frictional action of the air rubbing against the surface, acts tangentially to the surface at that point. The pressure and shear stress act at each point on the body surface, and hence the distribution of pressure and shear stress represent a distributed load over the surface. The net aerodynamic force on the body is due to the net imbalance between these distributed loads as they are summed (integrated) over the entire surface. *See* BOUNDARY LAYER FLOW; FLUID FLOW; WIND STRESS.

Lift and drag definitions. For purposes of discussion, it is convenient to consider the aerodynamic force on an airfoil (**Fig. 1**). The net resultant aerodynamic force *R* acting through the center of pressure

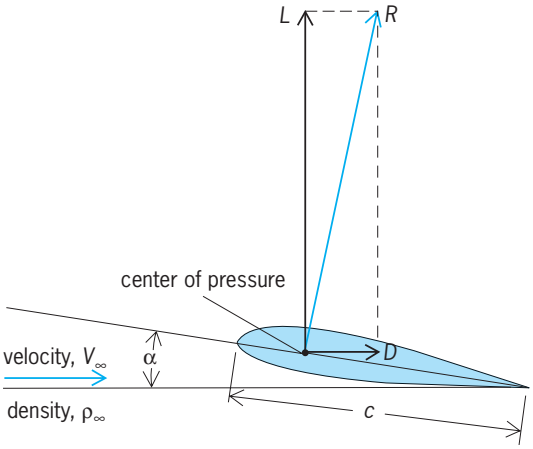


Fig. 1. Resultant aerodynamic force (*R*), and its resolution into lift (*L*) and drag (*D*) components.

on the airfoil represents mechanically the same effect as that due to the actual pressure and shear stress loads distributed over the body surface. The velocity of the airflow V_∞ is called the free-stream velocity or the free-stream relative wind. By definition, the component of R perpendicular to the relative wind is the lift, L, and the component of R parallel to the relative wind is the drag D. *See* AIRFOIL.

Force coefficients. The orientation of the body with respect to the direction of the free stream is given by the angle of attack, α. The magnitude of the aerodynamic force R is governed by the density ρ_∞ and velocity of the free stream, the size of the body, and the angle of attack. An index of the size of the body is given by some characteristic area S of the body. If the body is a winglike shape, a suitable choice for S is the planform area (the area seen when the wing is viewed directly from above). If the body is shaped like a projectile, a suitable choice for S is the maximum cross-sectional area of the body. Since S is used simply as an index of the size of the body, the area chosen is somewhat arbitrary.

The aerodynamic force on a given body shape moving through the air at speeds considerably less than the speed of sound is directly proportional to the density, the area, and the square of the velocity; hence, the same proportionality holds for lift and drag. The proportionality constants are denoted by $C_L/2$ and $C_D/2$, respectively, so that the lift and drag are given by Eqs. (1) and (2). Here, C_L is defined as

$$L = {}^1/_2\, \rho_\infty V_\infty^2 S C_L \tag{1}$$

$$D = {}^1/_2\, \rho_\infty V_\infty^2 S C_D \tag{2}$$

the lift coefficient and C_D as the drag coefficient.

These coefficients vary with angle of attack. At higher speeds, especially near and above the speed of sound, C_L and C_D also become functions of the free-stream Mach number M_∞, defined as $M_\infty = V_\infty/a_\infty$, where a_∞ is the speed of sound in the free stream. Also, in flows where friction is an important aspect of the aerodynamic force, C_L and C_D become functions of the Reynolds number, defined as $\mathrm{Re} = \rho_\infty V_\infty c/\mu_\infty$, where c is some characteristic length of the body (such as the chord length c; Fig. 1), and μ_∞ is the coefficient of viscosity. Therefore, in general, the functional dependence of the lift and drag coefficients is given by Eqs. (3) and (4).

$$C_L = f_1(\alpha, M_\infty, \mathrm{Re}) \tag{3}$$

$$C_D = f_2(\alpha, M_\infty, \mathrm{Re}) \tag{4}$$

See MACH NUMBER; REYNOLDS NUMBER.

Separation and stall. The lift varies linearly with angle of attack, up to a point where the curve that graphs the lift as a function of angle of attack bends over (**Fig. 2**). Beyond this point the lift decreases with increasing α. In this region, the wing is said to be stalled. On the linear portion of the lift curve, the flow over the airfoil is smooth and attached (Fig. 2). In contrast, in the stall region the flow has separated from the top surface of the wing, creating a type of slowly recirculating dead-air region, which decreases the lift and substantially increases the drag. An important measure of aerodynamic efficiency is the ratio of lift to drag, L/D. The higher the value of L/D, the more efficient is the lifting action of the body. The value of L/D reaches a maximum, denoted by $(L/D)_{\max}$ (see **table**), at a relatively low angle of attack (Fig. 2). *See* FLIGHT CHARACTERISTICS.

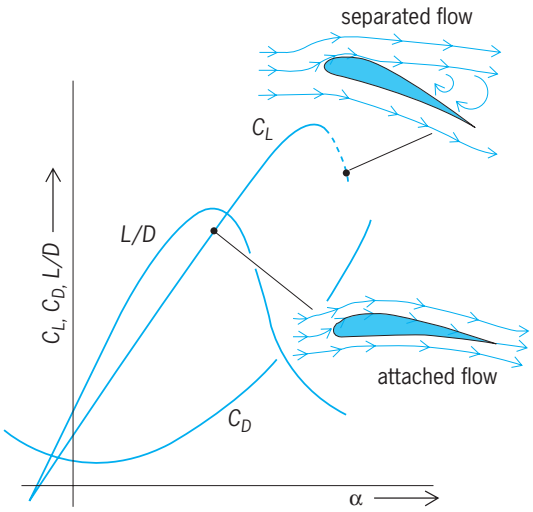


Fig. 2. Typical variation of lift coefficient (C_L), drag coefficient (C_D), and lift-to-drag ratio (L/D) for an airfoil as functions of angle of attack (α). Insets show physical nature of the flow field over the wing at the indicated points on the lift curve.

Maximum lift-to-drag ratios of representative aircraft

Aircraft	$(L/D)_{\max}$
Wright flyer	5.7
Sopwith Camel	7.7
Ryan NYP (*Spirit of St. Louis*)	10.1
Douglas DC-3	14.7
Boeing B-52	21.5
F-4 fighter	8.6

Drag components. The total aerodynamic drag on a body is often dissected into contributing parts.

Skin friction drag. This is due to the net integrated effect of the surface shear stress exerted over the body.

Pressure drag due to flow separation. When the flow separates from the body, the change in the surface pressure distribution in the dead-air separated region causes an increase in drag. Although the aerodynamic mechanism that causes the flow to separate from the surface involves the effect of friction in the flow, this drag is purely a pressure effect. This drag is sometimes called form drag.

Parasite and profile drag. The sum of skin-friction drag and pressure drag due to flow separation is denoted for a wing as profile drag, and for a complete airplane as parasite drag.

Induced drag. This is associated with the vortices that are produced at the tips of wings and trail downstream from the wings (**Fig. 3**). A wing produces lift because the pressure on its bottom surface is

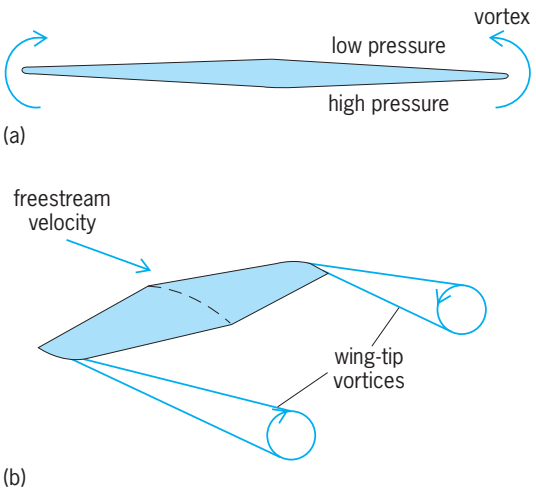


Fig. 3. Origin of wing-tip vortices on a finite wing as shown in (*a*) front view and (*b*) oblique view of wing.

higher than that on the top. At a wing tip, the higher-pressure air curls around the tip from bottom to top, setting up a circulatory motion which is superimposed on the main flow velocity and creates vortices which trail downstream from each tip. In turn, the presence of these vortices changes the pressure distribution over the wing surface in comparison to that which would exist if the wing had no tips (that is, if the wing were infinite in span). This change in the pressure distribution is in such a direction that it results in an increase in drag. This increment in drag is called induced drag; it is purely a pressure effect. Since the wing-tip vortices, which cause induced drag, in turn result from the pressure differential on the wing surfaces, and this pressure differential is also responsible for the creation of lift, induced drag is frequently called drag due to lift. *See* VORTEX.

Induced drag is the penalty paid for the production of lift. For an aircraft in steady, level flight, the lift produced must equal the weight of the aircraft. The induced drag caused by this lift must be overcome by increased thrust from the engines.

Wave drag. This is associated with the presence of shock waves on the aerodynamic vehicle. It is a pressure drag that is caused by severe increases in pressure that occur across shock waves. *See* AERODYNAMIC WAVE DRAG.

Drag coefficient expressions. In essence, the total drag on a body can be visualized as the sum of parasite drag, induced drag, and wave drag. For a wing at subsonic speeds, the total drag coefficient C_D is expressed as the sum of the profile drag coefficient $c_{d,p}$ and induced drag coefficient C_{Di}, as in Eq. (5), where the induced drag coefficient is given by Eq. (6). Here,

$$C_D = c_{d,p} + C_{Di} \tag{5}$$

$$C_{Di} = \frac{C_L^2}{\pi e \mathrm{AR}} \tag{6}$$

AR is the wing aspect ratio, defined as b^2/S, where b is the span of the wing (distance from one wing tip to the other), and e is a span effectiveness factor having to do with the shape of the planform of the wing; values of e typically range from 0.85 to 1.0. As seen in Eq. (6), induced drag increases as the square of the lift. *See* ASPECT RATIO.

For a complete airplane at subsonic speed, the total drag coefficient is the sum of the parasite drag coefficient C_{DP} and induced drag coefficient, as in Eq. (7). Since the pressure drag on an airplane depends on angle of attack, the total drag for an airplane is frequently expressed as a sum of the zero-lift drag and the drag due to lift. In coefficient form, the total airplane drag coefficient is given by Eq. (8). Here,

$$C_D = C_{DP} + \frac{C_L^2}{\pi e \mathrm{AR}} \tag{7}$$

$$C_D = C_{D,0} + \frac{C_L^2}{\pi e_1 \mathrm{AR}} \tag{8}$$

$C_{D,0}$ is the parasite drag coefficient at zero lift, and $C_L^2/\pi e_1 \mathrm{AR}$ is the drag coefficient due to lift, which is physically the sum of the induced drag coefficient plus the extra increment in parasite drag coefficient when the airplane is pitched to an angle of attack to produce lift. This extra parasite drag is reflected in the factor e_1, called Oswald's efficiency factor, whose value is typically of the order of 0.55–0.7. Equation (8) is called the drag polar for the airplane, and contains all the practical aerodynamic information needed to analyze the airplane's aerodynamic performance.

John D. Anderson, Jr.

Bibliography. J. D. Anderson, Jr., *Fundamentals of Aerodynamics*, 4th ed., 2005; J. D. Anderson, Jr., *Introduction to Flight*, 5th ed., 2005; J. J. Bertin, *Aerodynamics for Engineers*, 4th ed., 2002; S. F. Hoerner, *Fluid-Dynamic Drag*, 1965; S. F. Hoerner and H. V. Bortst, *Fluid-Dynamic Lift*, 2d ed., 1985.

Aerodynamic sound

Sound that is generated by the unsteady motion of a gas and its interaction with surrounding surfaces. Aerodynamic sound or noise may be pleasant, such as the sound generated by a flute, or unpleasant, such as the noise of an aircraft on landing or takeoff, or the impulsive noise of a helicopter in descending or forward flight. The typical human ear can detect a great range of sounds. A logarithmic scale, called the decibel (dB) scale, is used to describe the sound pressure level (SPL). The sound pressure level is given by the equation below, where p_{rms} = root-mean-square pressure level of the sound, and reference pressure $p_{\mathrm{ref}} = 2 \times 10^{-5}$ N/m^2. Zero decibels corresponds to the quietest whisper, whereas the threshold of pain occurs at 130–140 dB. *See* LOUDNESS; SOUND PRESSURE.

$$\mathrm{SPL} = 20 \log_{10} \left(\frac{p_{\mathrm{rms}}}{p_{\mathrm{ref}}} \right) \mathrm{dB}$$

Basic principles. Sources of aerodynamic sound may be classified according to their multipole order. Sources associated with unsteady mass addition to

the gas are called monopoles. These could be caused by the unsteady mass flux in a jet exhaust or the pulsation of a body. The sound power radiated by monopoles scales with the fourth power of a characteristic source velocity. Sources related to unsteady forces acting on the gas are called dipoles. The singing in telephone wires is related to the nearly periodic lift variations caused by vortex shedding from the wires. Such sources, called dipoles, generate a sound power that scales with the sixth power of the characteristic source velocity. *See* KÁRMÁN VORTEX STREET.

A turbulent fluid undergoes local compression and extension as well as shearing. These events are nearly random on the smallest scales of turbulent motion but may have more organization at the largest scales. The earliest theories of aerodynamic noise, called acoustic analogies, related these unsteady stresses in the fluid to the noise that they would generate in a uniform ambient medium with a constant speed of sound. Such sources are called quadrupoles. The sound power that they radiate scales with the eighth power of the characteristic source velocity. *See* TURBULENT FLOW.

Subsequent extensions of these theories allowed for the motion of these sources relative to a listener. This may result in a Doppler shift in frequency and a convective amplification of the sound if the sources are moving toward the listener. In addition, if the sources are embedded in a sheared flow, such as the exhaust plume of a jet engine, the sound is refracted away from the jet downstream axis. As sound propagates away from the source region, it experiences attenuation due to spherical spreading and real-gas and relaxational effects. The latter effects are usually important only for high-amplitude sound or sound propagation over large distances. *See* ATMOSPHERIC ACOUSTICS; DOPPLER EFFECT; SOUND ABSORPTION.

Aircraft noise. The main sources of aircraft noise are the jet engine and the airframe. Propeller-driven aircraft and helicopters have additional noise sources.

Engine noise. A modern commercial jet engine has a high bypass ratio, that is, the ratio of the mass flow rates through the fan and core (**Fig. 1**). An alternative design mixes the fan and core streams internally and has a single exhaust nozzle. The sound power generated by the jet exhaust plume is proportional to the eighth power of the exit velocity and the square of the exit area. The high bypass ratio reduces the amount of thrust generated by the high-speed flow and replaces it with thrust delivered by the relatively low-speed fan flow. This reduces the overall radiated noise power. For jets with a subsonic exit velocity, turbulent mixing generates stresses in the exhaust that in turn generate the radiated noise. The directivity of the jet noise peaks at a relatively small angle to the jet downstream direction. This is a result of the convective amplification and refraction effects. At high jet exit velocities, the turbulence may convect supersonically with respect to the ambient speed of sound. This results in highly directional and efficient noise radiation called Mach wave radiation. These are all examples of jet mixing noise. *See* AIRCRAFT ENGINE PERFORMANCE.

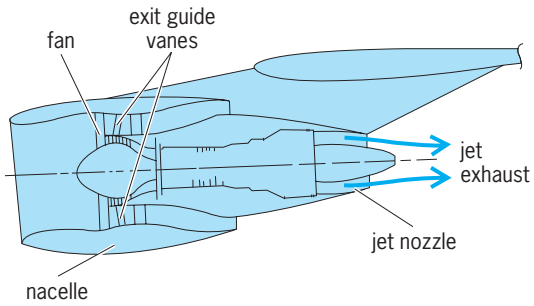


Fig. 1. Separate-flow high-bypass-ratio jet engine.

If the jet is operating off-design, that is, the jet exit static pressure is different from the ambient pressure, a shock cell structure forms in the jet exhaust. Shock cells are diamond-shaped regions of alternating high and low pressure in the jet exhaust through which the jet exit pressure adjusts to the ambient pressure. The interaction between the turbulence in the jet and the shock cell structure generates broadband shock-associated noise and screech. The former noise radiates primarily in the upstream direction, is broadband, and peaks at frequencies higher than the jet mixing noise. Screech is a feedback phenomenon in which sound generated by the turbulence interacting with the shock cells propagates upstream and triggers turbulence at the jet exit. This convects downstream, interacting again with the shock cells and completing the cycle. The feedback cycle generates an intense tone known as jet screech.

Noise is generated by the fan in the high-bypass-ratio engine due to interactions between the wakes shed by the fan blades and the exit guide vanes. This interaction generates spinning acoustic modes that may propagate in the fan duct, either upstream toward the engine inlet or downstream through the fan exhaust duct. (Spinning acoustic modes are the natural acoustic waveforms in a circular duct. They consist of pressure disturbances with wave fronts that follow helical paths.) The noise consists of a tonal component proportional to the fan's blade passage frequency and a broadband component associated with the turbulent wake interactions as well as turbulence ingested into the fan duct. The turbulent interactions cause unsteady loading on the fan blades and the exit or outlet guide vanes. The reaction forces acting on the fluid generate noise. Acoustic liners, usually made of perforated plates with a honeycomb backing, can attenuate the duct modes. In addition, the choice of the relative number of fan blades to exit guide vanes, as well as the tip speed of the fan, affects the noise generation efficiency. *See* AIRCRAFT ENGINE PERFORMANCE; JET PROPULSION; TURBOFAN; WAKE FLOW.

Airframe noise. Airframe noise is generated by the nonpropulsive components of an aircraft. With continued reduction in engine noise, airframe noise can be a significant component of the overall aircraft noise, especially during landing. The major sources

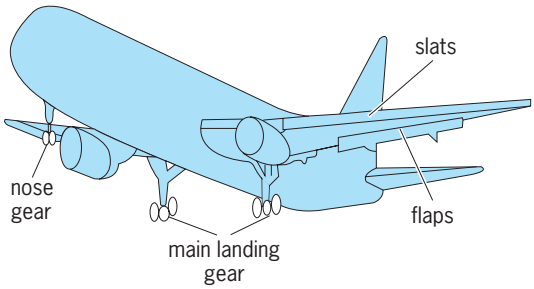


Fig. 2. Sources of airframe noise.

of airframe noise are associated with the high-lift devices, namely the trailing-edge flaps and leading-edge slats, and the landing gear—both the nose and main gear (**Fig. 2**). *See* FLIGHT CONTROLS; LANDING GEAR.

Flap systems may consist of multiple elements. Also, they may not extend over the entire wingspan. Gaps or gates may be used to limit engine exhaust impingement on the flaps. Vortices are shed from the flap side edges. The strength of the vortices depends on the loading distribution on the flap. The vortex moves over the flap side edge, inducing unsteady surface pressures on the flap, resulting in dipole noise radiation. In addition, the turbulence in the vortex generates noise that can scatter from the side edge corners or flap trailing edge. Since the loading noise scales with the sixth power of the local Mach number, and corners and edges scatter noise that scales with the fifth power, the latter mechanism can dominate at low Mach number. *See* MACH NUMBER; VORTEX.

Noise is generated at the trailing edge of the leading-edge slat. This can be tonal and is believed to be associated with vortex shedding from the slat trailing edge. This vortex shedding may be reinforced by a resonant feedback in which the radiated noise reflects from the main element of the wing. The vortex shedding depends on the slat trailing-edge thickness as well as the boundary layer thicknesses on the upper and lower surfaces of the slat.

The landing gear acts as a bluff body and generates noise associated with unsteady loading in a similar manner to the vortex-shedding noise of a circular cylinder. If any sharp edges are in the vicinity of the landing gear, such as a door that opens with the gear, acoustic scattering may occur, which increases the efficiency of the noise radiation. In addition, small-scale fittings, such as hydraulic lines attached to the gear, generate noise at a higher frequency, due to their smaller scale, than that generated by the wheels and the main gear components.

Propeller and rotorcraft noise. The noise generated by propellers contains a broadband and a harmonic component. The harmonic component occurs at multiples of the blade passage frequency. It consists of thickness noise and loading noise. Thickness noise is caused by the displacement of the air by the rotating propeller blade. It is a monopolelike source acting at the blade surface. Loading noise is related to the lift and drag forces acting on the blade. The source is of dipole type. For a propeller not in forward motion, the noise behind the plane of the propeller blades is greater than that ahead of the propeller. Forward flight causes convective amplification so that noise levels are greater when the aircraft is approaching a stationary observer.

As is the case for fan noise, broadband propeller noise is associated with turbulence ingestion and blade self-noise. The latter sources include the surface pressure fluctuations caused by the turbulent boundary layer on the blade and vortex shedding from the blade trailing edge. If the propeller tip speed is supersonic, sound generated at several points on the blade surface can reach a listener at the same time. The reason is that the source region is moving faster relative to the listener than the local speed of sound. The listener hears an intense, crackling sound during a short fraction of the propeller's rotation period. *See* PROPELLER (AIRCRAFT).

The same harmonic and broadband sources arise in helicopter noise generation. The noise levels for a helicopter in hover are greater below the plane of the rotor. In addition, in descending flight the rotor blades may encounter tip vortices shed from the passage of previous blades. This gives a transient loading to the blade and an impulsive noise radiation called blade vortex interaction (BVI) noise. The characteristic signatures of this type of noise differ, depending on whether the interaction occurs on the advancing or retreating sides of the rotor plane. In forward flight, the flow at the tip of the advancing rotor can become locally supersonic. The spatial signature of the shock that closes the supersonic region can radiate to the acoustic far field. This is referred to as high speed impulsive (HSI) noise. *See* HELICOPTER; SOUND.

Philip J. Morris

Bibliography. H. H. Hubbard, *Aeroacoustics of Flight Vehicles*, vols. 1 and 2, Acoustical Society of America, 1995.

Aerodynamic wave drag

The force retarding an airplane, especially in supersonic flight, as a consequence of the formation of shock waves. Although the physical laws governing flight at speeds in excess of the speed of sound are the same as those for subsonic flight, the nature of the flow about an airplane and, as a consequence, the various aerodynamic forces and moments acting on the vehicle at these higher speeds differ substantially from those at subsonic speeds. Basically, these variations result from the fact that at supersonic speeds the airplane moves faster than the disturbances of the air produced by the passage of the airplane. These disturbances are propagated at roughly the speed of sound and, as a result, primarily influence only a region behind the vehicle.

Causes of wave drag. The primary effect of the change in the nature of the flow at supersonic speeds is a marked increase in the drag, resulting from the formation of shock waves about the configuration. These strong disturbances, which may extend for many miles from the airplane, cause significant

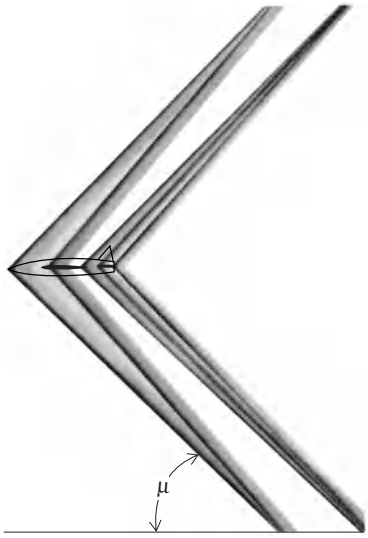


Fig. 1. Shock waves about an airplane at supersonic speeds. μ = angle of obliqueness.

energy losses in the air, the energy being drawn from the airplane. At supersonic flight speeds these waves are swept back obliquely, the angle of obliqueness decreasing with speed (**Fig. 1**). For the major parts of the shock waves from a well-designed airplane, the angle of obliqueness is equal to $\sin^{-1}(1/M)$, where M is the Mach number, the ratio of the flight velocity to the speed of sound. *See* SHOCK WAVE; SUPERSONIC FLIGHT.

The shock waves are associated with outward diversions of the airflow by the various elements of the airplane. This diversion is caused by the leading and trailing edges of the wing and control surfaces, the nose and aft end of the fuselage, and other parts of the vehicle. Major proportions of these effects also result from the wing incidence required to provide lift.

Magnitude of wave drag. Usually the aerodynamic forces acting on a vehicle, such as drag, are considered as coefficients. The wave-drag coefficient for a vehicle depends on many parameters, including the thickness-to-chord ratios and shapes of the airfoil sections of wings and fins, the planform shapes of such surfaces, the length-to-diameter ratio of the body, and the shape of the engine housing. It also depends on the Mach number. For an unswept surface with a thin airfoil whose shape is designed for wave-drag reduction (as discussed below), the coefficient of wave drag is given approximately by Eq. (1), where t/c is

$$C_d = \frac{4}{\sqrt{M^2 - 1}} \left[(t/c)^2 + \alpha^2 \right] \qquad (1)$$

the airfoil thickness-to-chord ratio and α is the angle of attack in radians. For surfaces with more complex airfoil shapes and planforms, the computation of wave drag is more complex.

For complete vehicles, the wave drag at zero lift is given by the area rule. For supersonic flight, the cross-sectional areas used are obtained in planes inclined at the Mach angle. At the lower supersonic speeds, the wave drag at the zero lift condition is usually more significant than the drag due to wing incidence. But when the Mach number is increased, the relative magnitude of wave drag at the zero lift condition gradually decreases, and the drag associated with wing incidence progressively increases, so at the higher supersonic speeds wave drag due to lift is usually more important than the zero lift value. For a well-designed vehicle, wave drag is usually roughly equal to the sum of the basic skin friction and the induced drag due to lift. *See* AERODYNAMIC FORCE; AIRFOIL; TRANSONIC FLIGHT.

Reduction of wave drag. The wave drag at the zero lift condition is reduced primarily by decreasing the thickness-chord ratios for the wings and control surfaces and by increasing the length-diameter ratios for the fuselage and bodies. Also, the leading edge of the wing and the nose of the fuselage are made relatively sharp (**Fig. 2**). With such changes, the severity of the diversions of the flow by these elements is reduced, with a resulting reduction of the strength of the associated shock waves. Also, the supersonic drag wave can be reduced by shaping the fuselage and arranging the components on the basis of the area rule. *See* WING; WING STRUCTURE.

The wave drag can also be reduced by sweeping the wing panels (**Fig. 3***a*). Some wings intended for supersonic flight have large amounts of leading-edge sweep and little or no trailing-edge sweep (Fig. 3*b*). Such planforms are referred to as delta or modified delta. For a simple infinite-span, constant-chord airfoil, the effective Mach number determining the aerodynamic characteristics is the component of the flight Mach number normal to the swept elements (**Fig. 4**). This Mach number is defined by

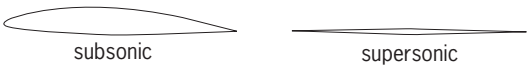


Fig. 2. Comparison of airfoil sections for subsonic flight and supersonic flight.

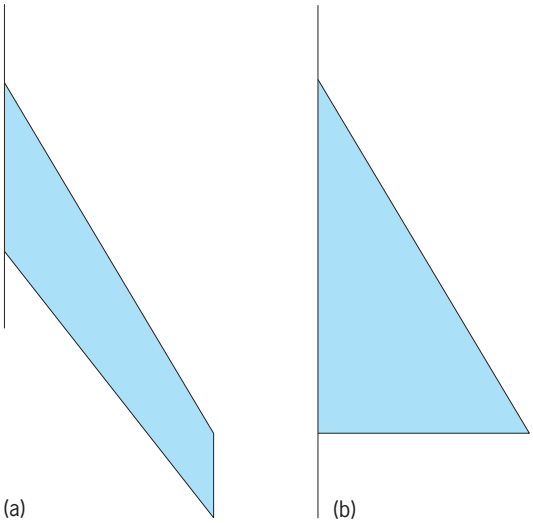


Fig. 3. Wing panels. (*a*) Sweep-back. (*b*) Delta.

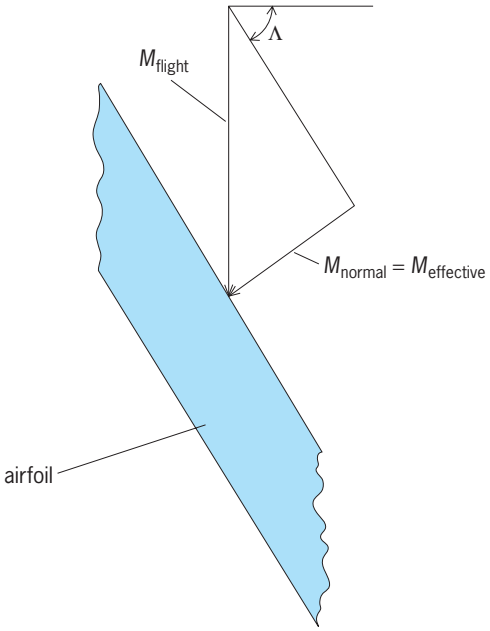


Fig. 4. Geometry that determines the effective Mach number for an infinite-span swept airfoil.

Eq. (2), where Λ is the angle of the sweep. If the

$$M_{normal} = M_{flight} \times \cos \Lambda \qquad (2)$$

sweep is such that M_{normal} is less than the subsonic Mach number at which the initial onset of wave drag occurs for the unswept airfoil, the swept airfoil will have no wave drag for the flight Mach number. For flight Mach numbers for which the normal component is greater than that for the onset of a shock on the unswept airfoil, the wave drag is not eliminated but is significantly reduced. For practical finite-span swept or delta wings, the reductions of wave drag are usually less than for the ideal infinite-span case. These reductions can be substantially improved by the proper chordwise and spanwise variations of thickness, camber, and incidence. The shape changes required are now determined using very complex fluid-dynamic relationships and supercomputers. *See* COMPUTATIONAL FLUID DYNAMICS.

When the speed is increased to supersonic values, an airplane at a given attitude and altitude experiences large increases in drag, in addition to those associated with the different nature of the flow, because of the higher dynamic pressure at these higher speeds. To offset this effect, supersonic airplanes usually fly at considerably higher altitudes than subsonic vehicles. For example, for efficient flight at Mach 2, an airplane must fly at an altitude of about 60,000 ft (18,000 m).

A major problem associated with supersonic flight, particularly at the higher supersonic speeds, is that of taking air into the engines. This air must be decelerated from the flight velocity to a relatively low speed at the compressor of the engine, without excessive energy losses. With a simple inlet, such as that used on subsonic and transonic airplanes, a strong normal shock wave forms ahead of the forward face at supersonic speeds. This shock causes severe loss of energy in the air reaching the engine, and consequent losses of engine performance. In addition, the drag of the airplane is increased. To reduce these losses, special inlets and diffusers which decelerate the airflow to the engine by a series of weak disturbances are used. *See* SUPERSONIC DIFFUSER.

Richard T. Whitcomb

Bibliography. J. D. Anderson, Jr., *Fundamentals of Aerodynamics*, 4th ed., 2005; J. J. Bertin, *Aerodynamics for Engineers*, 4th ed., 2002; A. M. Kuethe and C.-Y. Chow, *Foundations of Aerodynamics: Bases of Aerodynamic Design*, 5th ed., 1997.

Aerodynamics

The applied science that deals with the dynamics of airflow and the resulting interactions between this airflow and solid boundaries. The solid boundaries may be a body immersed in the airflow, or a duct of some shape through which the air is flowing. Although, strictly speaking, aerodynamics is concerned with the flow of air, now the term has been liberally interpreted as dealing with the flow of gases in general.

Depending on its practical objectives, aerodynamics can be subdivided into external and internal aerodynamics. External aerodynamics is concerned with the forces and moments on, and heat transfer to, bodies moving through a fluid (usually air). Examples are the generation of lift, drag, and moments on airfoils, wings, fuselages, engine nacelles, and whole airplane configurations; wind forces on buildings; the lift and drag on automobiles; and the aerodynamic heating of high-speed aerospace vehicles such as the space shuttle. Internal aerodynamics involves the study of flows moving internally through ducts. Examples are the flow properties inside wind tunnels, jet engines, rocket engines, and pipes. In short, aerodynamics is concerned with the detailed physical properties of a flow field and also with the net effect of these properties in generating an aerodynamic force on a body immersed in the flow, as well as heat transfer to the body. *See* AERODYNAMIC FORCE; AEROTHERMODYNAMICS.

Types of aerodynamic flow. Aerodynamics can also be subdivided into various categories depending on the dominant physical aspects of a given flow. In low-density flow the characteristic size of the flow field, or a body immersed in the flow, is of the order of a molecular mean free path (the average distance that a molecule moves between collisions with neighboring molecules); while in continuum flow the characteristic size is much greater than the molecular mean free path. More than 99% of all practical aerodynamic flow problems fall within the continuum category. *See* RAREFIED GAS FLOW.

Continuum flow can be subdivided into viscous flow, which is dominated by the dissipative effects of viscosity (friction), thermal conduction, and mass

diffusion; and inviscid flow, which is, by definition, a flow in which these dissipative effects are negligible. Both viscous and inviscid flows can be subdivided into incompressible flow, in which the density is constant, and compressible flow, in which the density is a variable. In low-speed gas flow, the density variation is small and can be ignored. In contrast, in a high-speed flow the density variation is keyed to temperature and pressure variations, which can be large, so the flow must be treated as compressible. *See* COMPRESSIBLE FLOW; FLUID FLOW; INCOMPRESSIBLE FLOW; VISCOSITY.

Compressible flow regimes. In turn, compressible flow is subdivided into four speed regimes: subsonic flow, transonic flow, supersonic flow, and hypersonic flow. These regimes are distinguished by the value of the Mach number, which is the ratio of the local flow velocity to the local speed of sound (**Fig. 1**). *See* MACH NUMBER.

Subsonic flow. A flow is subsonic if the Mach number is less than 1 at every point. Subsonic flows are characterized by smooth streamlines with no discontinuity in slope (Fig. 1*a*). Disturbances (caused by, say, the sudden deflection of the trailing edge of the airfoil) propagate both upstream and downstream and are felt throughout the entire flow field. The flow over light, general-aviation airplanes is subsonic. *See* SUBSONIC FLIGHT.

Transonic flow. A transonic flow is a mixed region of locally subsonic and supersonic flow. The flow far upstream of the airfoil can be subsonic (Fig. 1*b*), but as the flow moves around the airfoil surface it speeds up, and there can be pockets of locally supersonic flow over both the top and bottom surfaces of the airfoil. If the flow far upstream of the airfoil is at a low supersonic value, say with a Mach number of 1.1 (Fig. 1*c*), a bow shock wave is formed in front of the body. Behind the nearly normal portion of the shock wave, the flow is locally subsonic. This subsonic flow subsequently expands to a low supersonic value over the airfoil. *See* SHOCK WAVE; TRANSONIC FLIGHT.

Supersonic flow. In a supersonic flow, the local Mach number is greater than 1 everywhere in the flow. Supersonic flows are frequently characterized by the presence of shock waves. Across shock waves, the flow properties and the directions of streamlines change discontinuously (Fig. 1*d*), in contrast to the smooth, continuous variations in subsonic flow. If a disturbance is introduced into a steady supersonic flow at some point, the effects of this disturbance are not felt upstream. Instead, these disturbances pile up along a front just in front of the nose, creating a shock wave in the flow. *See* SUPERSONIC FLIGHT.

Hypersonic flow. This is a regime of very high supersonic speeds. A conventional rule is that any flow with a Mach number equal to or greater than 5 is hypersonic (Fig. 1*e*). Examples include the space shuttle during ascent and reentry into the atmosphere, and the flight of the X-15 experimental vehicle. The kinetic energy of many hypersonic flows is so high that, in regions where the flow velocity decreases, kinetic energy is traded for internal energy of the gas, creating high temperatures. Aerodynamic heating is a particularly severe problem for bodies immersed in a hypersonic flow. *See* ATMOSPHERIC ENTRY; HYPERSONIC FLIGHT.

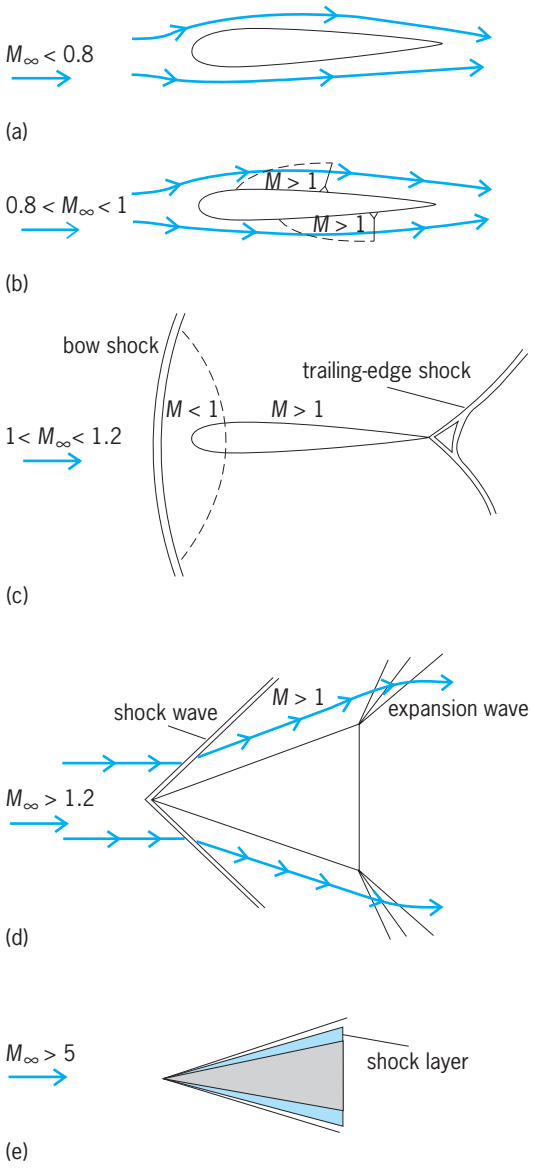


Fig. 1. Regimes for aerodynamic flows; M_∞ = Mach number for flow far upstream of the airfoil: (*a*) subsonic flow; (*b*) transonic flow with $M_\infty < 1$; (*c*) transonic flow with $M_\infty > 1$; (*d*) supersonic flow; and (*e*) hypersonic flow, where the thin, hot shock layer has viscous interaction and chemical reactions.

Nonaeronautical applications. Applications of aerodynamics cut across various disciplines. For example, automobile aerodynamics has taken on new importance in light of the incentive for energy-efficient vehicles. External aerodynamics is important in the reduction of aerodynamic drag on an automobile. With wind-tunnel testing (**Fig. 2**) and computational aerodynamic analyses, this drag has gradually been reduced by up to 70%, and similar improvements have been reported for trucks and trailers. Internal aerodynamics is important in understanding flows through engine and exhaust systems, thereby contributing to the

Fig. 2. Aerodynamic flow over an automobile. (*Ford Motor Co.*)

design of engines with improved gas mileage. In another field, environmental aerodynamics focuses on the study of airflows in heating and air-conditioning systems, and on the circulation patterns of air inside buildings and homes. Studies of the external airflow around buildings, especially under hurricane conditions, are of great interest. *See* COMPUTATIONAL FLUID DYNAMICS; GAS DYNAMICS; WIND TUNNEL.

John D. Anderson, Jr.

Bibliography. J. D. Anderson, Jr., *Fundamentals of Aerodynamics*, 4th ed., 2005; J. D. Anderson, Jr., *Introduction to Flight*, 5th ed., 2005; J. J. Bertin, *Aerodynamics for Engineers*, 4th ed., 2002; A. M. Kuethe and C. Y. Chow, *Foundations of Aerodynamics*, 5th ed., 1997.

Aeroelasticity

The branch of applied mechanics which deals with the interaction of aerodynamic, inertial, and structural forces. It is important in the design of airplanes, helicopters, missiles, suspension bridges, power lines, tall chimneys, and even stop signs. Variations on the term aeroelasticity have been coined to denote additional significant interactions. Aerothermoelasticity is concerned with effects of aerodynamic heating on aeroelastic behavior in high-speed flight. Aeroservoelasticity deals with the interaction of automatic controls and aeroelastic response and stability.

The primary concerns of aeroelasticity include flying qualities (that is, stability and control), flutter, and structural loads arising from maneuvers and atmospheric turbulence. Methods of aeroelastic analysis differ according to the time dependence of the inertial and aerodynamic forces that are involved. For the analysis of flying qualities and maneuvering loads wherein the aerodynamic loads vary relatively slowly, quasi-static methods are applicable, although autopilot interaction could require more general methods. The remaining problems are dynamic, and methods of analysis differ according to whether the time dependence is arbitrary (that is, transient or random) or simply oscillatory in the steady state.

Lift distribution and divergence. The redistribution of airloads caused by structural deformation will change the lifting effectiveness on the aerodynamic surfaces from that of a rigid vehicle. The simultaneous analysis of equilibrium and compatibility among the external airloads, the internal structural and inertial loads, and the total flow disturbance, including the disturbance resulting from structural deformation, leads to a determination of the equilibrium aeroelastic state. If the airloads tend to increase the total flow disturbance, the lift effectiveness is increased; if the airloads decrease the total flow disturbance, the effectiveness decreases. In a limiting case of increasing lift effectiveness, there is a critical speed at which the rate of change of airload with deformation equals the rate of change of structural reaction, and no statically stable equilibrium condition exists; at a higher speed the deformation will increase to a point of structural failure. This critical speed is called the divergence speed.

The effectiveness of a lifting surface influences the selection of planform size. Optimum design for strength requires design for the actual loads applied, taking due account of the redistribution of loading resulting from flexibility. The extent of changes in effectiveness and load distribution on a straight wing depends on its torsional stiffness and the chordwise distance between the aerodynamic center of the airfoil and its center of twist. Divergence can occur on a straight wing if the airfoil aerodynamic center is forward of the center of twist. Sweep has a significant influence on effectiveness and redistribution of loads to an extent also determined by the surface bending stiffness, while the effectiveness of stabilizers is also influenced by the stiffness of the fuselage. The bending of a swept-back wing has a substantial stabilizing effect on divergence to the extent that only a slight amount of sweep is required to eliminate the possibility of divergence. The redistribution of wing loading has a slight effect on the induced drag, but if the wing is swept it may have a more significant effect on static longitudinal stability. If the wing is swept-back, the redistribution of loading usually shifts the spanwise center of pressure location inboard, and correspondingly the chordwise center of pressure location moves forward. This forward movement of center of pressure reduces the static stability. The same effect occurs on swept-back stabilizing surfaces and reduces effective tail lengths. This can be an extremely critical problem on a vertical stabilizer since it provides the primary source of static directional stability, whereas it is the combination of wing and horizontal stabilizer that determines the static longitudinal stability. The bending of a wing will also introduce a change in dihedral that may affect the dynamic lateral-directional stability of the aircraft. *See* WING STRUCTURE.

On each of the attempts to attain crewed, powered flight in October and December of 1903, a wing of Samuel Langley's aerodome collapsed, and the failures may have been caused by torsional divergence. The first successful flight of a power-driven heavier-than-air machine was achieved the week following

Langley's second failure, when the Wright brothers' biplane flew at Kitty Hawk, North Carolina, on December 17. The Wright biplane experienced no aeroelastic difficulties and, in fact, made use of an aeroelastic control system in which warping of the wing gave rise to the necessary controlling aerodynamic forces. The Wright brothers had also noted a reduction in thrust due to twisting in their experimental work on propeller blades, although it had no significant effect on the blades' performance. During World War I the Fokker D-8 had just been put into service when three failures occurred in high-speed maneuvers. As a result of static strength and deflection measurements, it became apparent that the failures were due to torsional divergence; the load resulting from the air pressure in a steep dive would increase faster at the wing tips than at the middle.

Control effectiveness and reversal. The airloads induced by means of a control-surface deflection also induce an aeroelastic loading of the entire system. Equilibrium is determined as above in the analysis of load redistribution. Again, the effectiveness will differ from that of a rigid system, and may increase or decrease depending on the relationship between the net external loading and the deformation. In a situation of decreasing control-surface effectiveness, the speed at which the effectiveness vanishes is called the reversal speed. At this speed the control-surface–induced air loads are exactly offset by the resulting aeroelastic loads. At higher speeds the aeroelastic loads will exceed the control-surface loading, and the resultant load will act in the reverse direction from the control-surface loading, causing the control system to act in a direction opposite to that desired.

Control effectiveness determines the selection of control-surface size and maximum travel required. It may determine the need for all-movable horizontal or vertical stabilizers. Loss in control-surface effectiveness on a straight planform depends on the torsional stiffness of the forward primary surface, to some extent on the torsional stiffness of the control surface as well as its actuation system, and on the distance between the airfoil aerodynamic center and the aerodynamic center of the loading induced by the control surface. Sweepback, which is favorable in the case of divergence, has the unfavorable effect of decreasing control-surface effectiveness. The effectiveness of a tail or canard control surface is also influenced by fuselage bending and, in the case of a vertical stabilizer, torsion of the fuselage. Control reversal is an extremely critical problem with ailerons, and may involve severe weight penalties in providing sufficient torsional stiffness and, in the case of a swept-back wing, bending stiffness. The problem is less critical with elevators and rudders and, of course, is eliminated by the use of all-movable stabilizers for control. Unlike divergence, the phenomenon of reversal is not destructive in itself, but it can lead to structural failure when a marginally stable vehicle cannot be controlled. *See* AILERON; AIRCRAFT RUDDER; ELEVATOR (AIRCRAFT).

Aileron reversal first became a crucial problem for World War II fighters. High roll rates at high speeds were essential in combat maneuvers, and any loss in rolling ability put a fighter pilot at a serious disadvantage. The designers of the Spitfire, the FW-190, and the P-51 sought very high roll rates, and this led to shorter wingspans, aerodynamically balanced ailerons for larger deflection at high speeds, and higher torsional stiffness to increase the reversal speed well beyond the performance limits of the aircraft. The Japanese Zero, on the other hand, was designed for lightness and minimum-radius turns which resulted in a low reversal speed. The U.S. Navy pilots soon discovered this weakness and avoided circling combat but established high-speed, single-pass techniques, where their superior dive speeds and high roll rates could not be followed by the Zero operating close to its aileron reversal speed.

Flutter. A self-excited vibration is possible if a disturbance to an aeroelastic system gives rise to unsteady aerodynamic loads such that the ensuing motion can be sustained. At the flutter speed a critical phasing between the motion and the loading permits extraction of an amount of energy from the airstream equal to that dissipated by internal damping during each cycle and thereby sustains a neutrally stable periodic motion. At lower speeds any disturbance will be damped, while at higher speeds, or at least in a range of higher speeds, disturbances will be amplified. The simplest type of flutter occurs when a single motion induces an aerodynamic force having a component in the direction of the motion and in phase with the velocity. This is described as a case of negative aerodynamic damping or single-degree-of-freedom flutter. An example is a power line that "gallops" as a result of ice formation which gives the cross section of the cable something of an airfoil shape. The term classical flutter is used to denote the more complicated instability that typically arises from a critical coupling of two or more modes of motion, each of which is stable by itself.

The mechanism for classical flutter may be described in terms of the bending deflection, the angle of twist, and the disturbance in lift. If the structural characteristics (that is, stiffness and inertia) are such that an upward incremental lift, corresponding to a leading-edge upward twist, occurs when the airfoil is bending upward, and, when the motion reverses, the incremental lift acts downward as the airfoil is bending downward, energy will be extracted from the airstream. If the energy extracted in each cycle exceeds that dissipated by internal structural damping, the motion will oscillate with an increasing amplitude until a structural failure occurs.

The parameters of the flutter problem are those of the divergence problem with the addition of the dynamic inertial characteristics. The flutter speed is basically proportional to the fundamental torsion frequency so that torsional stiffness is a primary design variable. A forward center of gravity is stabilizing because it tends to drive the bending and twisting motions out of phase, which results in the incremental lift acting in the opposite direction from the bending motion. Mass balance is therefore

another primary design variable. Flutter speed is usually reduced at high altitudes because aerodynamic damping decreases faster than aerodynamic stiffness (that is, lift effectiveness) with air density. Flutter is essentially a resonance phenomenon between two motions, so that another design consideration is separation of natural vibration frequencies. The effect of sweepback is not as significant on flutter characteristics as it is with divergence, although increasing the sweepback angle does increase the flutter speed. An interesting design possibility is to suppress flutter by automatic control systems, and development of such systems has been undertaken. In order to control the aerodynamic forces, either an additional control surface or a rapidly responding actuator for an existing control surface is required, and an example configuration is a primary surface with both leading- and trailing-edge control surfaces. For practical applications the automatic control system will need multiple redundancies since a component failure could result in violent flutter.

The first known case of aerodynamic flutter occurred during World War I on the horizontal tail of a Handley-Page 400, twin-engine bomber. The flutter apparently was caused by coupling of the fuselage torsion mode and the antisymmetrical rotations of the independently actuated right and left elevators. The coupling was eliminated by connecting the elevators to a common actuating torque tube. Toward the end of the war, tail flutter difficulties were overcome by the same redesign, and the attachment of elevators to the same torque tube became a standard feature of subsequent designs. Soon after the war, static mass balancing of ailerons about their hinge lines was found to be an effective means of avoiding wing-aileron flutter. This technique has proved to be both a design and redesign procedure of permanent value, for, ever since, there have been occasional instances of coupled control-surface flutter involving both wings and tail surfaces that did not lead to destruction but could be eliminated merely by increasing the static mass balance of the control surface. *See* FLUTTER (AERONAUTICS).

Gust response. Transient meteorological conditions such as wind shears, vertical drafts, mountain waves, and clear air or storm turbulence impose significant dynamic loads on aircraft. So does buffeting during flight at high angles of attack or at transonic speeds. The response of the aircraft determines the stresses in the structure and the comfort of the occupants. Aeroelastic behavior makes a condition of dynamic overstress possible; in many instances, the amplified stresses can be substantially higher than those that would occur if the structure were much stiffer. *See* CLEAR-AIR TURBULENCE; LOADS, DYNAMIC; TRANSONIC FLIGHT.

Structural design considerations for gusts include both strength and fatigue life. Atmospheric turbulence and buffeting are random phenomena and can be described only statistically, but wind shears and mountain waves take on the appearance of discrete gusts. Since statistical data on all atmospheric conditions are limited, recourse to arbitrary design criteria is necessary. Strength has generally been determined for response to a discrete gust with a specified profile and maximum velocity normal to the flight path, while fatigue life is estimated by statistical methods based on the power spectral characteristics of the atmosphere. However, strength is frequently also calculated by the statistical methods since static strength may be regarded as the limiting case of fatigue under a single loading.

The gust loading condition is a critical one in the design of the wings of large aircraft, and the amount of sweep and the location of wing-mounted engines influence the dynamic response. The basic parameters that determine the response are the longitudinal dynamic flight characteristics of the aircraft, primarily its short-period frequency and damping, and the bending stiffness and frequency of the wing. These same factors, coupled with fuselage flexibility, also determine the aircraft ride qualities. Both gust response loads and ride qualities have been improved by use of automatic control systems.

The gustiness of the atmosphere has obviously been a concern throughout the history of flight. The first report of the National Advisory Committee for Aeronautics [now the National Aeronautics and Space Administration (NASA)], published in 1915, was an experimental and theoretical investigation of aircraft response to gusts. The first practical applications of automatic controls for gust alleviation and structural dynamic stability augmentation were made to the B-52 and B-70 bombers. Successful demonstrations of the load alleviation systems were begun with B-52 flights in late 1967 and B-70 flights in 1968. *See* AERODYNAMICS; FLIGHT CHARACTERISTICS; SUBSONIC FLIGHT.

William P. Rodden

Bibliography. R. L. Bisplinghoff and H. Ashley, *Principles of Aeroelasticity*, 2d ed., 1962, reprint 2002; E. H. Dowell (ed.), *A Modern Course in Aeroelasticity*, 4th ed., 2004; Y.-C. B. Fung, *An Introduction to the Theory of Aeroelasticity*, 1955, reprint 2002; W. P. Jones (ed.), *AGARD Manual on Aeroelasticity*, 1960–1968; D. McRuer, I. Ashkenas, and D. Graham, *Aircraft Dynamics and Automatic Control*, 1973.

Aeromonas

A bacterial genus in the family Vibrionaceae comprising oxidase-positive, facultatively anaerobic, monotrichously flagellated gram-negative rods. The mesophilic species are *A. hydrophila, A. caviae* and *A. sobria*; the psychrophilic one is *A. salmonicida*. Aeromonads are of aquatic origin and are found in surface and wastewater but not in seawater. They infect chiefly cold-blooded animals such as fishes, reptiles, and amphibians and only occasionally warm-blooded animals and humans.

Human wound infections may occur following contact with contaminated water. Septicemia has been observed mostly in patients with abnormally low white blood counts or liver disease. There is evidence of intestinal carriers. The three mesophilic species are also associated with diarrheal disease (enteritis and colitis) worldwide. A heat-labile

enterotoxin and cytotoxin(s) are produced only by *A. hydrophila* and *A. sobria*, but experimental diarrhea could not be produced in humans. *See* DIARRHEA.

A related lophotrichous genus, *Plesiomonas* (single species, *P. shigelloides*), is also known as an aquatic bacterium and is associated with diarrhea chiefly in subtropical and tropical areas. It is also found in many warm-blooded animals. Systemic disease in humans is rare. *See* MEDICAL BACTERIOLOGY.

Alexander von Graevenitz

Bibliography. E. H. Lennette et al. (eds.), *Manual of Clinical Microbiology*, 4th ed., 1985.

Aeronautical engineering

That branch of engineering concerned primarily with the special problems of flight and other modes of transportation involving a heavy reliance on aerodynamics or fluid mechanics. The main emphasis is on airplane and missile flight, but aeronautical engineers work in many related fields such as hydrofoils, which have many problems in common with aircraft wings, and with such devices as air-cushion vehicles, which make use of airflow around the base to lift the vehicle a few feet off the ground, whereupon it is propelled forward by use of propellers or gas turbines. *See* AERODYNAMICS; AIRPLANE; FLUID MECHANICS; HYDROFOIL CRAFT.

Aeronautical engineering expanded dramatically after 1940. Flight speeds increased from a few hundred miles per hour to satellite and space-vehicle velocities. The common means of propulsion changed from propellers to turboprops, turbojets, ramjets, and rockets. This change gave rise to new applications of basic science to the field and a higher reliance on theory and high-speed computers in design and testing, since it was often not feasible to proceed by experimental methods only. *See* JET PROPULSION; PROPULSION; ROCKET PROPULSION; SPACE TECHNOLOGY.

Aeronautical engineers frequently serve as system integrators of important parts of a design. For example, the control system of an aircraft involves, among other considerations, aerodynamic input from flow calculations and wind-tunnel tests; the structural design of the aircraft (since the flexibility and strength of the structure must be allowed for); the mechanical design of the control system itself; electrical components, such as servo-mechanisms; hydraulic components, such as hydraulic boosters; and interactions with other systems that affect the control of the aircraft, such as the propulsion system. The aeronautical engineer is responsible for ensuring that all of these factors operate smoothly together.

The advent of long-range ballistic missiles and spacecraft presented new challenges for aeronautical engineering. Flight velocities increased up to 35,000 ft/s (10,668 m/s), and the resulting heating problems required new kinds of materials and cooling-system concepts to protect the spacecraft or reentry vehicle. For example, materials were developed by ground testing in high-temperature electric arcs and rocket exhausts and then were tested by flights of crewless models at full speed. The resulting heat shields were successfully used to return astronauts from the Moon and to protect the warheads of ballistic missiles while maintaining missile accuracy. *See* BALLISTIC MISSILE.

The techniques that aeronautical engineers had developed for aircraft were applied to help develop such devices as the air-cushion vehicle, which is of much interest to the military for landing troops on a beach. Helicopters and vertical takeoff and landing aircraft are other developments in which aeronautical engineering has played an important role. Hydrofoils are basically boats that fly on wings under the water. Because the water is so dense, these wings are usually small compared to aircraft wings, but the same principles apply. The control of such wings is very sensitive, and careful attention must be paid to the control system used lest the high forces involved cause severe damage to the wings or to the boat itself. *See* AIR-CUSHION VEHICLE; HELICOPTER; VERTICAL TAKEOFF AND LANDING (VTOL).

Aerodynamics has always been one of the main tools of aeronautical engineering. As noted above, aerodynamics interacts closely with control design. For example, as an aircraft approaches the ground to land, the effect of the solid barrier of the ground affects the aerodynamics of the aircraft. This effect can be simulated in a computer model by placing another wing in a mirror image under the ground. The resulting flow pattern corresponds to the real wing and the solid ground. In any but the simplest cases, this type of calculation cannot be carried out except with the aid of high-speed computers, which have made possible the calculation of entire flow fields around aircraft and spacecraft as they pass through the atmosphere. Wind-tunnel testing of scale models is another important source of aerodynamic data. *See* COMPUTATIONAL FLUID DYNAMICS; WIND TUNNEL.

Aircraft and missile structural engineers have raised the technique of designing complex structures to a level never considered possible before the advent of high-speed computers. Structures can now be analyzed in great detail and the results incorporated directly into computer-aided design (CAD) programs. These programs are now so complete that in some cases no drafting work with paper and pencil is used at all and all drawings are printed directly from the design devised by the engineer working only on the computer screen. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

These examples illustrate the scope of aeronautical engineering. In general, it overlaps many other disciplines, and the aeronautical engineer may concentrate on aerodynamics, structures, control systems, propulsion systems, or the overall design of the vehicle.

John R. Sellars

Bibliography. J. E. Allen, *Aerodynamics,* 1982; W. Biddle, *Barons of the Sky: From Early Flight to Strategic Warfare,* 1991; P. A. Hanle, *Bringing Aerodynamics to America: A History of Aerodynamics Research,* 1982; F. E. Weick, *From the Ground Up: The Autobiography of an Aeronautical Engineer,* 1988.

Aeronautical meteorology

The branch of meteorology that deals with atmospheric effects on the operation of vehicles in the atmosphere, including winged aircraft, lighter-than-air devices such as dirigibles, rockets, missiles, and projectiles. The air which supports flight or is traversed on the way to outer space contains many potential hazards. Nine major hazards are considered in this article, which is principally concerned with commercial and general aviation.

Low visibility. Poor visibility caused by fog, snow, dust, and rain is a major cause of aircraft accidents and the principal cause of flight cancellations or delays. For the average private pilot with a minimum of equipment in the aircraft, an unexpected encounter with low-visibility conditions can be a serious matter, although technology is bringing aviation closer to the hitherto elusive goal of all-weather flying.

The weather conditions of ceiling and visibility required by regulations for crewed aircraft during landing or takeoff are determined by electronic and visual aids operated by the airport and installed in the aircraft. There are three categories of airports as related to the installation of various levels of specialized aids to navigation. At category 1 airports, landings and takeoffs are permitted when the ceiling is 200 ft (60 m) or more and the runway visual range (RVR) is 1800 ft (550 m) or more. The pilot operates the controls at touchdown, although the initial approach may be controlled by instruments. At category 2 airports, landings and takeoffs are authorized down to a ceiling height of about 150 ft (45 m) and runway visual range of about 500 ft (150 m). At category 3 airports, sophisticated aids to navigation, such as Global Positioning System (GPS), provide for safe, fully automated, and electronically coupled landings under "zero, zero" conditions. Airports smaller than the major city airports may have higher ceiling requirements, as may airports with tall buildings or mountains nearby. *See* AIR NAVIGATION; AIRCRAFT INSTRUMENTATION; INSTRUMENT LANDING SYSTEM (ILS).

The accurate forecasting of terminal conditions is critical to flight economy, and to safety where sophisticated landing aids are not available. Improved prediction methods are under continuing investigation and development, and are based on mesoscale and microscale meteorological analyses, electronic computer calculations, radar observations of precipitation areas, and observations of fog trends. *See* MESOMETEOROLOGY; MICROMETEOROLOGY; RADAR METEOROLOGY.

Turbulence and low-altitude wind shear. Atmospheric turbulence is principally represented in vertical currents and their departures from steady, horizontal airflow. When encountered by an aircraft, turbulence produces abrupt excursions in aircraft position, sometimes resulting in discomfort or injury to passengers, and sometimes even structural damage or failure. Major origins of turbulence are (1) mechanical, caused by irregular terrain below the flow of air; (2) thermal, associated with vertical currents produced by heating of air in contact with the Earth's surface; (3) thunderstorms and other convective clouds (**Fig. 1**); (4) mountain wave, a regime of disturbed airflow leeward of mountains or hills, often comprising both smooth and breaking waves formed when stable air is forced to ascend over the mountains; and (5) wind shear, usually variations of horizontal wind in the vertical direction, occurring along air-mass boundaries, temperature inversions (including the tropopause), and in and near the jet stream.

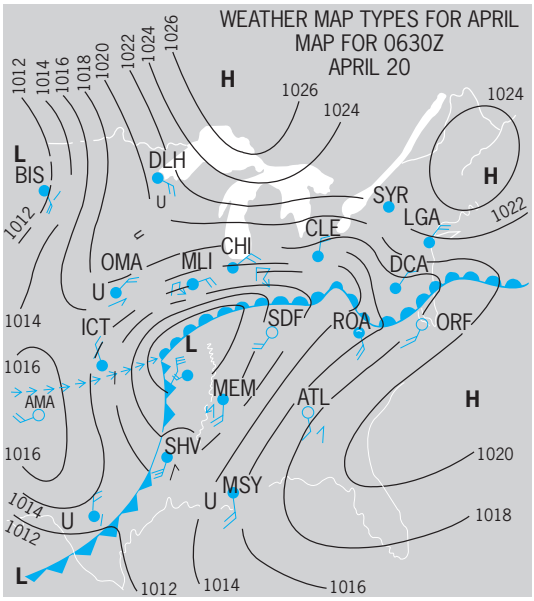


Fig. 1. Severe warm-front thunderstorm situation over the eastern United States. Open or solid circles with three letters represent reporting stations, such as OMA for Omaha and CHI for Chicago. Contour lines express millibar pressure. (*United Air Lines*)

Vertical and horizontal wind gusts in the upper troposphere and lower stratosphere, as encountered by jet transport aircraft that frequent these altitudes, are called clear-air turbulence. Turbulence encounters are the principal cause of nonground impact incidences in commercial aviation. The distinction between clear air and thunderstorm turbulence becomes uncertain as aircraft are routinely operated at altitudes above 30,000 ft (9 km) and encounter thunderstorms growing to more than 60,000 ft (18 km).

While encounters with strong turbulence anywhere in the atmosphere represent substantial inconvenience, encounters with rapid changes in wind speed and direction at low altitude can be catastrophic. Generally, wind shear is most dangerous when encountered below 1000 ft (300 m) above the ground, where it is identified as low-altitude wind shear. Intense convective microbursts, downdrafts usually associated with thunderstorms (**Fig. 2**), have caused many aircraft accidents often resulting in a great loss of life. The downdraft emanating from convective clouds, when nearing the Earth's surface, spreads horizontally as outrushing rain-cooled air. When entering a microburst outflow, an aircraft first meets a headwind that produces increased

performance by way of increased airspeed over the wings. Then within about 5 s, the aircraft encounters a downdraft and then a tailwind with decreased performance. A large proportion of microburst accidents, both after takeoff and on approach to landing, are caused by this performance decrease, which can result in rapid descent. *See* THUNDERSTORM.

Turbulence and low-altitude wind shear can readily be detected by a special type of weather radar, termed Doppler radar. By measuring the phase shift of radiation backscattered by hydrometeors and other targets in the atmosphere, both turbulence and wind shear can be clearly identified. It is anticipated that Doppler radars located at airports, combined with more thorough pilot training regarding the need to avoid microburst wind shear, will provide desired protection from this dangerous aviation weather phenomenon. *See* DOPPLER RADAR; METEOROLOGICAL RADAR.

Hail. Fabric-covered aircraft are particularly susceptible to hail damage. The advent of all-metal aircraft reduced the probability of damage from small (0.2 in. or 5 mm diameter) hail, but larger hail (1–2 in. or 25–50 mm diameter) can be very destructive—shattering the wind screen, dimpling skin surfaces, and cracking radomes. Avoidance of thunderstorms is the best preventive when in flight; on-board radar is a valuable detection device, especially at night and for storms embedded in widespread clouds. *See* HAIL.

Turbulence. Turbulence encounters in and near thunderstorms are increasing with attendant increasing reports of injury to passengers. Part of the increase is because more aircraft fly above 30,000 ft (9 km) and sky space is more crowded. Thunderstorms extend above 60,000 ft (18 km) with their upper portions and anvils spreading out for tens of miles (or kilometers). These storms also compete for space with air traffic. Turbulence at these levels is often located in the regions between the high reflectivity volumes—that is, on the edges of the larger scale vertical motion cores—and in areas where lightning activity is less frequent (charge separation tends to be destroyed by turbulence). Severe turbulence is encountered as much as 20 miles (37 km) from the thundersorm cores.

Turbulence associated with mountains and wind shears is frequently encountered in clear air, with little forewarning to the pilot. The unexpected nature of the clear-air turbulence adds to its danger, especially with respect to passenger safety. The intensity of clear-air turbulence depends on the size and speed of the aircraft. Generally, the accelerations are larger at faster speeds. Clear-air turbulence exists even at the 49,000–65,000 ft (15,000–20,000 m) flight levels of supersonic transport aircraft, and the paths of such aircraft must be planned to avoid areas where strong jet streams are located over tall mountains.

Since clear-air turbulence is a small-scale phenomenon (patches of turbulence frequently measure only a few miles in diameter and a few hundred to a few thousand feet in thickness) the upper-air sounding network is insufficient for locating or predicting the precise locations of clear-air turbulence zones. Remote sensors, such as the Doppler lidar and the infrared radiometer, are able to detect zones of wind shear and temperature inversions ahead of the aircraft, permitting evasive actions to be taken or passenger warning signs to be activated. *See* CLEAR-AIR TURBULENCE; LIDAR.

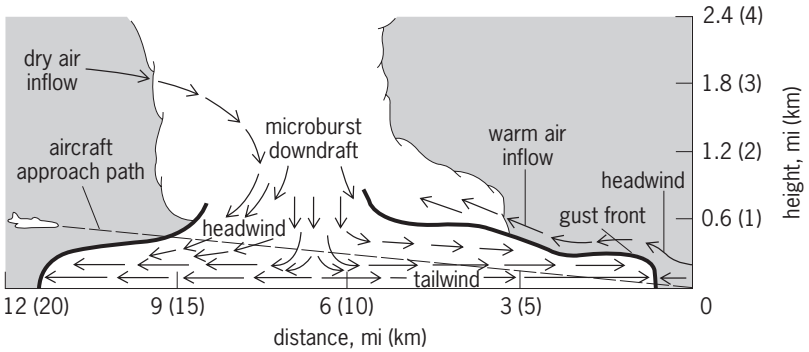


Fig. 2. Schematic of a thunderstorm intersecting the approach path to an airport. Arrows indicate wind-flow pattern. Broken line is 3° glide slope approach. Heavy lines outline the leading edge of the thunderstorm's cold air outflow (gust front). Note rapid change in wind in center of microburst downdraft.

Upper winds and temperature. Since an aircraft's speed is given by a propulsive component plus the speed of the air current bearing the aircraft, there are aiding or retarding effects depending on wind direction in relation to the track flown. Wind direction and speed vary only moderately from day to day and from winter to summer in certain parts of the world, but fluctuations of the vector wind at middle and high latitudes in the troposphere and lower stratosphere can exceed 200 knots (100 $m \cdot s^{-1}$). Careful planning of long-range flights is completed with the aid of prognostic upper-air charts. These charts are prepared through the use of computers to solve equations that model atmospheric motion. The selection of flight tracks and cruising altitudes having relatively favorable winds and temperature results in lower elapsed flight times, in some cases even with considerable added ground distance. The role of the aeronautical meteorologist is to provide accurate forecasts of the wind and temperature field, in space and time, through the operational ranges of each aircraft involved. For civil jet-powered aircraft, the optimum flight plan must always represent a compromise among wind, temperature, and turbulence conditions. *See* UPPER-ATMOSPHERE DYNAMICS; WIND.

Jet stream. This is a meandering, shifting current of relatively swift wind flow which is embedded in the general westerly circulation at upper levels. Sometimes girdling the globe at middle and subtropical latitudes, where the strongest jets are found, this band of strong winds, generally 180–300 mi (300–500 km) in width, has great operational significance for aircraft flying at cruising levels of 4–9 mi (6–15 km). It causes some serious flight-schedule delays, and unexpected encounters may force unscheduled fuel stops.

Average speed in the core of a well-developed jet near the tropopause at middle latitudes in winter is

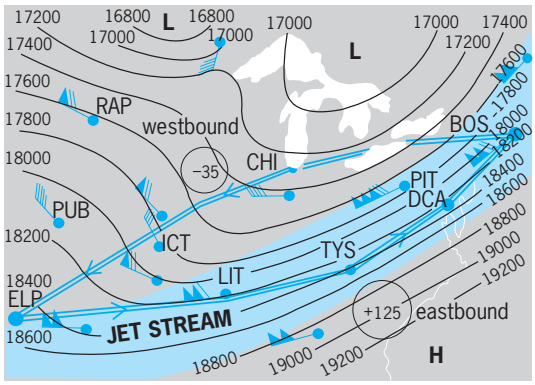


Fig. 3. Use of minimal time track in flight planning over the eastern United States. The solid circles represent weather-reporting places, landing places, or both, such as CHI for Chicago and BOS for Boston. The plus or minus figures shown in the large circles indicate the relation of air speed to ground speed (mi/h) to be expected in the direction of flight. Contour lines express height of the 500-millibar pressure surface in feet. *(United Air Lines)*

close to 100 knots (50 m · s^{-1}), but speeds nearly twice that are fairly common, and extremes may be three times larger. The jet stream challenges the forecaster and the flight planner to utilize tailwinds to the greatest extent possible on downwind flights and to avoid retarding headwinds as much as practicable on upwind flights (**Fig. 3**). As with considerations of wind and temperature, altitude and horizontal coordinates are considered in flight planning for jet-stream conditions. Turbulence in the vicinity of the jet stream is also a forecasting problem.

There is also the "low-level" jet, a relatively narrow (100-nautical-mile or 185-km) stream of air having wind speeds of more than 40 knots (20 m · s^{-1}) through a depth of around 1000 ft (300 m) at an altitude of 1000–3000 ft (300–900 m) above the surface. The northward-flowing jet is usually found from Texas on into the Great Lakes Region. When it occurs, the jet brings moist air rapidly northward and greatly influences the formation of convective storms. *See* JET STREAM.

Tropopause. This boundary between troposphere and stratosphere is generally defined in terms of the zone in the vertical air-mass soundings where the temperature lapse rate becomes less than 3°F/mi (2°C/km). It is sometimes defined as a discontinuity in the wind shear associated with horizontal temperature gradients. Sometimes the tropopause appears as a layer of irregular temperature variations, caused by overlapping "tropical" and "polar" leaves.

The significance of the tropopause to aviation arises from its frequent association with mild forms of clear-air turbulence and change of vertical temperature gradient with altitude. In midlatitudes, the tropopause is commonly located near 6–7 mi (10–12 km), which is about the cruising altitude of most jet aircraft, while in tropical regions it is around 9–10 mi (16–17 km), that is, a little below the cruising altitude of supersonic aircraft. The tropopause marks the usual vertical limit of clouds and storms; however, thunderstorms are known to punch through the surface to heights near 12 mi (20 km) at times, and lenticular clouds showing wave motions strongly developed over mountains have been observed at similar heights. *See* TROPOPAUSE.

Lightning. An electrical discharge to or from an aircraft is experienced as a blinding flash and a muffled explosive sound, usually audible above the roar of the engines. Structural damage to the metal portions of the craft is commonly limited to melting of small spots in the outer skin at the point of entry or exit, and fusing of antenna wires. Nonmetallic surfaces such as radomes and composite tail surfaces may suffer punctures or delaminations. Atmospheric conditions favorable for lightning strikes follow a consistent pattern, characterized by solid clouds or enough clouds for the aircraft to be flying intermittently on instruments; active precipitation of an icy character; and ambient air temperature near or below 32°F (0°C). Saint Elmo's fire, radio static, and choppy air often precede the strike. However, the charge separation processes necessary for the production of strong electrical fields is destroyed by strong turbulence. Thus turbulence and lightning usually do not coexist in the same space.

An aircraft usually triggers the lightning stroke, and evasive action by the pilot usually consists of air-speed reduction, change of course as suggested by radar echoes, or change of altitude. Pilots unable to avoid a typical strike situation should turn up cockpit lights and avert their eyes in order to decrease the risk of being blinded. As metal becomes replaced by composite semiconducting materials in aircraft, there is increased risk of damage from lightning, both to the composite section itself and to inadequately shielded microelectric components. *See* ATMOSPHERIC ELECTRICITY; LIGHTNING; SAINT ELMO'S FIRE.

Icing. Modern aircraft operation finds icing to be a major factor in the safe conduction of a flight both in the takeoff and landing phases as well as the in-flight portion of the trip. Ice and frost formation on an aircraft result in increased drag, decreased lift due to disruption of airflow over the wing and tail surfaces, and increased weight. Icing usually occurs when the air temperature is near or below freezing (32°F or 0°C) and the relative humidity is 80% or more. Clear ice is most likely to form when the air temperature is between 32 and −4°F (0 and −20°C) and the liquid water content of the air is high (large drops or many small drops). As these drops impinge on the skin of an aircraft, the surface temperature of which is 32°F (0°C) or less, the water freezes into a hard, high-density solid. When the liquid water content is small and when snow or ice pellets may also be present, the resulting rime ice formation is composed of drops and encapsulated air, producing an ice that is less dense and opaque in appearance.

The accumulation of ice can occur in a very short time interval. In flight, the use of deicing boots (inflatable rubber sleeves) or heated air of the leading edge of the wing or tail surfaces can reduce the problem. On the ground, ice and snow can be removed either by mechanical means or by application of

deicing solutions. Commercial operations use deicing solutions before takeoff to remove and temporally retard the formation of these increased drag-producing and decreased lift-producing hazards.

Carburetor-equipped aircraft may also experience carburetor icing. The sudden cooling of moist air caused by the vaporization of fuel and the expansion of air as it passes through the carburetor can be as much as 60°F (33°C) in a few seconds. Water thus condensed is deposited as frost or ice inside the carburetor, thereby restricting the airflow. The result is a power decrease and in some cases even engine failure. Carburetor heating judicially used can reduce this problem. Icy or deeply snow-covered airport runways offer operational problems for aviation, but these factors are reduced in importance through the use of modern snow removal equipment. However, accurate forecasts and accurate delineation of freezing conditions are essential for safe aircraft operations. J. T. Lee; J. McCarthy

Bibliography. D. R. MacGorman and W. D. Rust, *The Electoral Nature of Storms*, 1998.

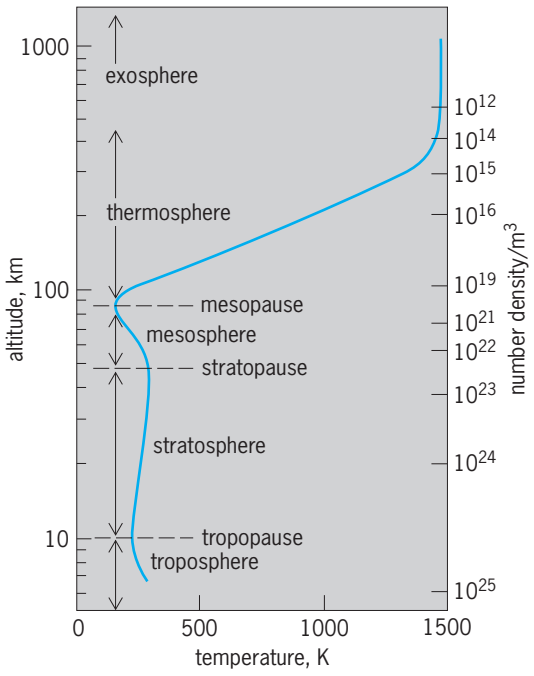
Aeronomy

Aeronomy concentrates on the regions above the tropopause or upper part of the atmosphere. The region of the atmosphere below the tropopause is the site of most of the weather phenomena that so directly affect all life on the planet; this region has primarily been the domain of meteorology.

Aeronomy describes the chemical and physical properties of planetary atmospheres and the changes that result from external and internal forces. The results of this interaction impact all altitudes and global distributions of atoms, molecules, ions, and electrons, both in composition and in density. Dynamical effects are seen in vertical and horizontal atmospheric motion, and energy is transferred through radiation, chemistry, conduction, convection, and wave propagation. Thus aeronomy is a truly interdisciplinary field encompassing chemistry, physics, electrical engineering, and mathematics and dealing with all aspects of theory, simulations, modeling, and laboratory and field experiments. Research may involve the use of specialized ground-based optical and radar facilities; satellite, rocket, and balloon instrumentation; and state-of-the-art computational facilities, locally, regionally, and globally. Since aeronomy is not restricted to earth science, optical and radar remote-sensing techniques are also used to study the atmospheres of other planets in the solar system. *See* METEOROLOGICAL RADAR; REMOTE SENSING.

The atmosphere of the Earth is separated into regions defined by the variation of temperature with height (see **illustration**). Above the surface, temperature minima occur near 10 km (6 mi; the tropopause) and 90 km (54 mi; the mesopause). In between these two minima, the maximum in temperature near 50 km (30 mi) is called the stratopause. The region below the tropopause is the troposphere; the regions between the minimum and maximum values are the stratosphere and mesosphere. Above the mesopause the region is called the thermosphere. The temperature gradually increases above the mesopause throughout the thermosphere to a constant value near 300–500 km (180–300 mi). Above this altitude the region is called the exosphere, which extends out to the limit of the detectable terrestrial atmosphere. The region between the tropopause and the mesopause (roughly 10–100 km or 6–60 mi) is called the middle atmosphere, and the region above (>100 km or 60 mi) is called the upper atmosphere. *See* MESOSPHERE; METEOROLOGY; STRATOSPHERE; TROPOSPHERE.

Middle atmosphere. In the middle atmosphere, that region of the atmosphere between the tropopause and the mesopause (10–100 km or 6–60 mi), the temperature varies from 250 K (−9.7°F) at the tropopause to 300 K (80°F) at the stratopause and back down to 200 K (−100°F) at the mesopause (see illus.). These temperatures are average values, and they vary with season and heat and winds due to the effect of the Sun on the atmosphere. Over this same height interval the atmospheric density varies by over five orders of magnitude. Although there is a constant mean molecular weight over this region, that is, a constant relative abundance of the major atmospheric constituents of molecular oxygen and nitrogen, there are a number of minor constituents that have a profound influence on the biosphere and an increasing influence on the change in the atmosphere below the tropopause associated with the general topic of global change. These constituents, called the greenhouse gases (water vapor, ozone, carbon dioxide, methane, chlorine compounds,



Typical atmospheric temperature variation with altitude and constituent number density (number of atoms and molecules per cubic meter) at high solar activity. °F = (K × 1.8) − 459.67. 1 km = 0.6 mi.

nitrogen oxides, chlorofluorocarbons, and others), are all within the middle atmosphere. To fully understand and model this region of the atmosphere requires consideration of the solar variability, horizontal and vertical wind systems, and the reaction rates and time constants involved with all the chemical and radiative processes that control these constituent distributions. This information needs to be obtained simultaneously at all altitudes and all global locations. The magnitude of this task is huge, and thus aeronomy requires national and international coordinated programs. *See* GREENHOUSE EFFECT.

Solar radiation. Solar radiation is the primary energy source that drives all of the various processes in middle-atmospheric aeronomy. Although the solar spectrum can at times be approximated by a body radiating at a temperature around 6000 K (10,300°F), a detailed study of the spectrum as received at different altitudes in the middle atmosphere shows that absorption by the constituents of the atmosphere modifies the solar radiation at specific wavelengths. Since ozone in the stratosphere strongly absorbs solar radiation below 300 nanometers, the amount of that radiation reaching the ground is very dependent on the total amount of ozone. The harmful relationship between radiation below 300 nm and the Earth's biosystems is the driving concern about the destruction of ozone through the addition of new chemicals that will modify the historical chemical balance of the stratosphere. *See* SOLAR RADIATION; STRATOSPHERIC OZONE.

Dynamics. Understanding the aeronomy of the middle atmosphere requires the study of the physical motion of the atmosphere. The particular composition and chemistry at any given time, location, or altitude depends on how the various constituents are transported from one region to another. Thus, global circulation models for both horizontal and vertical motions are needed to completely specify the chemical state of the atmosphere. In understanding the dynamics of the middle atmosphere, internal gravity waves and acoustic gravity waves play significant roles at different altitudes, depending on whether the particle motion associated with the wave is purely transverse to the direction of propagation or has some longitudinal component. These waves originate primarily in meteorological events such as wind shears, turbulent storms, and weather fronts; and their magnitude can also depend on orographic features on the Earth's surface. One example of coupling between regions in the middle atmosphere is seen in the events known as stratospheric warmings, which are produced by gravity waves generated from the large tropospheric waves that warm the stratosphere and cool the mesosphere. In the polar regions the winter circulation tends to be very stable and circumpolar, and it is usually referred to as the polar vortex. In the Antarctic this region is so stable that photolysis of many of the middle atmospheric species does not occur, thus allowing the buildup of species that would otherwise never develop significant concentrations. Upon reexposure to sunlight during the spring, there is a dramatic photolytic conversion to high concentrations of other species that combine in the presence of polar stratospheric clouds to produce the observed spring ozone depletion (ozone hole) in the lower stratosphere. In the summer polar regions the mesopause temperature becomes very low (<140 K or −208°F) and leads to the formation of very high altitude (84 km or 50 mi) clouds. Historically seen from the ground since 1885 and known as noctilucent clouds, they have now been seen from satellites and are also known as polar mesospheric clouds. *See* ATMOSPHERIC CHEMISTRY; MIDDLE-ATMOSPHERE DYNAMICS.

Upper atmosphere. The upper atmosphere is that region above the middle atmosphere that extends from roughly 100 km (60 mi) to the limit of the detectable atmosphere of the planet. This region is characterized by an increasing temperature until it reaches a constant exospheric temperature. There is a slow transition from the region of constant mean molecular weight associated with the middle atmosphere to that of almost pure atomic hydrogen at high altitudes of the exosphere. This is also the region of transition between transport dominated by collision and diffusion, and transport influenced by plasma convection in the magnetic field. The neutral density varies by over ten orders of magnitude from one end to the other and is dominated by molecular processes in the high-density region and an increasing importance of atomic, electron, and ion processes as the density decreases with altitude.

The ionosphere is formed primarily by solar ultraviolet radiation at wavelengths below that produced by the hydrogen atom emission line (Lyman alpha) at 121.6 nm. This radiation ionizes a fraction, variable with altitude, of the atoms and molecules of the upper atmosphere. Basic studies are concerned with energy sources and sinks, dynamics as a function of altitude, chemistry and particle precipitation leading to optical emissions from the atmospheric constituents, coupling between the low altitudes and the higher altitudes, and the impact of heating by both ions and electrons and their reaction to variable electric fields (related to solar activity) and subsequent variation in the Earth's magnetic field. *See* IONOSPHERE; ULTRAVIOLET RADIATION.

Dynamics. Studies of the movement of the neutral atmosphere are complicated because of the variation in atmospheric density and various other influences. The major structure of the upper-atmospheric wind field is produced by the pressure gradient caused by the difference in solar heating on the day and night sides of the Earth. Comparison of observations and this type of simple model shows that the interaction of the ionosphere with the neutral atmosphere needs to be considered on a global basis, even though most of the major effects are associated with particle precipitation and effects of the configuration of the electric field in the polar regions. Only when the impact of ion drag (the slowing down of moving ions because of collisions with atmosphere gases) is taken into consideration do the models begin to reproduce the observations. Thus topics including Joule

heating, particle precipitation, magnetospheric electric fields, conductivity, plasma convection, and gravity tidal- and planetary-wave activity need to be included in the coupled Ionospheric-thermospheric models. *See* UPPER-ATMOSPHERE DYNAMICS.

Dayglow. While the visible solar light plays an important role in the heating of the middle atmosphere during the daytime, the incident solar far-ultraviolet light at wavelengths below 200 nm ionizes, dissociates, and excites to radiative states the various atmospheric molecules and atoms that make up the upper atmosphere. The light given off by these atmospheric constituents makes up what is called the dayglow spectrum of the atmosphere. *See* AIRGLOW.

Nightglow. As the Sun sets on the atmosphere, dramatic changes due to the loss of the solar radiation occur. The excitation of the atmosphere declines, and chemical reactions between the constituents become more and more dominant as the night progresses. When observed with very sensitive instruments, the night sky appears to glow at various colors of light. Prominent green and red atomic emissions due to oxygen at 555.7 and 630 nm, respectively, and yellow sodium light at 589 nm appear, while molecular bands of the hydroxyl radical and molecular oxygen even further in the red collectively contribute most of the total intensity of the nightglow spectrum.

Aurora. The aurora that appears in the southern and northern polar regions is the optical manifestation of the energy loss of energetic particles precipitating into the atmosphere. The region of highest probability of occurrence is called the auroral oval. At high altitudes, electrons and ions present in the magnetosphere are accelerated along magnetic field lines into the atmosphere at high polar latitudes. The energy of most of the auroral particles (<15 keV) is lost before penetration very far into the atmosphere; hence most of the energy is deposited between 100 and 120 km (60 and 72 mi), although optical emissions can be detected throughout the 80–300-km (48–180-mi) region. In the visible part of the spectrum the most intense feature is the 557.7-nm atomic oxygen line. The bands of molecular nitrogen ions fill the region between 380 and 530 nm, and the molecular oxygen and nitrogen bands occur above 590 nm. However, optical emissions from the various atmospheric constituents occur at discrete wavelengths over the whole region from 30 nm to the far-infrared (15,000 nm). The primary auroral energy resides in the flux of electrons; however, at times some regions of the auroral oval are dominated by energetic proton precipitation, with some evidence of significant amounts of energetic oxygen ions and helium ions at other times. *See* AURORA; MAGNETOSPHERE.

Other planets. Most of the topics that are of concern in middle- and upper-atmospheric aeronomy apply to atmospheres of other planets as well. The remote-sensing techniques applied to the study of the Earth can also be used to study various aspects of the aeronomy of other planets. Mars and Venus have atmospheres dominated by carbon dioxide, traces of oxygen, some molecular nitrogen, and very little water. However, there is very little ozone, so without the heat input from absorption of solar radiation as in the Earth's atmosphere, the middle atmospheres of these planets remain cold. The outer planets have atmospheres mainly of hydrogen and helium, and thus they do not absorb much radiation and remain very cold. However, some ionization does take place, and ionospheres form and vary with altitude much like that on Earth. Energetic particle precipitation on Jupiter and Saturn and their satellites Io and Titan produce aurora similar to those on Earth, except with emissions related to the characteristics of their particular atmospheres and magnetic fields. *See* ATMOSPHERE; JUPITER; MARS; SATURN; VENUS.

Gerald J. Romick

Bibliography. S. J. Bauer and H. Lammer, *Planetary Aeronomy: Atmosphere Environments in Planetary System*, 2004; G. Brasseur and S. Solomon, *Aeronomy of the Middle Atmosphere: Chemistry and Physics of the Stratosphere and Mesosphere*, 3d ed., 2005; R. M. Goody and Y. L. Yung, *Atmospheric Radiation: Theoretical Basis*, 2d ed., 1995; M. C. Kelley, *The Earth's Ionosphere, Plasma Physics and Electrodynamics*, 1989; R. R. Meier, Ultraviolet spectroscopy and remote sensing of the upper atmosphere, *Space Sci. Rev.*, 58:1–185, 1991; M. H. Rees, *Physics and Chemistry of the Upper Atmosphere*, 1989.

Aerosol

A suspension of small particles in a gas. The particles may be solid or liquid or a mixture of both. Aerosols are formed by the conversion of gases to particles, the disintegration of liquids or solids, or the resuspension of powdered material. Aerosol formation from a gas results in much finer particles than disintegration processes (except when condensation takes place directly on existing large particles). Dust, smoke, fume, haze, and mist are common terms for aerosols. Dust usually refers to solid particles produced by disintegration, while smoke and fume particles are generally smaller and formed from the gas phase. Mists are composed of liquid droplets. These special terms are helpful but are difficult to define exactly.

Aerosol particles range in size from molecular clusters on the order of 1 nanometer to 100 micrometers (**Fig. 1**). The stable clusters formed by homogeneous nucleation and the smallest solid particles that compose agglomerates have a significant fraction of their molecules in the surface layer.

Role in science and technology. Aerosols are important in the atmospheric sciences and air pollution; inhalation therapy and industrial hygiene, manufacture of pigments, fillers, and metal powders; and fabrication of optical fibers. Atmospheric aerosols influence climate directly and indirectly. They directly affect radiation transfer on global and regional scales. Indirect effects result from their role as cloud condensation nuclei in changing droplet-size distributions that affect the optical properties of clouds and

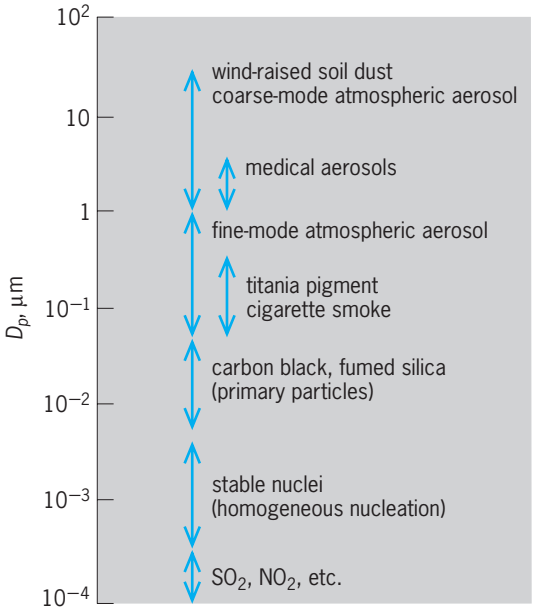


Fig. 1. Examples of aerosol particle-size ranges.

precipitation. There is evidence that the stratospheric aerosol is significant in ozone destruction. *See* METEOROLOGICAL OPTICS; RADIATIVE TRANSFER.

The atmospheric aerosol consists of material emitted directly from sources (primary component) and material formed by gas-to-particle conversion in the atmosphere (secondary component). The secondary component is usually the result of chemical reactions in either the gas or aerosol phases. Contributions to the atmospheric aerosol come from both natural and anthropogenic sources. The effects of the atmospheric aerosol are largely determined by the size and chemical composition of the individual particles and their morphology (shape or fractal character). For many applications, the aerosol can be characterized sufficiently by measuring the particle-size distribution function and the average distribution of chemical components with respect to particle size. The chemical composition of the atmospheric aerosol can be used to resolve its sources, natural or anthropogenic, by a method based on chemical signatures. Particle-to-particle variations in chemical composition and particle structural characteristics can also be measured; they probably affect the biochemical behavior and nucleating properties of aerosols.

Aerosol optical properties depend on particle-size distribution and refractive index, and the wavelength of the light. These are determining factors in atmospheric visibility and the radiation balance.

It is possible to generate aerosols composed of particles that are nearly the same size. These monodisperse (homogeneous) aerosols are normally used for the calibration of aerosol instruments. In industry and nature, aerosols are usually composed of particles of many different sizes; that is, they are polydisperse. The most important physical characteristic of polydisperse aerosols is their particle-size distribution. It is defined, for spherical particles of diameter D_p, as $dN = n(D_p)dD_p$. Here dN is the concentration of particles in a volume element of gas in the particle size range D_p to $D_p + dD_p$, and $n(D_p)$ is the particle-size distribution function. Aerosol size distribution data are often shown by plotting the volume distribution, a weighted function of $n(D_p)$ as a function of particle diameter. Volume distributions for the atmospheric aerosol are frequently bimodal (**Fig. 2**). The lower mode is usually composed of the products of atmospheric gas-to-particle conversion processes, principally ammonium sulfate and nitrate, and of combustion products such as soot and organics. The upper mode includes soil dust and coarse fly-ash particles.

Effects of the atmospheric aerosol on human health have led to the establishment of ambient air-quality standards by the United States and other industrialized nations. Adverse health effects have stimulated many controlled studies of aerosol inhalation by humans and animals. There is much uncertainty concerning the chemical components of the atmospheric aerosol that produce adverse health effects detected in epidemiological studies. *See* AIR POLLUTION.

Aerosols containing pharmaceutical agents have long been used in the treatment of lung diseases such as asthma. Current efforts are directed toward systemic delivery of drugs, such as aerosolized insulin, which are transported across the alveolar walls into the blood. Three main types of medical aerosol inhalers have been developed: (1) nebulizers in which aqueous solutions are aerosolized, (2) pressurized metered-dose inhalers that use chlorofluorocarbon propellant to aerosolize a suspension of the drug, and (3) dry-powder inhalers (DPIs), which depend on inspiration by the user for dispersion. The phase-out of chlorofluorocarbons has led to increased use of dry-powder inhalers; however, a serious technical problem is the need to generate a fine aerosol from a powder that is increasingly cohesive as the particle size decreases.

Aerosol processes are used routinely in the manufacture of fine particles. Aerosol reaction engineering refers to the design of such processes, with the goal of relating product properties to the properties of

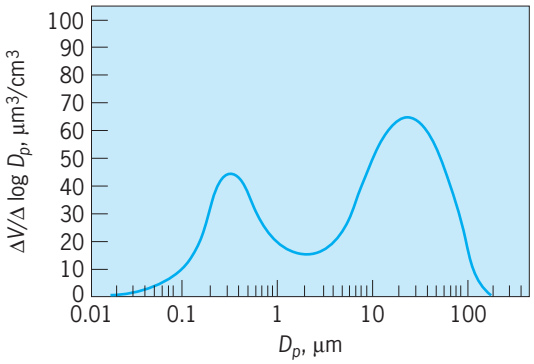


Fig. 2. Bimodal volume distribution of a type frequently found for atmospheric and combustion aerosols. Such distributions usually result from two different generation processes, the smaller mode from molecular condensation and the larger from breakup or redispersion.

the aerosol precursors and the process conditions. The most important large-scale commercial systems are flame reactors for production of pigments and powdered materials such as titania and fumed silica. Optical fibers are fabricated by an aerosol process in which a combustion-generated silica fume is deposited on the inside walls of a quartz tube a few centimeters in diameter, along with suitable dopant aerosols to control refractive index. The original 1-m (3.3-ft) tube is then pulled out to produce a 100-μm fiber 50–100 km (30–60 mi) long. Pyrolysis reactors are used in carbon black manufacture. Micrometer-size iron and nickel powders are produced industrially by the thermal decomposition of their carbonyls. Large pilot-scale aerosol reactors are operated using high-energy electron beams to irradiate flue gases from fossil fuel combustion. The goal is to convert sulfur oxides and nitrogen oxides to ammonium sulfate and nitrate that can be sold as a fertilizer.

Chemical composition. Atmospheric aerosols and aerosols emitted from industrial sources are normally composed of mixtures of chemical compounds. Each chemical species is distributed with respect to particle size in a way that depends on its source and past history; hence, different substances tend to accumulate in different particle-size ranges. This effect has been observed for emissions from pulverized coal combustion and municipal waste incinerators, and it undoubtedly occurs in emissions from other sources. Chemical segregation with respect to size has important implications for the effects of aerosols on public health and the environment, because particle transport and deposition depend strongly on particle size.

Aerosol composition is used routinely to determine the sources of atmospheric aerosols. For example, the chemical mass-balance method permits quantitative estimates of sources based on measured atmospheric aerosol chemical compositions at a given site and a set of known (or assumed) source compositions. There are many variations on this approach, which has been adopted by pollution control agencies.

Transport and deposition. An understanding of particle transport is basic to the design of particle collection equipment and aerosol sampling instruments as well as to the scavenging of particulate matter from the atmosphere. For particles smaller than a few tenths of a micrometer, at normal temperatures and pressures the controlling mechanism of particle transport is Brownian diffusion. Other important driving forces for particle transport are the temperature gradient and, for charged particles, electrical potential gradients. Thermophoresis, which refers to the motion of particles in a temperature gradient, plays a key role in the manufacture of optical fibers. Particles are always driven from high to low temperatures; for particles smaller than the mean free path of the gas, the thermophoretic migration velocity is nearly independent of particle size and chemical composition. Electrical migration is the controlling mechanism for particle deposition in electrostatic precipitators, widely used in industry for gas cleaning. The electrical migration velocity normally passes through a minimum with respect to particle diameters in the range 0.1–1.0 μm. *See* BROWNIAN MOVEMENT; ELECTROSTATIC PRECIPITATOR.

Deposition from flowing gases. Diffusional transport in flowing gases is called convective diffusion. Particle deposition on surfaces by this mechanism can be predicted from the usual mass-transfer correlations, provided that the particle size is small compared with the characteristic length of the collecting elements. The tendency for deposition by diffusion to decrease with increasing particle size is countered by the fact that the particles need to diffuse to only one particle radius from the surface on which they will be deposited. This effect, direct interception, results in an increase in particle deposition with size, usually for particle diameters in the range 0.3–3 μm. *See* DIFFUSION.

Particles unable to follow the motion of an accelerating gas because of their inertia may deposit on surfaces by a process known as inertial deposition or impaction. Inertial deposition is often the controlling removal mechanism for the collection of particles larger than a few micrometers. Inertial deposition rates can be calculated from a force balance on a particle, and depend on the dimensionless group known as the Stokes number, which is related to the particle density, the gas velocity, and the gas viscosity. Inertial deposition often increases sharply over a narrow range of values for the Stokes number.

The efficiency of particle removal from a gas flowing over a set of collecting elements such as those in a filter or packed bed is a function of particle size for a fixed gas velocity (**Fig. 3**). The removal efficiency is high for very small particles because of their intense Brownian motion: it declines until the direct interception mechanism becomes operative, and it continues to rise as impaction and sedimentation become important. The minimum in the efficiency usually occurs in the particle-size range near 0.3 μm—the range normally used to test filter performance.

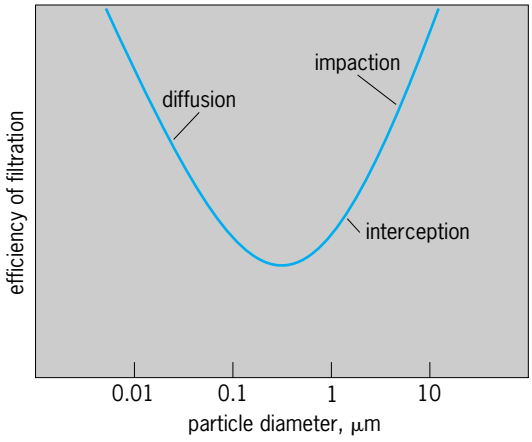


Fig. 3. Variation in the collection efficiency of a fibrous filter with particle size, showing the mechanisms controlling deposition in various particle-size ranges. Leaks or pinholes in the filter material lead to markedly different behavior.

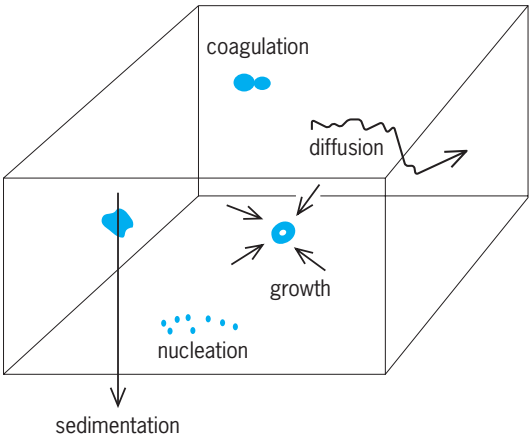


Fig. 4. Processes shaping the particle-size distribution function in a small volume element of gas. Diffusion and sedimentation involve transport across the walls of the element. Coagulation, nucleation, and growth take place within the element.

Aerosol dynamics. The aerosol-size distribution function changes as a gas flows through industrial process equipment or the atmosphere. Aerosol dynamics, which refers to the study of the change in the particle-size distribution function with position and time, gives a distinctive theoretical and experimental character to aerosol science and technology. As an example, consider a small volume element of gas in which the aerosol size distribution is modified by various processes (**Fig. 4**). Particle collisions lead to coagulation. Cooling or chemical reaction may convert molecules in the gas to the aerosol phase by the formation of many new particles smaller than 1–10 nm in diameter or by condensation on existing particles. The formation of new particles, homogeneous nucleation, leads to a particle-size distribution quite different from the one resulting when the same amount of material deposits on existing particles (heterogeneous condensation).

Both gas-to-particle conversion and coagulation take place within the volume element and do not change the total mass of material within the element. Particles may be transported across the element boundary by diffusion, inertia, or an external force field. These processes also change the particle-size distribution function; the change in the distribution function is the sum of the rates for the internal and external processes. A general dynamic equation can be set up for the particle-size distribution function. Also known as a population balance equation, it serves as the starting point for the analysis of aerosol behavior in industrial reactors, pollution sources, and the atmosphere.

Coagulation is a key internal process shaping the aerosol size distribution. The change in the size distribution function resulting from coagulation is given by the Smoluchowski equation,

$$\left(\frac{\partial n}{\partial t}\right)_{\text{coag}} = \frac{1}{2}\int_0^v \beta(\tilde{v}, v-\tilde{v})n(\tilde{v})n(v-\tilde{v})\,d\tilde{v} - \int_0^\infty \beta(v,\tilde{v})n(v)n(\tilde{v})\,d\tilde{v}$$

in which coagulation is related to formation by collision of smaller particles and loss by collision with other particles. Collision kernels have been derived for Brownian coagulation of particles and for many other collision mechanisms, where the particle-size distribution, $n(v,\vec{r},t)$ is a function of position $\vec{r}$, time t, and particle volume v. The collision kernel $\beta(v,\tilde{v})$ is the collision frequency function for particles of volumes v and $\tilde{v}$.

In classical coagulation theory, it is assumed that colliding particles coalesce instantaneously to form larger spherical particles. A striking aspect of Brownian coagulation is that the particle-size distribution tends to approach an asymptotic form independent of the initial size distribution. This asymptotic distribution is referred to as self-preserving, because it does not change its shape when plotted in reduced form as a function of the ratio of particle volume to the average particle volume.

Agglomerates. Aerosol agglomerates are dendritic structures composed of small solid particles that collide but do not coalesce. Industrial aerosols such as pyrogenic silica, titania, and carbon black are often produced in this form.

The individual subunits (primary particles) composing the agglomerate structures are usually 2–100 nm in diameter, much larger than the stable clusters that form in homogeneous nucleation. Indeed, for many metal salt or oxide particles the vapor pressure is so low that the individual molecules serve as stable nuclei. Under such circumstances, the size of the primary particles is determined by the relative rates of the collisions among the particles and their coalescence (or sintering). Coalescence may result from solid-state diffusion or viscous flow (for liquid particles). Both are activated processes and depend sensitively on the temperature of the system.

When coalescence is rapid, collisions produce larger spherical particles. As the temperature falls, coalescence ceases, resulting in solid particles that continue to collide and assemble as agglomerate structures. The size of the primary particles depends on the maximum temperature in the system, decreasing rapidly as the temperature increases. Since agglomerate structures have much larger collision diameters than spheres of the same mass, agglomerates collide much more rapidly than coalescing spheres, leading to the formation of large flocs of low concentration and mass density.

Measurement methods. Aerosol measurement systems are used to monitor atmospheric pollution or industrial gas streams or to test gas-cleaning equipment such as filters and scrubbers. Most instruments used for measuring aerosol properties depend on particle transport or light scattering. Theory provides useful guidelines for instrument design, but it is rarely possible to predict performance from first principles; it is necessary to calibrate the instruments by using monodisperse aerosols. *See* LIGHT-SCATTERING TECHNIQUES.

Quantities often measured include the mass concentration, number concentration, various ranges of the particle-size distribution function, extinction coefficient, average chemical composition, and chemical composition over certain size ranges. There have been rapid developments in the design of instruments for the measurement of size distribution down to the single-particle level.

The aerosol instruments selected for a particular application depend on several factors. Most important is the type of information sought. Other factors include cost, portability, and reliability under the conditions of operation. Stack-gas monitoring poses particularly difficult demands because of extreme conditions of temperature and humidity. In the case of measurement systems designed for routine monitoring, maintenance is an important consideration.

Sampling. In some cases, optical methods can be used to measure aerosol characteristics in the original gas stream, particularly the light extinction, without withdrawing a sample. In most cases, however, it is necessary to sample from a flowing gas through a tube into an instrument to characterize the aerosol. The sampling stream intake should be designed to minimize preferential withdrawal of particles with respect to size. Deposition on the inside walls of the sampling tube and subsequent reentrainment must be minimized or taken into account. Precautions are also necessary to prevent condensation and other gas-to-particle conversion processes. This problem is particularly acute in the sampling of hot, humid process gases.

In general, redispersing deposited aerosols is not a good way to determine the original particle-size distribution. Unless special care is taken in sampling, deposited particulate matter loses its original morphological properties. For microscopic observations, particles can be deposited on electron micrograph grids and optical microscope slides, taking precautions to avoid sampling artifacts.

Physical properties. Special methods have been developed for measuring physical properties of aerosols. Condensation particle counters are used to determine the total particle concentration. The particles entering the counter are mixed with an alcohol vapor, then cooled to produce condensation on the particles. The particle concentration can be determined from light-scattering measurements. Such counters usually respond to particles larger than 5–10 nm. Smaller particles may not be detected because of enhanced vapor pressure (Kelvin effect).

Another integral aerosol property, the mass density, can be determined by weighing the material that accumulates in a filter or, on-line, by measuring beta-ray attenuation. Sampling artifacts include the adsorption and reaction of polar gases in the filter and evaporative losses from the deposited particles.

There is no instrument that is capable of measuring size distributions over the entire size range of practical interest. The single-particle optical counter is widely used for on-line measurement of particles ranging about 0.1–10 μm. Such instruments consist of a light source (which may be a laser), a sensitive volume in which the individual particle and the light interact, and a sensor for the light scattered, usually in some special configuration designed to ensure a unique relationship between the signal and the particle diameter. For particles much smaller than about 0.1 μm, the signal is difficult to distinguish from instrument noise.

In the size range 0.01–0.1 μm, particle-size distributions can be measured with an electrostatic classifier. The particles are charged in an ionizer of which there are several types, by mixing with negative ions generated by a corona discharge. The aerosol then flows through the annular space between a tube and a collecting rod down its center, with an electrical potential between the tube and rod. The charged aerosol enters near the outer circumference of the annular space, and particles with a narrow range in electrical mobility exit through a slit near the end of the rod. The distribution in so-called mobility diameters is determined by measuring the number concentration exiting the slit (usually with a condensation particle counter) as a function of the applied potential. Simultaneous operation of the mobility analyzer and optical particle counter has provided the most complete data available on aerosol size distributions.

Chemical properties. The average aerosol chemical composition is usually determined by collecting particles on a filter and applying a variety of analytical techniques. The chemical speciation and valence states for the collected aerosol may differ significantly from the aerosol in its airborne state because of chemical changes during filtration or storage. Variations in average chemical composition with particle size can be determined by separating the particles into different size fractions. The cascade impactor makes use of inertial deposition to collect particles of progressively smaller aerodynamic diameter on a series of stages. The material on each stage is then analyzed chemically. The method gives the average composition in each size range, but it does not give information on variations in composition from particle to particle; methods based on mass spectrometry have been developed for nearly real-time measurement of the chemical composition of individual particles. *See* CHEMOMETRICS; MASS SPECTROMETRY.

Sheldon K. Friedlander

Bibliography. P. A. Baron and K. Willeke (eds.), *Aerosol Measurement: Principles, Techniques, and Applications*, 2d ed., 2001; Environmental Protection Agency, Office of Research and Development, *Air Quality Criteria for Particulate Matter*, vols. 1 (EPA/600/P-99/002aF) and 2 (EPA/600/P-99/002bF), 2004; S. K. Friedlander, *Smoke, Dust, and Haze: Fundamentals of Aerosol Dynamics*, 2d ed., 2000; W. C. Hinds, *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*, 2d ed., 1999; T. T. Kodas and M. J. Hampden-Smith, *Aerosol Processing of Materials*, 1999; S. E. Pratsinis, The role of aerosols in materials processing, *J. Aerosol Sci. Technol.*, special issue, vol. 19, no. 4, 1993; R. K. Wolff (ed.), Advances in medical aerosol delivery and diagnostics, *Aerosol Sci. Technol.*, vol. 22, 1995.

Aerospace medicine

The special field of medicine that deals with humans in environments encountered beyond the surface of the Earth. It includes both aviation medicine and space medicine and is concerned with humans, their environment, and the vehicles in which they fly. Its objective is to ensure human health, safety, well-being, and effective performance through careful selection and training of flight personnel, protection from the unique flight environment and its physiological and psychological effects, and understanding of the flight vehicle and humans' interaction with it.

The environment encountered in air or space flight is very different from that on Earth. Consequently aerospace medicine must deal with the physics of the atmosphere and space and with the conditions and influences introduced by the flight vehicle. Aerospace medicine is therefore concerned with the physiological effects of changes in barometric pressure, atmospheric constituents, toxic substances, acceleration, weightlessness, noise, vibration, ionizing radiation, and thermal and other environmental stresses, as well as psychological stresses and their effects on behavior and performance.

Environment of flight. The flight conditions of advanced supersonic aircraft introduce extreme variations in temperature, extensive and sudden changes in pressure, and rapid acceleration. The unaided human body cannot compensate for these conditions and must rely upon equipment to survive.

Three gas laws apply directly to physiological problems at altitude. Boyle's law, which states that the volume of a mass of gas is inversely proportional to the pressure exerted on it, providing the temperature remains constant, explains expansion of intestinal gases, and the earaches that can be caused by trapped gases in the middle ear, sinuses, and mastoid cells. Henry's law, that the weight of a given gas dissolved in a liquid varies directly with the partial pressure of the gas, explains how exposure to the low pressure at altitudes may release nitrogen gas normally held in the blood in solution and thus produce bends symptoms. Dalton's law states that in a mixture of gases the part of the total pressure or partial pressure exerted by each gas is proportional to its volume percent. As altitude increases, oxygen continues to constitute 21% of the atmosphere, but the decreased pressure of the upper atmosphere and consequently of the oxygen results in reduced oxygenation of the blood. Physical well-being and the ability to think and to reason are therefore inherently dependent upon the pressure of oxygen breathed. *See* BOYLE'S LAW; DALTON'S LAW; GAS.

The physiological problems of acceleration are an integral part of aerospace medicine because humans are exposed to forces of acceleration almost constantly throughout flight that are different from the 1-*g* environment of Earth. Acceleration is expressed in *g* units, where 1 *g* represents the acceleration of falling bodies due to the force of gravity at 32 ft/s^2 (9.8 m/s^2). Positive *g* means increased acceleration and is accompanied by a physical feeling of heaviness, while negative *g* means reduced acceleration and imparts a sensation of reduced weight. When positive *g* forces are applied to the body, the blood is forced downward away from the head and heart, and blackout and unconsciousness may occur. Under negative *g* conditions, the blood is forced upward so that the blood vessels of the head are engorged. This may result in redout, a condition in which the visual field reddens due to engorged eye blood vessels, and unconsciousness. *See* ACCELERATION.

Effects of weightlessness. All of the environmental hazards incumbent in flight in aircraft are present in space flight, but space also has unique hazards, of which weightlessness is paramount. The equilibrium of many biological systems is disrupted by extended exposure to weightlessness.

Biomedical data collected on humans exposed to long and short periods of weightlessness on space shuttle, Skylab, and Salyut missions have provided information on the time course of physiological acclimation to space and readaptation to Earth. The susceptibility, rate of change, and reversibility of the physiological effects vary with each affected physiological system. Thus neurovestibular effects associated with space motion sickness occur during the first few days in orbit. At the same time there is a shift in body fluids toward the head, and faces become puffy. This is followed by a loss of fluids and electrolytes. With decreased fluids, red blood cell mass slowly decreases for about 60 days into flight. Cardiovascular deconditioning also occurs within the first month. These systems in general appear to acclimate to the weightless environment in 4–6 weeks.

Postflight symptoms include orthostatic intolerance associated with shifts in body fluid and resultant cardiopulmonary neuroreceptor reflex responses. Difficulties in postural equilibrium and occasionally motion sickness accompany neurovestibular readaptation. A return to preflight levels of body function is reached within 1–3 months after return to Earth in a time course similar to the initial adjustment of the various systems to 0 *g*. This is not so with the musculoskeletal system. Many believe there is a gradual progressive loss of calcium and lean body mass regardless of flight duration. While Y. Romanenko spent 430 days in flight, including one flight of 326 days, and 16 other cosmonauts have spent more than 100 days in space, there is as yet no documented residual pathology in astronauts or cosmonauts. In fact, the Russians have reported that, with the use of their current countermeasures, bone loss reaches a plateau or is prevented, while antigravity muscles are reduced by only 10–15% after 237 days of flight. Yet it is still surmised that the normal morphology and function of bone and possibly muscle are not fully regained after continuous exposure of over 6 months to weightlessness.

Extensive data collected on the physiological effects of space flight on humans have led to a number of hypotheses to explain the responses, but the underlying mechanisms are still generally unknown. The various countermeasures used to prevent adverse effects of space flight, coupled with the small

sample size of individuals who have flown in space, the marked variation in mission profiles, and the limited capabilities for scientific observation, have compromised scientific investigations. *See* WEIGHTLESSNESS.

Countermeasure. The United States and Russia use many similar countermeasures including physical, psychological, pharmacological, and nutritional means to prevent or control deleterious physiological responses to space flight. Specifically, these include exercise, especially of the lower extremities, by using unique bicycles and treadmills; a vacuum suit that applies negative pressure to the lower body to stress the cardiovascular system; salt water loading on the last day of flight to increase blood volume and prevent orthostatic intolerance; and nutritional supplements including calcium. Pharmacologic agents are used by the United States to curb space sickness symptoms and by Russia for radiation protection.

Radiation hazards. Radiation hazards have been circumvented in both the United States and Russian space programs through the relatively short duration of flights in carefully selected low-Earth orbits in the absence of solar flare activity. This will not be possible when longer space flights and interplanetary missions will expose spacecraft to increased radiation levels. Space radiation includes heavy ions, electrons, protons, and neutrons. The different types of radiation produce different amounts of biological damage. Charged particles such as heavy ions [abbreviated HZE, for high charge (Z) and energy] and low-energy protons, with high rates of energy loss per unit length of path (high linear energy transfer), are more deleterious than electrons and high-energy protons with low rates of energy loss. *See* LINEAR ENERGY TRANSFER (BIOLOGY).

Radiation is monitored on United States flights with passive dosimeters for precise radiation information and active dosimeters worn by the crew to determine current radiation danger. Dosimetry data from United States flights showed a variation from 11 mrad/day (0.11 milligray/day) for *Gemini 4* to almost 90 mrad/day (0.9 mGy/day) for *Skylab 4*. *See* DOSIMETER.

The potential biological effects of galactic cosmic radiation are categorized as either early, including damage to bone marrow and lymphopoietic, intestinal, and gonadal tissues, or late, including infertility, cancer induction, and heritable effects. Both shielding and radioprotective chemicals afford some protection from space radiation. Total shielding is impossible because of the excess weight it would impart to the spacecraft and the ability of heavy ions to penetrate even heavy shielding. Space medicine consequently has the responsibility to identify appropriate protective procedures, define exposure limits, and develop therapeutic measures. *See* RADIATION BIOLOGY; RADIATION INJURY (BIOLOGY); SPACE BIOLOGY.

Thora Waters Halstead

Bibliography. S. Bonting (ed.), *Advances in Space Biology and Medicine*, vol. 1, 1991; R. L. DeHart (ed.), *Fundamentals of Aerospace Medicine*, 2d ed., 1996; F. J. Del Vecchio, *Physiological Aspects of Flight*, 1985; D. Lorr and V. Garshnek (eds.), *Working in Orbit and Beyond*: *The Challenges for Space Medicine*, 1989; A. Nicogossian, C. Leach-Huntoon, and S. Pool (eds.), *Space Physiology and Medicine*, 3d ed., 1993.

Aerospike engine

The aerospike engine (**Fig. 1***a*) is an advanced liquid-propellant rocket engine with unique operating characteristics and performance advantages over conventional rocket engines. It combines a contoured axisymmetric plug nozzle (**Fig. 2**), an annular torus-shaped combustion chamber, conventional turbopumps, a turbine exhaust system that injects the turbine drive gases into the base of the plug nozzle, and a simple combustion tap-off engine cycle. The aerospike is one-quarter the length of a conventional rocket engine, yet it delivers comparable performance (efficiency) at high altitude and superior performance at low altitude. The low-altitude performance advantage is primarily due to the fact that the plug nozzle compensates for altitude whereas the nozzle of a conventional rocket engine does not. While the plug nozzle and its benefits are not new to the field of air-breathing propulsion, the aerospike represents the first application of this type of nozzle to the field of rocket propulsion. Typical propellants are liquid hydrogen (fuel) and liquid oxygen (oxidizer). A variation of the aerospike engine is the linear aerospike engine.

Altitude compensation. For a given set of operating conditions in the combustion chamber (pressure, temperature, and mass-flow rate) and ambient pressure, the thrust of a rocket engine is governed by the nozzle exit area, or the nozzle area ratio if the exit area is divided by the nozzle throat area. When the nozzle area ratio is specified such that the pressure at the nozzle exit is equal to ambient pressure, the flow through the nozzle is optimally expanded and the engine is operating at its maximum efficiency. This is the nozzle design point.

When the area ratio is too large, the nozzle exit pressure is lower than ambient pressure and the flow through the nozzle is overexpanded. This situation is accompanied by a decrease in thrust (efficiency). More thrust could be produced if the area ratio of the nozzle were reduced, yielding a higher exit pressure. When the area ratio is too small, the nozzle exit pressure is greater than ambient pressure and the nozzle is underexpanded. Again, thrust is reduced relative to the optimally expanded condition. Under this condition, a nozzle with a higher area ratio would produce more thrust.

Because rocket engines are required to operate over a range of ambient pressure, or altitude, and because engine efficiency is greatest at only a single altitude, selection of the nozzle area ratio (that is, the nozzle design point) for a particular application introduces a compromise. Operation at high altitude where ambient pressure is lower calls for a high area ratio to extract maximum thrust from the

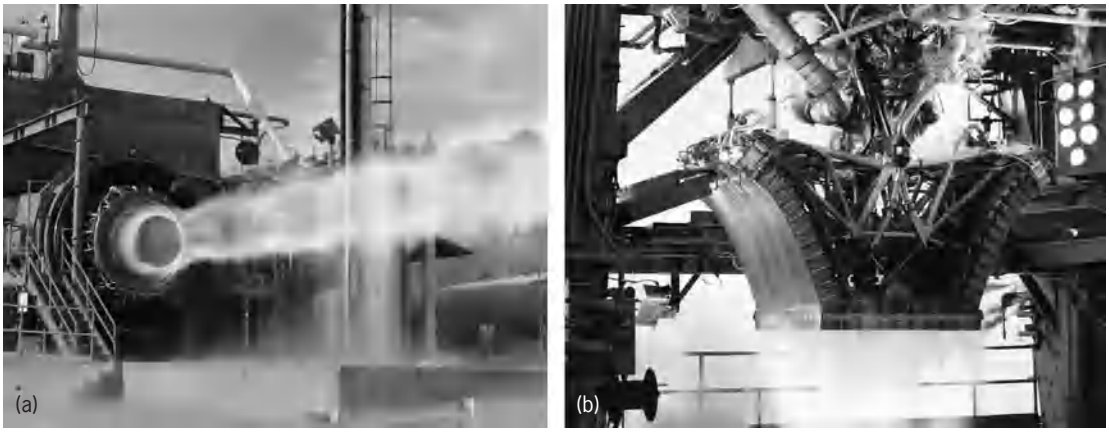


Fig. 1. Static firing tests of (*a*) aerospike engine with 250,000 pounds (1,112,000 newtons) of thrust, and (*b*) linear aerospike engine with 125,000 pounds (556,000 newtons) of thrust. Both engines use hydrogen/oxygen propellants. *(Boeing Company, Rocketdyne Division)*

high-pressure, hot combustion gases, while a lower area ratio is better during low-altitude operation to avoid the reduction in thrust caused by overexpansion. Typically, the selected area ratio falls between these two limits.

Of course, if the geometry of the nozzle were not fixed, and area ratio were allowed to increase with increasing altitude, optimum performance would be obtained from the rocket engine throughout the entire flight trajectory, resulting in dramatic gains in overall vehicle performance. Since the nozzle area ratio is dependent on the nozzle exit area, changes in area ratio could be effected by simply varying nozzle exit area. This effect is referred to as altitude compensation.

Comparison with conventional engines. In a very general sense, the aerospike engine is similar to a conventional liquid-propellant rocket engine. In both cases, turbopumps are used to pressurize liquid propellants. These high-pressure propellants are then injected into a combustion chamber, where they rapidly mix and chemically react to form high-temperature, high-pressure combustion gases. Finally, these hot gases are accelerated to high velocities by expanding them through a nozzle to produce thrust.

In more specific terms, however, several key differences distinguish the aerospike engine from the conventional rocket engine. The most striking difference is the nozzle. The conventional rocket engine uses a converging-diverging nozzle, sometimes referred to as a bell nozzle owing to its shape. The hot combustion gases expand inside the nozzle in the general direction of the nozzle axis, completely contained by the nozzle wall (**Fig. 3***a*). Because the wall of a conventional nozzle forms a physical boundary between the expanding combustion gases and the surrounding atmosphere, the expansion process in the nozzle is isolated from the effect of ambient pressure. Consequently, the combustion gases will not necessarily be optimally expanded, but will be under- or overexpanded depending upon the ambient pressure. *See* NOZZLE.

The aerospike engine features a plug nozzle that is characterized by the combustion gases expanding on the outside of the nozzle (Fig. 3*b*). Initially, the gas flow is directed radially inward, toward the axis or center of the nozzle. The expansion process occurs about the engine cowling, controlled by the plug wall on the inside and the effect of ambient pressure on the outer jet boundary. Because the amount of expansion is controlled by ambient pressure, the aerospike engine has the ability to altitude-compensate in the absence of variable nozzle geometry. Variations in area ratio are achieved automatically by the influence of ambient pressure acting on the jet boundary, causing the effective flow area at the nozzle exit to change with altitude. Thus, ambient pressure effectively provides passive control of the expansion process in the aerospike engine, preventing both under- and overexpansion and the associated loss in performance.

The plug nozzle of the aerospike engine is obtained by truncating a full-length spike nozzle. Since

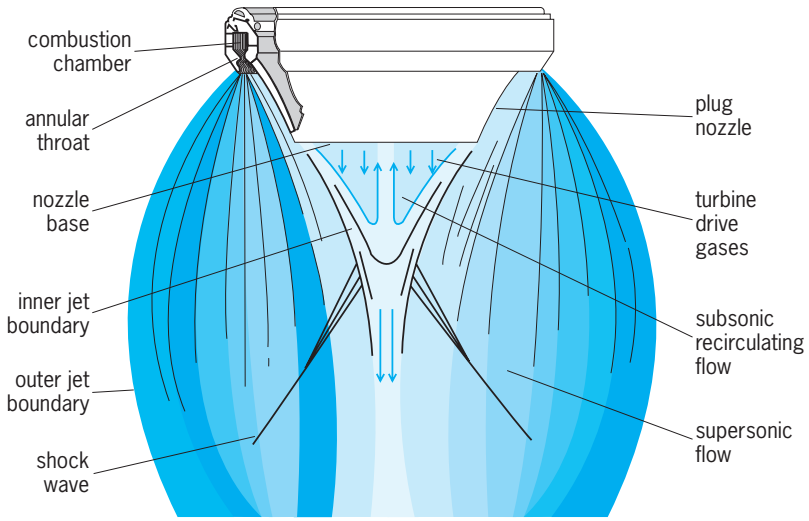


Fig. 2. Aerospike engine with plug nozzle, torus-shaped combustion chamber, and combustion gases expanding outside the nozzle. *(Boeing Company, Rocketdyne Division)*

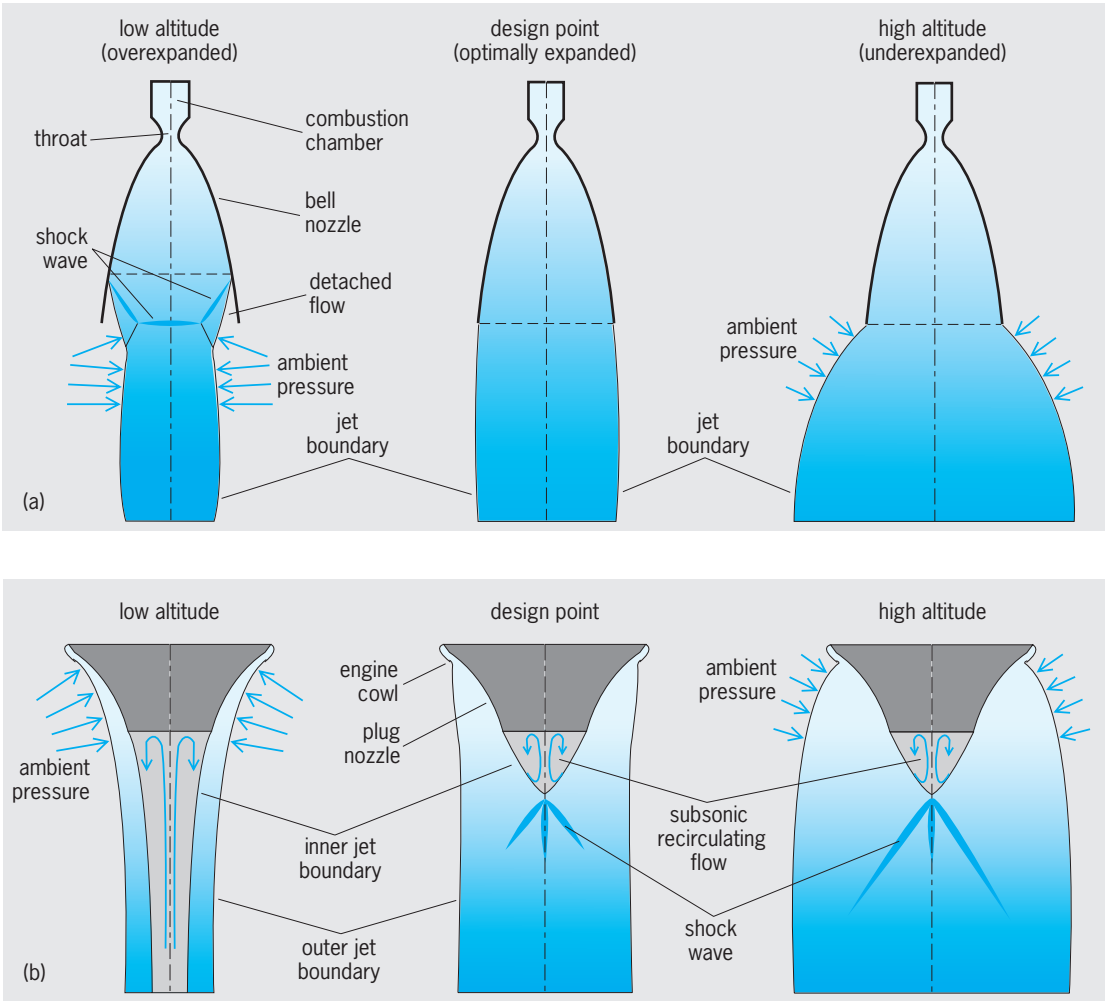


Fig. 3. Nozzle flow comparison. (*a*) Conventional rocket engine (bell nozzle). (*b*) Aerospike engine (plug nozzle). *(After R. A. O'Leary and J. E. Beck, Nozzle design, Threshold no. 8, pp. 34–43. Rockwell International Corporation, Rocketdyne Division, Spring 1992)*

the amount of thrust produced by the spike nozzle drops off very rapidly toward the aft end, a significant portion of the spike nozzle can be eliminated without undue loss in thrust. By injecting the turbine drive gases into the resulting base (Fig. 2), an aerodynamic spike (thus the name aerospike) is formed.

In the combustion tap-off cycle, the turbine that powers the pumps is driven by a small amount of hot gases that are tapped from the combustion chamber. This procedure eliminates the need for an additional gas generator. Once expanded through the turbine, the gases are exhausted into the blunt base of the plug nozzle, increasing the base pressure, thrust, and overall engine efficiency. This arrangement contrasts with that of a conventional rocket engine, where the turbine drive gases are usually ejected through a small secondary nozzle to produce additional thrust.

To match the unique shape of the plug nozzle, the shape of the combustion chamber and the arrangement of the injectors into the aerospike engine are very different from those of the conventional rocket engine. In the conventional engine the combustion chamber is typically cylindrical and the injectors are arranged in the circular forward end of the chamber. In the aerospike engine the combustion chamber is torus shaped (Fig. 2) and the injectors are arranged in an annular pattern.

Advantages. The aerospike engine offers performance, operational, and configurational advantages over the conventional rocket engine equipped with a bell nozzle. First, altitude compensation yields higher performance at low altitude while producing comparable high performance at high altitude. Second, for the same thrust the aerospike is considerably shorter, by as much as 75%. This allows for a shorter-length, lighter-weight vehicle or, for the same-length vehicle, greater volume for higher-capacity propellant tanks or payload. For reusable winged or lifting-body vehicles, the short length moves the center of gravity forward, improving flight characteristics. Third, the altitude-compensating feature of the plug nozzle permits the safe operation of a nozzle with high area ratio at low altitude (sea level) without undesirable flow detachment in the nozzle (Fig. 3*a*) and the accompanying asymmetric lateral forces, or side loads, on the nozzle. Fourth, the aerospike integrates well into the base of a conventional cylindrical launch vehicle with high aspect ratio, reducing

vehicle base drag and eliminating the need for a base heat shield, since the engine fills the entire base. Finally, the thrust load of the aerospike is distributed at the maximum diameter of the engine instead of concentrated at a single point (the gimbal), eliminating the need for a heavy thrust structure to distribute the load through the vehicle, and resulting in a lighter vehicle.

Linear aerospike. A variation is the linear aerospike engine (Fig. 1*b*). This rocket engine concept offers the same performance advantages as the annular aerospike while offering some unique configurational advantages owing to its linear shape. The combustion chamber is made up of a series of modular chamber segments, and the gas generator engine cycle is used in place of the combustion tap-off cycle.

Advanced launch vehicles . During the 1990s, interest has been renewed in single-stage-to-orbit reusable launch vehicles. Numerous studies have shown that reduced launch costs will be best achieved through the development of a fully reusable single-stage-to-orbit vehicle.

Unlike multistage launch vehicles that depend upon one rocket propulsion system for boost and others for high-altitude operation, future single-stage-to-orbit vehicles will be dependent on a single rocket propulsion system from boost to orbit insertion. While each rocket engine of a multistage vehicle can be individually tailored (for example, the nozzle area ratio) to meet the requirements of its portion of the trajectory, rocket engines for single-stage-to-orbit vehicles must provide high performance over the entire flight trajectory. Thus, advanced rocket propulsion technologies that further increase the performance of liquid-propellant rocket engines will be required. The aerospike engine is one of these advanced propulsion concepts. *See* ROCKET PROPULSION; SPACECRAFT PROPULSION. James E. Beck

Bibliography. R. I. Baumgartner and J. D. Elvin, *Lifting Body: An Innovative RLV Concept*, AIAA Pap. 95-3531, 1995; T. K. Mattingly, A simple ride into space, *Sci. Amer.*, 277(4):120–125, October 1997; I. Parker and F. Colucci, Inside-out engine, *Space*, 8(4):38–40, August–September 1992; G. P. Sutton, *Rocket Propulsion Elements*, 6th ed., 1986.

Aerothermodynamics

Flow of gases in which heat exchanges produce a significant effect on the flow. Traditionally, aerodynamics treats the flow of gases, usually air, in which the thermodynamic state is not far different from standard atmospheric conditions at sea level. In such a case the pressure, temperature, and density are related by the simple equation of state for a perfect gas; and the rest of the gas's properties, such as specific heat, viscosity, and thermal conductivity, are assumed constant. Because fluid properties of a gas depend upon its temperature and composition, analysis of flow systems in which temperatures are high or in which the composition of the gas varies (as it does at high velocities) requires simultaneous examination of thermal and dynamic phenomena. For instance, at hypersonic flight speed the characteristic temperature in the shock layer of a blunted body or in the boundary layer of a slender body is proportional to the square of the Mach number. These are aerothermodynamic phenomena.

Two problems of particular importance require aerothermodynamic considerations: combustion and high-speed flight. Chemical reactions sustained by combustion flow systems produce high temperatures and variable gas composition. Because of oxidation (combustion) and in some cases dissociation and ionization processes, these systems are sometimes described as aerothermochemical. In high-speed flight the kinetic energy used by a vehicle to overcome drag forces is converted into compression work on the surrounding gas and thereby raises the gas temperature. Temperature of the gas may become high enough to cause dissociation (at Mach number ≥ 7) and ionization (at Mach number ≥ 12); thus the gas becomes chemically active and electrically conducting. *See* COMBUSTION; HYPERSONIC FLIGHT; JET PROPULSION; MACH NUMBER; ROCKET PROPULSION.

In order to describe aerothermodynamic problems more fully, three specific phenomena are discussed in the following sections: internal flow, external flow, and aerodynamic heating.

Internal flow. In internal flow the gas is confined by the walls of a duct. Aerothermodynamic effects in this case are caused either by gases, such as air at high temperatures, or by combustion. The internal flow of gases at high temperature is a phenomenon largely confined to laboratory equipment, such as the hypersonic wind tunnel, the shock tube, the hot-shot tunnel, or the plasma jet used to simulate flight conditions for testing models. The rocket, ramjet, and the turbojet engine also involve combustion processes in which aerothermodynamic effects are important.

The shock tube consists basically of a long pipe, divided by a diaphragm into two compartments and closed at both ends. Gases at different pressures are placed in the two sections of pipe, the diaphragm is ruptured, and a shock wave propagates into the quiescent low-pressure gas. The gas behind the shock wave is accelerated, compressed, and heated to a high temperature. This region of high-velocity and high-temperature air can be used to simulate high-speed flight conditions. The radiation from the hot gas, chemical kinetics, heat transfer to simple shapes, forces on simple shapes, and boundary-layer transition can all be studied by this experimental device. *See* SHOCK TUBE; SHOCK WAVE.

For the propulsion units normally used in various vehicles that rely on ambient air as a source of oxygen and as a working fluid, the free-stream tube of air entering an inlet must be decelerated to a low subsonic velocity before entering the combustion chamber. In practical applications the deceleration of a supersonic stream (flow traveling faster than the local speed of sound) is not possible without the formation of discontinuities, or shock waves. The formation of shock waves in a stream always results in

an increase in entropy; that is, the available energy in the stream is diminished as the flow proceeds across the shock wave. *See* SUPERSONIC DIFFUSER.

Deceleration of the supersonic airstream to subsonic speed before it enters the combustion chamber is accomplished most efficiently by an oblique shock diffuser. The oblique shock diffuser, which consists of a cone or a wedge, produces a shock wave system of one or more oblique shock waves followed by one normal shock wave. Because the Mach number of the air following a normal shock wave is always subsonic, the downstream air is at the required low velocity. As the upstream Mach number increases, the subsonic Mach number following the normal shock decreases.

In combustion engines aerothermodynamic effects are important in several respects. First there is the diffusion and mixing process of the fuel and oxidizer. When fuel is introduced in liquid form, as in liquid-propellant rocket engines, the combustors of ramjet and turbojet or fanjet engines, and afterburners, the process involves the breakup of the fuel spray, evaporation of the liquid drops, and mixing of fuel and oxidizer. *See* AFTERBURNER; ROCKET ENGINE; SUPERCHARGER; TURBOFAN; TURBOJET.

In the combustion process itself chemical reactions are complex and turbulence is high. One important problem is to stabilize the flame. To produce useful work, the high-temperature products of combustion are expanded, either in a turbine or in a nozzle. *See* GAS TURBINE.

As the flight speed increases into the high supersonic and hypersonic regimes, the diffusion process in the inlet becomes very inefficient. The flow that is slowed down by the normal shock causes an increase in both static temperature and static pressure. As a result of the high static temperature, air begins to dissociate appreciably, causing a substantial energy loss. In addition, both high static temperatures and static pressures may, in turn, require a heavier structure. These problems can be considerably minimized if the combustion process is allowed to occur in the supersonic stream. However, the energy losses due to heat addition processes in a supersonic stream can be appreciable. *See* SUPERSONIC FLIGHT.

In a turbine the design problems are twofold: the analysis of flow over airfoils in cascade, and heat transfer to the turbine blades. Flow is usually treated by ideal-gas techniques. Because the turbine blades are constantly exposed to hot combustion gases, it is necessary to develop materials able to withstand higher temperatures and to devise ways of cooling the turbine blades. *See* TURBINE.

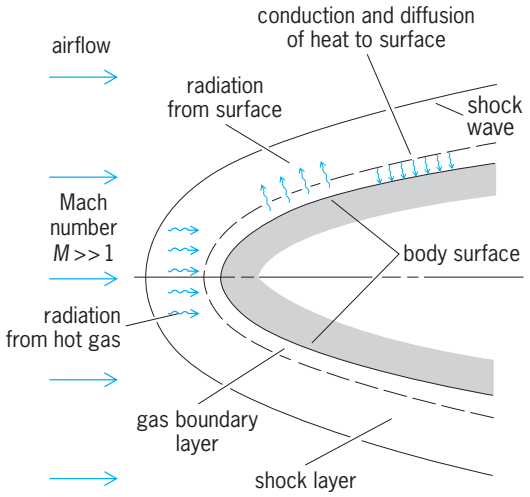
In the nozzle, high-temperature and high-pressure products of combustion are expanded to a high velocity as the temperature and pressure decrease. The flow is complicated by the complex chemical composition of the gas. The efficiency of the nozzle is affected by its contour, the amount of heat lost through its wall, the degree of incomplete combustion, and the presence of nonequilibrium thermal and chemical states, flow separation, shock waves, and turbulence. Heat transfer to the nozzle walls, especially at the throat, is important because of its effect on the structural integrity of the walls. *See* NOZZLE.

External flow. Aerothermodynamic effects in the external flow about bodies occur during high supersonic or hypersonic flight. Any real body has a finite radius of curvature at the leading edge or nose. This radius results in a bow shock, which is detached and nearly normal to the flow at the stagnation region. The air between the normal shock and the stagnation region of the body is compressed to a high temperature. This compressed region is the primary source of high-temperature gas in the external flow about a body. For the purpose of discussion and analysis, the external flow field about a body (between the shock wave and the body surface) is divided into three regions, namely, the inviscid external flow, the viscous boundary layer, and the wake or jet in the base region.

Both the compression and the temperature rise that a gas experiences as it passes through the bow shock increase with Mach number and shock wave angle. The high-temperature gas transfers heat from the boundary layer to the body. A large amount of heat is transferred in the stagnation region of a nose (see **illustration**) and on the leading edge of a wing, not only because the gas pressure is maximum, but also because the boundary layer has a minimum thickness in such regions. *See* NOSE CONE.

In designing a vehicle, it is necessary to apply information obtained from the solution of the equations of the external flow field to such problems as structural design (in the form of applied forces and heat input), guidance and control (in the form of applied forces), and communication and detection (in the form of electromagnetic disturbances). The applied forces are usually determined by solving the aerodynamic equations for an ideal inviscid gas and correcting for real-gas effects. Heat transfer, on the other hand, is estimated by a consideration of the thin viscous boundary layer.

At high temperatures the gas in the shock layer emits electromagnetic radiation. While this phenomenon increases heat transfer, it is more important



Heat transfer in stagnation region of blunted nose cap.

as a source of electromagnetic noise, which can also adversely affect communication with the vehicle. At still higher temperatures the gas ionizes, producing electromagnetic forces in addition to the radiation. *See* MAGNETOHYDRODYNAMICS.

Aerodynamic heating. Aerodynamic heating is a severe problem at high flight speeds and at all altitudes at which air forms a continuum, that is, up to approximately the altitude at which the mean free path is of the same order of magnitude as the diameter of the vehicle's nose. Such an altitude might be about 350,000 ft (107 km) for a nose having a diameter of about 1 ft (0.3 m).

To investigate aerodynamic heating on a missile, aircraft, spacecraft, or other flying object, one needs detailed information about flight profiles, such as the altitude, velocity, angle of attack, and bank-angle history of the vehicle as a function of time. For most supersonic aircraft, the thermal design can be based on the fixed-equilibrium skin temperature (the temperature at which heat input equals heat output). For a missile, equilibrium is not reached, and the transient nature of missile flight trajectories complicates the temperature analysis. For an atmospheric-entry spacecraft, skin temperature will in some cases reach equilibrium because of its longer flight time, but the equilibrium skin temperature may change with time because of changes in altitude, speed, angle of attack, and bank angle. *See* ATMOSPHERIC ENTRY.

The atmospheric-entry characteristics of spacecraft and of ballistic missiles differ in both the total heat load experienced and in maximum heating rates. Total heat load is greater for the spacecraft, whereas the maximum heating rate is greater for the ballistic missile because of its steeper entry path. For spacecraft with a high lift-to-drag ratio ($L/D \approx 3$), the severity of heating can be reduced by trajectory selection, configuration design, and radiation cooling. For spacecraft with medium and low lift-to-drag ratios ($L/D \approx 2$ or less) and for the ballistic entry vehicle, ablation cooling techniques have been used extensively to protect interior structure. Changing the nose configuration alone cannot completely solve the structural heating problem for a ballistic missile. *See* SPACECRAFT STRUCTURE.

In general, heating analysis is conducted by dividing the vehicle into components; appropriate heat-transfer formulas are then applied to each component to estimate its heating rate. Analysis can be further complicated by boundary-layer transition, boundary-layer interaction, separation flow, shock impingement, mass transfer, nonequilibrium phenomena, gap effects, and rough or wavy surfaces.

Accurate estimates of aerodynamic heating also require knowledge of fluid properties of the gas, such as composition, thermal and transport properties, reaction rates, relaxation times, and radiation characteristics within the fluid around the hypersonic vehicle. Under standard conditions air is composed chiefly of diatomic molecules. Motions of translation and rotation are excited at room temperature. As the temperature of air increases, vibration begins, increasing in amplitude as temperature increases, until eventually the intramolecular bond is broken. The molecule is then said to be dissociated. Still higher energy levels are excited as the temperature increases further; eventually, electrons are separated from their parent atoms and the gas becomes ionized.

Shih-Yuan Chen

Bibliography. J. D. Anderson, Jr., *Hypersonic and High Temperature Gas Dynamics*, 1989, reprint, 2000; J. D. Anderson, Jr., *Modern Compressible Flow*, 3d ed., 2003; J. J. Bertin, *Hypersonic Aerothermodynamics*, 1994; G. Oates, *Aerothermodynamics of Gas Turbine and Rocket Propulsion*, 3d ed., 1997; C. Park, *Nonequilibrium Hypersonic Aerothermodynamics*, 1990.

Affective disorders

A group of psychiatric conditions, also known as mood disorders, that are characterized by disturbances of affect, emotion, thinking, and behavior. Depression is the most common of these disorders; about 10–20% of those affected by depression also experience manic episodes (hence this condition is known as manic-depression or bipolar affective disorder). The affective disorders are not distinct diseases but psychiatric syndromes that likely have multiple or complex etiologies.

Clinical Syndromes

Affective disorders are distinguished from states of normal sadness or elation by several features. Their clinical syndromes consist of characteristic changes in brain and psychological function that affect sleep, appetite, speed of movement, energy, concentration, self-esteem, motivation, and hedonic (pleasure) capacity. During an episode, these features are persistent, occurring day after day for at least several weeks. Affective syndromes also differ from normal mood states by the degree of associated impairment of functioning. The most severe episodes include hallucinations or delusions (unshakable false beliefs of obvious personal significance, typically reflecting themes that are congruent with the mood disturbance).

A natural parallel to the clinical syndrome of depression is the state of grief. Grief is associated with many of the symptoms of depression, although the changes of mood and behavior that are associated with grief are less persistent and more circumscribed. Conversely, new romantic love or an infatuation is a natural parallel of a hypomanic episode (that is, a mild episode of mania), although again the normal "affliction" is less persistent and pervasive and should not cause functional impairment. Unresolved grief can progress into clinical depression, however, and there is a pathological form of delusional romantic love that is called erotomania.

Aside from the death of a loved one, other types of losses, including divorce, the breakup of a romantic relationship, or setbacks in vocational pursuits, commonly precede the onset of an initial depressive episode. A transitional diagnosis of an adjustment disorder is used when the duration or severity of

depressive symptoms has not crossed the threshold necessary for diagnosis of a depressive disorder. A majority of adjustment disorders, like normal grief, remit as the individual adapts to the stressor.

Major depressive disorders. The most common form of affective disorder is a major depressive episode. The episode is defined by a pervasively depressed or low mood (experienced most of the day over a period of 2 weeks or longer) and at least four associated symptoms affecting sleep, appetite, hedonic capacity, interest, and behavior. Major depressive episodes have several clinical forms.

Melancholia is a severe episode characterized by anhedonia (inability to experience pleasure), markedly diminished appetite (anorexia; note the anorexia of depression is *not* related to anorexia nervosa), with weight loss, early morning awakening, observable motor disturbances (extreme slowing, or retardation, or pacing and stereotypic agitated behaviors), and diurnal mood variation (mood is worse in the morning). Some episodes of melancholia are associated with delusions, typically of guilt, sin, illness, punishment, or nihilism.

Common among young patients, especially women, is a milder syndrome historically referred to as atypical depression. Atypical depression is characterized by a less pervasive mood disturbance, with mood reactivity preserved (that is, the patient's spirits go up or down in response to day-to-day events) and by so-called reverse symptoms: oversleeping, overeating, or gaining weight. Significant anxiety symptoms, including phobias and panic attacks, also are common in atypical depression.

Both atypical and other milder forms of major depressive disorder were formerly referred to as reactive or neurotic because psychosocial factors often appeared to play a prominent role in their onset. Melancholia, by contrast, is more autonomous and less clearly linked to life stress. In fact, melancholia was formerly called endogenous depression because the cause was presumed to be within the brain. Current thinking on the roles of life events and brain-related processes is less simplistic: it is likely that both sets of factors are relevant to all forms of affective disorder.

Dysthymia. A more chronic, insidious form of depression known as dysthymia "smolders" at a subsyndromal level (that is, there are three or four daily symptoms rather than five or more) for at least 2 years. Dysthymia often begins early in life and, historically, has been intertwined with atypical and neurotic characteristics of depression. There may be a depressive personality or temperament that underlies childhood-onset dysthymia. Most people with untreated dysthymia eventually develop major depressive episodes and, despite its milder symptoms, dysthymia causes considerable impairment.

Mania and bipolar affective disorder. About 10–20% of people who experience major depressive episodes will also suffer from abnormal periods of elation, which are referred to as manias (or in less severe form, hypomanias). A manic episode is heralded by euphoric or irritable mood and at least four of the following symptoms: increased energy, activity, self-esteem, or speed of thought; decreased sleep; poor judgment; and increased risk-taking. The energy, drive, and optimism of a hypomanic individual can seem captivating to others. About one-half of manic episodes are psychotic. The delusions of mania typically reflect grandiose or paranoid themes.

Almost everyone who experiences manic episodes also will suffer from recurrent depressive episodes. Since there is a 50-50 chance that an episode of depression may precede the onset of the first mania, changes in diagnosis are common.

The term bipolar affective disorder has largely replaced the former term manic-depression, although both names convey the cyclical nature of this illness. The classical presentation (which includes full-blown manic episodes) is known as type I disorder. The diagnosis of bipolar type II disorder is used when there are recurrent depressive episodes and at least one hypomania. A third diagnosis, cyclothymia, is used when neither hypomanias nor depressions have reached syndromal levels. For some, bipolar II or cyclothymic disorders subsequently evolve into the more severe type I form of illness.

Several important variations of bipolar disorder are increasingly recognized.

Mixed episode. A mixed episode is diagnosed when the symptoms of mania and depression coexist. Mixed episodes are diagnostically and therapeutically challenging, and their existence suggests that mania and depression are not caused by polar-opposite abnormalities of brain functioning. Mixed episodes are often associated with sleep deprivation or drug and alcohol use.

Rapid cycling. The term rapid cycling is used when there have been four or more episodes within a time frame of 1 year. Cases of ultrarapid cycling (mood cycles occur within hours or days) also have been reported. Rapid cycling is more common among women, and may be provoked by antidepressant therapies. Thyroid disease and the bipolar II subtype also increase the likelihood of rapid cycling.

Seasonal affective disorder. A number of affective disorders follow a seasonal pattern. A pattern of recurrent fall/winter depressions (also known as seasonal affective disorder) has generated considerable interest because it may be treated with bright white (full-spectrum) light, which artificially lengthens the photoperiod.

Disorders associated with general medical conditions. Both depression and mania can be caused by general medical illnesses and medications that affect brain function (such as antihypertensives, hormonal therapies, steroids, and stimulant drugs). The diagnosis "mood disorder associated with a general medical condition" is applied to these conditions. Recognition that the mood disorder is linked to another general medical illness can have important treatment implications.

Pathophysiology

Affective disorders have diverse biopsychosocial underpinnings that result, at least in part, in extreme or distorted responses of several neurobehavioral systems. The neurobehavioral systems of greatest

relevance regulate a person's drives and pursuits, responses to acute stress, and capacity to dampen or quiet pain or distress. The system that regulates the daily rhythms of sleeping and waking and associated patterns of hormonal secretion likewise helps to ensure well-being. The coordinated function of these response and homeostatic systems conveys important and, from an evolutionary standpoint, ancient advantages for survival. Despite obvious overlap with how stress can adversely affect other primates and lower mammals, the affective disorders are distinctly human conditions. They may be thought of as one disadvantageous consequence of the complexity of affective, cognitive, and social functioning that distinguishes humans from their phylogenetic relatives.

Although there is considerable evidence that affective disorders are heritable, vulnerability is unlikely to be caused by a single gene. Bipolar disorder is clearly more commonly inherited than major depressive disorders, and among the major depressive disorders the more severe, psychotic, and early-onset forms are more heritable than the remainder. Such risks are relative, however, and even identical twins have no more than a 50–60% risk of concordance of bipolar disorder. It is likely that some combination of genes conveys greater risk and, like an amplifier, distorts the neural signals evoked by stress and distress. *See* BEHAVIOR GENETICS; HUMAN GENETICS.

Research permits several firm conclusions about brain neurochemistry in stress and depression. Acute stress affects the release of three vital brain monoamines—serotonin, norepinephrine, and dopamine—as well as glucocorticoids such as cortisol. Cortisol release is "driven" by the neuropeptide corticotrophin releasing hormone, which is released from neurons in both the hypothalamus and limbic cortex. Sustained unresolvable stress eventually depletes the neurotransmitters (cortisol levels remain high), inducing a behavioral state that has been called "learned helplessness." In severe recurrent episodes of depression, especially those with psychotic features, cortisol elevations approach those seen in Cushing's disease. Sustained elevations of cortisol have been shown to inhibit the growth of new neurons (neurogenesis) and ultimately may cause atrophy of some regions of the brain. *See* ADRENAL CORTEX HORMONE; BRAIN; STRESS (PSYCHOLOGY).

As described earlier, both depression and mania are characterized by alterations in the sleep-wake cycle, including changes in nocturnal body temperature, growth hormone secretion, and the electroencephalographic patterns that define the stages of sleep. In depression, core body temperature tends to be elevated, growth hormone release is blunted, and there are increased awakenings. The brain's patterns of electrical activity are altered, with decreased deep (slow-wave) sleep and increased rapid eye movement (REM) sleep in individuals with mania and depression. These changes suggest increased nocturnal arousal, perhaps because of decreased inhibitory serotoninergic tone. *See* SLEEP AND DREAMING.

Positron emission tomography (PET) studies of brain function in depression document decreased prefrontal cortex metabolism and increased activity of deeper, paralimbic structures such as the amygdala. The former disturbance may explain the cognitive abnormalities of severe depression, whereas the latter may reflect intense affective arousal.

Mania, by virtue of the increased activity and characteristic changes in judgment and cooperativeness, has proved more difficult to study. In the prevailing view, it is a state of increased catecholamine activity, perhaps superimposed on a deficit of inhibitory serotoninergic input. Repeated studies of patients with bipolar affective disorder suggest that cerebral blood flow and glucose metabolism increase in prefrontal regions following a shift from depression to hypomania.

Psychosocial and neurobiologic vulnerabilities, no doubt, intersect. For example, harsh early maltreatment, neglect, or other abuses can have lasting effects on both self-concept and brain responses to stress. Low self-esteem, in turn, can amplify the impact of a particular adverse event for an individual. Conversely, blunted hedonic capacity, perhaps mediated by low serotonin and catecholamine levels, decreases coping responses and the willingness to seek out new rewards. Similarly, chronic hyperarousal undoubtedly reinforces caution and more avoidant and dependent behaviors.

Epidemiology

The lifetime rates of affective disorders are increasing, with an earlier age of onset in more recent generations. The risks in the United States are for major depressive disorder, 10–15% of individuals; dysthymia, 3–5%; bipolar disorder type I, 1%; and bipolar disorder type II, 1–3%. Among these conditions, only bipolar disorder type I has a comparable incidence among men and women. For the remainder, female preponderance (1.5 or 2 to 1 ratios) is observed in most western cultures, which probably reflects both psychosocial and biological vulnerabilities. Minor depressive symptoms, including insomnia, constitute a major longitudinal risk factor. Other risk factors include divorced or single marital status and concomitant psychiatric, general medical, or substance-abuse disorders.

The onset of major depression most often occurs in those in their late 20s to mid-30s; dysthymia and bipolar disorder typically begin about a decade earlier. However, no age group is immune to an affective disorder. Vulnerability is not strongly related to social class or race, although the affluent are more likely to receive treatment.

Worldwide, the affective disorders are underrecognized and undertreated. In the United States, less than one-third of depressed people are receiving proper treatment, and only 60% of those with bipolar disorder are treated at any given time. Untreated depressive episodes tend to last months longer, increasing the risk of chronicity. Once chronicity is established, spontaneous remission is uncommon and episodes may last decades.

About 70% of the first-onset major depressive episodes eventually become recurrent. After three or more episodes, the risk of further episodes (without

preventive treatment) exceeds 90%. This apparently increasing risk of recurrence is presumed to result from changes in brain stress responses referred to as "kindling."

The affective disorders, on average, cause more day-to-day disability than arthritis, diabetes, and most other common chronic illnesses. Depression's effects on the quality of life essentially match that of congestive heart failure or chronic obstructive pulmonary disease. The World Health Organization ranks depression as the world's fourth greatest health problem and estimates a second-place ranking in 2020. Bipolar disorder is ranked as the eighth greatest cause of illness burden. The lifetime risk of suicide is about 8–10% for the depressive disorders and 10–15% for bipolar disorder.

Treatment

Most episodes of dysthymia and major depressive disorder respond to treatment with either psychotherapy or antidepressant medication, either singly or in combination. Many experts now recommend the newer forms of psychotherapy, including cognitive-behavioral therapy and interpersonal therapy, because they have been better studied than more traditional psychoanalytic therapies and because they have been found to be as effective as medications.

Nearly 30 antidepressant medications are available worldwide, with most falling into three classes: tricyclic antidepressants (TCAs), selective serotonin reuptake inhibitors (SSRIs), and monoamine oxidase reuptake inhibitors (MAOIs). Most classes of antidepressants enhance the efficiency of serotonin or norepinephrine neurotransmission. Antidepressants are not habit-forming and have no mood-elevating effects for nondepressed people. *See* MONOAMINE OXIDASE; NORADRENERGIC SYSTEM; SEROTONIN.

The classes of antidepressants are not interchangeable, and a poor response or intolerable side effects with one medication do not preclude a good response to another. The SSRIs [in the United States, fluoxetine (Prozac®), sertraline (Zoloft®), paroxetine (Paxil®), citalopram (Celexa®), and escitalopram (Lexapro®)], venlafaxine (Effexor®), and bupropion (Wellbutrin®) are most often prescribed as first-line treatments because of their tolerability, safety, and relative simplicity. It is hard to predict which depressed patient will respond best to which particular antidepressant. When treatment is effective, an improvement should be apparent within 4–6 weeks. An effective antidepressant should be maintained for 6–9 months to reduce the risk of relapse. Even longer prophylactic treatment is recommended for prevention of highly recurrent depressions. *See* PSYCHOPHARMACOLOGY; PSYCHOTHERAPY.

There has been recent controversy about the possibility that antidepressants might paradoxically increase the likelihood of suicidality, especially early in the course of therapy. Although a cause-and-effect relationship has not been established, it is true that a small percentage of people who begin taking antidepressants (on the order of 1–3%) do feel worse and report either the onset, or worsening, of suicidal thoughts or actions. Accordingly, regulatory authorities such as the U.S. Food and Drug Administration have required that the manufacturers of antidepressants include warnings about the possibility of treatment-emergent suicidality and the need for careful monitoring. A related issue concerns the use of antidepressants in children and teenagers. In brief, younger depressed people may be at greater risk for treatment-emergent suicidality and, in youth, the therapeutic benefits of antidepressant medications are not as well established.

Acute manic episodes are usually treated with lithium salts, divalproex sodium, or one of the newer atypical antipsychotic medications, singly or in combination. The newer class of antipsychotic medication includes olanzapine (Zyprexa®), risperidone (Risperdal®), quetiapine (Seroquel®), Ziprasidone (Geodon®), and aripiperazole (Ability®). The potent sedative benzodiazepines, such as lorazepam or clonazepam, also are used to control agitation and insomnia. Several other drugs, notably carbamazepine, are used when the primary agents are ineffective.

Successful treatment of a manic episode should usually lead to preventive therapy. In addition to the medications used to treat mania, the anticonvulsant lamotrigine (Lamictal®) is used for preventive care. Relapses that occur despite preventive treatment may require the use of more complex, multidrug regimens. Although psychotherapy does not have a major role in the acute treatment of mania, it may help people come to terms with their illness, cope more effectively with stress, or curb minor depressive episodes. Antidepressants also are used to treat "breakthrough" major depressive episodes.

When pharmacotherapies are not effective, the oldest proven treatment of the affective disorders, electroconvulsive therapy (ECT), still provides a powerful alternative. Today, ECT is a highly modified and carefully monitored treatment that has little in common with its depictions in the movies. Nevertheless, confusion and transient amnesia are still common short-term side effects. Vigorous pharmacotherapy is needed after successful treatment to lessen the risk of relapse. *See* ELECTROCONVULSIVE THERAPY.

Michael E. Thase

Bibliography. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. (DSM-IV), 1994; American Psychiatric Association, Practice guideline for the treatment of patients with major depressive disorder (revision), *Amer. J. Psychiat.*, 157 (4 suppl.):1–45, 2000; F. K. Goodwin and K. R. Jamison, *Manic-Depressive Illness*, Oxford University Press, 1990; R. M. A. Hirschfeld et al., Practice guideline for the treatment of patients with bipolar disorder (revision), *Amer. J. Psychiat.*, 159(suppl. 4): 4:4–50, 2002; S. D. Hollon, M. E. Thase, and J. C. Markowitz, Treatment and prevention of depression, *Psychol. Sci. Public Interest*, 3(2):39–77, 2002; M. E. Thase, R. Jindal, and R. H. Howland, Biological aspects of depression, in I. H. Gotlib and C. L. Hammen (eds.), *Handbook of Depression*, pp. 192–218, Guilford Press, New York, 2002.

Aflatoxin

Any of a group of secondary metabolites produced by the common molds *Aspergillus flavus* and *A. parasiticus* that cause a toxic response in vertebrates when introduced in low concentration by a natural route. The group constitutes a type of mycotoxin. Discovered in 1960 after a massive poisoning of turkey poults fed moldy peanut meal, aflatoxin has become the focus of a highly interdisciplinary research field involving chemists, mycologists, veterinarians, agriculturalists, toxicologists, and other basic and applied scientists. *See* TOXIN.

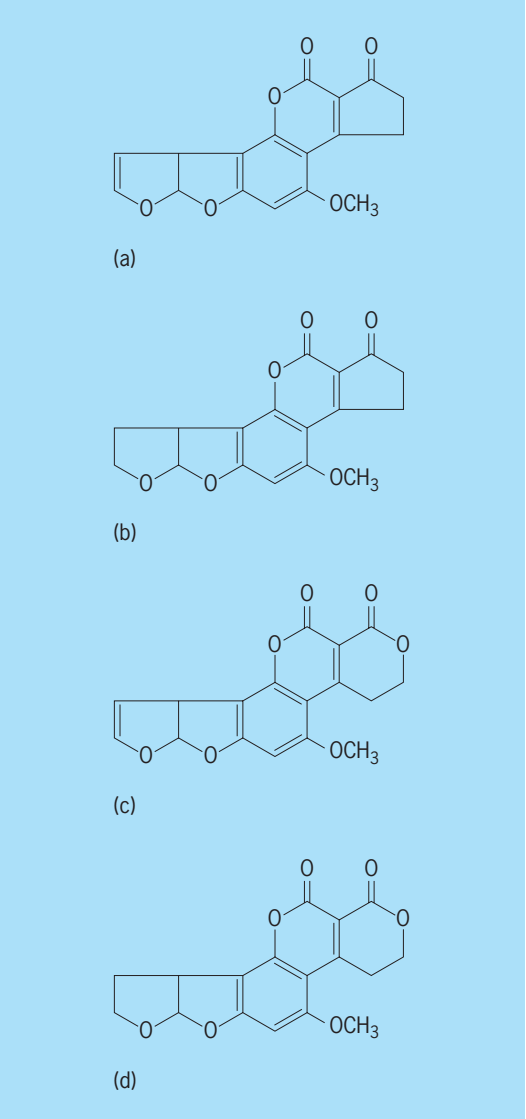
Chemistry. The naturally occurring aflatoxins are identified in physicochemical assays as intensely blue (aflatoxins B_1 and B_2) or blue-green (aflatoxins G_1 and G_2) fluorescent compounds under long-wave ultraviolet light. The common structural feature of the four major aflatoxins is a dihydrodifurano or tetrahydrodifurano group fused to a substituted coumarin group (see **illus.**). The relative proportions of the four major aflatoxins synthesized by *Aspergillus* reflect the genetic constitution of the producing strain and the parameters associated with fungal growth. In addition, derivative aflatoxins are produced as metabolic or environmental products.

Aflatoxins are formed through a polyketide pathway involving a series of enzymatically catalyzed reactions. In laboratory cultures, aflatoxins are biosynthesized after active growth has ceased, as is typical for secondary metabolites. By using blocked mutants and metabolic inhibitors, many of the intermediates have been identified as brightly colored anthraquinones. Working with these colored metabolites has made studies of aflatoxin biosynthesis into a model system for polyketide research and has also facilitated molecular genetics. Several genes encoding enzymes that catalyze steps in the pathway or affect more global regulatory mechanisms have been cloned by using complementation studies in which transformants are selected based on pigmentation.

Biological effects. Aflatoxins are potent molecules with many biological effects. They are toxigenic, carcinogenic, mutagenic, and teratogenic in various animal species. Aflatoxin B_1 is usually the most abundant naturally occurring member of the family, and most studies on the pharmacological activity of aflatoxin have been conducted with this congener.

Interspecific and interstrain variation is the rule for both acute and chronic effects. Trout and ducklings are among the most sensitive species known. The toxic properties manifest themselves in different ways depending on the test system used, the dose given, and the duration of exposure. In addition, age, nutritional status, and sex affect the toxicological response; in general, the young of a species are the most sensitive.

In veterinary medicine, aflatoxicosis is often difficult to detect. Inhibited growth, weight loss, and immune response breakdown are observed, and subtle synergies with both infectious and nutritional diseases complicate the diagnostic profile. The main target organ for aflatoxin is the liver. Aflatoxin B_1 is the most potent hepatocarcinogenic agent known, although the liver by no means is the only organ susceptible to aflatoxin carcinogenesis.



Structures of major naturally occurring aflatoxins. (*a*) B_1. (*b*) B_2. (*c*) G_1. (*d*) G_2.

Epidemiological studies in Africa and Asia link ingestion of mold-contaminated food with incidence of primary liver cancer. Moreover, independent studies on liver cancer from southern Africa and Qidong, China, revealed guanine to thymine transversions in a tumor suppressor gene. Identification of this molecular hot spot constitutes some of the most convincing evidence that a natural product in a nonindustrial setting can induce a specific form of cancer. Aflatoxin is listed as a probable human carcinogen by the International Agency for Research on Cancer. *See* BOTULISM; LIVER DISORDERS; MUTAGENS AND CARCINOGENS; PLANT PATHOLOGY.

Economics. Aflatoxins are a major agricultural problem. Contamination can occur in the field, during harvest, or in storage and processing. Corn, rice, cottonseed, and peanuts are the major crops

regularly displaying high levels of aflatoxin contamination. Since *A. flavus* and *A. parasiticus* are nearly ubiquitous in the natural environment, numerous other grain, legume, nut, and spice crops, as well as coffee and cocoa, have been reported to contain aflatoxins. Given the potential of aflatoxins as human carcinogens and their known activity as toxins in animal feeds, many international regulatory agencies monitor aflatoxin levels in susceptible crops.

Contaminated agricultural commodities typically contain low amounts of aflatoxins (in the parts per billion or parts per million range) so numerous methods have been devised for extraction, concentration, and purification. Since agricultural commodities are diverse, no single extraction procedure can be applied uniformly to every product. Most of the original analytical methods employ thin-layer chromatography or high-performance liquid chromatography. More recently, immunological assays have become popular because they require fewer extractions and less complicated equipment.

Prevention is the main line of defense against aflatoxins entering the food chain. Moisture, temperature, and composition of the substrate are the chief factors affecting fungal growth and toxin production. In the field, insect damage is often involved. Good agronomic techniques that minimize damage to crops, storage under conditions unfavorable to mold growth, and screening of crops in order to exclude those which have been contaminated are the major methods for protecting the food supply. In some cases, portions of crops may be salvaged by mechanical separation. Detoxification is a last line of defense. Several commercially feasible methods of ammoniation have been developed for reducing levels of aflatoxin contamination in animal feeds.

Safety. Laboratory workers handling aflatoxins, especially in purified forms, should maintain strict safety precautions. Aflatoxin-contaminated foods should not be eaten or fed to domesticated animals. *See* AGRONOMY; MYCOTOXIN. J. W. Bennett

Bibliography. J. W. Bennett and M. A. Klich (eds.), *Aspergillus: Biology and Industrial Applications*, 1992; R. J. Cole and R. H. Cox, *Handbook of Toxic Fungal Metabolites*, 1981; J. E. Smith and M. S. Moss, *Mycotoxins: Formation Analysis, and Significance*, 1985.

Africa

A continent that straddles the Equator, extending between 37°N and 35°S. It is the second largest continent, exceeded by Eurasia. The area, shared by 55 countries, is 11,700,000 mi^2 (30,300,00 km^2), approximately 20% of the world's total land area. Despite its large area, it has a simple geological structure, a compact shape with a smooth outline, and a symmetrical distribution of climate and vegetation.

Structure and geology. The structural makeup is partially explained by the continental drift theory, in which all of the Earth's continents were previously assembled into one supercontinent called Pangaea (or Pangea). Over time, Pangaea broke into two landmasses, Laurasia and Gondwana, and subsequently into several continents. Africa occupied the core of Pangaea, and as neighboring landmasses drifted away, the present structure and configuration of the continent evolved. *See* CONTINENTAL DRIFT; CONTINENTS, EVOLUTION OF; PLATE TECTONICS; SUPERCONTINENT.

Three major rock formations can be identified. The first is a massive crystalline shield (Precambrian) that underlies most of the continent, with occasional outcrops in about one-third of the area. In some places, these rocks have been subject to metamorphic processes, resulting in a rich variety of mineral resources. The second set of significant rocks is the Karoo series in southern and eastern Africa. These are sedimentary rocks varying in age from the Carboniferous to the lower Jurassic Period. They are presently vital sources of coal. A third set of rock formations comprises marine sediments stretching in a discontinuous belt from the northwest to the Gulf of Guinea. These were formed during the Cretaceous Period, when parts of the continent were submerged in water. The sediments deposited at the time contain hydrocarbons that now account for oil production in Libya, Algeria, Nigeria, Gabon, and Angola. *See* CARBONIFEROUS; CRETACEOUS; JURASSIC; PRECAMBRIAN.

Africa has few inlets or natural harbors and a small number of offshore islands that are largely volcanic in origin. Madagascar is the largest island, with an area of 250,000 mi^2 (650,000 km^2). It is separated from the mainland by the Mozambique Channel, which is about 250 mi (400 km) wide. Close to Madagascar is the Mascarene, a volcanic ridge submerged in the Indian Ocean. Its highest peaks are exposed above the sea surface to form the islands of Seychelles, Mauritius, and Réunion. North of Madagascar, there is another group of volcanic islands, the Comoros. The largest of these is Njazide (Grand Comore), with an active volcano, Mount Kartala (7677 ft; 2340 m). *See* INDIAN OCEAN.

Surface features. Africa is primarily a high interior plateau bounded by steep escarpments. These features show evidence of the giant faults created during the drift of neighboring continents. The surface of the plateau ranges from 4000–5000 ft (1200–1500 m) in the south to about 1000 ft (300 m) in the Sahara. These differences in elevation are particularly apparent in the Great Escarpment region in southern Africa, where the land suddenly drops from 5000 ft (1500 m) to a narrow coastal belt. Although most of the continent is classified as plateau, not all of its surface is flat. Rather, most of its physiographic features have been differentially shaped by processes such as folding, faulting, volcanism, erosion, and deposition. *See* ESCAPEMENT; FAULT AND FAULT STRUCTURES; PLATEAU.

Basin surfaces. In several areas the plateau has sagged under the weight of accumulated riverine sediments to form large structural basins. These include the Sudan, Chad, and Djouf basins in the north, Congo (Zaire) basin in central Africa, and the Kalahari basin

in the south. These basins play a significant role in the hydrography of the continent. *See* BASIN; HYDROGRAPHY.

Rift valley system. The old crustal surfaces in the eastern part of the continent have been disrupted by a major structural feature known as the East African rift valley system that consists of a complex set of troughs formed by the downward displacement of land blocks against one another. The origin of the rift valley system can be traced to movements associated with separation of the Arabian plate from the African plate. The resulting landscape emanates from the Red Sea and passes through the Ethiopian highlands, along the course of the Awash river and a number of lakes (including Zwai, Shala, Abaya, and Turkana) into western Kenya. From there, it cuts across the Kenyan highlands and a series of lakes from Lake Banyo to Lake Magadi. It then enters Tanzania via Lake Natron, after which it loses its trough-like appearance as a result of the accumulation of sediments and volcanic ash. The system reappears in its original form as it passes through Lake Malawi and the Shire river into the lower Zambezi river and finally enters the Indian Ocean. *See* RED SEA.

The western arm of the rift valley starts north of Lake Albert (Lake Mobutu) in Uganda and passes through Lakes Edward (Amin), Kivu, and Tanganyika into Lake Nyasa, with a trench running southwest as the Luangwa valley of Zambia. Lake Tanganyika is the continent's deepest lake (5100 ft; 1500 m) with a bottom that is roughly 0.5 mi (800 m) below sea level.

The rift valley system is one of the most striking features of the African landscape. Sliding blocks have created wide valleys 20–50 mi (30–80 km) wide bounded by steep walls of variable depth and height. Within the eastern and western branches of the system, there is a large but shallow depression occupied by Lake Victoria. Between Lake Edward and Lake Albert on the western extension is the ice-capped Ruwenzori range, a block mountain, that rises over 16,732 ft (5100 m). These areas, along with other parts of the continent, occasionally experience earthquakes; the largest earthquake in the twentieth century occurred near the southern end of Lake Tanganyika. *See* RIFT VALLEY.

Volcanic landforms. Several volcanic features are associated with the rift valley system. The most extensive of these are the great basalt highlands that bound either side of the rift system in Ethiopia. These mountains rise over 10,000 ft (3000 m), with the highest peak, Ras Dashan, reaching 15,158 ft (4500 m). There are also several volcanic cones, including the most renowned at Mount Elgon (14,175 ft; 4321 m); Mount Kenya (17,040 ft; 5194 m); and Mount Kilimanjaro, reaching its highest point at Mount Kibo (19,320 ft; 5889 m). Mounts Kenya and Kilimanjaro are permanently snowcapped. *See* BASALT.

There are spectacular volcanic landforms in other parts of the continent such as the Ahaggar (Hoggar) mountains in central Sahara, which has over 300 lava plug domes rising above 1000 ft (305 m). There are also huge craters or calderas scattered in various places such as Trou au Natron in the Tibesti mountains and the Ngorongoro in Tanzania. Present-day active volcanoes include Oldoinyo Lengai in Tanzania; Nyamlagira and Nyriragongo in the Virunga range on the Congo (Zaire)–Uganda border; Erta Ale in the Afar Triangle of Ethiopia and on several offshore islands such as San Juan on Las Palmas islands; Fogo on Cape Verde islands; Kartala on Comoros islands; and Piton de la Fournaise on Réunion island. The largest active volcano is Mount Cameroon (13,350 ft; 4069 m). Although it lies on the mainland, it is part of a chain of volcanoes extending into the Atlantic Ocean. The islands of Fernando Po, Principe, and São Tomé are also part of this system. *See* OCEANIC ISLANDS; VOLCANO.

Fold mountains. The northwestern and southern margins of the continent have been subject to folding and subsequent erosion, resulting in a complex chain of mountains and valleys. The Atlas ranges in the northwest extend in a belt up to 200 mi (320 km) wide from southern Morocco to Tunisia, a distance of 1400 mi (2250 km). The system consists of a series of parallel mountains built by folding and uplifting processes during the Tertiary Period. Geologically they are an extension of the Alpine system of Europe, now separated by the Mediterranean Sea. The two principal chains are found along the coast and in the interior. The coastal chain starts from Tangiers in Morocco as the Rif, extends eastward into Algeria as the Tell, and ends in Tunisia. The interior chain starts west of Morocco as the High Atlas and Anti-Atlas. It continues eastward into Algeria as the Sahara Atlas and terminates in Tunisia. Between the two chains lies the Plateau of Shott, a salty shallow basin with internal drainage. Another plateau, the Moroccan Meseta, is found in the western extreme, sloping toward the Atlantic Ocean. *See* TERTIARY.

The Cape ranges in southern Africa are less complex, consisting of folded sedimentary rocks that correspond in age and topography to Appalachia in the eastern United States. The mountains run in a north-south direction along the west coast for about 150 mi (240 km) and then make right angles to run east-west for another 600 mi (900 km). *See* OROGENY; SEDIMENTARY ROCKS.

Climatic factors. Since the Equator transects the continent, the climatic conditions in the Northern Hemisphere are mirrored in the Southern Hemisphere. Nearly three-quarters of the continent lies within the tropics and therefore has high temperatures throughout the year. Frost is uncommon except in mountainous areas or some desert areas where nighttime temperatures occasionally drop below freezing. These desert areas also record some of the world's highest daytime temperatures, including an unconfirmed record of 136.4°F (58°C) at Azizia, Tripoli. *See* EQUATOR.

Africa is largely under the influence of subtropical high-pressure belts separated by the equatorial low-pressure zone known as the Inter-Tropical Convergence Zone (ITCZ). The position of this zone across the tropics shifts with the seasonal movement of the Sun. In July, it moves to North Africa along the

margins of the Sahara. In January, it moves south, taking up a clearly defined position in West Africa. Its position in eastern and southern Africa is not well defined because of the elevational variations there, but it tends to slope along the eastern margins of the Democratic Republic of Congo (formerly Zaire) and eastward through Mozambique.

The Inter-Tropical Convergence Zone plays a fundamental role in the circulation of the airstreams and subsequent climatic conditions within the continent. The associated air masses are the tropical maritime and the tropical continental air with notable differences in humidity due to their source areas. The tropical maritime air is warm and moist, originating from the Atlantic Ocean. The tropical continental air is warm and dry because it develops over the desert areas. In January, following the southern movement of the Inter-Tropical Convergence Zone, most of the Northern Hemisphere falls under the influence of tropical continental air, bringing in stable dry weather and dusty haze known as the harmattan in West Africa. In July, the shifting Inter-Tropical Convergence Zone and the tropical maritime air take over the region, bringing in heavy rainfall. *See* AIR MASS; CLIMATOLOGY; TROPICAL METEOROLOGY.

The wind reversals associated with the Inter-Tropical Convergence Zone produce distinct monsoonal seasons along the coastal margins of eastern and western Africa. The conditions observed along the eastern margins (including the island of Madagascar) are part of the larger Asian monsoonal system. During the wet summer monsoon, the northern shift of the Inter-Tropical Convergence Zone results in moist onshore winds from the warm tropical Indian Ocean. The atmospheric conditions are reversed in January, when dry offshore winds prevail following the southern shift of the Inter-Tropical Convergence Zone. *See* MONSOON METEOROLOGY.

The West African monsoon is smaller in magnitude than the Asian monsoonal system observed in eastern Africa. It dominates within 400 mi (650 km) of the coastline, particularly along the Gulf of Guinea. The causal factors are similar, however: the shifting wind belts produce alternate wet and humid conditions in the summer and dry and dusty conditions in winter.

Finally, the convergence of surface winds along the Inter-Tropical Convergence Zone is also responsible for the Easterly Wave, which occurs when the Inter-Tropical Convergence Zone is displaced away from the Equator, usually between 5° and 20° latitude. The weak wave is an area of low pressure that is propelled westward within the trade wind belt. Given the appropriate surface and weather conditions, these systems can develop into powerful, well-organized tropical disturbances that eventually grow into hurricanes. Most of the tropical storms and hurricanes in the northern Atlantic are linked to the Easterly Wave originating off the western coast of Africa. However, not all storms grow into hurricanes. Hurricanes typically develop where surface water temperatures are warm, 79°F (26°C) or higher, over a large area with sustained winds exceeding 74 mi/h (119 km/h).

Regional distribution of climate, vegetation, and soils. Africa can be classified into broad regions based on the climatic conditions and their associated vegetation and soil types. The tropical rainforest climate starts at the Equator and extends toward western Africa. The region has rainfall up to 200 in. (500 cm) per year and continuously high temperatures averaging 79°F (26°C). The eastern equatorial region does not experience these conditions because of the highlands and the presence of strong seasonal winds that originate from southern Asia.

The tropical rainforest vegetation consists of tall broadleaf trees that are categorized into different layers. The topmost layer, or emergents, consists of tall, straight trees that average 165 ft (50 m) in height. The middle layer, with trees of 80–155 ft (25–35 m), forms a dense interlacing canopy that allows little sunlight to penetrate to the forest floor. The lowest layer consists of young trees that average 50 ft (15 m) in height. The forest floor is typically damp and relatively open except for dead leaves, fallen branches, logs, and woody vines that connect the the tree trunks to the canopy.

The areal extent of the African rainforest region (originally 18%) has dwindled to less than 7% as a result of high rates of deforestation. Despite these reductions, the region is still one of the most diverse ecological zones in the continent. It has not only supported the timber industry in several countries but also served as an excellent reservoir for wildlife and medicines for traditional herbalists. Estimates from the International Union for Conservation of Nature show that the forested regions along the West African coast and within the central African basin contain more than 8000 plant species, including a large variety of hard and soft woods. This region also has about 84% of Africa's primate species and 68% of its passerine birds. Most of the mammals are arboreal, but there are several ground-dwelling animals such as rodents, wild pigs, and deer. Insects generally outnumber all other animal species combined. Typical are lateritic soils or oxisols with limited plant nutrients due to accelerated rates of decomposition and severe leaching. *See* RAINFOREST.

The tropical rainforest gradually merges northward and southward into the tropical savanna. This savanna encompasses a more extensive area and is best characterized by its distinct wet and dry seasons. Annual rainfall is lower (40 in.; 100 cm), and the mean annual temperature is higher than the equatorial zone. Vegetation includes savanna woodlands and shrubs interspersed with coarse tall grasses.

Unlike the forested zones, there is no continuous tree canopy. The woodlands consist of open stands of trees that are at least 16 ft (5 m) tall. The two major zones are the Guinea savanna region in West Africa and the Zambezian region in eastern Angola, the Democratic Republic of Congo, and Zambia. Dominant tree species include the African fan palm (*Borassus aethiopicum*), silk cotton tree (*Bombax petandrum*), boabab (*Adansonia digitata*), African

rubber (*Landolphia* spp.), and mopane and miombo tree species. In moving farther north or south of the Equator, tree size and overall density gradually decrease, with an increasing coverage of herbaceous plants and grassland.

Extensive savanna grasslands are found along the Sudanian zone of West Africa, within the Zambezian region and the Somalia-Masai plains. Large areas such as the Serengeti plains in the Somalia-Masai plains are home to a diverse range of wild animals. The soils are well developed and rich in nutrients because they are not subject to severe erosion. However, in some areas, especially where the topsoil has been removed, they have developed into thick crusts or pans during the dry seasons. The volcanic regions in eastern Africa are an exception, having productive soils.

The tropical steppe forms a transition zone between the humid areas and the deserts. This includes the area bordering the south of the Sahara that is known as the Sahel, the margins of the Kalahari basin, and the Karoo grasslands in the south. Average temperatures range from a daily high of 87°F (31°C) in August to 107°F (42°C) in April. Annual rainfall is highly variable, averaging 10–20 in. (25–50 cm). Inadequate rainfall coupled with poor soil chemistry (such as lack of nutrients) limits plant growth and species diversity. Nonetheless, this region has been used extensively for pastoral farming and for growing staples such as millet and sorghum. It is the most vulnerable ecological zone in the continent. In particular, the Sahelian zone, stretching from Mauritania and Senegal in the west to southern Sudan and Ethiopia in the east, has gained worldwide attention as a result of a series of devastating droughts that have affected millions of lives since the 1970s. The underlying causes for these problems are complex and intricately tied to both physical environmental and human-made factors.

Climatically, the most significant cause of desertification is the variability and marked reduction in rainfall beginning in the 1970s. This has been partly attributed to global warming trends. Specifically, higher sea surface temperatures over the Atlantic tend to hinder the northward penetration of the southwest monsoonal winds, thus reducing the likelihood for precipitation. An alternative explanation links the observed drought conditions to dust storms that frequently occur in the Sahara. These tend to heat up the atmosphere and prevent the upward flow of air needed for precipitation to occur. *See* DESERTIFICATION.

Aside from physical factors, the desertlike conditions in the Sahel are partly the result of specific human activities that degrade land. These activities include the removal of stabilizing natural vegetation for fuel wood and other land uses, overcultivation of the marginal areas, and overgrazing. Also contributing are a host of socioeconomic and political problems in the affected countries.

As one moves farther away from these marginal areas, the steppe trends into the desert landscape in the southwestern and northern parts of the continent. The Namib desert borders the Atlantic Ocean in southern Africa. Inland, it merges into the Kalahari, more accurately described as a semidesert instead of a true desert. In the north the world's largest desert, the Sahara, covers about 30% of the continent. The desert consists of mostly flat, dry plains and plateau with a few mountains such as the Ahaggar, the Air, and the Tibesti. Unlike most deserts around the world, more than 20% of the desert surface is covered with ergs (rolling plains of sand) such as the Libyan Erg (close to 200,000 mi^2 or 80,940 km^2) and the eastern and western ergs of Algeria (each about 20,000 mi^2 or 8094 km^2). Given this vast areal extent, the landscape takes on an undulating appearance with ripples of sand occasionally interrupted by small groups of dunes. In other areas, the desert landscape is characterized by regs (plains covered by pebbles and boulders), hammadas (polished rock surfaces), and basins of interior drainage (some containing salt pans). *See* DESERT.

Average annual rainfall in the desert is consistently below 10 in. (25 cm) with extreme variability from year to year. The relative humidity is very low (under 5% and rarely above 20%). There are wide annual temperature ranges and extreme diurnal ranges. The most common weather occurrences are sandstorms, particularly in the winter and early spring. Depending on the velocity and the volume of sand being transported, these storms can be disastrous for crops, livestock, and the inhabitants of this region. Extensive studies conducted in the last few decades have focused on the characteristics of these storms in relation to the materials deposited elsewhere (such as the Americas, Asia, and Europe). Specifically, airborne dust from the Sahara negatively impacts the ecology and climate within the region itself but provides valuable loess deposits in areas westward across the Atlantic Ocean, northward over the Mediterranean Sea, and eastward across the Red Sea. *See* LOESS.

Land-use patterns are extremely variable in the deserts. They range from extensive nomadic grazing of livestock to intensive cultivation around oases. These activities, however, depend on the local variation in climate, soil moisture, and the availability of water through traditional irrigation systems. The most important staple is the date palm. Fruits, vegetables, and cereals are also grown.

The northwestern and southwestern margins of the continent have a mediterranean climate, with dry summers and cool wet winters. The latter are caused by the polar air mass which brings relatively low temperatures and rainfall. Annual rainfall is about 30 in. (75 cm) or less. Mostly drought-resistant trees and scrubs are found here. The soils are variable, depending on the underlying bedrock, but can be generally classified as acidic and less fertile. *See* PLANT GEOGRAPHY; SOIL.

Drainage systems. The structural evolution of the continent has much to do with the drainage patterns. Originally, most of the rivers did not drain into the oceans, and many flowed into the large structural basins of the continent. However, as the continental

drift occurred and coasts became more defined, the rivers were forced to change courses, and flow over the escarpments in order to reach the sea. Several outlets were formed, including deep canyons, waterfalls, cataracts, and rapids as the rivers carved out new drainage patterns across the landscape. Most of the rivers continue to flow through or receive some of their drainage from the basins, but about 48% of them now have a direct access into the surrounding oceans. The major rivers are the Nile, Congo (Zaire), Niger, and Zambezi. *See* RIVER.

Nile. This is the only perennial river that enters the Mediterranean Sea from Africa. With an overall length of 4157 mi (6690 km), it is the longest river. The water originates from three major sources. About 56% comes from the Blue Nile, 22% from the Atbara River, and the rest from the White Nile. At Khartoum the White Nile joins the Blue Nile, and the plain between the two rivers, the Gezira, is the site of a large irrigation project. The Nile is controlled by several dams. The largest is the Aswan High Dam, which created Lake Nasser, over 300 mi (480 km) long. *See* DAM; MEDITERRANEAN SEA.

Congo. The Congo (Zaire), which is 3000 mi (4800 km) long, is Africa's largest river in terms of discharge. Its flow is stable throughout the year because it receives water from tributaries that lie on both sides of the Equator. The river originates near Lake Tanganyika and descends over a series of rapids and falls on its way down the Great Escarpment to the Atlantic Ocean. The Inga Dam, built between Kinshasa and Matadi, was completed in 1972. Large sections of the river are also navigable, particularly its middle course, which provides a 1085-mi (1746-km) stretch from Kinshasa to Kisangani. *See* ATLANTIC OCEAN.

Niger. The Niger rises in the Guinean highlands, an area of heavy rainfall in West Africa. After a total course of 2600 mi (4160 km), it enters the Atlantic through a large delta in Nigeria. It was dammed in 1969 at Kainji, Nigeria, creating a large artificial lake. Other large rivers in this region include the Senegal and the Volta. *See* DELTA.

Zambezi. The Zambezi begins in Angola, and its 2200-mi (3540-km) course drains most of the area south of the Congo system into the Indian Ocean. There are several obstructions in its middle course, including Victoria Falls. Here the Zambezi plunges 350 ft (107 m) from the inland Kalahari basin onto the edge of hard resistant rocks, cutting deep canyons in its path. A dam was built downstream from the falls in 1959, creating Lake Kariba. Also in Mozambique, the Caborra Bassa dam was completed in 1976 at the point where the river leaves the Kebrabasa gorge on its way to coastal lowlands. Other major rivers within southern Africa include the Limpopo and the Orange.

Fauna. The tremendous diversity in wildlife continues to be one of the primary attractions of this continent. Africa is one of the few remaining places where one can view game fauna in a natural setting. There is a tremendous diversity in species, including birds, reptiles, and large mammals. Wildlife are concentrated in central and eastern Africa because of the different types of vegetation which provide a wide range of habitats. Until recently, these regions were the least disrupted areas. Human population densities were relatively low except within highland and fertile volcanic areas. However, the situation is changing rapidly as the animals face constant threats, including disease, frequent drought, poaching, and population encroachment. Efforts to save them include the development of game reserves, bans on hunting or poaching, controlling exports of wildlife (or their products), and wildlife management. *See* CONSERVATION OF RESOURCES.

Population and environmental hazards. Africa is not a densely populated continent. With an estimated population of 743,000,000, its average density is 64 per square mile (26 per square kilometer). However, some areas have large concentrations, including the Nile valley, the coastal areas of northern and western Africa, the highland and volcanic regions of eastern Africa, and parts of southern Africa. These are mostly areas of economic or political significance.

In several regions, the human population is exposed to a variety of hazards such as riverine flooding, drought, disease, and pests. These hazards are largely meteorological in origin, such as the long periods of drought in the Sahel and eastern and southern Africa. The climate also provides a habitat for harmful insects and other vectors of endemic diseases such as malaria, yellow fever, trypanosomiasis, onchocerciasis, and bilharzia. Crops are occasionally attacked by swarms of locusts and other pests. *See* CONTINENT; EPIDEMIOLOGY.

Florence Lansana Margai

Bibliography. S. Aryeetey-Attoh (ed.), *Geography of Sub-Saharan Africa*, Prentice Hall, 1997; A. Grainger, *The Threatening Desert: Controlling Desertification*, Guersney Press, 1982; I. L. Griffiths, *An Atlas of African Affairs*, University of Witwatersrand Press, 1994; J. Gritzner, *The West African Sahel*, University of Chicago, Committee on Geographic Series, 1988; International Union for Conservation of Nature, *The Conservation Atlas of Tropical Forests in Africa*, Simon and Schuster, 1992; L. A. Lewis and L. Berry, *African Environments and Resources*, 1988; A. B. Mountjoy and D. Hilling, *Africa: Geography and Development*, Barnes and Noble, 1987; S. E. Nicholson, Climatic variations in the Sahel and other African regions during the past five centuries, *J. Arid Environ.*, 1:3–24, 1987; J. M. Pritchard, *Landform and Landscape in Africa*, Arnold, 1979; P. Richards, The environmental factor in African studies, *Prog. Human Geog.*, 4:589–600, 1980; C. Stager, Africa's Great Rift, *Nat. Geog.*, pp. 2–14, May 1990.

African horsesickness

A highly fatal insect-borne viral disease of horses and mules, and a mild subclinical disease in donkeys and zebras. It normally occurs in sub-Saharan Africa but occasionally spreads to North Africa, the Iberian Peninsula, and Asia Minor.

Etiologic agent. The African horsesickness virus is an orbivirus (family Reoviridae) measuring 68–70 nanometers in diameter. The outer layer of the double-layered protein shell is ill defined and diffuse and is formed by two polypeptides. The highly structured core consists of five structural proteins arranged in icosahedral symmetry. The viral genome is composed of 10 double-stranded ribonucleic acid (RNA) segments (genes) ranging in size from 240,000 to 2,530,000 daltons. Nine distinct serotypes which can be distinguished by neutralization tests are known. The virus can be cultivated in various cell cultures, in the brains of newborn mice, and in embryonated hen eggs by intravascular inoculation. *See* ANIMAL VIRUS.

Transmission. African horsesickness is a noncontagious disease that can readily be transmitted by the injection of infective blood or organ suspensions. In nature, the virus is biologically transmitted by midges of the genus *Culicoides*, such as *C. imicola*. The disease has a seasonal incidence in temperate regions (late summer to autumn), and its prevalence is influenced by climatic conditions favoring insect breeding (for example, warm, moist conditions in low-lying areas). Mechanical transmission by large biting flies is possible, but plays a much smaller role than biological transmission in the epidemiology of this disease.

Clinical signs. These indications result from selective replication of the virus in vascular endothelial cells, leading to increased vascular permeability in specific organs or regions of the body. Clinically, four forms of the disease are recognized. (1) In the peracute, pulmonary form, an incubation period of 3–5 days is followed by a febrile reaction and signs of respiratory distress. The breathing accelerates (75 breaths per minute or more), and the horse stands with its neck extended and the nostrils fully dilated. Profuse sweating is common, and terminally spasmodic coughing may be observed with frothy fluid exuding from the nostrils. Death usually occurs within hours after the first respiratory signs are observed. (2) In the subacute, edematous or cardiac form, the incubation period is about 7–14 days. The febrile reaction is followed by characteristic swellings of the hollows above the eyes, the eyelids, the lower lip, and the neck region. Swelling of the limbs is never observed. Some animals may show signs of colic, and death occurs 4–8 days after the onset of the febrile reaction. (3) Most horses show the acute or mixed form, which represents a combination of the pulmonary and edematous forms. (4) African donkeys, zebras, and partially immune horses may exhibit the fever form, where a mild febrile reaction is the only clinical sign.

The mortality rate in fully susceptible horses varies from about 70 to 95% and is about 50% in mules. In European and Asian donkeys, the fatality rate may be about 10% while African donkeys and zebras never succumb to the disease.

Pathology. The gross autopsy findings depend largely on the clinical form of disease manifested by the animal before death. In the pulmonary form, large amounts of fluid are found in the lung tissue or in the thoracic cavity, and death results from suffocation. Other lesions commonly observed are intense inflammation of the glandular fundus of the stomach (which can easily be confused with poisoning) and hemorrhage or swelling of the large and small intestines.

In the subacute form, there is extensive infiltration of subcutaneous and intermuscular tissues with yellow gelatinous fluid. In mild cases just the head and neck are involved, but in more severe cases the edema also involves the brisket and shoulders. Accumulation of fluid in the pericardial sac is a constant and very significant finding. Extensive hemorrhages are found on the surface of the heart as well as in the left ventricle. The lymph nodes are generally enlarged and edematous, and lesions in the gastrointestinal tract are largely similar to those found in the pulmonary form except that submucosal edema of the large intestine tends to be more conspicuous.

In the mixed form the lesions represent a combination of those observed in the pulmonary and edematous forms.

Treatment and control. There is no specific treatment for this disease. Infected animals should be given complete rest as the slightest exertion may result in death.

Stabling of horses at night during the African horsesickness season reduces exposure to insect vectors and hence the risk of disease transmission. Prophylactic vaccination is the most practical and effective control measure. In epidemic situations outside Africa, the causal virus should be serotyped as soon as possible, allowing the use of a monovalent vaccine. However, in endemic regions it is imperative to use a polyvalent vaccine, which should render protection against all nine serotypes of African horsesickness virus. *See* DISEASE; EPIDEMIC; VACCINATION.

Baltus Erasmus

Afterburner

A device in a turbojet aircraft engine, between turbine and nozzle, which improves thrust by the burning of additional fuel (**Fig. 1**). To develop additional

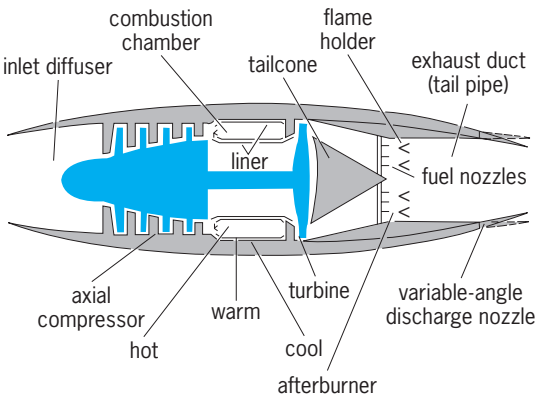


Fig. 1. Diagram of turbojet engine showing afterburner.

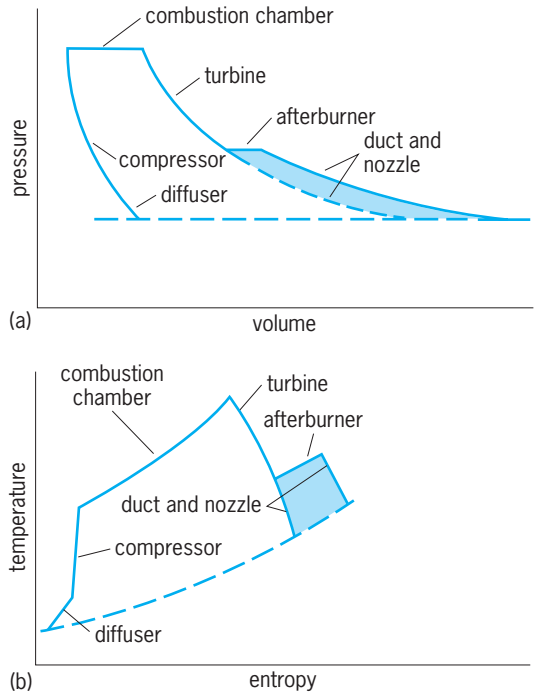


Fig. 2. Thermocycle diagrams for turbojet engine. (*a*) Pressure vs volume. (*b*) Temperature vs entropy. Shaded areas show increased output produced by afterburner.

thrust for takeoff and climb and for periods of dash of military aircraft, it is advantageous to augment the engine thrust. This is done by afterburning, also called reheating, tail-pipe burning, or postcombustion. **Figure 2** shows the effect on exhaust-nozzle gas pressure and temperatures. The augmentation of thrust obtained by afterburning may be well over 40% of the normal thrust and at supersonic flight can exceed 100% of normal thrust.

When a turbojet is developing the maximum thrust for which it is built, the compressor is operating at its maximum compression ratio and the gases enter the turbine at maximum allowable temperature. The most feasible means to increase the thrust further is to reheat the gases after they leave the turbine. In a conventional turbojet the gases discharge from the turbine at approximately 1500°R (800 K) and with sufficient oxygen to permit burning liquid hydrocarbon fuel. By afterburning, the gases may be heated to approximately 3500°R (2000 K) before they enter the discharge nozzle.

The augmented thrust depends directly on the increase in temperature and also on aircraft speed, afterburning being more effective at supersonic speeds. At subsonic speeds specific fuel consumption is approximately doubled so that the technique is suitable only for brief periods. At supersonic speeds the use of continuous afterburning is feasible. An engine with an afterburner also requires a variable-area nozzle because of the large difference in nozzle area required during the afterburning and nonafterburning conditions. *See* AIRCRAFT ENGINE; TURBOJET.

Benjamin Pinkel

Agar

A major constituent of the cell walls of certain red algae, especially members of the families Gelidiaceae and Gracilariaceae. Extracted for its gelling properties, it is one of three algal polysaccharides of major economic importance, the others being alginate and carrageenan.

Agar is composed of two similar fractions, agarose and agaropectin, in which the basic unit is galactose, linked alternately α-1,3-(D-galactose) and β-1,4-(α-L-galactose). Agaropectin has substitutions by sulfate and pyruvic acid. In the closely related carrageenans, all of the galactose is in the D form and usually at least one unit of each pair is sulfated. The gel strength of the agar is proportional to its agarose content. Agar is insoluble in cold water, but swells by absorbing up to 20 times its weight of water. Dissolved in boiling water and cooled, it sets to a firm gel at concentrations as low as 0.5%. A 1% solution solidifies when cooled to 90–102°F (32–39°C), while the gel does not melt at a temperature below 185°F (85°C).

Agar is prepared by boiling the algae (agarophytes) in water, after which the filtered solution is cooled, purified, and dried. It is an amorphous, translucent material that is packaged in granules, flakes, bricks, or sheets. Before World War II, Japan monopolized the industry, but important production facilities have subsequently been established in Spain, Portugal, Argentina, and China.

Agar derives its name from the Malayan agar-agar, a traditional algal jelly now known to be a carrageenan. One of its chief uses is as a gelling agent in media for culturing microorganisms, a use begun by Robert Koch in 1882. Earlier it had been used (and continues to be used) in making confections. It is also used as an emulsifier in cosmetics and food products, as a sizing agent, as an inert carrier of drugs in medicine, and as a laxative. *See* CULTURE; RHODOPHYCEAE.

Paul C. Silva; Richard L. Moe

Agate

A variety of chalcedonic quartz that is distinguished by the presence of color banding in curved or irregular patterns (**Fig. 1**). Most agate used for ornamental purposes is composed of two or more tones or intensities of brownish-red, often interlayered with white, but is also commonly composed of various shades of gray and white. Since agate is relatively porous, it can be dyed permanently in red, green, blue, and a variety of other colors. The difference in porosity of the adjacent layers permits the dye to penetrate unevenly and preserves marked differences in appearance between layers. The properties of agate are those of chalcedony: refractive indices 1.535 and 1.539, hardness 6.5 to 7, and specific gravity about 2.60.

The term agate is also used with prefixes to describe certain types of chalcedony in which banding is not evident. Moss agate is a milky or almost transparent chalcedony containing dark inclusions in a

Fig. 1. Section of polished agate showing the characteristic banding. (*Field Museum of Natural History, Chicago*)

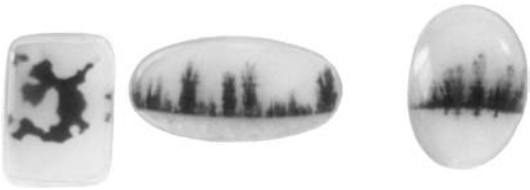


Fig. 2. Moss agate cut as gems, showing the dendritic patterns. (*The Mineralogist*)

dendritic pattern (**Fig. 2**). Iris agate exhibits an iridescent color effect. Fortification, or landscape, agate is translucent and contains inclusions that give it an appearance reminiscent of familiar natural scenes. Banded agate is distinguished from onyx by the fact that its banding is curved or irregular, in contrast to the straight, parallel layers of onyx. *See* CHALCEDONY; GEM; QUARTZ.

Richard T. Liddicoat, Jr.

Agglutination reaction

A reaction in which suspended particles are aggregated or clumped. It occurs upon the admixture of another type of particle, a change in the composition of the suspending fluid, or the addition of a soluble agent that acts as a bridge between two or more particles. The reaction is a secondary one in that the process resulting in agglutination occurs after the primary antigen-antibody linkage has taken place.

The particles undergoing agglutination may be either unicellular or microscopic multicellular organisms (such as bacteria and parasites), individual cells of multicellular organisms (such as erythrocytes and lymphocytes), or artificial particles (such as beads of plastic, glass, or polysaccharide). The immunological specificity of agglutination depends upon the uniqueness of the reaction between a marker substance on one type of particle and a receptor on either another type of particle or a specific antibody in solution. The marker can be a usual biological component of the surface of the particle (for example, a specific polysaccharide on bacteria) or blood group substance on red cells. It can be an enzymatically or a chemically modified chemical group on the surface of biological particles. It can also be an adsorbed or a chemically attached substance. The attachment can be to biological particles (for example, bacteria, cells, and subcellular particle components) or artificial ones (for example, spheres of plastic such as latex, glass, or polysaccharide such as agarose). The receptor can be a biological component of the particle, an attached antibody, or antibody in solution. A reverse reaction is one in which the antibody is attached to a particle (say, latex) and the addition of the antigen (such as bacterial product) causes the mixture to clump. The addition of substances, typically polymeric (for example, polyethylene glycol, polyvinylpyrrolidone, and polyions), that affect the mobility or charge of the particles tends to increase the propensity of a mixture to agglutinate. Increased adherence can also occur if charged groups are removed by treatment of biological particles (such as erythrocytes) with enzymes (bromelin, papain, neuraminidase, and so on). Inhibition of agglutination can also be used to test for antigens, especially of low molecular weight, in a manner similar to that for agglutination itself. *See* IMMUNOASSAY.

Alexander Baumgarten

Bibliography. C. P. Engelfriet et al., *Blood Transfusion in Clinical Medicine*, 10th ed., 1997; N. R. Rose, E. C. De MacArio, and J. D. Folds, *Manual of Clinical Laboratory Immunology*, 5th ed., 1997.

Agglutinin

A substance that will cause a clumping of particles such as bacteria or erythrocytes. Of major importance are the specific or immune agglutinins, which are antibodies that will agglutinate bacteria containing the corresponding antigens on their surfaces. Agglutinin activity is frequently displayed by purified antibody preparations that also precipitate or give other serological reactions. Agglutinins are readily determined, and their presence is of diagnostic value to indicate present or past host contact with the microbial agent sufficient to result in antibody formation. *See* AGGLUTINATION REACTION; ANTIBODY; RICKETTSIOSES.

Analogous reactions involve erythrocytes and their corresponding antibodies, the hemagglutinins. Hemagglutinins to a variety of erthyrocytes occur in many normal sera, and their amounts may be increased by immunization. The blood group isoagglutinins of humans and animals are important special cases which must be considered in all proposed blood transfusions lest transfusion reactions result. Some, but not all, hemagglutinins may display further lytic activities when complement is added. *See* BLOOD GROUPS; COMPLEMENT.

Certain agglutinating systems agglutinate weakly, if at all, unless a conglutinin is added. Bacteria and erythrocytes may also, at times, be nonspecifically agglutinated by other substances, such as acids and plant extracts. Unfortunately, these latter agents are also frequently termed agglutinins. *See* CONGLUTINATION.

Henry P. Treffers

Bibliography. E. L. Cooper, *General Immunology*, 1982; D. M. Weir (ed.), *Handbook of Experimental Immunology*, 5th ed., 1996.

Aggression

Behavior that is intended to threaten or inflict physical injury on another person or organism; a broader definition may include such categories as verbal attack, discriminatory behavior, and economic exploitation. The inclusion of intention in defining aggression makes it difficult to apply the term unequivocally to animals in which there is no clear means of determining the presence or absence of intention. As a result, animal violence is usually equated with aggression.

Causes. There are four main approaches to understanding the causes or origins of human aggression. First, the basis may be differences among people, due either to physiological difference or to early childhood experiences. Second, there are sociological approaches which seek the causes of aggression in social factors such as economic deprivation and social (including family) conflicts. Third, causes may be found in the power relations of society as whole, where aggression arises as a function of control of one group by another. Fourth, aggression may be viewed as an inevitable (genetic) part of human nature; this approach has a long history and has produced extensive arguments. Earlier suggestions that aggression operated as a drive (for example, by S. Freud) and was generated in part by increases in drive level due to deprivation do not have any support. Some support for the role of individual genetic factors derives from breeding experiments which have shown that selective breeding for animal aggression is possible. *See* DEVELOPMENTAL PSYCHOLOGY.

Given the wide variation in aggressive behavior in different societies and the occasional absence of such behavior in some groups and some individuals, a general human genetic factor is unlikely. However, some genetic disposition to react with force when an individual is blocked from reaching a goal may provide an evolutionary basis for the widespread occurrence of violence and aggression. The existence of different kinds of aggression suggests that different evolutionary scenarios need to be invoked and that aggression is not due to a single evolutionary event. All of these approaches may have some degree of validity, and as with many explanations of complex human behavior, it is likely that aggression is multidetermined and rarely, if ever, due to a single factor. *See* BEHAVIOR GENETICS.

Types. Aggressive behavior can occur in a large number of different situations, and is often identified by the occasions when it occurs as well as by the behavior that is produced. Aggression in humans ranges through fear-induced aggression, parental disciplinary aggression, maternal aggression, and sexual aggression. E. Fromm identified one clearly biologically adaptive type, defensive aggression, which occurs when fight responses are mobilized in defense of an organism's vital interests, such as obtaining food or the protection of its young. The aim of defensive aggression is not destruction but the preservation of life. Thus, aggression can serve both destructive and constructive purposes. Among animals, the varieties of aggression include most of the human types as well as predatory aggression, territorial defense, and sexually related aggression in competition for a mate. *See* REPRODUCTIVE BEHAVIOR.

Humans and other animals. In most animals, aggression and violence is stimulus- and response-specific. That is, the appearance of a threat or a hunting object produces the aggressive response, but responses such as specific attack behavior are very delimited. In contrast, aggression in humans can be produced by a wide variety of situations, and aggression may incorporate any number of physical, social, or psychological forms. In addition, humans produce much more intraspecific aggression than other animals. It was once believed that only humans could show evidence of intraspecific aggression and that all seriously aggressive behavior in other animals was directed toward other species. The fighting and dominance behavior shown by the males of many species is usually placed in a different category, since it very rarely results in the death or severe incapacitation of either participant. However, it has been shown that intraspecific aggression also occurs in some animal species, although much less frequently than in humans.

Cultural factors. The incidence of aggression in humans varies widely as a function of cultural background and societal factors in general. Aggression in different cultures varies, for example, from some of the Yanomamo people of the Amazon Basin, who consider murder and rape natural and acceptable occurrences, to the Semai Senoi of West Malaysia, who are embarrassed by shows of violence and successfully teach their children to avoid violent or aggressive behavior. Among industrial societies, also, the incidence of violence and aggression varies widely.

Physiology. A number of different physiological bases of aggression have been proposed. Among them are a prevalence of parasympathetic over sympathetic nervous system activity, leading to very low levels of arousal and thrill-seeking aggressive behavior; left cerebral hemisphere dysfunction, which leads to a loss of control over impulsive behavior; and a high level of the hormone testosterone, which is related to some indices of aggression. *See* EMOTION.

George Mandler

Bibliography. E. Fromm, *The Anatomy of Human Destructiveness,* 1974; R. G. Geen, *Human Aggression,* 1990; G. Mandler, *Mind and Body: Psychology of Emotion and Stress,* 1984; A. Montagu, *The Nature of Human Aggression,* 1976.

Aging

Definitions of aging differ between biologists and behavioral scientists. The biologists regard aging as reflecting the sum of multiple and typical biological

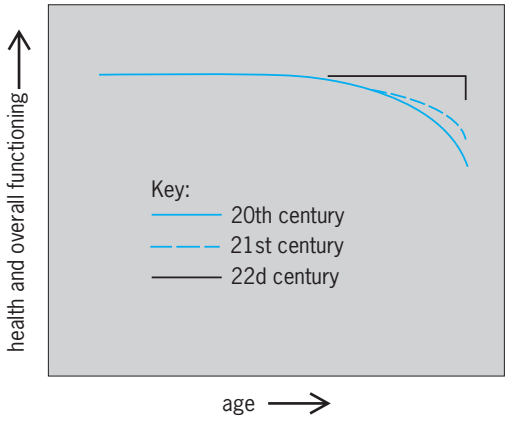
decrements occurring after sexual maturation; the behavioral scientists view it as reflecting regular and expected changes occurring in genetically representative organisms advancing through the life cycle under normal environmental conditions. It is difficult to define normal aging, since many changes observed in older adults and previously perceived as concomitants of normal aging are now recognized as effects of disease in later life. The behavioral science view allows for incremental as well as decremental changes with aging. Senescence is not always equated with aging; it is viewed as the increasing vulnerability or decreasing capacity of an organism to maintain homeostasis as it progresses through its life span. Gerontology refers to the study of aging. Geriatrics refers to the clinical science that is concerned with health and illness in the elderly.

Demography. The dramatic growth of the older age group is a rather recent phenomenon. In the United States, for example, in 1900 approximately 3 million people (4% of the total population) were 65 years of age or older. By the mid-1980s, more than 28 million (12%) were in this age group. Moreover, by 2050, this age group is expected to increase to over 60 million (over 20% of the total population). The group 85 years of age and older is actually the fastest-growing segment of the population; it is expected to multiply more than sixfold between 1985 and 2050, to more than 16 million.

Not only is life expectancy generally increasing, but it is also increasing from age 65. In 1900 the average life expectancy in the United States was 47 years; by the mid-1980s it was 74.5. In Rome during the height of the Roman empire, average life expectancy was only in the low 20s (because of war, disease, famine, and childbirth). In contemporary America, average life expectancy from age 65 is approaching 20 years.

Scientific progress in gerontology and geriatrics has been substantial since the mid-1960s, but scientists are confronting a potential dilemma: progress in understanding the mechanisms of aging leading to increased longevity should not outpace progress in improving the quality of life in later years.

Aging versus illness. There are many myths and stereotypes about what changes in later life reflect normal aging as opposed to illness. For example, it was long thought that senility was the inevitable outcome of growing old and that most people over age 65 ended up in nursing homes. But research has revealed that only about 6% of the group 65 and older develop Alzheimer's disease, the most common form of "senility'' or dementia (marked impairment of memory and intellectual functioning). Under conditions of generally good health, older people maintain a high level of mental functioning. This is not to say that decrements do not normally occur in later life; some do, such as reaction time, which diminishes with aging. If the person is healthy, however, some increments also occur with aging. Vocabulary, for example, has been shown to continue to expand in studies of individuals followed well into their 80s. *See* ALZHEIMER'S DISEASE.



Hypothetical "curve" illustrating impact of reduced disability of the aging population in the twentieth century. The goal is to keep the interval of disability prior to death as high (rectangular) as possible at the end of the life cycle.

Similarly, most older people continue to maintain a high level of independent functioning. Only about 5% of the elderly are in nursing homes or hospitals; the remainder reside in the community. More and more people are maintaining stable functioning longer into the life cycle, resulting in the rectangularization of the curve representing their level of functioning as a factor of age (see **illus.**). The years of good health have been extended for these individuals, while the interval of disability has been reduced, in comparison with individuals of the same age at the beginning of the twentieth century. But there is another group, small in percentage but large in number, for whom increased longevity has brought increased years of disability.

Theories of aging. Inquiries into why organisms age involve both the purpose of aging and the process of aging. There are theories, but no conclusive evidence, in both areas. The most common explanations of the purpose of aging are based on theories that aging is adaptive for a species: aging leading to death is a mechanism by which species keep from overcrowding their environments; and finite life spans permit greater turnover of generations and thereby greater opportunities for a species to adapt, in a Darwinian sense. However, findings of very high accidental mortality among various species, resulting in few of that species having the opportunity to grow old, suggest that aging may not be necessary in an evolutionary sense. Still, the adaptive theories of why people age remain popular.

Theories about the process of aging concern how people age. These biological theories address two sets of factors—intrinsic and extrinsic to the organism. Intrinsic factors are influences operating on the body from within, such as the impact of genetic programming. Extrinsic factors are influences on the body from the environment, such as the impact of cumulative stresses. There are at least seven different major theories to explain the biological basis of aging: (1) Genetic programming: the organism's deoxyribonucleic acid (DNA) contains an intrinsic, built-in genetic program for aging leading to death. (2) DNA damage or genetic mutation: extrinsic

factors from the environment (such as radiation, environmental toxins, and viruses) damage the organism's DNA or cause the genetic material to mutate, bringing about aging. (3) Free radical: highly reactive chemicals (known as free radicals) within the body, produced by exposure to various extrinsic or intrinsic influences, interfere with normal cellular events, resulting in aging. (4) Cross linkage: improper diet or excessive caloric intake results in abnormal protein cross linkages between various cells of the body, resulting in age-associated changes like cataracts or skin problems. (5) Senescent substances: certain naturally produced substances (such as the pigment lipofuscin) or a protein (such as statin) in cells build up over time, altering the function of the cell and inducing age-related changes. (6) Immunologic: a combination of intrinsic or extrinsic factors cause cumulative alterations in the immune system over time, leading to problems associated with aging. (7) Hormones as clocks of aging: certain hormones may function as pacemakers or clocks of aging, affecting development and death. Despite an abundance of theories to explain the process of aging, its mechanisms remain a mystery.

Life extension. It is important to differentiate between life expectancy and life span. Life expectancy is the average number of years of life in a given species; it is significantly influenced by factors beyond aging alone, such as famine and disease. Life span is the maximum number of years of life possible for that species; it is more fundamentally linked to the process of aging itself. Over the centuries, life expectancy has increased (due to improved sanitation and health care practices); life span has not. Approximately 115 years appears to be the upper limit of life span in humans.

Life extension efforts really aim to increase life span. But just as the causes of aging have been elusive, sure formulas for life extension in humans have not been found. Efforts to achieve life extension have involved a wide diversity of approaches, including special diets and dietary restriction, exercise programs, vitamins and minerals, and drugs.

Many drugs have been used to reduce the age pigment lipofuscin, antioxidants have been used to combat free radicals, and hormones have been used to improve immune functioning. None of these approaches has had an unequivocal impact on life span in humans. At the same time, several of them, such as improvements in diet and exercise and the cessation of consumption of certain substances (for example, tobacco) and drugs (for example, excessive alcohol) have led to increase in life expectancy.

Expectations. Long overlooked but now increasingly recognized are the capacities for change, developmental growth, and creativity throughout the life cycle. This is apparent both in the response of aging brain neurons to environmental challenge and in the behavior of aging individuals. For example, Verdi composed operas in his 80s, Picasso painted into his 90s, and some people embark on a second career after initial retirement. *See* DEATH; HUMAN GENETICS.

Gene D. Cohen

Bibliography. R. H. Binstock and E. Shanas, *Handbook of Aging and the Social Sciences*, 1990; J. E. Birren, S. Birren, and K. Warren (eds.), *Handbook of the Psychology of Aging*, 1990; G. D. Cohen, *The Brain in Human Aging*, 1990.

Agnosia

An impairment in the recognition of stimuli in a particular sensory modality. True agnosias are associative defects, where the perceived stimulus fails to arouse a meaningful state. An unequivocal diagnosis of agnosia requires that the recognition failure not be due to sensory-perceptual deficits, to generalized intellectual impairment, or to impaired naming (as in aphasia). Because one or more of these conditions frequently occur with agnosia, some clinical scientists have questioned whether pure recognition disturbances genuinely exist; but careful investigation of appropriate cases has affirmed agnosia as an independent entity which may occur in the visual, auditory, or somesthetic modalities. *See* APHASIA.

Visual agnosias. The patient with visual object agnosia, though quite able to identify objects presented auditorily or tactually, cannot name or give other evidence of recognizing visually presented objects. Perceptual intactness is indicated by retained ability to draw such objects or to perform match-to-sample tasks. Because visual object agnosia is a rather rare disorder, knowledge of its underlying neuropathology is incomplete. Most reported cases have shown bilateral occipital lobe lesions, with the lesion extending deep into the white matter and often involving the corpus callosum. In the few cases where the reported lesion has been unilateral, the left occipital lobe has been the focus. *See* BRAIN.

Prosopagnosia is the inability to recognize familiar faces. Persons well known to the individual before onset of the condition, including members of the immediate family, are not recognized. In many instances, individuals fail to recognize picture or mirror images of themselves. Some individuals with prosopagnosia utilize nonvisual cues, such as voice, in order to identify a person. Specific visual features (a beard, facial scar, or hairstyle, for example) may also be used. The ability to utilize specific facial details for recognition, even though the overall facial configuration remains elusive, suggests that prosopagnosia might be understood as a highly specific breakdown in holistic visual perception (considered to be a function of the right hemisphere) though the ability to utilize perceived detail (a left-hemisphere function) remains intact. This view is consistent with the association of prosopagnosia with right posterior lesions, though bilateral lesions are common. *See* HEMISPHERIC LATERALITY; VISION.

Isolated impairment of reading is frequently considered to be an exotic form of aphasia. Logically, however, it may be considered as a visual-verbal agnosia (also referred to as pure word blindness or alexia without agraphia). Individuals with this disorder show a marked reduction in their ability to read

the printed word, though their writing and other language modalities remain essentially intact. Particularly dramatic is the failure of these persons to read material which they have written. The specific visual nature of the deficit is indicated by retained ability to decode words spelled aloud. Visual-verbal agnosia is typically associated with right homonymous hemianopsia (blindness in one-half of the visual field), indicating damage to the left occipital lobe. This lesion is typically coupled with damage to the corpus callosum, which connects the right and left visual association zones. The disorder has been interpreted as a disconnection syndrome. The patient's general visual functioning is good because of an intact right visual cortex. But in order for linguistic materials to be decoded, activity of the language-dominant left hemisphere is required, and this is precluded by the callosal lesion.

Auditory agnosias. The term auditory agnosia is most often used to indicate failure to recognize nonverbal acoustic stimuli despite adequate hearing sensitivity and discrimination. The individual fails to recognize a telephone by its ringing sound, for example, though it is identified immediately by sight or touch. In most well-documented cases of agnosia for sounds, the subjects have had bilateral temporal lobe lesions. In those instances where the damage has been unilateral, the right temporal lobe has been involved, suggesting a special role for this brain region in the recognition of nonverbal acoustic stimuli.

Auditory-verbal agnosia (or pure word deafness) is a disturbance in comprehension of spoken language, in the presence of otherwise intact auditory functioning and essentially normal performance in other language modalities. The person's speech expression is remarkably intact in comparison with the gross impairment in understanding speech. Like its visual analog, visual-verbal agnosia, this is a disconnection syndrome. It is produced by damage to the left primary auditory cortex (or the tracts leading to it) coupled with a lesion to the corpus callosum. The individual is thus able to carry out most auditory functions because of an intact right hemisphere; but auditory-linguistic material cannot be interpreted because the information fails to reach the dominant left hemisphere.

Phonagnosia is a disturbance in the recognition of familiar voices. The person has good comprehension of what is spoken, but the speaker cannot be identified. Phonagnosia, like its visual analog, prosopagnosia, is associated with right-hemisphere damage; in fact, these two disorders frequently coexist. *See* HEARING (HUMAN); PSYCHOACOUSTICS.

Astereognosis. Astereognosis (or tactile agnosia) is the inability to recognize objects by touch and manipulation. One or both hands may be affected. The deficit is at the associative level and is not explainable on the basis of impaired somesthetic sensation or perception. A diagnosis of astereognosis thus requires that the patient show reasonably intact sensitivity and discrimination for somesthetic stimuli. In addition, the perceptual features of the object (such as size, shape, or texture) should be recognizable. The classical view concerning localization is that the lesion responsible for the impairment is in the somesthetic association cortex opposite the affected side of the body. The evidence for this is tenuous, however, because astereognosis is so often complicated by sensory-perceptual impairment that few unequivocal cases have been satisfactorily studied. *See* PERCEPTION; SENSATION.

Gerald J. Canter

Bibliography. G. J. Beaumont, *Understanding Neuropsychology*, 1988; K. M. Heilman and E. Valenstein (eds.), *Clinical Neuropsychology*, 3d ed., 1993; M. H. Johnson (ed.), *Brain Development and Cognition: A Reader*, 1993.

Agonomycetes

Fungi which usually produce neither sexual (meiotic) nor asexual (mitotic) spores. These fungi have been denoted by various names. The most common names are Mycelia Sterilia (having sterile mycelium), Agonomycetales (used when classifying these fungi at the level of order), or Agonomycetes (used when assigning their rank at the level of class). These fungi make up an artificial (nonphylogenetic) group consisting of both ascomycetes and basidiomycetes. A few species have associated sexual states, but these sexual states are infrequently produced. Although true spores are mostly lacking in this group, some species produce sporelike structures which function effectively in survival and dispersal of the species. Some of these sporelike structures mimic true spores so well in form and function that the genera producing them (for example, *Beverwykella*, *Cancellidium*, and *Tretopileus*) have occasionally been classified with spore-forming fungi. *See* ASCOMYCOTA; BASIDIOMYCOTA; DEUTEROMYCOTINA; FUNGI.

Morphology and function. The most common sporelike structures are allocysts, bulbils, chlamydospores, papulospores, and sclerotia. These structures are not formed in the same manner as sexual spores (via meiosis), nor are they formed like true asexual spores, because the sporelike structures are simply modified hyphae. Chlamydospores originate from terminal or intercalary hyphae, have thickened walls, and are primarily a survival structure; they tend to be unicellular but often occur in pairs or chains. Allocysts resemble chlamydospores but store metabolic by-products and do not germinate. Bulbils are compact, multicellular, and internally undifferentiated propagules (somatic structures of propagation or survival); they resemble papulospores except that the latter have a core of large cells and a sheath of smaller cells. There are propagules intermediate between bulbils and papulospores as well. Sclerotia are larger, multicellular aggregations of hyphae and are often differentiated into layers at maturity. Outer layers may have dark, thick-walled cells and the inner layers may have thin-walled cells, but such differentiation is not universal. Sclerotia occur in a range of sizes and shapes, depending on the species forming them, and the smaller ones are termed microsclerotia. These various structures formed from hyphae

may also be present in species of fungi producing true spores (sexual and/or asexual). Chlamydospores and sclerotia are especially common. In general, the larger propagules serve for survival more than dissemination, but any propagule may be transported in soil particles by machinery or tools, movement of animals or humans, irrigation water, or strong winds. Sclerotia and chlamydospores may also be disseminated with plant seed, vegetative plant parts, and nursery stock.

Identification, pathology, and economic significance. The absence of true spores and structures producing true spores renders identification difficult. However, morphology-based keys to genera have been constructed using such characters as whether hyphae are simple or aggregated, the manner in which hyphal aggregations are formed, the presence of dolipore septa or clamp connections (indicating affiliation with basidiomycetes), and the manners in which chlamydospores, bulbils, sclerotia, papulospores, or other propagules are formed and positioned. Commercially produced kits are available for identification of *Rhizoctonia* and *Sclerotinia* (by ELISA [enzyme-linked immunosorbent assay] or PCR [polymerase chain reaction]), and experimental identification of selected genera of Agonomycetes has been published using techniques such as PCR-RFLPs [restriction fragment length polymorphisms], RAPDs [random amplifications of polymorphic DNA], and DNA (deoxyribonucleic acid) sequence analyses.

Several members of the Agonomycetes are of extreme importance as plant pathogens. *Rhizoctonia* species (which have infrequently produced sexual states in *Ceratobasidum* or *Thanatephorus*) cause root diseases. *Rhizoctonia solani* is a pathogen with an exceptionally broad host range encompassing dozens of genera, but is especially important on beans, carrots, cabbage, and other vegetables, and *R. cerealis* is important on wheat and other small grains. Interestingly, some *Rhizoctonia* species are beneficial mycorrhizal symbionts with orchid roots. In general, a nonsporulating fungus can be identified as *Rhizoctonia* by the thickness, septation, growth rate, and branching patterns of its hyphae. However, the taxonomy and identification of *Rhizoctonia* as to species and important subspecific groups is complex, and depends to considerable extent on anastomosis groupings, that is, how pairs of isolates behave when confronted with each other on artificial laboratory media.

Also important is the form-genus *Sclerotium*, species of which produce sclerotia. Some *Sclerotium* isolates will produce sexual states in *Athelia* (a basidiomycete) or other genera. The most important plant pathogens include *S. rolfsii* (corresponding to *Athelia rolfsii*) which is a pathogen of many crops, especially vegetables, and *S. cepivorum*, the causal agent of white rot of onion and garlic. A fungus sometimes classified and keyed with the Mycelia Sterilia is *Sclerotinia sclerotiorum*. Many isolates of this fungus reproduce sexually in nature or under proper, carefully controlled laboratory conditions, and some also produce small, nongerminating asexual spores, but on artificial media in the laboratory isolates often produce only sclerotia. *Sclerotinia sclerotiorum* has an exceptionally broad host range but is especially important on vegetables. *Sclerotinia trifoliorum* and *S. minor* are important on legumes and lettuce, respectively. Prior to the advent of molecular genetics, and still today in routine diagnoses, species of *Sclerotinia* are separated on the basis of characters of the colonies and sclerotia on artificial media. Once established in soil, sclerotia may survive for years, even decades, rendering disease control extremely difficult.

Other notable members of the Agonomycetes are *Papulaspora byssina*, troublesome in commercial mushroom operations, and *Cenococcum geophilum*, an ectomycorrhizal fungus symbiotic with many species of higher plants. *See* PLANT PATHOLOGY.

Frank M. Dugan

Bibliography. H. L. Barnett and B. B. Hunter, *Illustrated Genera of Imperfect Fungi*, 1998; E. Kiffer and M. Morelet, *The Deuteromycetes: Mitosporic Fungi, Classification and Generic Keys*, 2000; J. A. von Arx, *Plant Pathogenic Fungi: Beihefte zur Nova Hedwigia*, vol. 87, 1987.

Agouti

A large rodent that resembles the rabbit or hare in size and shape as well as in the elongated hindlegs, which make them well adapted for speed (see **illustration**). The agouti and the closely related smaller acouchi are inhabitants of clearings in forested

Agouti (*Dasyprocta aguti*), a rodent found in Mexico, South America, and the West Indies. (*Courtesy of Lloyd Glenn Ingles; © California Academy of Sciences*)

areas of the Amazon region. Some range into Central America and as far as the Guianas.

Thirteen species of agouti have been described, the common one being *Dasyprocta aguti*. Some authorities are of the opinion that all of these are varieties or subspecies of *D. aguti*. The acouchi is represented by two species, the green acouchi (*Myoprocta pratti*) and the red acouchi (*M. acouchy*). Behaviorally the agouti and acouchi have many similarities; thus, some authorities believe that the acouchi may be a dwarf variety of the agouti.

Dentition is typical of the cavimorph rodents, and the dental formula is I 1/1 C 0/0 Pm 1/1 M 3/3. Like most rodents, the agouti is a prolific breeder, and may breed at any time during the year. These animals feed primarily on vegetation, such as roots, berries, and fruits. Ecologically they form a part of the food chain for a variety of carnivorous animals, including humans, because they are the basic available animals common to the area. *See* DENTITION; FOOD WEB; RODENTIA.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Agricultural aircraft

Aircraft designed, or adapted from a general utility airframe, for use in agriculture and forestry and for control of insect vectors of human, animal, and plant diseases. Agricultural aircraft have become an indispensable tool for high-productivity agriculture and have contributed to the worldwide crop production revolution.

Agricultural aircraft are widely used in developed countries and, increasingly, in developing countries, particularly on high-value export crop and forest products. Aircraft use covers a wide range of agricultural crop and pest applications, including control of competing weeds and unwanted brush and trees, control of insect and disease pests, application of plant nutrients, and broadcast seeding of many crops. Other, less widely used applications include defoliation to aid in cleaner crop harvesting, frost protection, rain removal from sensitive horticultural crops, and control of insect vectors or disease carriers affecting plants, animals, or humans. Large-area applications, frequently with twin- and four-engine transport as well as smaller-type aircraft, are used to control outbreaks of locust and grasshoppers and reduce or eradicate specific insect pests, such as fruit flies. Releases of sterile male insects and predacious spiders are made by aircraft, and other nonchemical biopesticides, such as bacteria, fungi, and various insect hormones that interfere with mating or metamorphosis, are applied as capsules, as baits, or in liquid carriers. *See* DEFOLIANT AND DESICCANT; FERTILIZER; PESTICIDE.

The principal advantages of either fixed- or rotary-wing aircraft for the treatment of crops lies with their ability to rapidly cover large crop acreages and to travel over rough terrain, irrigation structures, and wet fields. This timeliness factor is often considered critical to optimum pest control and effective crop protection. The primary disadvantage is their inability to direct the released material onto the target crop with the precision that can be accomplished with ground-based applications. This lack of precision, together with the losses in the form of drift or movement away from the target area, constitutes an area of concern. Increasing limitations are being placed on aircraft use, especially when highly toxic crop chemicals are to be applied to small fields and where sensitive nontarget crops are grown nearby.

Types. The two basic types, fixed- and rotary-wing aircraft, are both used for all of the applications noted above. In the United States, there are approximately 6000 aircraft in use, of which about 8% (480) are helicopters, flying an estimated total of 1.875×10^6 h per year. Each aircraft averages approximately 150 acres (60 hectares) per flight hour; thus, approximately 300×10^6 crop acres (120×10^6 hectares) are treated each year. There has been a gradual change to larger-capacity aircraft powered by engines of 1000–1500 horsepower (750–1125 kW), 10–15% of which are more powerful gas turbine-propeller types. Average fixed-wing aircraft carry 250–400 gallons (946–1514 liters), with the largest capable of lifting 800 gallons (3028 liters). Helicopters in agricultural use carry loads of about one-half these amounts.

While larger, faster-flying (up to 140 mi/h or 225 km/h) aircraft can accomplish greater hourly productivity, other factors such as size and location of crop fields, distance to fields from suitable landing strips, and time lost in field turns also affect field productivity. Helicopters, which can turn quickly and can land and be serviced close to crop fields or even on landing docks on top of the service trucks, can thus gain back some of the lowered productivity of these aircraft due to smaller load capacity and lower field speeds. However, it is the greater application precision and downblast from their rotary wing (at reduced forward speeds) that has generated the increased use of helicopters, in spite of their greater initial and operating costs.

Dispersion wake. The aerodynamic wake or air dispersion produced by the action of the wing or helicopter rotor and the mass of air displaced by the aircraft in motion are utilized to aid in spreading either liquid or dry materials. The interaction of the wing or rotor length and the length of the spray boom, along with the strength of the wake and height of flight, controls the usable swath width laid down by the aircraft.

Application equipment. The application equipment for either fixed- or rotary-wing aircraft is customarily an integral part of the aircraft. The spray equipment for a fixed-wing aircraft (**Fig. 1**) may be detached from the bottom of the tank or hopper, and a ram-air-type spreader for dry materials may then be brought into place. Spray equipment for helicopters (**Fig. 2***a*) is similar to that for fixed-wing aircraft. However, since helicopters are usually operated at lower speeds (60–90 mi/h or 96–145 km/h), windmill-powered pumps and ram-air spreaders do

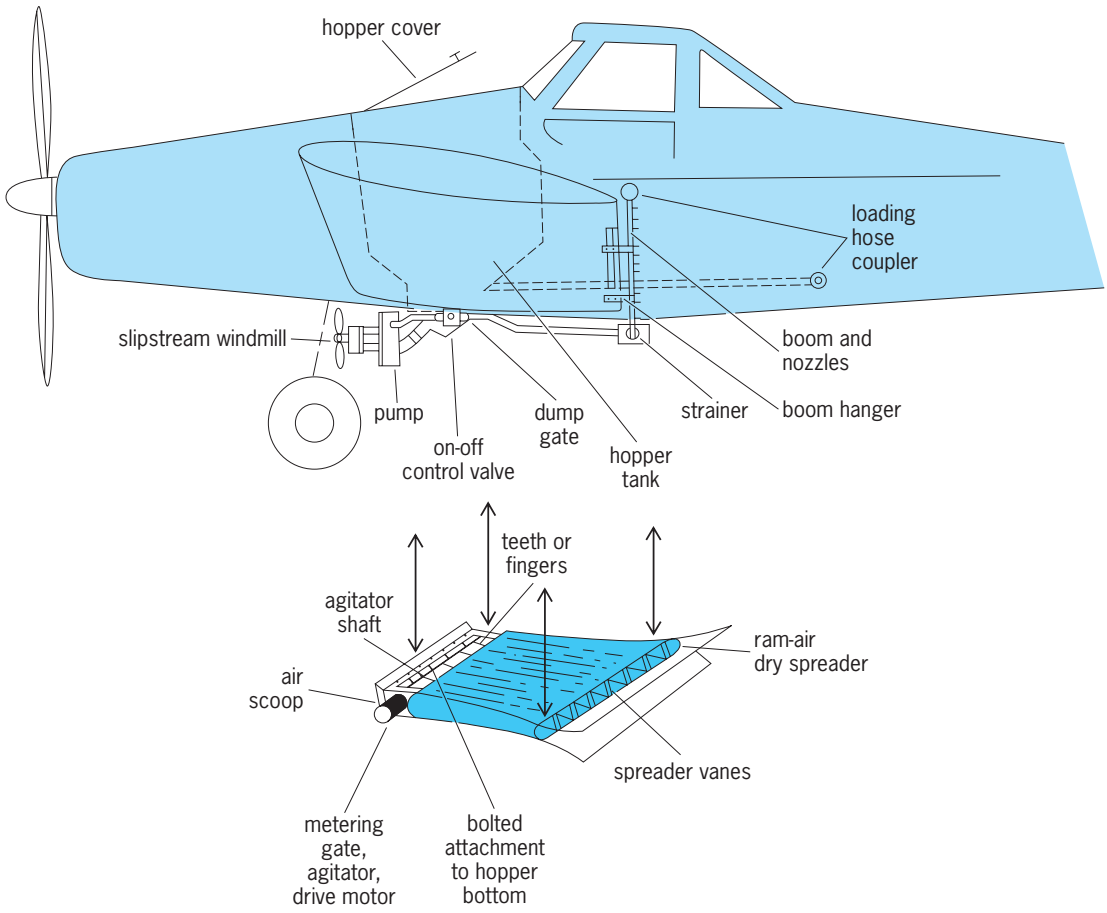


Fig. 1. Application equipment for fixed-wing agricultural aircraft. The spraying equipment is shown on the aircraft, and a ram-air-type spreader for dry materials is shown below the aircraft.

not function well on these aircraft. Another primary difference between aircraft types is the requirement for lateral as well as fore-and-aft balance on the helicopter. This necessitates the use of split hoppers, one on each side of the helicopter cockpit. A large crossover pipe connects the bottoms of the two hoppers. A pump moves the liquid formulation from its connection at the crossover to the spray boom. For dry materials, each hopper has a controlled feeder and usually a spinning device (Fig. 2*b*) to distribute the dry granules, seeds, fertilizers, or baits. Another helicopter system for applying both spray and dry materials is the sling unit, which can be suspended below the helicopter as a single hopper. Application materials from this unit are distributed in much the same way as from the hopper used on fixed-wing aircraft. Power for the pump or dry-materials spinner may be supplied by electric or hydraulic motors, these in turn being powered from generators or hydraulic pumps on the main engine. Auxiliary gasoline engines may also be used to furnish power directly or through electric or hydraulic coupling to the pumps or spinners.

Spraying equipment. The spraying equipment for a fixed-wing aircraft (Fig. 1) consists of a hopper bottom assembly bolted to the hopper, which supports the pump, the slipstream windmill with brake, and the on-off control valve. The spray boom is customarily attached to the trailing edge of the wing by hangers fastened to the wing frame. Nozzles for atomizing the spray along with their diaphragm cutoff valves are placed along the boom on each wing. Pump pressure forces liquid against the diaphragms, which open and permit flow to the atomizers for a hollow-cone nozzle. Stationary whirl plates behind the nozzle orifices spin the liquid, causing it to form circular or conical sheets on discharge. A liquid-tight cover is located on the top of the hopper tank, but loading is customarily done by means of hoses that are quick-coupled to loading couplers located either at the end of the boom or at the end of a special pipe leading directly into the hoppers. *See* ATOMIZATION; NOZZLE.

Centrifugal pumps are customarily used for high volume rates of 50–100 gal/min (190–387 liters/min) at relatively low pressures of 40–60 lb/in.^{2} (276–415 kilopascals). On fixed-wing aircraft these may be powered by a windmill (Fig. 1), or by electric or hydraulic motors powered though electric generators or hydraulic pumps directly driven from the propulsion engine. The latter approach is customarily usedfor the helicopter.

Spreading dry materials. The ram-air spreader is used on the fixed-wing plane for applying dry materials such as dusts or granules (Fig. 1). The liquid system is unbolted and dropped, and the ram-air dry

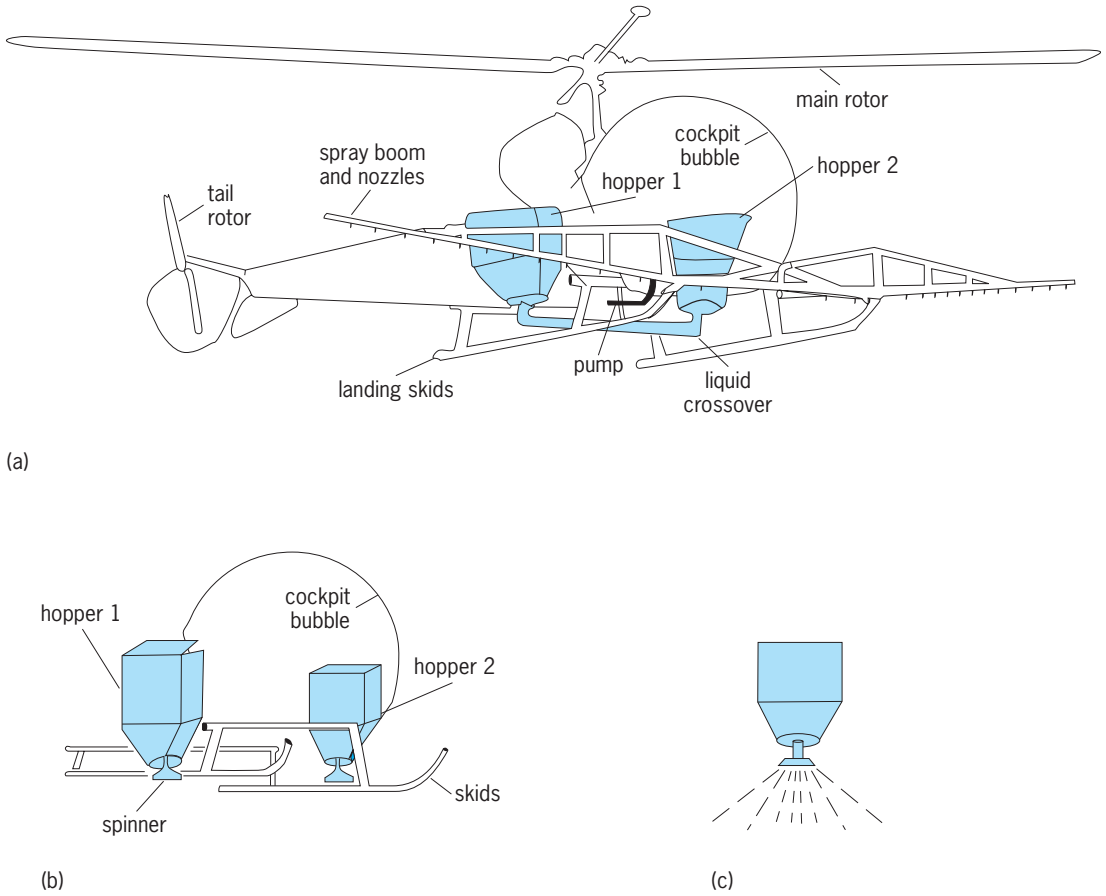


Fig. 2. Application equipment for rotary-wing agricultural aircraft. (*a*) Helicopter with spray equipment attached. (*b*) Dry equipment. (*c*) Detail of hopper and spinner.

spreader is then attached. This operates on the principle of a venturi device with air gathered by a large frontal opening then forced through a narrow section directly beneath the hopper and metering gate, which then widens by a factor of 1.5–2 at the discharge end. Longitudinal vanes inside the spreader guide the flow of material and help to direct the material outward from the aircraft. An agitator, either windmill- or motor-driven, may be located in the throat of the spreader, just below the hopper. On helicopters, detachable spinner-type spreaders for dry materials can be placed below each of the twin hoppers (Fig. 2*c*). A metering valve is placed between each hopper outlet and the spinner to control the flow of material to each spinner.

Spray applications. The characteristic spray swath laid down by the aircraft can be determined by catching and analyzing the amount of spray caught on plastic or paper tapes or sheets laid perpendicular to the aircraft flight pattern. A rapid spray deposit evaluation technique uses a fiber string or plastic monofilament stretched across the flight line to catch the spray. Analysis is performed by feeding the string through a special washing unit which strips the string of tracer dye and evaluates the amount of collected spray caught by the string. Another system uses a video camera and dyed spray to visually indicate high- and low-density portions of the spray pattern. The amount of drift loss, however, can only be measured by air samples.

A single pass of a fixed-wing aircraft can spread deposits over a width of about 180 ft (55 m). However, the useful or flagged swath width that can produce a uniformly deposited pattern from successive swaths across a field is considerably less, depending on spray drop size.

The spray pattern from a helicopter application is much the same as that from a fixed-wing aircraft. The swath width is a function of the rotor width, which creates much the same type and strength of downwash as is produced by a fixed-wing aircraft of similar wing span.

A typical spray drop spectrum produced by hollow cone nozzles on the spray boom of a fixed-wing aircraft, directed with a 100-mi/h (160-km/h) slipstream, has a volume median diameter (50% of the spray volume in drops above the volume mean diameter, and 50% below this size) of 454 micrometers. The drop size for 0–10% by volume is 207 μm, and the size for 0–90% by volume is 896 μm. These values indicate that drop sizes ranging from 20 to 1100 μm in diameter are produced by this atomizer.

Dry applications. The swath patterns from applications of dry materials such as granules, seeds, baits, or other dry particles can be patterned by collecting the materials in buckets placed across the aircraft

flight pattern. The graphs produced will appear similar to the spray distribution patterns, whether the materials are applied by ram-air or rotary spreaders, on fixed- or rotary-wing aircraft.

Airborne transport or drift. The movement of particles or spray drops less than 100 μm in diameter constitutes the burden of the contamination or drift problem that has plagued the aircraft application industry since its inception. After release from the aircraft, the distances that these small particles can travel are dependent on atmospheric conditions, which control their rate of diffusion or dispersion in the downwind atmosphere. Principal among these dispersion factors is atmospheric stability, or the rate of air mixing, vertically and horizontally. A temperature inversion with warmer air overhead at 100–200-ft (30–60-m) height will prevent vertical mixing, and minimal winds will reduce lateral diffusion or spreading of the pesticide-laden cloud.

The first step taken to reduce and control the drift loss of pesticide, sprays, or dry particles is to increase the size of the particles or spray drops as much as is practical. Spray releases with a volume median diameter over 1000 μm will minimize the percentage of drops present below 100 μm diameter, reducing the volume of these to 1–2% of the total. This control will reduce drift losses beyond displaced swath distances of 200–300 ft (60–90 m) but will not eliminate all drift loss. However, sprays with such large drop sizes are suitable only for translocated-type herbicides and crop defoliants, or for applications that target the top surfaces of crops or are applied to the soil surface. For fungicides and insecticides to be effective, the median drop size must be reduced to 200–400 μm; the airborne or drift losses will then increase to 25% or more of the applied active pesticide. *See* AGRICULTURE; AIRPLANE; HELICOPTER. Norman B. Akesson

Bibliography. B. N. Devisetty, D. G. Chasin, and P. D. Berger (eds.), *Pesticide Formulation and Application Systems*, ASTM STP 1146, American Society for Testing and Materials, 1993; C. G. McWhorter and M. R. Gebhardt (eds.), *Methods of Applying Herbicides*, Weed Science Society of America, 1987; H. R. Quantick, *Handbook for Agricultural Pilots*, 4th ed., 1985.

Agricultural buildings

All buildings used on the farm for housing livestock and the farm family and for storing grain, livestock feed, and farm machinery. In the United States, today's farms are larger, more specialized, and more mechanized than those of the past, and these changes are reflected in the buildings. For example, the two-story barns with haylofts are seldom built any more, in favor of one-story buildings without interior columns and with trussed roofs. The building frames are usually constructed of steel or wood and are covered with prepainted metal sheets. This type of construction is low in cost and has a useful life of about 25 years, which results in a nearly simultaneous wearout and obsolescence of the building. Some buildings have an open front facing the south or east. Others are completely enclosed and insulated so that a warm environment may be maintained to protect young animals or to maximize animal growth and production.

Housing of livestock. Livestock farms usually contain one or two species of animals. Buildings are designed for specific uses which depend upon the species and age of animal, climatic conditions, and the management practices followed. Buildings are the shelter portion of the housing system, which also includes a ventilation system and feed, produce, and manure-handling equipment.

Dairy cattle. These animals are housed in two types of barn. Stall barns have individual stalls in which the cows are tied. Feed is conveyed to the manger at the front of the stalls. The manure is mechanically scraped out of the barn. The barn temperature is held above 45°F (7°C) by conserving the body heat given off by the cows through the use of insulation and thermostatic controlling of the rate of ventilation. Milking is performed in the stalls; therefore, the barn and the cows must be kept very clean. The cows are let out of the barn daily for exercise or pasturing.

The second type, the free-stall barn (**Fig. 1**), also provides a separate stall for each animal to rest; however, they are not tied and are free to leave the stall at any time. A mechanical feeder is installed in the feeding area. The manure is collected and transported by mechanical scrapers, by flushing with water, or by passing through slots in the floor. The cows are milked in a separate milking parlor with elevated stalls. The milk is weighed and conveyed through stainless steel pipes into a milk cooler located in the milk room. *See* DAIRY CATTLE PRODUCTION.

Swine. Swine housing changed dramatically between the late 1950s and 1980s. Better facilities were provided to reduce labor and to allow for farrowing of piglets throughout the year. Thus a more uniform supply of pork is available over time.

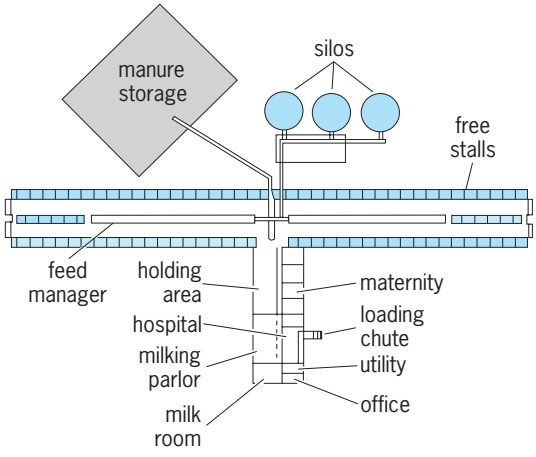


Fig. 1. Typical arrangement of a 200-cow free-stall dairy barn. Cows are held in the holding area before entering the double-8 herringbone milking parlor. (*After Midwest Plan Service, Structures and Environment Handbook, 11th ed., 1983*)

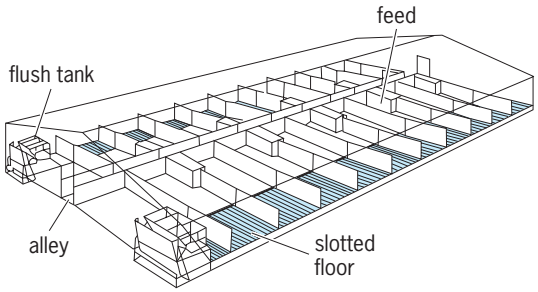


Fig. 2. Environmentally controlled swine growing-finishing house. (*After Midwest Plan Service, Structures and Environment Handbook, 11th ed., 1983*)

There are special buildings for each age group so that an optimum arrangement and thermal environment can be provided for good health and growth. Two types of systems are used to house swine. With the pasture system, minimal buildings are utilized and the pigs are raised on pasture during the growing season. The confinement system confines the pigs to a number of specialized buildings.

Farrowing houses shelter the piglets from birth to about 4 weeks of age. These buildings are heated and well insulated to provide an air temperature of 95°F (35°C) for the piglets and 60°F (16°C) for the sows. After weaning, the pigs are moved to a nursery building for about 8 weeks, where the room temperature is first held at 85°F (29°C) and gradually lowered to 65°F (18°C). A growing house is used next for about 6 weeks until the pigs reach about 150 lb (68 kg) in weight with the room temperature continued at about 65°F (18°C). Lastly, the pigs are moved to a finishing house and kept there until they reach a market weight of about 220 lb (100 kg). Some farmers use a combination of these buildings (for example, nursery plus growing, or growing plus finishing; **Fig. 2**). Additional buildings are required for the gestating sows and boars.

Much labor-saving equipment is used in swine production: conveyors transport the feed from storage to the feeders; automatic waterers supply drinking water; and thermostatically controlled fans regulate the barn temperature. Slotted floors are commonly used for all ages to allow the waste to pass into an underfloor tank where the manure can be easily pumped out into large tank wagons for field spreading. *See* SWINE PRODUCTION.

Beef cattle. These animals require minimal shelter for survival and growth. In the western states, the cattle are kept on the open range or in an open feed lot with only a windbreak for protection. In the midwestern states, more shelter is required because of higher amounts of precipitation during cold weather. The housing consists of an open front shed with access to a yard. The shed contains a layer of bedding and the yard is for exercise and feeding. Mounds of earth are often built in the yards to give well-drained resting areas (**Fig. 3**). A less common type, confinement housing, confines the animals in pens in an open-front building with a slotted or flume-type floor for the handling of manure. Mechanical feeders are commonly used in all kinds of housing systems. *See* BEEF CATTLE PRODUCTION.

Sheep. Sheep require minimum shelter, similar to beef cattle, to minimize lamb losses and increase labor efficiency. An open-front barn with an outdoor exercise yard or a confinement building can be used. Labor-saving equipment is available for feeding, sorting, treating, and shearing animals and for manure disposal. *See* SHEEP.

Poultry. Poultry buildings for egg production usually confine the birds in cages arranged in single to triple decks. Feeding, watering, and egg collection are completely mechanized. The buildings are insulated and ventilated to control the thermal environment. Broiler production also utilizes enclosed buildings where the birds are raised on a floor covered with litter such as straw or wood shavings. *See* POULTRY PRODUCTION.

Other specialized buildings are used to shelter such animals as horses, goats, or rabbits. *See* GOAT PRODUCTION; HORSE PRODUCTION.

Livestock feed and commercial grain. Grass and corn silage is a major feed for dairy and beef cattle. Silos provide an oxygen-deficient environment for the fermentation of grass or corn with a moisture content of 45–65%. The end product of the fermentation process is silage.

Vertical silos are recommended for small and medium-size farms. They consist of cylinders 14–30 ft (4.3–9.1 m) in diameter and up to 90 ft (27.4 m) in height. Silos can be of the conventional type, in which the top surface is exposed to the atmosphere with some spoilage occurring, or the sealed type, with an oxygen-free environment throughout with no spoilage. Mechanical unloaders installed on the top or the bottom of the silo remove the silage and drop it into a conveyor leading to a mechanical feeding system.

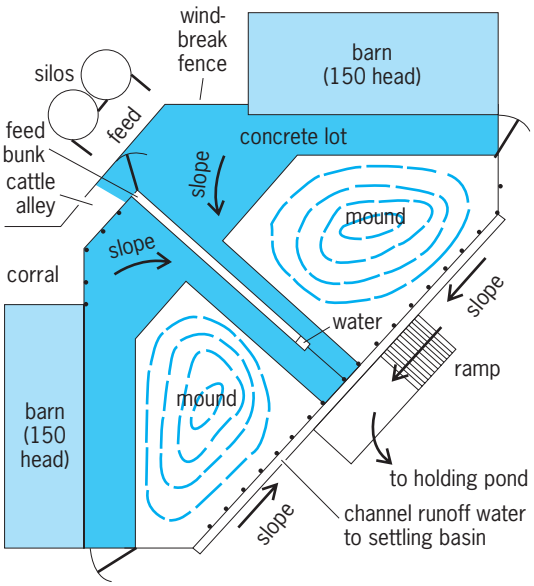


Fig. 3. Beef cattle barn with open front, mechanical bunk feeder, and mounds in the lot. (*After Midwest Plan Service, Structures and Environmental Handbook, 11th ed.,1983*)

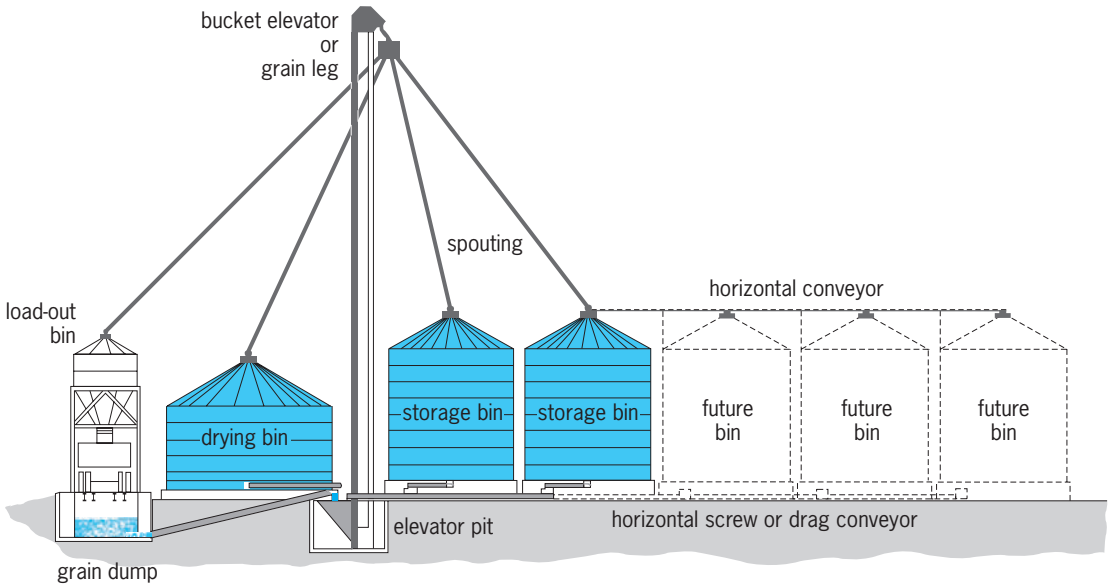


Fig. 4. Farm grain center with drying, conveying, and storage facilities. (*After Midwest Plan Service, Structures and Environment Handbook, 11th ed., 1983*)

Horizontal silos are more economical for large farms and consist of concrete-lined bunkers. A large top surface is exposed and is usually covered with plastic sheet to limit spoilage. The silage is generally removed with front-end loaders on tractors.

Grain fed to livestock need not be dried and is often stored as high-moisture grain in silos. Grain produced for sale must be dried and stored in a dry, insect-free environment to maintain its quality. A grain-handling system consists of a number of components for transporting, drying, and storage. These components include a grain dump to catch the grain discharged from a truck, and conveyors or grain legs to elevate and move it to the dryers and then to storage. Round metal bins and special buildings, called flat storages, are used for grain storage. An aeration system of fans is needed to blow air through the grain and keep it from spoiling. A farm grain center is shown in **Fig. 4**.

Machinery storage. Buildings to house the field machinery are provided on most farms. These buildings must have at least 14-ft-high (4-m) side walls, large doors, and clear span roofs to accommodate the large machines needed on today's farms. A heated machinery repair shop is often located in a portion of this building. *See* AGRICULTURAL MACHINERY.

Farm houses. These homes have all the modern conveniences of urban houses and are constructed of similar materials. The house design includes special features to accommodate the farm activities. Near the rear entrance are facilities to allow the workers to clean up and to store their farm clothing. An office is also desirable for record keeping.

Farmstead planning. The various buildings, fences, and roads composing the farmstead are laid out for efficient use. The farm home, the center of family living, is located to provide a view of the yards. The other buildings are located so as to maintain a high quality of living around the house and efficient operation. The large livestock buildings are at the greatest distance from the house and on its downwind side so as to minimize the effect of objectionable odors and insects. Machinery storage and repair and grain storage buildings are located closer to the house. *See* AGRICULTURE. C. O. Cramer

Bibliography. American Society of Agricultural Engineers, *Agricultural Engineers Yearbook*, 1983–1984; Midwest Plan Service, *Structures and Environment Handbook*, 11th ed., 1983; Midwest Plan Service, *Swine Housing and Equipment Handbook*, 4th ed., 1983; J. H. Whitaker, *Agricultural Buildings and Structures*, 1979.

Agricultural chemistry

The science of chemical compositions and changes involved in the production, protection, and use of crops and livestock. As a basic science, it embraces, in addition to test-tube chemistry, all the life processes through which humans obtain food and fiber for themselves and feed for their animals. As an applied science or technology, it is directed toward control of those processes to increase yields, improve quality, and reduce costs. One important branch of it, chemurgy, is concerned chiefly with utilization of agricultural products as chemical raw materials.

Scope of field. The goals of agricultural chemistry are to expand understanding of the causes and effects of biochemical reactions related to plant and animal growth, to reveal opportunities for controlling those reactions, and to develop chemical products that will provide the desired assistance or control. So rapid has progress been that chemicalization of agriculture has come to be regarded as a twentieth century revolution. Augmenting the benefits of mechanization (a revolution begun in the mid-nineteenth

century and still under way), the chemical revolution has advanced farming much further in its transition from art to science.

Every scientific discipline that contributes to agricultural progress depends in some way on chemistry. Hence agricultural chemistry is not a distinct discipline, but a common thread that ties together genetics, physiology, microbiology, entomology, and numerous other sciences that impinge on agriculture. Chemical techniques help the geneticist to evolve hardier and more productive plant and animal strains; they enable the plant physiologist and animal nutritionist to determine the kinds and amounts of nutrients needed for optimum growth; they permit the soil scientist to determine a soil's ability to provide essential nutrients for the support of crops or livestock, and to prescribe chemical amendments where deficiencies exist. *See* FERTILIZER; SOIL.

Chemical materials developed to assist in the production of food, feed, and fiber include scores of herbicides, insecticides, fungicides, and other pesticides, plant growth regulators, fertilizers, and animal feed supplements. Chief among these groups from the commercial point of view are manufactured fertilizers, synthetic pesticides (including herbicides), and supplements for feeds. The latter include both nutritional supplements (for example, minerals) and medicinal compounds for the prevention or control of disease. *See* HERBICIDE; PESTICIDE.

Important chemicals. Chemical supplements for animal feeds may be added in amounts as small as a few grams or less per ton of feed or in large quantities, for example, urea, which is used as an economical nitrogen for ruminant animals. The tremendous tonnage of processed feeds sold, coupled with the high unit value of some of the chemical supplements, makes this a large market. *See* ANIMAL FEEDS.

Of increasing importance since their commercial introduction have been chemical regulators of plant growth. Besides herbicides (some of which kill plants through overstimulation rather than direct chemical necrosis), the plant growth regulators include chemicals used to thin fruit blossoms, to assist in fruit set, to defoliate plants as an aid to mechanical harvesting, to speed root development on plant cuttings, and to prevent unwanted growth, such as sprouting of potatoes in storage. *See* DEFOLIANT AND DESICCANT.

Striking effects on growth have been observed in plants treated with gibberellins. These compounds, virtually ignored for two decades after they were first isolated from diseased rice in Japan, attracted widespread research attention in the United States in 1956; first significant commercial use began in 1958. The gibberellins are produced commercially by fermentation in a process similar to that used to manufacture penicillin. New product classes include the cytokinins and the polyamines. *See* GIBBERELLIN.

In the perennial battle with insect pests, chemicals that attract or repel insects are increasingly important weapons. Attractants (usually associated with the insect's sexual drive) may be used along with insecticides, attracting pests to poisoned bait to improve pesticidal effectiveness. Often highly specific, they are also useful in insect surveys; they attract specimens to strategically located traps, permitting reliable estimates of the extent and intensity of insect infestations. Additional approaches involve the use of antifeedants, juvenile hormones, and other life-cycle-altering chemicals, and research focused on natural compounds that are more pest-specific and biodegradable. *See* INSECT CONTROL, BIOLOGICAL.

Repellents have proved valuable, especially in the dairy industry. Milk production is increased when cows are protected from the annoyance of biting flies. Repellents also show promise as aids to weight gain in meat animals and as deterrents to the spread of insect-borne disease. If sufficiently selective, they may protect desirable insect species (bees, for instance) by repelling them from insecticide-treated orchards or fields.

Agricultural chemistry as a whole is constantly changing. It becomes more effective as the total store of knowledge is expanded. Synthetic chemicals alone, however, are not likely to solve all the problems humans face in satisfying food and fiber needs. Indeed, many experts are coming to the view that the greatest hope for achieving maximum production and protection of crops and livestock lies in combining the best features of chemical, biological, and cultural approaches to farming, for example, integrated pest management. *See* AGRICULTURAL SCIENCE (ANIMAL); AGRICULTURAL SCIENCE (PLANT); AGRICULTURE.

Rodney N. Hader

Agricultural engineering

A discipline concerned with solving the engineering problems of providing food and fiber for the people of the world. These problems include designing improved tools to work the soil and harvest the crops, as well as developing water supplies for agriculture and systems for irrigating and draining the land where necessary. Agricultural engineers design buildings in which to house animals or store grains. They also work on myriad problems of processing, packaging, transporting, and distributing the food and fiber products.

Agricultural engineering combines the disciplines of mechanical, civil, electrical, and chemical engineering with a basic understanding of biological sciences and agricultural practices. Concern for the properties of agricultural materials and the way they influence engineering designs characterizes agricultural engineers.

Some agricultural engineers work directly with farmers. Most, however, work with the companies that manufacture and supply equipment, feeds, fertilizers, and pesticides. Others work for companies that provide services to farmers, such as developing irrigation and drainage systems or erecting buildings and facilities. Still others work with food-processing companies.

Training. Some engineers, trained in mechanical, civil, electrical, or chemical engineering, have become agricultural engineers through experience in working on agricultural problems. Since 1906, however, special agricultural engineering curricula have been available. The United States has 49 universities offering such curricula; most of those are approved by the Accreditation Board for Engineering and Technology. Several of the universities also offer advanced education at the master of science and doctorate level.

The American Society of Agricultural Engineers, founded in 1907, provides continuing education in the form of local, regional, and national conferences, workshops, and technical meetings. It establishes standards for the profession and publishes technical information.

Applications. A few examples will serve to illustrate the broad general classes of problems with which agricultural engineers work, and the almost infinite variety of special problems they encounter.

Crop production. Soil and water are the basic resources of agriculture, and agricultural engineers design farming systems that include methods of preparing land for cultivation and of controlling water. Land-grading equipment has been designed to follow a beam of laser light. These machines can produce surfaces free of standing pools of water and with slopes so gentle that runoff water will not erode the soil. *See* IRRIGATION (AGRICULTURE); LAND DRAINAGE (AGRICULTURE).

The tillage needed for seedbed preparation depends on the soil type, the crop, and the climate, requiring agricultural engineers to produce a variety of special tools. Chisels deep-rip the subsoil. Bed shapers with rotary tillers produce fine seedbeds for small seeds. Planters with special furrow openers plant directly through the trash from the previous crop for no-till planting. *See* AGRICULTURAL SOIL AND CROP PRACTICES.

Equipment for applying chemicals for insect and weed control must apply the materials uniformly, minimize drift to neighboring regions, and not damage the crop. For weeds which grow taller than the crop, a wick-type applicator has been developed which wipes herbicides onto the weeds. *See* HERBICIDE.

Harvesting equipment is particularly dependent on the nature of the crop. The combined harvester, which cuts and threshes cereal grains, has been adapted to handle corn and fine grass seeds. For horticultural crops, agricultural engineers work closely with plant breeders to find combinations of machine principles and special cultivars that can be harvested mechanically. Processing of tomatoes provides a success story. Plant breeders found varieties for which most of the fruit ripened at one time, and then bred these varieties for high skin strength. Agricultural engineers designed machines to cut and elevate the plants, remove the tomatoes from the vines, sort the red fruit from the green, and deliver the fruit to bulk containers for transport to a cannery. *See* AGRICULTURAL MACHINERY; AGRICULTURAL SCIENCE (PLANT).

Processing. The value of agricultural products depends on their quality and location with respect to time. Wheat is cleaned to eliminate chaff, weed seeds, and lightweight wheat kernels. It is dried for long-term storage and stored with ventilation to remove heat of respiration. Conveyors, hopper-bottom trailers, and barges are used to transport it. It is cracked, rolled, or ground for animal feed; it is milled and packaged for human food. Agricultural engineers design the processes and equipment, the storage structures, and the handling systems. The design requirements are different for each commodity. *See* FOOD ENGINEERING.

Animal production. Growth rate and productivity of animals depend to a large extent on the environment in which they are housed. Air temperature, humidity, speed of air movement, light level, day length, and available floor space are all important. Agricultural engineers control these factors in designs of buildings. *See* AGRICULTURAL BUILDINGS.

Fresh feed and clean water are automatically supplied to laying hens in special cages. Eggs roll out of the cages automatically to collection points. *See* POULTRY PRODUCTION.

In some dairies, data on milk production are stored in computers. Transponders hung from the necks of the cows signal the identity of each cow as it enters the milking parlor. The computer measures an amount of feed based on each cow's production record. The milk is weighed and the computer record is updated—all automatically. Heat recovered from cooling the milk is used to heat water for washing the cows and sterilizing the milking equipment. *See* AGRICULTURAL SCIENCE (ANIMAL); DAIRY CATTLE PRODUCTION.

Waste management. When animals graze the range, their waste material is recycled, and few problems result. When cows are brought to the barn for milking or chickens are raised in confinement, agricultural engineers are faced with the problems of designing equipment, facilities, and processes for managing the concentrated wastes. Processing plants and even field crops produce wastes, and the problems of collecting, treating, storing, and transporting the wastes need specific solutions.

New directions. As agriculture adopts new products, agricultural engineers encounter a new array of production, harvesting, handling, processing, storage, and waste management problems. Greenhouse culture of horticultural and floricultural crops requires structural design, environmental control, irrigation systems, and handling schemes. Forest culture requires equipment for seed collection, seedling culture and transplanting, control of forest weeds and stand thinning, and harvest system design. The culture and harvest of aquatic animals and plants offers new opportunities for agricultural engineers. Processes and systems for controlling water quality are crucial. Feeding, harvesting, and processing equipment are also needed. *See* AQUACULTURE; FLORICULTURE; FOREST AND FORESTRY; HORTICULTURAL CROPS.

Computer technology is rapidly being adopted by

agriculture. The primary bottleneck to fuller utilization of the technology is a shortage of low-cost, reliable sensors to provide data about such factors as soil moisture, plant growth and health, weather, and product quality. Agricultural engineers will need to develop these sensors.

It is reasonable to expect that robots capable of performing some agricultural operations will be important developments of agricultural engineers. The lack of well-controlled conditions and well-defined operations in agriculture suggests, however, that the human operator will remain the key controller in most systems. Agricultural engineers will design equipment and systems more to enhance human capabilities than to replace them. *See* AGRICULTURE; AGRONOMY.

Roger E. Garrett

Bibliography. American Society of Agricultural Engineers, *Agricultural Engineers Yearbook*; S. M. Henderson and R. L. Perry, *Agricultural Process Engineering*, 3d ed., 1976; R. A. Kepner, R. Bainer, and E. L. Barger, *Principles of Farm Machinery*, 3d ed., 1978; Midwest Plan Service, *Structures and Environment Handbook,* 11th ed., 1983; L. O. Roth, F. R. Crow, and G. W. A. Mahoney, *An Introduction to Agricultural Engineering*, 2d ed., 1991; R. E. Stewart, *Seven Decades that Changed America*, American Society of Agricultural Engineers, 1979.

Agricultural machinery

Mechanized systems of food and fiber production used in agriculture. These systems extend from initial tillage of the soil through planting, cultural practices during the growing season, protection from pests, harvesting, conditioning, livestock feeding, and delivery for processing. The trend has been to larger self-propelled special-purpose machines, except for tillage where the trend has been to large four-, six-, or eight-wheel or crawler tractors used in conjunction with high-capacity plows, disks, and deep rippers.

The use of hydraulic power has made possible highly specialized mechanisms to perform intricate operations. Hydraulic power offers the advantages of being easily controlled and automated. Sophisticated technology is used to increase the precision needed in modern agriculture: lasers for laying out fields for surface irrigation systems; microprocessors for sensing and controlling intricate operations, such as controlling feed mixtures for dairy cows and grading fruits and vegetables; and electronic devices in the automation of many harvesters.

Tillage equipment. Primary and secondary tillage equipment, such as plows, disks, and harrows, are designed to prepare the seedbed and root zones for crop production. Multipurpose machines are used where a high degree of precision and specialization is needed. **Figure 1** shows a machine that may be used to simultaneously rototill the soil, form beds and irrigation furrows, incorporate a herbicide, and plant the seed in one tripacross the field.

Laser land-leveling machinery allows the farmer to prepare a field for surface irrigation by grading the field to an exact slope. A laser beam, transmitted from a rotating command post at a preset slope, is received on a mobile scraper that generates a signal to automatically keep the scraper blade on the desired slope. These techniques allow land to be leveled with greater accuracy than can be achieved by traditional methods. Irrigation water savings of 20–30% have been reported after laser land leveling. *See* IRRIGATION (AGRICULTURE).

Planting and cultivation equipment. Agricultural planting equipment is commonly used for the planting of raw uncoated seed. Some of the more expensive or irregularly shaped seeds can be coated to form a uniformly shaped object that can be handled and planted with extreme accuracy and efficiency. For some crops, other types of planting equipment may be used to plant tubers, such as potatoes, and transplants, such as tomatoes.

No-till planters have been developed for the planting of seed into undisturbed soil; no-till farming eliminates all primary and secondary tillage normally conducted between cropping years, and therefore the soil remains undisturbed. Conditions for no-till planting are not at all the same as for conventional tillage farming in that the soil is covered with residue or sod

Fig. 1. Multipurpose machine used for precision tillage, planting, bed shaping, and fertilizing in one pass, and for later bed shaping, cultivating, and fertilizing. (*Johnson Farm Machinery Co.*)

Fig. 2. Agricultural aircraft applying insecticide on a field.

Fig. 3. Portable feed grinder-mixer delivering processed feed to feed bunk. (*Sperry New Holland, Division of Sperry Rand Corp.*)

and is likely wetter, firmer, and rougher than conventionally prepared soils. Generally, a no-till planter is built heavier and stronger than conventional planters and is normally equipped with a cutting disk in front of each planting unit. The no-till farming concept is widely used in the Corn Belt states and generally benefits farmers in terms of soil and water conservation, lower production costs, and greater production efficiencies.

Crop cultivation is accomplished primarily to rid the crop of competing weeds. Cultivation equipment is normally quite simple and is designed to cut, slice, bury, or rip out weeds.

Chemical application equipment. Crop chemicals, such as fertilizers and pesticides, are routinely used in agricultural production. Application equipment for crop chemicals should apply a designated chemical in the proper amount, at the correct time, and to the correct target. Liquid pesticides are commonly applied with high- or low-pressure sprayers mounted on tractors, trailers, trucks, and aircraft. Air-blast or orchard-type sprayers utilize a high-speed air stream to carry the liquid chemical to the surface being treated.

Dry chemicals, such as fertilizers, are applied with either broadcast or band-type applicators. Broadcast applicators spread the chemical over the entire field surface at a prescribed rate, while band-type applicators apply the chemical at a prescribed rate in strips or bands across the field. *See* FERTILIZING.

Fig. 4. Hay cuber compressing alfalfa hay into 1.6-in.-square (4-cm-square) cubes. (*University of California*)

Fig. 5. Self-propelled combined harvester-thresher. (*Sperry New Holland, Division of Sperry Rand Corp.*)

The use of aircraft has revolutionized many farming operations. Rice is sown in flooded fields, and fertilizers and pesticides are applied from the air (**Fig. 2**). Herbicides and insecticides are also applied by air on other crops, such as grains, vegetables, and tree fruits. *See* AGRICULTURAL AIRCRAFT; HERBICIDE; INSECTICIDE.

Farmstead machinery and equipment. Mechanized operations around farmsteads vary from very little on subsistence-type farms to nearly complete on larger commercial farms. These operations are for either crop conditioning or materials handling. The equipment is powered by electric motors if it is stationary, and by tractor hydraulics or power takeoff shafts if it is portable (**Fig. 3**). Stationary equipment includes conveyors and grinders, and portable equipment includes feed wagons, mixers, and manure spreaders.

Hay, grain, and cotton harvesting equipment. Self-propelled hay cubers compress hay into small cubes to make it easy to handle with conveyors (**Fig. 4**). Practically all small grains are harvested by self-propelled combines (**Fig. 5**) that cut the crop and deliver the grain to a truck. While some corn for cattle and dairy cows is chopped for silage (stalk and grain), most of it is harvested as grain by harvesters similar to combines but fitted with harvester heads that strip the ears from six or more rows at a time and shell them. Cotton harvesting is done by stripping the fiber-containing bolls from the plant or by rotating spindles that collect the fibers with a twisting action and release them in an airstream that carries them into a hopper.

Fruit and vegetable harvesting. The biggest demand for hand labor has been in the harvest of fruits and vegetables. Much research has been done on breeding of plants and on machine development to reduce hand labor, primarily during harvesting, since that is the time of peak labor demand. Most successful attempts have utilized shaking of trees or vines to remove fruits. For some vegetables, such as cabbage and celery, the head or stalk is cut off by disks or knives. Most of these machines are nonselective, that is, all of the crop is harvested in one operation whether it is mature or not. Selection for color, maturity, and size is usually done by people on machines or at grading stations. Electronics are used for color maturity sorting on tomato harvesters (**Fig. 6**); each tomato passing through the harvester is inspected individually at a high rate of speed by an optical sensor. The optical sensor determines whether the tomato is red (mature) or green (not mature); if the tomato is not mature a mechanical system is triggered that removes it from the harvester.

There are several selective harvesters in use, but on a very limited basis. For example, an asparagus harvester has been developed that selects only spears of sufficient length, a cucumber harvester removes only the mature fruits, and a lettuce harvester electronically selects and removes those heads that are

Fig. 6. This tomato harvester cuts the vines, removes the fruit, and then loads it into a truck after the rejected fruit has been removed by the sorters. (*Blackwelder Mfg. Co.*)

Fig. 7. Pea pods being harvested and threshed by a self-propelled pea viner machine. (*FMC Co.*)

Fig. 8. These grapes are being harvested by beaters which impact the vines. The berries and bunches are collected on conveyors and then delivered to a dump trailer. (*Up-Right Inc.*)

large or dense enough. Harvesters for root crops such as potatoes, carrots, turnips, beets, and parsnips cut the tops and then perform a lifting action with a blade. This is followed by an open-chain conveyor that retains the roots while sifting out the soil. Sizing on the machine is usually done mechanically, while grading for quality is done by individuals who ride on the machines. Some of these harvesters have color sorters that remove some rejects as well as clods. While dried beans have been mechanically harvested for many years, it was only in the late 1970s that a green bean havester was developed. Harvesters for canning peas (**Fig. 7**) and peas for freezing are also available.

Some fruits and vegetables for the fresh produce market, such as citrus fruits, melons, grapes, eggplant, broccoli, and cauliflower, are still harvested by hand.

Machinery development. The increased mechanization in agriculture has resulted from economic and social factors. Economic forces have lowered profits per unit of production. In the free-market situation,

the only way for a farmer to improve total income has been to produce more units. This has led to larger farms and fewer farmers. The uncertainty of available seasonal or migrant labor for performing farm tasks has also speeded the farmers' steady move to mechanization.

Precision has received maximum emphasis in the development of new machinery. Plant scientists and design engineers have worked together to produce plant–machine compatibility in production and harvesting. For example, grain breeders have produced hybrid varieties that have their heads at uniform height and are easily threshed, and grape producers have trained the vines so that mechanical harvesters can clip or shake off the berries and retrieve them on conveyors (**Fig. 8**). *See* AGRICULTURAL SOIL AND CROP PRACTICES; AGRICULTURE; DAIRY MACHINERY.

James W. Rumsey; Michael O'Brien

Bibliography. D. Hunt, *Farm Power and Machinery Management*, 9th ed., 1995; R. A. Kepner et al., *Principles of Farm Machinery*, 1978; M. O'Brien et al., *Principles and Practices in Harvesting and Handling Fruits and Nuts*, 1980.

Agricultural meteorology

A branch of meteorology that examines the effects and impacts of weather and climate on crops, rangeland, livestock, and various agricultural operations. The branch of agricultural meteorology dealing with atmospheric-biospheric processes occurring at small spatial scales and over relatively short time periods is known as micrometeorology, sometimes called crop micrometeorology for managed vegetative ecosystems and animal biometeorology for livestock operations. The branch that studies the processes and impacts of climatic factors over larger time and spatial scales is often referred to as agricultural climatology. *See* CLIMATOLOGY; MICROMETEOROLOGY.

Agricultural meteorology, or agrometeorology, addresses topics that often require an understanding of biological, physical, and social sciences. It studies processes that occur from the soil depths where the deepest plant roots grow to the atmospheric levels where seeds, spores, pollen, and insects may be found. Agricultural meteorologists characteristically interact with scientists from many disciplines.

Role of meteorologists. Agricultural meteorologists collect and interpret weather and climate data needed to understand the interactions between vegetation and animals and their atmospheric environments. The climatic information developed by agricultural meteorologists is valuable in making proper decisions for managing resources consumed by agriculture, for optimizing agricultural production, and for adopting farming practices to minimize any adverse effects of agriculture on the environment. Such information is vital to ensure the economic and environmental sustainability of agriculture now and in the future. *See* WEATHER OBSERVATIONS.

Agricultural meteorologists also quantify, evaluate, and provide information on the impact and consequences of climate variability and change on agriculture. Increasingly, agricultural meteorologists assist policy makers in developing strategies to deal with climatic events such as floods, hail, or droughts and climatic changes such as global warming and climate variability.

Agricultural meteorologists are involved in many aspects of agriculture, ranging from the production of agronomic and horticultural crops, trees, and livestock to the final delivery of agricultural products to market. They study the energy and mass exchange processes of heat, carbon dioxide, water vapor, and trace gases such as methane, nitrous oxide, and ammonia, within the biosphere on spatial scales ranging from a leaf to a watershed and even to a continent. They study, for example, the photosynthesis, productivity, and water use of individual leaves, whole plants, and fields. They also examine climatic processes at time scales ranging from less than a second to more than a decade.

Crop micrometeorology. This branch of agricultural meteorology deals with the study of mass, radiation, and energy exchanges between plants and their surrounding physical environment. Fundamental is an understanding of the physical laws that govern the interactions between vegetation and radiation in the various wavelength regions of the solar spectrum—including ultraviolet radiation, photosynthetically active radiation, and near-infrared radiation, and thermal infrared radiation. Likewise, the laws that govern the conduction and convection of heat to plants, soil, and the air along with the latent heat exchange and the turbulent transport of water vapor and carbon dioxide between the plants and the atmosphere are central to crop micrometeorology.

Crop micrometeorologists, for example, cooperate with other scientists to change plant and canopy architecture and improve management practices to optimize radiation use efficiency, defined as the amount of plant material produced per unit of solar radiation intercepted; and to increase water use efficiency, defined as the amount of plant material produced per unit of water used. They often work with agronomists and biological engineers to develop methods to promote the efficient use of irrigation and to enhance the capture and storage of precipitation. They also assist in the study and development of practices such as windbreaks and frost protection schemes to modify the plant environment to benefit crop production.

Crop ecology and phenology. The specific types of crops and crop cultivars that survive and grow well in a given area are determined primarily by climatic factors such as daylength, precipitation during the growing season, maximum and minimum temperatures, and the length of the frost-free period. Agricultural meteorologists develop an understanding of the relationships between the climate and the growth, yield, and phenological (growth stage) development of crops. They then use such relationships to evaluate areas for sustaining the growth and survival of specific crop types.

Relationships between the accumulation of thermal time (also referred to as heat units or growing degree days) and the phenological stages of various crops have been developed. These relationships provide information valuable for planning of critical farm operations during the proper growth stage. For example, by using thermal time to predict the period of corn pollination, a farmer might irrigate to provide adequate soil water to prevent water stress during this critical growth stage.

Crop production and growing season. Crop productivity is affected by the weather. The weather or climate of an area determines the crop growing season, defined as the period of time when climatic conditions are favorable for plant growth. Typically the growing season is defined as the length of time between the first and last killing frost in an area, but in some locations the length of the growing season is set by the period when adequate water is available to the crops. During the growing season the most important variables affecting crop production are precipitation, temperature, and solar radiation. Wind and the carbon dioxide concentration in the air are also important. Sometimes it is possible to adopt practices that will alter the microclimate to maintain a more favorable atmospheric environment for crop production. Agricultural meteorologists have helped develop practices such as irrigation and the establishment of windbreaks, which modify and improve the crop microclimate.

Important climatic factors that characterize the growth environment for crops are the length of the growing season; the amount, timing, and intensity of rainfall; and the patterns of plant water use (evapotranspiration). Agricultural meteorologists often collaborate with agronomists, horticulturists, and agricultural economists to evaluate the effects of changing and variable climatic patterns on crop production and the economic value of crops. For example, the estimated levels of crop production at various growth stages made by agricultural meteorologists are often used to plan worldwide food reserves and to estimate future commodity prices. Such information can be used to influence crop planting and production patterns. *See* AGRICULTURAL SOIL AND CROP PRACTICES.

Climate change and agricultural production. Any change in climate will affect agricultural production, either beneficially or deleteriously. Because agricultural meteorologists study the interactions between climatic factors and crop growth and yield, they are able to make a major contribution to understanding the impact of potential climate change on agriculture. For several decades the concentration of carbon dioxide and other radiatively active gases has been increasing, due primarily to the combustion of fossil fuels and secondarily to the clearing of forests for agricultural purposes. Many climatologists believe that these increasing concentrations of carbon dioxide, methane, nitrous oxide, and the chloro- fluorohydrocarbons and other radiatively active trace gases are increasing the global temperature and altering regional precipitation patterns. There is considerable disagreement among scientists about the magnitude of the change and the impact that these changes will have on agriculture. Many believe that the increasing concentration of carbon dioxide will increase crop production and improve the water use efficiency of crops. Changing temperature patterns may cause a shift in areas where certain crops are grown. Agricultural meteorologists face numerous challenges in assembling and interpreting the facts about climatic change to assess the future real impacts on agriculture and to help agriculturists adapt to climate change. *See* CLIMATE HISTORY.

Water use efficiency. Agricultural meteorologists conduct research aimed at improving the water use efficiency of crops. Throughout much of the world, crop productivity is limited by inadequate precipitation. Irrigation has been widely used around the world to supplement natural rainfall. However, crops require large amounts of water, and in many areas agriculture has competed for increasingly limited water resources. Agricultural meteorologists play an important role in ensuring that water is used as efficiently as possible by agricultural producers. They work with plant breeders to select crops that improve plant water use efficiency or, in other words, that are more conservative in the use of water. They develop methods for providing reliable and accurate estimates of the amount of water consumed by crops. This information is needed to determine when to schedule irrigation and how much water to apply. They also assist in the development of other practices to promote more efficient use of water. These practices include selection of optimum row widths, use of mulches, and adoption of minimum tillage methods. *See* IRRIGATION (AGRICULTURE).

Plant and animal diseases and pests. The total productivity of agriculture is often considerably reduced by plant and animal diseases and pests. Outbreaks of diseases and insect infestations are controlled in large measure by climatic variables, the most important of which are wind, temperature, humidity, and dew. In some cases, insects and disease-causing organisms are carried long distances by the wind, so synoptic weather patterns are studied to predict where disease or pest outbreaks are expected.

In general, there is sufficient inoculum for plant pathogens and adequate numbers of harmful insects always present in the agricultural environment to cause severe economically damaging outbreaks if favorable microclimatic conditions occur. Agricultural meteorologists work with plant and animal pathologists and entomologists to construct mathematical models to describe the dynamic processes of disease and pest outbreaks. They develop models that describe the conditions favorable for the development of beneficial organisms useful in the biological control of some pests and diseases. Accurate prediction of disease and pest outbreaks facilitates the ability of agriculturists to control diseases and pests through proper use of cultural practices, pesticides, and biological control methods. Teams of agricultural meteorologists, entomologists, and plant pathologists will increasingly find ways to reduce the destructive

influence of diseases and insects on agriculture, while reducing pesticide use. *See* ENTOMOLOGY, ECONOMIC; INSECT CONTROL, BIOLOGICAL; PLANT PATHOLOGY.

Pesticide application. In addition to predicting the outbreaks of diseases and insects, agricultural meteorologists play an important role in defining the environmental conditions for the application of pesticides for maximum benefit. They also describe the environmental conditions when pesticides should or should not be applied. Information on the effects of wind and temperature gradients on the airborne drift of a pesticide after it is released from an aircraft is required for its safe aerial application. Agricultural meteorologists in cooperation with weed scientists, entomologists, aircraft operators, and pesticide manufacturers help define conditions for effective pesticide application. *See* AGRICULTURAL AIRCRAFT.

Agricultural meteorologists help to determine the fate of pesticides and fertilizers. Many chemicals break down in the soil or are taken up by plants and do not harm in any way the environment. However, some chemicals are carried to the ground water or into water bodies, and others volatize and are carried into the atmosphere by wind and turbulence. The fate of many agricultural chemicals is dependent on climatic factors. For example, excessive rains can carry chemicals from the soil into streams, rivers, ponds, lakes, oceans, or ground water. Winds and high temperature may act on agricultural chemicals to cause air-pollution problems. By defining clearly the role that climate plays on the fate of chemicals, agricultural meteorologists, working with other scientists, can assist in developing appropriate management strategies to reduce water and air pollution. *See* AIR POLLUTION; FERTILIZER; GROUND-WATER HYDROLOGY; PESTICIDE; WATER POLLUTION.

Frost protection. Production of some agronomic and horticultural crops may be severely limited by frost, when frost comes before the crops are ready to be harvested. Protecting agronomic crops from frost damage can rarely be economically justified. However, agricultural meteorologists can define optimal planting dates for a given site that will reduce the chance of frost damage.

By determining weather conditions that cause frost, agricultural meteorologists can suggest the best frost protection methods for a particular type of frost and discourage the use of the least successful methods. Effective forecasting of frost events and the use of effective methods for altering the microclimate that reduce the occurrence or intensity of a frost can help prevent millions of dollars in crop losses. For high-value crops such as vegetables and fruits, especially citrus fruits, agricultural meteorologists play important roles in predicting the occurrence of frost and in developing methods that protect against frost damage. The best method of frost protection is proper site selection. Agricultural meteorologists can help in selecting the most appropriate site for growing high-value crops by defining the microclimate of certain areas and evaluating the effects of large bodies of water and topographic position on these microclimates. They can determine the appropriate sites on hillsides for orchards to take advantage of a warm layer of air that often exists on hillsides. They can also determine sites where cold air pools as it drains to lower elevations. *See* FROST.

Wind damage and erosion. Wind sometimes causes mechanical damage to trees and crops, and often contributes to an environment that promotes plant stress. Windbreaks are an effective way to reduce mechanical damage by wind and to provide a modified microclimate that reduces plant water use, improves plant water use efficiency, and increases crop productivity. Agricultural meteorologists play an important role in designing windbreaks that are most effective in bringing about the desired microclimate modification. They help to determine and quantify the climatic parameters that have been modified.

Wind causes soil erosion, which results in very serious damage and depletes a vital natural agricultural resource. Working with soil scientists, agricultural meteorologists develop equations to describe the amount of soil movement under given wind and surface conditions. They also suggest approaches to reduce erosion, such as roughing the soil surface, leaving crop residue on the surface, and constructing windbreaks. *See* SOIL CONSERVATION.

Simulation models. Since the late 1970s, agricultural meteorologists have assisted in the development of mathematical models of crop and livestock behavior based on various climatological and physiological parameters. The earliest models were statistical and were developed by correlating climatic data with yields of various crops or with livestock performance. These empirical models generally included temperature and precipitation as the most important parameters. They work reasonably well, except when the weather deviates significantly from normal. Of course, estimates of the deviation of yields are most valuable for those years with abnormal temperature and precipitation patterns. Since the advent of computers, more mechanistic models have been developed. Mechanistic models are based on sound physical and physiological theories, and they perform quite satisfactorily. Because these models account for changing weather during a growing season, they can help improve management practices for crop and livestock production. Expert systems and artificial intelligence are useful for suggesting management practices that achieve increased crop and animal production. *See* ARTIFICIAL INTELLIGENCE; CLIMATE MODELING; EXPERT SYSTEMS; SIMULATION.

Livestock production. The performance of livestock is influenced by weather and climate. In many cases, the microclimate has been modified and controlled by confining animals to buildings. Certain agricultural meteorologists, known as biometeorologists, work with animal scientists to design buildings with an environment favorable for efficient feed utilization and improved production and efficiency. They also develop indices to evaluate the degree of stress that animals experience under conditions of high heat and humidity. With the use

of weather forecasts, these indices alert livestock producers to initiate stress-relieving procedures and reduce production losses by avoiding shipment of animals under these conditions. Agricultural meteorologists also study the effects of cold stress on animals in confined environments. Results may suggest, for example, the need to establish windbreaks for animals to use for shelter to reduce cold stress in unconfined environments. *See* AGRICULTURAL SCIENCE (ANIMAL).

Developing strategies and policy. Agricultural meteorologists work with governmental officials and other policy makers to develop, formulate, and adopt reasonable strategies and policies for dealing with extreme climatic events such as drought and floods. Agricultural meteorologists, working with social scientists, also determine the likely impacts of extreme climatic events and climate change on agriculture and all elements of society and identify rational response options to mitigate those impacts. *See* METEOROLOGY.

Blaine L. Blad

Bibliography. B. J. Barfield and J. F. Gerber (eds.), *Modification of the Aerial Environment of Crops*, 1979; G. S. Campbell and J. M. Norman, *An Introduction to Environmental Biophysics*, 1988; D. M. Gates, *Biophysical Ecology*, 1980; I. F. Griffiths (ed.), *Handbook of Agricultural Meteorology*, 1994; W. P. Lowry, *Atmospheric Ecology for Designers and Planners*, 1988; J. L. Monteith (ed.), *Vegetation and the Atmosphere*, vols. 1 and 2, 1977; N. J. Rosenberg, B. L. Blad, and S. B. Verma, *Microclimate: The Biological Environment*, 1983; D. A. Wilhite (ed.), *Drought Assessment, Management and Planning: Theory and Case Studies*, 1993.

Agricultural science (animal)

The science which deals with the selection, breeding, nutrition, and management of domestic animals for economical production of meat, milk, eggs, wool, hides, and other animal products. Horses for draft and pleasure, dogs and cats for pets, rabbits for meat production, and bees for honey production may also be included in this group. *See* BEEF CATTLE PRODUCTION; BEEKEEPING; DAIRY CATTLE PRODUCTION; GOAT PRODUCTION; HORSE PRODUCTION; MULE PRODUCTION; POULTRY PRODUCTION; SHEEP; SWINE PRODUCTION.

When primitive societies first domesticated animals, they were kept as means of meeting the immediate needs for food, transportation, and clothing. Sheep probably were the first and most useful animals to be domesticated, furnishing milk and meat for food, and hides and wool for clothing.

As chemistry, physiology, anatomy, genetics, nutrition, parasitology, pathology, and other sciences developed, their principles were applied to the field of animal science. Since the beginning of the twentieth century, great strides have been made in livestock production. Today, farm animals fill a highly important place in human existence. They convert raw materials, such as pasture herbage, which are of little use to people as food, into animal products having nutritional values not directly available in plant products.

Ruminant animals (those with four compartments or stomachs in the fore portion of their digestive tract, such as cattle and sheep) have the ability to consume large quantities of roughages because of their particular type of digestive system. They also consume large tonnages of grains, as well as mill feeds, oil seed meals, industrial and agricultural byproducts, and other materials not suitable for human food.

Products of the animal industry furnish raw materials for many important processing industries, such as meat packing, dairy manufacturing, poultry processing, textile production, and tanning. Many services are based on the needs of the animal industry, including livestock marketing, milk deliveries, poultry and egg marketing, poultry hatcheries, artificial insemination services, feed manufacturing, pharmaceutical industry, and veterinary services. Thus, animal science involves the application of scientific principles to all phases of animal production, furnishing animal products efficiently and abundantly to consumers. Products from animals are often used for consumer products other than food, for example, hides for leather, and organ meats for preparation of drugs and hormones.

Livestock breeding. The breeding of animals began thousands of years ago. During the last half of the nineteenth century, livestock breeders made increasing progress in producing animals better suited to the needs of humans by simply mating the best to the best. However, in the twentieth century animal breeders began to apply the scientific principles of genetics and reproductive physiology. Some progress made in the improvement of farm animals resulted from selected matings based on knowledge of body type or conformation. This method of selection became confusing due to the use of subjective standards which were not always related to economic traits. This error was corrected, however, and many breeders of dairy cattle, poultry, beef cattle, sheep, and swine in the mid-1970s made use of production records or records of performance. Some of their breeding plans were based on milk fat production or egg production, as well as on body type or conformation. The keeping of poultry and dairy cow production records began in a very limited way late in the nineteenth century. The first Cow-Testing Association in the United States was organized in Michigan in 1906. Now over 1,500,000 cows are tested regularly in the United States. *See* BREEDING (ANIMAL).

Many states have production testing for beef cattle, sheep, and swine, in which records of rate of gain, efficiency of feed utilization, incidence of twinning, yield of economically important carcass cuts, and other characteristics of production are maintained on part or all of the herd or flock. These records serve as valuable information in the selection of animals for breeding or sale.

Breeding terminology. A breed is a group of animals that has a common origin and possesses

characteristics that are not common to other individuals of the same species.

A purebred breed is a group that possesses certain fixed characteristics, such as color or markings, which are transmitted to the offspring. A record, or pedigree, is kept which describes their ancestry for five generations. Associations were formed by breeders primarily to keep records, or registry books, of individual animals of the various breeds. Purebred associations take active roles in promoting and improving the breeds.

A purebred is one that has a pedigree recorded in a breed association or is eligible for registry by such an association. A crossbred is an individual produced by utilizing two or more purebred lines in a breeding program. A grade is an individual having one parent, usually the sire, a purebred and the other parent a grade or scrub. A scrub is an inferior animal of nondescript breeding. A hybrid is one produced by crossing parents that are genetically pure for different specific characteristics. The mule is an example of a hybrid animal produced by crossing two different species, the American jack, *Equus asinus*, with a mare, *E. caballus*.

Systems of breeding. The modern animal breeder has genetic tools such as selection and breeding, and inbreeding and outbreeding. Selection involves the retaining or rejecting of a particular animal for breeding purposes, based largely on qualitative characteristics. Inbreeding is a system of breeding related animals. Outbreeding is a system of breeding unrelated animals. When these unrelated animals are of different breeds, the term crossbreeding is usually applied. Crossbreeding is in common use by commercial swine producers. About 80–90% of the hogs produced in the Corn Belt states are crossbred. Crossbreeding is also used extensively by commercial beef and sheep producers.

Grading-up is the process of breeding purebred sires of a given breed to grade females and their female offspring for generation after generation. Grading-up offers the possibility of transforming a nondescript population into one resembling the purebred sires used in the process. It is an expedient and economical way of improving large numbers of animals.

Formation of new breeds. New breeds of farm animals have been developed from crossbred foundation animals. Montadale, Columbia, and Targhee are examples of sheep breeds so developed. The Santa Gertrudis breed of beef cattle was produced by crossing Brahman and Shorthorn breeds on the King Ranch in Texas. In poultry, advantage has been taken of superior genetic ability through the development of hybrid lines. *See* GENETICS.

Selection of animals for breeding. The evaluation or selection of animals for breeding purposes is of importance to the commercial as well as to the purebred breeder. In selecting animals for breeding, desirable conformation or body type is given careful attention. The animals are also examined carefully for visible physical defects, such as blindness, crooked legs, jaw distortions, and abnormal udders. Animals known to be carriers of genes for heritable defects, such as dwarfism in cattle, should be discriminated against.

When they are available, records of performance or production should be considered in the selection of breeding animals. Some purebred livestock record associations now record production performance of individual animals on pedigrees.

Genetic engineering is a new frontier in animal breeding. For example, mature eggs are collected and fertilized in test tubes, and then a fertilized egg is placed in a uterus where it develops into a fetus and eventually an offspring. There are other techniques, largely experimental, that may be practical, such as embryo splitting for the production of identical twins. Identical twins are exceedingly important to the researcher, since experiments can be conducted without involving a large number of animals. They are also important when saving rare and endangered species. Another genetic engineering technique is fusing embryos of different species, and this method can also be used to save the genetic material of endangered species. *See* GENETIC ENGINEERING.

Artificial insemination. In this process spermatozoa are collected from the male and deposited in the female genitalia by instruments rather than by natural service. In the United States this practice was first used for breeding horses. Artificial insemination in dairy cattle was first begun on a large scale in New Jersey in 1938. Freezing techniques for preserving and storing spermatozoa have been applied with great success to bull semen, and it is now possible for outstanding bulls to sire calves years after the bulls have died. The use of artificial insemination for beef cattle and poultry (turkeys) is common practice. There are also drugs which stimulate beef cow herds to come into heat (ovulate) at approximately the same time. This permits the insemination of large cow herds without the individual handling and inspection which is used with dairy cattle. Although some horses are bred by artificial insemination, many horse breed associations allow only natural breeding.

Livestock feeding. Scientific livestock feeding involves the systematic application of the principles of animal nutrition to the feeding of farm animals. The science of animal nutrition has advanced rapidly since 1930, and the discoveries are being utilized by most of those concerned with the feeding of livestock. The nutritional needs and responses of the different farm animals vary according to the functions they perform and to differences in the anatomy and physiology of their digestive systems. Likewise, feedstuffs vary in usefulness depending upon the time and method of harvesting the crop, the methods employed in drying, preserving, or processing, and the forms in which they are offered to the animals. *See* ANIMAL FEEDS.

Chemical composition of feedstuffs. The various chemical compounds that are contained in animal feeds are divided into groups called nutrients. These include proteins, fats, carbohydrates, vitamins, and mineral matter. Proteins are made up of amino acids. Twelve amino acids are essential for all nonruminant animals and must be supplied in their diets. Fats and

carbohydrates provide mainly energy. In most cases they are interchangeable as energy sources for farm animals. Fats furnish 2.25 times as much energy per pound as do carbohydrates because of their higher proportion of carbon and hydrogen to oxygen. Thus the energy concentration in poultry and swine diets can be increased by inclusion of considerable portions of fat. Ruminants cannot tolerate large quantities of fat in their diets, however. *See* CARBOHYDRATE; LIPID; PROTEIN.

Vitamins essential for health and growth include fat-soluble A, D, E, and K, and water-soluble vitamins thiamine, riboflavin, niacin, pyrodoxine, pantothenic acid, and cobalamin. *See* VITAMIN.

Mineral salts that supply calcium, phosphorus, sodium, chlorine, and iron are often needed as supplements, and those containing iodine and cobalt may be required in certain deficient areas. Zinc may also be needed in some swine rations. Many conditions of mineral deficiency have been noted by using rations that were not necessarily deficient in a particular mineral but in which the mineral was unavailable to the animal because of other factors in the ration or imbalances with other minerals. For example, copper deficiency can be caused by excess molybdenum in the diet.

By a system known as the proximate analysis, developed prior to 1895 in Germany, feeds have long been divided into six fractions including moisture, ether extract, crude fiber, crude protein, ash, and nitrogen-free extract. The first five fractions are determined in the laboratory. The nitrogen-free extract is what remains after the percentage sum of these five has been subtracted from 100%. Although proximate analysis serves as a guide in the classification, evaluation, and use of feeds, it gives very little specific information about particular chemical compounds in the feed.

The ether extract fraction includes true fats and certain plant pigments, many of which are of little nutritional value.

The crude fiber fraction is made up of celluloses and lignin. This fraction, together with the nitrogen-free extract, makes up the total carbohydrate content of a feed. *See* CELLULOSE.

The crude protein is estimated by multiplying the total Kjeldahl nitrogen content of the feed by the factor 6.25. This nitrogen includes many forms of nonprotein as well as protein nitrogen. *See* NITROGEN.

The ash, or mineral matter fraction, is determined by burning a sample and weighing the residue. In addition to calcium and other essential mineral elements, it includes silicon and other nonessential elements.

The nitrogen-free extract (NFE) includes the more soluble and the more digestible carbohydrates, such as sugars, starches, and hemicelluloses. Unfortunately, most of the lignin, which is not digestible, is included in this fraction. *See* HEMICELLULOSE.

A much better system of analysis was developed at U.S. Department of Agriculture laboratories for the crude fiber fraction of feedstuffs. This system separates more specifically the highly digestible cell-soluble portion and the less digestible fibrous portion of plant cell walls.

Digestibility of feeds. In addition to their chemical composition or nutrient content, the nutritionist and livestock feeder should know the availability or digestibility of the different nutrients in feeds. The digestibility of a feed is measured by determining the quantities of nutrients eaten by an animal over a period of time and those recoverable in the fecal matter. By assigning appropriate energy values to the nutrients, total digestible nutrients (TDN) may be calculated. These values have been determined and recorded for a large number of feeds.

Formulation of animal feeds. The nutritionist and livestock feeder finds TDN values of great use in the formulation of animal feeds. The TDN requirements for various classes of livestock have been calculated for maintenance and for various productive capacities. However, systems have been developed for expressing energy requirements of animals or energy values of feeds in units which are more closely related to the body process being supported (such as maintenance, growth, and milk or egg production). Tables of feeding standards have been published using the units of metabolizable energy and net energy, which are measurements of energy available for essential body processes. Recommended allowances of nutrients for all species of livestock and some small animals (rabbit, dogs, and mink) are published by the National Academy of Sciences–National Research Council.

Nutritional needs of different animals. The nutritional requirements of different classes of animals are partially dependent on the anatomy and physiology of their digestive systems. Ruminants can digest large amounts of roughages, whereas hogs and poultry, with simple stomachs, can digest only limited amounts and require more concentrated feeds, such as cereal grains. In simple-stomached animals the complex carbohydrate starch is broken down to simple sugars which are absorbed into the blood and utilized by the body for energy.

Microorganisms found in the rumen of ruminant animals break down not only starch but the fibrous carbohydrates of roughages, namely, cellulose and hemicellulose, to organic acids which are absorbed into the blood and utilized as energy. Animals with simple stomachs require high-quality proteins in their diets to meet their requirements for essential amino acids. On the other hand, the microorganisms in ruminants can utilize considerable amounts of simple forms of nitrogen to synthesize high-quality microbial protein which is, in turn, utilized to meet the ruminant's requirement for amino acids. Thus, many ruminant feeds now contain varying portions of urea, an economical simple form of nitrogen, which is synthesized commercially from nonfeed sources. Simple-stomached animals require most of the vitamins in the diet. The microorganisms in the rumen synthesize adequate quantities of the water-soluble vitamins to supply the requirement for the ruminant animal. The fat-soluble vitamins A, D,

and E must be supplied as needed to all farm animals. Horses and mules have simple stomachs but they also have an enlargement of the cecum (part of the large intestine), in which bacterial action takes place similar to that in the rumen of ruminants. The requirements for most nutrients do not remain the same throughout the life of an animal but relate to the productive function being performed. Therefore, the requirements are much higher for growth and lactation than they are for pregnancy or maintenance.

Livestock judging. The evaluation, or judging, of livestock is important to both the purebred and the commercial producer. The purebred producer usually is much more interested in show ring judging, or placings, than is the commercial producer. Because of the short time they are in the show ring, the animals must be placed on the basis of type or appearance by the judge who evaluates them. The show ring has been an important influence in the improvement of livestock by keeping the breeders aware of what judges consider to be desirable types. The shows have also brought breeders together for exchange of ideas and breeding stock and have helped to advertise breeds of livestock and the livestock industry. The demand for better meat-animal carcasses has brought about more shows in which beef cattle and swine are judged, both on foot and in the carcass. This trend helps to promote development of meat animals of greater carcass value and has a desirable influence upon show-ring standards for meat animals. The standards used in show rings have shifted toward traits in live animals which are highly related to both desirable carcass traits and high production efficiency.

Grading of market animals. The grading on foot of hogs or cattle for market purposes requires special skill. In many modern livestock markets, hogs are graded as no. 1, 2, 3, or 4 according to the estimated values of the carcasses. Those hogs grading no. 3 are used to establish the base price, and sellers are paid a premium for better animals. In some cases the grade is used also to place a value on the finished market product. For example, the primal cuts from beef cattle are labeled prime, choice, or good on the basis of the grade which the rail carcass received.

Livestock disease and pest control. The numerous diseases of farm livestock require expert diagnosis and treatment by qualified veterinarians. The emphasis on intensive animal production has increased stresses on animals and generally increased the need for close surveillance of herds or flocks for disease outbreaks.

The transmission of infectious diseases is a three-stage process involving the reservoir of the disease organisms, the mode of transmission, and the susceptible animal. A disease can be thwarted at each stage: at the reservoir by isolation and quarantine, at the transmission stage by good hygiene, and in the susceptible animal by immunization and antibiotics. The most efficacious point for disease control varies with the particular disease etiology.

Some diseases are common to several classes of livestock. For example, the following diseases may affect cattle, sheep, and goats: actinomycosis, anthrax, blackleg, brucellosis, leptospirosis, listeriosis, mastitis, pinkeye, and shipping fever. Other diseases generally affect only a particular species of livestock.

Both external and internal parasites are common afflictions of livestock but can be controlled by proper management of the animals. Sanitation is of utmost importance in the control of these pests, but under most circumstances sanitation must be supplemented with effective insecticides, ascaricides, and fungicides. *See* FUNGISTAT AND FUNGICIDE; INSECTICIDE; PESTICIDE.

Internal parasites. Internal parasites, such as stomach and intestinal worms in sheep, cannot be controlled by sanitation alone under most farm conditions. They are a more critical problem under intensive management systems and in warm, humid climates. For many years the classic treatment for sheep was drenching with phenothiazine and continuous free choice feeding of one part phenothiazine mixed with nine parts of salt. Drugs have been developed which more effectively break the life cycle of the worm and have a broader spectrum against different classes of parasites.

Control of gastrointestinal parasites in cattle can be accomplished in many areas by sanitation and the rotational use of pastures. In areas of intensive grazing, animals, especially the young ones, may become infected. Regular and timely administration of antiparasite drugs is the customary means of controlling the pests. Otherwise their effects will seriously decrease the economic productivity of animals. *See* MEDICAL PARASITOLOGY.

The use of drugs to control gastrointestinal parasites and also certain enteric bacteria in hogs is commonplace. Control is also dependent on good sanitation and rotational use of nonsurfaced lots and pastures. Similarly, because of intensive housing systems, the opportunities for infection and spread of both parasitism and disease in poultry flocks are enhanced by poor management conditions. The producer has a large number of materials to choose from in preventing or treating these conditions, including sodium fluoride, piperazine salts, nitrofurans, and arsenicals and antibiotics such as penicillin, tetracyclines, and hygromycin.

External parasites. Control of horn flies, horseflies, stable flies, lice, mange mites, ticks, and fleas on farm animals has been changed with insecticides. Such compounds as methoxychlor, toxaphene, lindane, and malathion were very effective materials for the control of external parasites. The use of these materials has been restricted to certain conditions and classes of animals by law. Reliable information should be obtained before using these materials for the control of external parasites.

Control of cattle grubs, or the larvae of the heel fly, may be accomplished by dusting the backs of the animals with powders or by spraying them under high pressure. Systemic insecticides for grub control are approved if they are used according to the manufacturer's recommendation.

Fungus infections. Actinomycosis is a fungus disease commonly affecting cattle, swine, and horses. In cattle this infection is commonly known as lumpy jaw. The lumpy jaw lesion may be treated with tincture of iodine or by local injection of streptomycin in persistent cases. Most fungus infections, or mycoses, develop slowly and follow a prolonged course. A veterinarian should be consulted for diagnosis and treatment.

General animal health care. Numerous other disease organisms pose a constant threat to livestock. Although many of these organisms can be treated therapeutically, it is much more advisable economically to establish good preventive medicine and health care programs under the guidance of a veterinarian.

Ronald R. Johnson; W. A. Williams; R. Albaugh

Bibliography. D. Acker and M. Cunningham, *Animal Science and Industry*, 5th ed., 1997; J. R. Campbell and J. F. Lasley, *The Science of Animals That Serve Mankind*, 3d ed., 1985; J. R. Campbell and R. T. Marshall,*The Science of Providing Milk for Man*, 1975; M. E. Ensminger, *Beef Cattle Science*, 6th ed., 1987; M. E. Ensminger, *Horses and Horsemanship*, 7th ed., 1998; M. E. Ensminger, *Sheep and Goat Science*, 1998; J. L. Krider and W. E. Carroll, *Swine Production*, 5th ed., 1982; M. O. North, *Commercial Chicken Production Manual*, 3d ed., 1984; J. F. Timoney et al., *Hagan and Bruner's Microbiology and Infectious Diseases of Domestic Animals*, 8th ed., 1988.

Agricultural science (plant)

The pure and applied science that is concerned with botany and management of crop and ornamental plants for utilization by humankind. Crop plants include those grown and used directly for food, feed, or fiber, such as cereal grains, soybeans, citrus, and cotton; those converted biologically to products of utility, such as forage plants, hops, and mulberry; and those used for medicinal or special products, such as digitalis, opium poppy, coffee, and cinnamon. In addition, many plant products such as crambe oil and rubber are used in industry where synthetic products have not been satisfactory. Ornamental plants are cultured for their esthetic value. *See* FLORICULTURE; ORNAMENTAL PLANTS.

The ultimate objective of plant agriculture is to recognize the genetic potential of groups of plants and then to manipulate and utilize the environment to maximize that genetic expression for return of a desirable product. Great advancements in crop culture have occurred by applying knowledge of biochemistry, physiology, ecology, morphology, anatomy, taxonomy, pathology, and genetic engineering. Contributions of improved plant types by breeding, and the understanding and application of principles of atmospheric science, soil science, and animal and human nutrition, have increased the efficiency and decreased the risks of crop production. *See* GENETIC ENGINEERING.

Domestication of crop plants. All crops are believed to have been derived from wild species. However, cultivated plants as they are known today have undergone extensive modification from their wild prototypes as a result of the continual efforts to improve them. These wild types were apparently recognized as helpful to humans long before recorded history. Desirable plants were continually selected and replanted in order to improve their growth habit, fruiting characteristics, and growing season. Selection has progressed so far in cases such as cabbage and corn (maize) that wild ancestors have become obscure.

Centers of origins of most crop plants have been determined to be in Eurasia, but many exceptions exist. This area of early civilization apparently abounded with several diverse plant types that led to domestication of the crops known today as wheat, barley, oats, millet, sugarbeets, and most of the cultivated forage grasses and legumes. Soybeans, lettuce, onions, and peas originated in China and were domesticated as Chinese civilization developed. Similarly, many citrus fruits, banana, rice, and sugarcane originated in southern Asia. Sorghum and cowpeas are believed to have originated in Africa. Crops which were indigenous to Central and South America but were brought to North America by migrating peoples include corn, potato, sweet potato, pumpkin, sunflower, tobacco, and peanut. Thus, prior to 1492 there was little, if any, mixing of crop plants and cultural practices between the Old and the New Worlds. Most of the major agricultural crops of today in the United States awaited introduction by early settlers, and later by plant explorers. *See* ORIENTAL VEGETABLES.

During domestication of crop plants the ancient cultivators in their geographically separated civilizations must have had goals similar to present-day plant breeders. Once valuable attributes of a plant were recognized, efforts were made to select the best types for that purpose. Desirable characteristics most likely included improved yield, increased quality, extended range of adaptation, insect and disease resistance, and easier cultural and harvesting operations. About 350,000 species of plants exist in the world, yet only about 10,000 species can be classified as crops by using the broadest of definitions. Of these, about 150 are of major importance in world trade, and only 15 make up the majority of the world's food crops. On a world basis, wheat is grown on the most acreage, followed by rice, but rice has a higher yield per area than wheat so their total production is about equal. Other major crops are corn, sorghum, millet, barley; sugarcane and sugarbeet; potato, sweet potato, and cassava; bean, soybean, and peanut; and coconut and banana.

Redistribution of crop plants. Crop distribution is largely dictated by growth characteristics of the crop, climate of the region, soil resources, and social habits of the people. As plants were domesticated by civilizations in separate parts of the world, they were discovered by early travelers. This age of exploration led to the entrance of new food crops into European

agriculture. The potato, introduced into Spain from the New World before 1570, was to become one of the most important crops of Europe. When introduced into Ireland, it became virtually the backbone of the food source for that entire population. Corn was introduced to southern Europe and has become an important crop. Rice has been cultivated in Italy since the sixteenth century. Tobacco was also introduced to European culture, but remained a major source of export of the early colonial settlements in America.

European agriculture reciprocated by introducing wheat, barley, oats, and several other food and feed crops into the New World. In the new environment, plants were further adapted by selection to meet the local requirements. Cultural technology regarding seeding, harvesting, and storing was transferred along with the crops. This exchange of information oftentimes helped allow successful culture of these crops outside of their center of origin and domestication. Today the center of producton of a crop such as wheat in the United States and potato in Europe is often markedly different from its center of origin.

The United States recognized early the need for plant exploration to find desirable types that could be introduced. Thomas Jefferson wrote in 1790, "The greatest service which can be rendered to any country is to add a useful plant to its culture." Even today the U.S. Department of Agriculture conducts plant explorations and maintains plant introduction centers to evaluate newly found plants. Explorations are also serving a critical need for preservation of germplasm for plant breeders, as many of the centers of origin are becoming agriculturally intensive, and wild types necessary to increase genetic diversity will soon be extinct.

Adaptation of crop plants. Crop plants of some sort exist under almost all environments, but the major crops on a world basis tend to have rather specific environmental requirements. Furthermore, as crop plants are moved to new locations the new environment must be understood, and cultural or management changes often must be made to allow best performance. Varieties and strains of crops have been specifically developed that can better cope with cold weather, low rainfall, diseases, and insects to extend even further the natural zone of adaptation.

Temperature. Temperature has a dominant influence on crop adaptation. The reason is that enzyme activity is very temperature-dependent and almost all physiological processes associated with growth are enzymatically controlled. Crops differ widely in their adapted temperature range, but most crops grow best at temperatures of 59–88°F (15–32°C). Optimum day temperature for wheat, however, is 68–77°F (20–25°C), while for corn it is about 86°F (30°C) and for cotton about 95°F (35°C).

The frost-free period also influences crop adaptation by giving an indication of the duration of the growing season. Thus a growing season for corn or soybeans might be described as one with mean daily temperatures between 64 and 77°F (18 and 25°C), and with an average minimum temperature exceeding 50°F (10°C) for at least 3 months. Small grains such as wheat, barley, and oats tolerate a cooler climate with a period of only 2 months when minimum temperatures exceed 50°F (10°C). Because of optimum growth temperatures and frost-free periods, it is easily recognized why the spring Wheat Belt includes North Dakota and Montana, the Corn Belt Iowa and Illinois, and the Cotton Belt Misssissippi and Alabama. Farmers use planting dates and a range in maturity of varieties to match the crop to the growing season. *See* PLANT GROWTH.

In many winter annual (the plants are sown in fall, overwinter, and mature in early summer), biennial, and perennial crops, cold temperatures also influence distribution. The inherent ability of the crop to survive winter limits distribution of crops. As a generalized example, winter oats are less able to survive cold winters than winter barley, followed by winter wheat and then winter rye. Thus oats in the southern United States are mostly winter annual type, while from central Arkansas northward they are spring type. The dividing line for barley is about central Missouri and for wheat about central South Dakota. Winter rye can survive well into Canada. *See* COLD HARDINESS (PLANT).

A cold period may be required for flowering of winter annual and perennial crops. This cold requirement, termed vernalization, occurs naturally in cold environments where the dormant bud temperature is near 32°F (0°C) for 4 to 6 weeks during winter. Without this physiological response to change the hormonal composition of the terminal bud, flowering of winter wheat would not occur the following spring. Bud dormancy is also low-temperature-induced and keeps buds from beginning spring growth until a critical cold period has passed. The flower buds of many trees such as peach, cherry, and apple require less chilling than do vegetative buds, and therefore flower before the leaves emerge. The intensity and duration of cold treatment necessary to break dormancy differs with species and even within a species. For example, some peaches selected for southern areas require only 350 h below –12°F (8°C) to break dormancy, while some selected for northern areas may require as much as 1200 h. This cold requirement prevents production of temperate fruit crops in subtropical regions, but in temperate climates its survival value is clear. A physiological mechanism that prevents spring growth from occurring too early helps decrease the possibility of cold temperature damage to the new succulent growth. *See* DORMANCY; VERNALIZATION.

Temperature of the crop environment can be altered by date of planting of summer annuals, by proper site selection, and by artificial means. In the Northern Hemisphere the south- and west-facing slopes are usually warmer than east- or north-facing slopes. Horticulturists have long used mulches to control soil temperature, and mists and smoke for short-term low-temperature protection.

Water. Water is essential for crop production, and natural rainfall often is supplemented by irrigation. Wheat is grown in the Central Plains states of the

United States because it matures early enough to avoid the water shortage of summer. In contrast, corn matures too late to avoid that drought condition and must be irrigated to make it productive. In the eastern United States where rainfall is higher, irrigation is usually not needed for good yields. *See* IRRIGATION (AGRICULTURE).

Crops transpire large amounts of water through their stomata. For example, corn transpires about 350 lb of water for each pound of dry weight produced. Wheat, oats, and barley transpire about 600 lb, and alfalfa about 1000 lb for each pound of dry weight produced. Fallowing (allowing land to be idle) every other season to store water in the soil for the succeeding crop has been used to overcome water limitations and extend crops further into dryland areas. *See* PLANT-WATER RELATIONS.

Light. Light intensity and duration also play dominant roles. Light is essential for photosynthesis. The yield of a crop plant is related to its efficiency in intercepting solar radiation by its leaf tissue, the efficiency of leaves in converting light energy into chemical energy, and the transfer and utilization of that chemical energy (usually sugars) for growth of an economic product. Crops differ markedly in the efficiency of their leaves. *See* PHOTOSYNTHESIS.

Crops also differ markedly in leaf area and leaf arrangement. Humans greatly influence the photosynthetic area by altering number of plants per area and by cutting and pruning. Corn producers have gradually increased plant population per area about 50% since about 1950 as improved varieties and cultural methods were developed. This has increased the leaf area of the crop canopy and the amount of solar energy captured. Defoliation and pruning practices also influence solar energy capture and growth rate. Continued defoliation of pastures by grazing may reduce the photosynthetic area to the point that yield is diminished. In these perennial plants, carbohydrates are stored in underground organs such as roots, rhizomes, and stem bases to furnish food for new shoots in spring and following cutting. Availability of water and nitrogen fertilizer also influences the amount of leaf area developed.

Photoperiodism, the response of plants to day length, also has a dramatic effect on plant distribution. Such important adaptive mechanisms as development of winter hardiness, initiation and maintenance of bud dormancy, and floral initiation are influenced by photoperiod. Plants have been classified as long-day, those that flower when day lengths exceed a critical level; short-day, which is opposite to long-day; and day-neutral, those that are not affected by day length. Photoperiod also has a controlling influence on formation of potato tubers, onion bulbs, strawberry runners, and tillers of many cereal grains and grasses. Farmers select varieties bred for specific areas of the country to ensure that they flower properly for their growing season. Horticulturists, in a more intensive effort, provide artificial lighting in greenhouses to lengthen the photoperiod, or shorten it by shading, to induce flowering and fruit production at will; for example, to ready poinsettias for the Christmas holiday trade or asters for Easter. Natural photoperiods are important in plant breeding when varieties of day-length-sensitive crops must be developed for specific localities. Soybeans are day-length-sensitive and have been classified into several maturity groups from north to south in latitude. *See* PHOTOPERIODISM.

Pathogens. Pathogens of plants that cause diseases include fungi, bacteria, viruses, and nematodes. These organisms are transmitted from plant to plant by wind, water, and insects and infect the plant tissue. Organisms infect the plant and interfere with the physiological functions to decrease yield. Further, they infect the economic product and decrease its quality. Pathogens are most economically controlled by breeding resistant varieties or by using selective pesticides. Insects decrease plant productivity and quality largely by mechanical damage to tissue and secretion of toxins. They are also usually controlled by resistant varieties or specific insecticides. *See* PLANT PATHOLOGY.

Soil. The soil constitutes an important facet of the plant environment. Soil physical properties such as particle size and pore space determine the water-holding capacity and influence the exchange of atmospheric gases with the root system. Soil chemical properties such as pH and the ability to supply nutrients have a direct influence on crop productivity. Farmers alter the chemical environment by addition of lime or sulfur to correct acidic or basic conditions, or by addition of manures and chemical fertilizers to alter nutrient supply status. Soil also is composed of an important microbiological component that assists in the cycling of organic matter and mineral nutrients. *See* SOIL.

Management. During and following domestication of crop plants, many cultural or management practices have been learned that enhance production or quality of the crop. This dynamic process is in operation today as the quest continues to improve the plant environment to take advantage of the genetic potential of the crop. These practices have made plant agriculture in the United States one of the most efficient in the world. Some major changes in technology are discussed in the following sections.

Mechanization. In 1860 an average United States farm worker produced enough food and fiber for fewer than 5 other persons. In 1950 it was enough for 25, and today exceeds 50 other persons. Mechanization, which allowed each farm worker to increase the area managed, is largely responsible for this dramatic change.

A model of a grain reaper in 1852 demonstrated that nine farmers with the reaper could do the work of 14 with cradles. By 1930 one farmer with a large combine had a daily capacity of 20–25 acres (8–10 hectares), and not only harvested but threshed the grain to give about a 75-fold gain over the cradle-and-flail methods of a century earlier. In 1975 one farmer with a faster-moving combine could harvest 50–75 acres (20–30 ha) per day. The mechanical cotton picker harvests a 500-lb (227-kg) bale in 75 min,

40–50 times the rate of a hand picker. The peanut harvester turns out about 300 lb (136 kg) of shelled peanuts per hour, a 300-h job if done by hand labor. Hand setting 7500 celery plants is a day's hard labor for one person. However, a modern transplanting machine with two people readily sets 40,000, reducing labor costs by 67%. Today crops such as cherries and tomatoes are machine-harvested. When machines have been difficult to adapt to crops such as tomatoes, special plant varieties that flower more uniformly and have tougher skins on the fruit have been developed. *See* AGRICULTURAL MACHINERY.

Fertilizers and plant nutrition. No one knows when or where the practice originated of burying a fish beneath the spot where a few seeds of corn were to be planted, but it was common among North American Indians when Columbus discovered America and is evidence that the value of fertilizers was known to primitive peoples. Farm manures were in common use by the Romans and have been utilized almost from the time animals were first domesticated and crops grown. It was not until centuries later, however, that animal fertilizers were supplemented by mineral forms of lime, phosphate, potassium, and nitrogen. Rational use of these substances began about 1850 as an outgrowth of soil and plant analyses.

Justus von Liebig published *Chemistry and Its Application to Agriculture and Physiology* in 1840, which led to concepts on modifying the soil by fertilizers and other amendments. These substances soon became the center of crop research. In 1842 John Bennet Lawes, who founded the famous Rothamsted Experiment Station in England, obtained a patent for manufacture of superphosphates and introduced chemical fertilizers to agriculture. As research progressed, crop responses to levels of phosphate, lime, and potassium were worked out, but nitrogen nutrition remained puzzling. The nitrogen problem was clarified in 1886, when H. Hellriegel and H. Wilfarth, two German chemists, reported that root nodules and their associated bacteria were responsible for the peculiar ability of legume plants to use atmospheric nitrogen. These findings collectively allowed rational decision-making regarding nutrition of crop plants and were understandably very significant in increasing crop productivity.

By the early twentieth century 10 elements had been identified as essential for proper nutrition. These were carbon, hydrogen, and oxygen, which are supplied by the atmosphere; and nitrogen, potassium, phosphorus, sulfur, magnesium, calcium, and iron, supplied by the soil. The first 40 years of the twentieth century witnessed the addition of manganese, boron, copper, zinc, molybdenum, and chlorine to the list of essential mineral nutrients. These 6 are required only in very small amounts as compared with the first 10 and have been classified as micronutrients. From a quantitative view they are truly minor, but in reality they are just as critical as the others for plant survival and productivity. Many early and puzzling plant disorders are now known to be due to insufficient supplies of micronutrients in the soil, or to their presence in forms unavailable to plants. *See* PLANT MINERAL NUTRITION.

An important result of discoveries relating to absorption and utilization of micronutrients is that they have served to emphasize the complexity of soil fertility and fertilizer problems. In sharp contrast to early thoughts that fertilizer practices should be largely a replacement in the soil of what was removed by the plant, it is now recognized that interaction and balance of mineral elements within both the soil and plant must be considered for efficient crop growth. Usage of chemical fertilizers has become more widespread with time. *See* FERTILIZER; FERTILIZING.

Pesticides. Total destruction of crops by swarms of locusts and subsequent starvation of many people and livestock have occurred throughout the world. Pioneers in the Plains states suffered disastrous crop losses from hordes of grasshoppers and marching army worms. An epidemic of potato blight brought hunger to much of western Europe and famine to Ireland in 1846–1847. The use of new insecticides and fungicides has done much to prevent such calamities. Various mixtures, really nothing more than nostrums (unscientific concoctions), were in use centuries ago, but the first really trustworthy insect control measure appeared in the United States in the mid-1860s, when paris green was used to halt the eastern spread of the Colorado potato beetle. This was followed by other arsenical compounds, culminating in lead arsenate in the 1890s.

A major development occurred during World War II, when the value of DDT (dichlorodiphenyltrichloroethane) for control of many insects was discovered. Although this compound was known to the chemist decades earlier, it was not until 1942 that its value as an insecticide was definitely established and a new chapter written in the continual contest between humankind and insects. Three or four applications of DDT gave better season control of many pests at lower cost than was afforded by a dozen materials used earlier. Furthermore, DDT afforded control of some kinds of pests that were practically immune to former materials. In the meantime other chemicals have been developed that are even more effective for specific pests, and safer to use from a human health and ecological viewpoint.

Dusts and solutions containing sulfur have long been used to control mildew on foliage, but the first highly effective fungicide was discovered accidentally in the early 1880s. To discourage theft, a combination of copper sulfate and lime was used near Bordeaux, France, to give grape vines a poisoned appearance. Bordeaux mixture remained the standard remedy for fungus diseases until the 1960s, when other materials with fewer harmful side effects were released.

New pesticides are constantly being developed by private industry and are carefully monitored by several agencies of the federal government. Besides evaluation for its ability to repel or destroy a certain pest, influence of the pesticide on physiological processes in plants, and especially long-range

implications for the environment and human health, are carefully documented before federal approval for use is granted. *See* PESTICIDE.

Herbicides. Because they are readily visible, weeds were recognized as crop competitors long before microscopic bacteria, fungi, and viruses. The time-honored methods of controlling weeds have been to use competitive crops or mulches to smother them, or to pull, dig, hoe, or cultivate them out. These methods are still effective and most practical in many instances. However, under many conditions weeds can be controlled chemically much more economically. A century ago regrowth of thistles and other large weeds was prevented by pouring salt on their cut stubs. Salt and ashes were placed along areas such as courtyards, roadsides, and fence rows where all vegetation needed to be controlled. However, until the 1940s selective weed control, where the crop is left unharmed, was used on a very limited scale. The first of these new selective herbicides was 2,4-D (2,4-dichlorophenoxyacetic acid), followed shortly by 2,4,5-T (2,4,5-trichlorophenoxyacetic acid) and other related compounds. This class of herbicides is usually sprayed directly on the crop and weeds and kills susceptible plants by upsetting normal physiological processes, causing abnormal increases in size, and distortions that eventually lead to death of the plant. As a group, these herbicides are much less toxic to grasses than to broad-leaved species, and each one has its own character of specificity. For instance, 2,4-D is more efficient for use against herbaceous weeds, whereas 2,4,5-T is best for woody and brushy species.

Today's herbicides have been developed by commercial companies to control a vast array of weeds in most crop management systems. Some herbicides are preemergence types that are sprayed on the soil at time of corn planting. This type of herbicide kills germinating weed seeds by interfering with their photosynthetic system so they starve to death. Meanwhile metabolism of the resistant corn seedlings alters the chemical slightly to cause it to lose its herbicidal activity. Such is the complexity that allows specificity of herbicides. The place that herbicides have come to occupy in agriculture is indicated by the fact that farmlands treated in the United States have increased from a few thousand acres in 1940 to over 150,000,000 acres (60,000,000 ha), not including large areas of swamp and overflow lands treated for aquatic plant control and thousands of miles of treated highways, railroad tracks, and drainage and irrigation ditches. *See* HERBICIDE.

Growth regulators. Many of the planters' practices since the midnineteenth century may be classified as methods of regulating growth. Use of specific substances to influence particular plant functions, however, has been a more recent development, though these modern uses, and even some of the substances applied, had their antecedents in century-old practices in certain parts of the world.

Since about 1930 many uses have been discovered for a considerable number of organic compounds having growth-regulating influences. For instance, several of them applied as sprays a few days to several weeks before normal harvest will prevent or markedly delay dropping of such fruits as apples and oranges. Runner development in strawberries and sucker (tiller) development in tobacco can be inhibited by sprays of maleic hydrazide. Another compound called CCC (2-chloroethyl trimethyl ammonium chloride) is used to shorten wheat plants in Europe to allow higher levels of fertilizer. Many greenhouse-grown flowers are kept short in stature by CCC and other growth regulators. Striking effects from gibberellins and fumaric acid have also been reported. The first greatly increase vegetative growth and the latter causes dwarfing. In practice, growth-inhibiting or growth-retarding agents are finding wider use than growth-stimulating ones. In higher concentration many growth-retarding agents become inhibiting agents. *See* GIBBERELLIN.

There are marked differences between plant species and even varieties in their response to most plant growth regulators. Many, if not most, growth regulators are highly selective; a concentration of even 100 times that effective for one species or variety is necessary to produce the same response in another. Furthermore, the specific formulation of the substances, for example, the kind or amount of wetting agent used with them, is important in determining their effectiveness. In brief, growth regulators are essentially new products, though there are century-old instances of the empirical use of a few of them. Some influence growth rate, others direction of growth, others plant structure, anatomy, or morphology. With the discovery of new ones, indications are that it is only a matter of time before many features of plant growth and development may be directly or indirectly controlled by them to a marked degree. Applications of these substances in intensive agriculture are unfolding rapidly, and their use is one of the many factors making farming more of a science and less of an art. *See* PLANT HORMONES.

Plant improvement. From earliest times humans have tried to improve plants by selection, but it was the discovery of hybridization (cross-mating of two genetically different plants) that eventually led to dramatic increases in genetic potential of the plants. Hybridization was recognized in the early 1800s, well before Mendel's classic genetic discoveries, and allowed the combination of desirable plants in a complementary manner to produce an improved progeny. Plant breeding had a dramatic flourish in the early twentieth century following the rediscovery of Mendel's research and its implications, and has had much to do with the increased productivity per area of present-day agriculture.

Corn provides the most vivid example of how improvement through genetic manipulation can occur. Following the commercial development of corn hybrids about 1930, only a few acres were planted, but by 1945 over 90% of the acreage was planted to hybrids, and today nearly 100% is planted. It has been conservatively estimated that hybrids of corn have 25% more yield potential than old-style varieties. Subsequently, plant breeders have utilized hybridization

for development of modern varieties of most major crop species.

While very significant changes through crop breeding have occurred in pest resistance and product quality, perhaps the character of most consequence was an increase in the lodging (falling over) resistance of major grain crops. Corn, wheat, and rice have all been bred to be shorter in stature and to have increased stem resistance to breaking. These changes have in turn allowed heavier fertilization of crops to increase photosynthetic area and yield. The impact of this was recognized when Norman Borlaug was awarded the Nobel Peace Prize in 1970 for his breeding contribution to the "green revolution." His higher-yielding wheats were shorter and stiff-strawed so they could utilize increased amounts of nitrogen fertilizer. They were also day-length-insensitive and thus had a wide adaptation. New varieties of rice such as IR-8 developed at the International Rice Research Institute in the Philippines are shorter, are more responsive to nitrogen fertilizer, and have much higher potential yield than conventional varieties. With those new wheat and rice varieties many countries gained time in their battle between population and food supply. *See* BREEDING (PLANT).

Often in tree crops and high-value crops it is not feasible to make improvements genetically, and other means are utilized. Grafting, or physically combining two or more separate plants, is used as a method of obtaining growth control in many fruit trees, and also has been used to provide disease resistance and better fruiting characteristics. This technique is largely limited to woody species of relatively high value. Usually tops are grafted to different rootstocks to obtain restricted vegetative growth, as in dwarf apple and pear trees which still bear normal-sized fruit. Alternatively, junipers and grapes are grafted to new rootstocks to provide a better root system. In both cases the desirability of the esthetic or economic portion warrants the cost and effort of making the plant better adapted to environmental or management conditions.

Curtis L. Nelson

Seed production. Seed production is the process of sexual plant propagation in which genetic recombination occurs, and it is an important step in crop improvement and distribution. Developing more productive crops through plant hybridization and genetic engineering is a primary method of increasing agricultural production. Improvement comes from greater crop yield, better quality, and improved pest resistance. Commercial seed production is the major method of propagating a new, improved variety. Genetic change must not occur during seed multiplication.

A prescribed system of maintenance and seed multiplication is established for individual crops. The methods used to increase a new variety will depend on the species and the breeding system. In the United States, state agencies supervise the multiplication to ensure that adequate safeguards have been used to protect the genetic purity through a certification program. Special tags are used to identify certified seed. Privately owned varieties are often produced and marketed under the supervision of the organization's quality control department, without public agency certification.

Seed production is a major economic enterprise of international scope. Regions with a soil and a climate favorable to seed production specialize in providing high-quality seed, and are often located far from the consuming area. For example, the dry climate in the irrigated regions in the western United States favors high seed yield, and harvest is not affected by rain; therefore much of the vegetable and forage legume seed is produced there. Cool season grass species are adapted to the mild climate of the Pacific Northwest, a major source of forage grass seed. When rapid seed multiplication is needed, stock seed from the Northern Hemisphere crops is shipped to the Southern Hemisphere to allow the harvest of two seed crops in 12 months.

Seed crop species are self- or cross-pollinated. Isolation distance and crop management are determined by pollination characteristics. Self-pollinated species include wheat, barley, soybeans, and lettuce. These crop varieties are homozygous when released, and the progeny will be true to type. Once a wheat variety has been released, for example, the breeder will maintain the variety by selecting heads which are threshed separately and grown in headrows. A row which fails to conform to the varietal type is discarded. Approved rows are harvested, bulked, and classified as breeder seed for further multiplication. *See* POLLINATION.

Cross-pollinated species are fertilized by pollen carried by wind or insects. Wind-borne pollen is important in crops such as corn, sugarbeet, and ryegrass. Insect pollination is valuable in the fertilization of alfalfa, clovers, and mustards. Cross-pollinated crops require isolation from other varieties of the same species for protection of purity.

Superior plant vigor is developed from F1 hybrid (first-generation cross) seed produced by crossing two or more inbred lines. Hybrid vigor is the increased performance over parental types. Although seed of hybrid varieties are more difficult to produce, they are important in commercial production of crops such as corn which is grown exclusively from hybrid varieties. Other important hybrid crops include cabbage, cantaloupe, sugarbeet, cucumber, onion, petunia, tomato, squash, and watermelon. Seed from the F1 hybrid will not produce the same yield or crop quality of the parent plants; therefore new hybrid seed must be purchased each year, creating a large demand for planting seed. Some hybrid seed crops such as corn are grown in large fields, while certain flower hybrids such as petunias require greenhouse culture.

Seed production has become a very specialized enterprise serving scientific agriculture. *See* REPRODUCTION (PLANT); SEED.

Harold Youngberg

Bibliography. J. A. Barden, R. G. Halfacre, and D. J. Parrish, *Plant Science*, 1987; R. J. Delorit, L. J. Greub, and H. L. Ahlgren, *Crop Production*, 4th ed., 1974; W. R. Fehr, *Principles of Cultivar*

Development, 1987; J. R. Harlan, *Crops and Man*, 2d ed., 1992; J. Janick et al., *Plant Science: An Introduction to World Crops*, 3d ed., 1981; J. H. Martin et al., *Principles of Field Crop Production*, 3d ed., 1989.

Agricultural soil and crop practices

The techniques and methods used in plowing, harrowing, planting, tilling, harvesting, drying, storing, and processing of agricultural crops.

Plowing

Cutting and turning a furrow with a plow is usually the first and most effective operation of tillage. There are two types of plow, the moldboard (**Fig. 1**) and the disk.

Moldboard plow. The working part of the moldboard plow is the plow bottom. The plow furrow is cut or broken loose on the land side by the shin of the plow or by a coulter or jointer attached to the plow beam, and on the bottom by the point and edge of the plow share. The advancing plow moldboard exerts pressure on the furrow slice, pulverizing it and causing it to move upward and to the side in a spiraling path so that it is partially or completely inverted. The shape of the moldboard may be described as a section cut from a warped cylinder, on the inside of which the soil spirals. The selection of design of the moldboard depends upon the physical reaction of the kind of soil and sod cover on which it is to be used. Because precise information of soil physical reaction is inadequate and the amount of pulverization and placement, or throw, of the furrow slice depends upon the speed of plowing, plows have been designed largely by the cut-and-try method. This has resulted in a great variety of shapes. Moldboard plows have largely been adapted to power farming by the selection of bottoms suitable for certain soils and combination of these bottoms into gangs sufficiently large to use the power of the tractor at the approximate speed for which the bottoms were designed. The power required at normal speed of 4 mi/h (6.5 km/h) varies from 2–3 lb/in.^{2} (14–21 kilopascals) of cross section of furrow slice for sand to 20 lb/in.^{2} (140 kPa) for tough clay. Modifications of shape have permitted a speed of 5–6 mi/h (8–9.5 km/h).

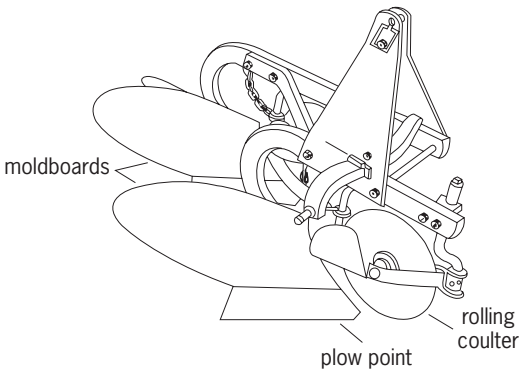


Fig. 1. One type of moldboard plow. (*Tractor and Implement Operations–North America, Ford Motor Co.*)

Moldboard plow bottoms are designated as right-hand or left-hand bottoms, depending upon the direction of throw. Both may be combined on one frame so that the plow can be pulled back and forth, always throwing the furrow downhill, or in one direction. Sizes range from 8 to 20 in. (20 to 50 cm) in width of cut; the 10–12-in. (25–30-cm) size is suitable for use with animal power, and the 14–16-in. (35–40-cm) size is largely used with power equipment. Although a plow may be used to plow at different depths, most have been designed to work at a depth of approximately one-half the width of the furrow.

The effectiveness of plowing may be materially increased by attachments. A jointer, or a jointer and coulter combination, which cuts a small furrow in front of the plow shin, permits complete coverage of sod. Where there are large amounts of rubbish or crop residue, a weed chain or wire can be used to drag the debris into the furrow and hold it there until covered. A furrow wheel may reduce land-side friction, and a depth gage on the beam helps secure uniform depth. A modified form of plow bottom, called a lister, is in effect a right and a left moldboard joined at the shin so as to throw soil both to the right and to the left. This produces a furrow or trough, called a list, in which seed is planted. Because it concentrates rainfall in the furrow, this method is used largely in areas of light rainfall.

Disk plow. The disk plow consists of a number of disk blades attached to one axle or gang bolt. This plow is used for rapid, shallow plowing. In fields where numerous rocks and roots are present, the disk plow, which rolls over obstacles, is substituted for the moldboard. The disk is also used for sticky soils that will not scour on a moldboard. The disk plow is manufactured in widths of from $2\frac{1}{2}$ to 20 ft (0.7 to 6 m). The disks are commonly spaced 8–10 in. (20–25 cm) apart. The angle between the gang bolt and the direction of travel is usually adjustable from 35 to 55°.

Harrowing

Soil preparation for planting usually involves the pulling of an implement called a harrow over the plowed soil to break clods, level the surface, and destroy weeds. A wide variety of implements are classified as harrows; the most common kinds are the disk harrow, the spike-tooth harrow, the spring-tooth harrow, and the knife harrow. Previously the function of seedbed preparation was performed almost entirely by the implements classified as harrows. With the introduction of power, farming is now performed in large part by field cultivators, rod weeders, rotary hoes, or treaders, subsurface packers, and various designs of rollers. Power-driven rotary tillers perform the function of both plowing and harrowing.

Kinds of harrows. The spike-tooth is the oldest form of commercial harrow and consists of spikes or teeth (usually adjustable) extending downward from

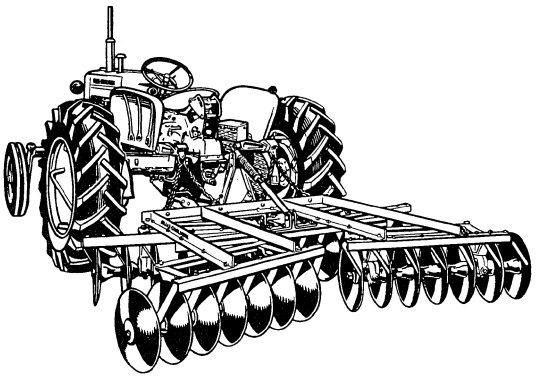


Fig. 2. One type of disk harrow. (*Allis-Chalmers*)

a frame. The teeth extend into the soil, and when the harrow is pulled forward, they cut through clods and break them. The teeth also stir and level the soil surface and kill weeds. This type of harrow has light draft and is built in sections, many of which may be joined together so that large areas can be covered quickly. This implement is most effective if used before clods dry; it is frequently attached behind the plow.

The spring-tooth harrow is similar to the spike-tooth type but has long curved teeth of spring steel. The spring action renders it suitable for rough or stony ground. It is particularly useful in bringing clods to the surface, where they can be pulverized. It is also used to bring the roots of weeds and obnoxious grasses to the surface for destruction, and to renovate and cultivate alfalfa fields. The knife harrow consists of a frame holding a number of knives which scrape and partly invert the surface to smooth it and destroy small weeds.

The disk harrow is probably the most universally used type (**Fig. 2**). It cuts clods and trash effectively, destroys weeds, cuts in cover crops, and smooths and prepares the surface for other farming operations. The penetration of the disk harrow depends largely upon weight. The disk blades are commonly 16–24 in. (40–60 cm) in diameter and are spaced 6–10 in. (15–25 cm) apart in gangs of 3–12 disks. Disk harrows can be obtained in widths up to 20 ft (6 m). A single-acting disk harrow has two opposed gangs throwing soil outward from the center; a tandem or double-acting disk has two additional gangs which throw the soil back toward the center. An important advancement in the design of the disk harrow is the offset disk. A right-hand offset disk harrow has a gang in front which throws to the right and a rear gang which throws to the left. It may be adjusted to pull to one side and to the rear of the tractor so as to harrow beneath the low limbs of orchard trees.

Other soil-preparation equipment. The field cultivator is used to perform many of the jobs of harrows before planting. It usually consists of a number of adjustable standards with sweeps or scrapes attached to tool bars in such a fashion that the soil is stirred from underneath, killing the weeds and creating a surface mulch for moisture conservation. The rod weeder is a power-driven rod, usually square in cross section, which also operates beneath the surface of loose soil, killing weeds and maintaining the soil in a loose mulched condition. It is adapted to large operations and is used in dry areas of the Northwest. A variety of rollers and packing wheels and clod crushers have been designed.

Planting

The practice of placing seed or vegetative propagating material in soil for multiplication through growth and reproduction is usually a seasonal operation. Its success depends upon soil preparation and placing the seed in an environment favorable to growth. The seed, which is an embryonic plant enclosed in a protective membrane, usually contains enough nutritional material to start growth. It must have suitable temperature, adequate air, and sufficient moisture to overcome its dormant condition and induce vigorous growth. In general, the seeding process consists of opening a furrow in properly prepared soil to the correct depth, metering and distributing the seed or planting material, depositing the seed in the furrow, and covering and compacting the soil around the seed to a degree suitable to the crop. Fertilizer is usually placed in the soil sufficiently near the seed that it will be available to the young plants after germination.

There are five general methods of planting based on five special types of machinery: (1) broadcasters; (2) seed drills used for small seed and grains; (3) planters for cultivated row crops such as corn or cotton; (4) special planters for parts of plants used for propagation, such as potato planters; and (5) transplanters used to set out small plants that have been grown in beds from small seed. The last is commonly used for tobacco, sweet potatoes, cabbage, trees, and many horticultural crops.

Broadcasters. Small grains, grasses, and clovers are planted by broadcasting or drilling. The broadcaster is usually a rotating fanlike distributor which throws the seed over a wide area by centrifugal force. Like hand seeding, this method requires the absence of gusty wind for most effective distribution. Under proper conditions, broadcasting can be done from airplanes. Broadcasting is especially suited to sowingseed in another crop without unduly disturbing the soil, such as sowing clover seed in wheat.

Drills. The grain drill opens a furrow and places the seed in it. Attachments, such as covering chains and wheels to press seed into the soil, are commonly used (**Fig. 3**). The seed is metered by a special apparatus into rows 6–14 in. (15–35 cm) apart. Several types of furrow openers adapted to different soil and crop conditions are available. Grain drills are also commonly equipped for fertilizer distribution and grass seeding.

Row-crop planters. Such crops as corn and cotton are planted with special planters in rows to simplify cultivation. Because yield may be greatly affected by the stand's being too thick or too thin, precision planting is important to avoid the cost of thinning or interplanting. Delinting of cotton-seed and sizing of seed corn and other seeds are important to precision

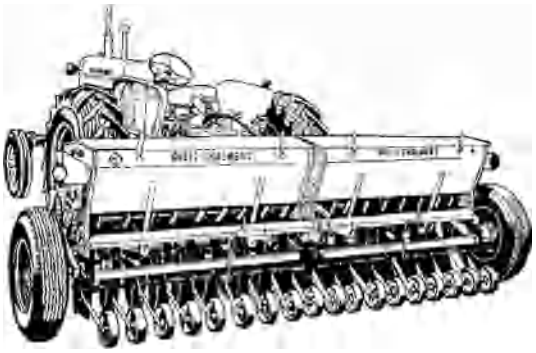


Fig. 3. All-crop drill with rubber press wheels to push seeds into soil. (*Allis-Chalmers*)

planting. Planters usually are equipped for dropping in hills or drilling in rows. The hills may be checkrowed, that is, spaced equally apart on the corners of squares so that the crop can be cultivated in two directions, thus avoiding hoeing. The precision necessary for this type of planting is secured by opening valves in the planter shank at measured intervals by means of buttons on a check wire.

Transplanters. Special kinds of equipment designed for the planting of cuttings or small plants are known as transplanters. Such machines usually transport one or more workers who assist the action of the machine in placing the plants in a furrow and properly covering them. Transplanters commonly supply a small quantity of water to each plant.

Tillage

The mechanical manipulation of the soil to improve its physical condition as a habitat for plants is called tillage. It includes plowing, inversion, loosening, harrowing, pulverization, packing, and rolling the soil, all to improve aeration and temperature conditions and to produce a firm seedbed (**Fig. 4**). Subsurface tillage is the loosening of soil by sweeps or blades pulled beneath the surface without inversion of the soil. This practice, especially adapted to dry areas, fragments the soil and leaves a mulch of stubble or other plant residues on the soil surface to conserve water and help control erosion.

Effective tillage eliminates competitive vegetation, such as weeds, and stimulates favorable soil microbiological activities. Natural forces of heating and cooling, swelling and shrinkage, wetting and drying, and freezing and thawing account for the major pulverization of soil and assist in the production of a favorable crumb structure. Wise practice dictates the avoidance of tillage when soil is so wet and plastic that its crumb structure is easily destroyed, as well as the use of those operations which put the soil in the most favorable condition for natural forces to act. This results in the minimum amount of time and power for soil preparation. Manipulation of the soil by machinery is an essential part of soil management, which includes such soil-building practices as grass and legume rotations, fertilization, and liming.

Mark L. Nichols

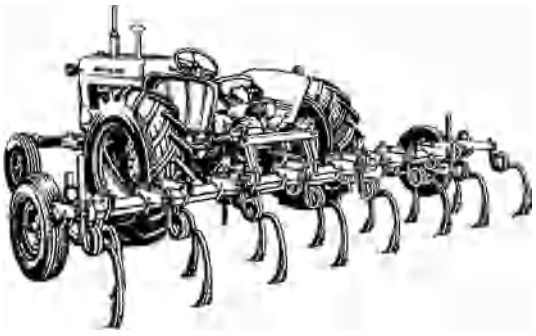


Fig. 4. Coil-shank field cultivator. (*Allis-Chalmers*)

Reduced tillage and no-tillage. These are land preparation practices that were used on over 31% of the cropland in the United States in 1992. These soil management systems involve limited or no physical disturbance of the soil prior to planting. Weed growth is controlled either by application of herbicides or by midseason cultivation. Residue from the previous crop remains on or near the soil surface; with more conventional clean tillage, essentially all crop residues are incorporated into the soil, usually by plowing or disking. The no-till planter is a multifunction device fitted with a disk that cuts through the stubble of a prior year's crops, making a furrow, a planter that drops a seed into the furrow, and wheels that close the furrow.

Advantages. Reduced and no-tillage practices are commonly used on highly erodible soils because they provide soil erosion control and greatly reduce fuel, time, and equipment costs, compared to clean tillage. Also, crops can be planted when the soil is wetter. Because of these advantages, the crop can be planted closer to the optimum planting date, potentially increasing crop yields. Leaving crop residues on the surface reduces water losses to runoff and evaporation, increasing infiltration rate and soil water storage by 1–2 in. (25–50 mm), which also improves crop yield potential. Improved soil erosion control maintains soil fertility and quality, thereby maintaining crop yield potential. Consequently, reduced-till and no-till production methods may result in greater net income because of lower production costs, greater yields, and diminished offsite damage.

Reduced-till and no-till systems have been used extensively only since about 1970, when suitable herbicides and equipment became widely available. Early research showed that when weed growth was controlled by herbicides and crop residues were left on the soil surface, soil water storage was improved by reduced weed competition and soil water evaporation rate. Also, soil productivity was maintained because erosion losses were controlled. These findings stimulated research to develop economical herbicide technology and suitable planting equipment. To a large extent, this objective was achieved during the 1970s and 1980s, permitting the widespread use of reduced-till and no-till management systems in agricultural production. *See* HERBICIDE.

Because of the presence of surface residues, the surface of no-till soil is moister and cooler than that

of bare soils. Surface residues also provide soluble carbon as an energy source for soil microorganisms. Consequently, because of the more favorable environment, the populations of all classes of microbes are greater on the surface of no-till soils than on bare soils. This results in more sustained and prolonged turnover of nitrogen, phosphorus, and other nutrients, which is often at a slower rate because of lower temperatures. Soil aggregation and soil porosity are usually improved by the enhanced biological activity resulting from the more favorable soil environment created by no-tillage. Greater earthworm activity in no-till soils also provides large-diameter soil pores that enhance infiltration and reduce runoff of water. Collectively, these advantages of reduced and no-tillage systems provide United States farmers with a useful tool for meeting the soil erosion limits on highly erodible land imposed by the conservation compliance requirements of the 1990 Farm Bill. As a consequence, acreage of reduced and no-tillage has increased greatly in recent years.

Disadvantages. Problems associated with reduced-till and no-till systems include herbicide costs, reduced soil temperature, inadequate weed control, poor drainage, soil compaction, disease and insect problems, and residual damages from herbicides. While these tillage systems require less investment in equipment, fuel, and labor than does clean tillage, they do require a substantial cash investment in herbicides. Each herbicide has its own spectrum of plant species for which it is effective; herbicide selection is therefore dictated by the crops to be produced and by the dominant weed species likely to be present. Most herbicides are degraded through microbiological activity or by volatilization. In warm wet years, the herbicide may rapidly degrade and later-germinating weeds may survive and grow, requiring additional expense for weed control. When the herbicide degrades slower than anticipated (cool or dry years), residual herbicidal activity may damage the following crop.

Special situations. Reduced-till and especially no-till production methods are poorly suited for slowly drained soils because no-tillage increases soil water content, thereby intensifying drainage problems. This is especially true in more northern climates where the lower soil temperature plus greater soil water content associated with no-tillage may result in delayed crop establishment. These problems can be offset to some extent by using ridge-till systems, wherein soil ridges 6–10 in. (15–25 cm) high are built during cultivation of the previous crop. At planting time, the top few inches of the ridge are leveled off, providing a well-drained, warm, weedfree, flat zone on the ridge top for the crop seedlings. Weeds between crop rows are controlled by tillage. Ridge-till systems have become very popular in the midwestern and southeastern United States.

Use of no-till systems may reduce need for phosphorus fertilizers because much of the organic phosphorus in crop residues is made available to the next crop through microbial processes. Fertilizer nitrogen requirements for no-tillage may be slightly greater than for clean tillage because the improved soil water supply may provide a greater crop growth potential. More nitrogen may also be temporarily tied up in the microbial biomass of no-till soils. This nitrogen often becomes available to the crop later in the growing season because the surface residues for no-till soil keep the surface soil moist longer, allowing microbial activity to continue longer into dry summer months. Although these systems conserve more soil water and increase water infiltration, the quantity of soluble nitrogen in no-till and reduced-till soils is usually less than in clean, cultivated soils. Thus, these tillage practices have little effect upon the leaching of nitrates to ground-water aquifers. *See* FERTILIZER.

J. F. Power

Soil and tillage interactions. Tillage is the use of a tool to change the physical condition of the soil. Many times the tool is part of a machine powered by a tractor, but it can be as simple as a hoe or shovel. A tillage tool affects the soil by cutting, shattering, packing, mixing, or inverting it. Normally, the objective of tillage is to improve the condition of the soil or to control weeds so that crop seeds will germinate quickly after planting and the resulting crop plants will produce to their potential.

Tillage can create short-term improvements in a soil condition by lowering bulk density by cutting and shattering the soil mass, but that may lead to long-term effects such as loss of stable structure. For example, a soil that has not been tilled recently often has a more stable structure than soil in the tilled field next to it. Better structure means that the basic particles (sand, silt, and clay) are grouped into aggregates with space around them for the movement of air, water, and roots. Stable structure means that the structure will resist forces such as tillage or wetting and drying without having the aggregates break into smaller pieces. The soil that is tilled may form a crust with these smaller pieces after a rain that will cause the next rain to run off or impede air transfer or plant growth. A fence row might maintain an open structure, allowing adequate air flow and water between its stable aggregates.

The use of a tillage tool depends on the soil type and the condition of the soil at the time it is tilled. Maps have been produced which cover much of the crop-producing areas of the United States and other developed countries. The areas of soil that are similar in properties are grouped and identified by names that specify texture as well as other attributes that tend not to change over time. Soil textures range from coarse (for example, sandy soil) to very fine (clay soil).

Soil tilth is a qualitative term describing the physical condition of soils at any time. It indicates the ease of tillage, seedbed preparation, seedling emergence, and root growth. It may also indicate the ability of the soil to resist erosion. The best condition for seeding wheat may not be the best condition for growing cranberries. Soil quality is dependent upon the ability of a soil to function in its immediate environment for a particular use and interact positively with the general environment.

A proposed Tilth Index describes the physical condition of the soil by combining numerical values for the density of the soil, the resistance of the soil to penetration by a steel cone, the organic matter content of the soil, the slipperiness of the soil when it is wet, and the size distribution of aggregates of soil particles. The Tilth Index combines soil physical condition indicators by multiplication after converting them to numbers between 0 and 1.

An ideal management system would use the Tilth Index, a soil quality index, or some other quantitative measure of the soil condition as a benchmark for deciding whether tillage is necessary. By using this approach, the condition of the soil could be determined at any time prior to planting, and tillage would be done only when benefits outweighed direct costs, such as fuel, and indirect costs, such as increased soil erosion potential. Tillage that has been done to change the physical condition of the soil may have been more important for its impact on the crop residue cover of the soil.

There is evidence that it may be more important to manage the residue cover than to change physical condition of the soil. In the northern regions of the United States corn belt, maintaining crop residues on the soil surface can reduce soil temperatures and negatively affect growth and development of subsequent crops. In Iowa and Minnesota, it has been difficult to prevent lower yields of corn in a rotation when a moldboard plow was not used to incorporate crop residue. In Ohio, on well-drained soils, there has been a yield advantage to leaving the crop residue on the surface without tillage prior to planting. The increased potential for harboring and transmitting plant pathogens in surface residues is a concern that is being evaluated through research programs because of the probability of increased area of crops produced with prior crop residue on the soil surface in the middle to late 1980s. This change is due to government efforts to reduce soil erosion in the United States.

The force required to perform tillage is often influenced by soil type and condition and by implement speed. When this force is provided as a pull by animals or tractors, it is often called draft. A plow equipped with high-speed moldboards, coulters, and a landside working at 4 mi/h (6.5 km/h) might require a 4-lb/in.^{2} (2.5-newton/cm^2) cross section of cut sandy loam soil. This cross section is also called the furrow slice. This would be compared to 6 lb/in.^{2} (4 N/cm^2) in loam or 13 lb/in.^{2} (9 N/cm^2) in silty clay. The draft decreases approximately 10% for each 1% increase in moisture content. Increasing soil strength by increasing the density by compaction or by planting a crop such as alfalfa, because of tough roots, will increase the draft. Increasing the speed of the plow by 50% will increase the draft by 25% in loam soil.

The draft of a disk plow is not affected by soil texture as much as that of a moldboard plow. The draft of subsoilers, chisel plows, and field cultivators depends on the depth and speed of the operation and is normally expressed as a value per shank. The reason is that the shanks may be spaced at different widths and the soil is not worked to the same depth across the width of the machine. The draft of a disk harrow is related to implement weight as well as soil texture. A disk harrow weighing 120 lb/ft (177 kg/m) and operating in sandy loam soil requires a draft force of 96 lb/ft (1380 N/m) of width. This requirement would increase by 50% if the disk were used in silt loam, or by 90% in clay.

Fuel requirements for tillage vary with implement draft and traction conditions for the tractor. Heavy draft primary-tillage tools, such as plows or subsoilers, require more fuel per unit of land area (acre or hectare) than other tillage tools. Fuel requirements for secondary tillage tools (such as disking) are influenced by soil conditions created by primary tillage in addition to soil type. If the tractor is on loose soil, it experiences more motion resistance than if it is operating on a firm soil. Properly matching tractor power to load size is important in conserving fuel. It is important to conserve fuel in crop production, but more energy is usually required in corn production to manufacture nitrogen-rich fertilizer and the fuel used for drying the crop and for irrigating some soils than the combined fuel used directly by machines to till, plant, and harvest.

Matching the tractor to the load includes consideration of whether the tractor is two- or four-wheel drive or uses tracks. The weight of the tractor as it is prepared for the field can influence slippage. Slippage of more than 10–15% with a wheeled tractor usually wastes fuel and can be corrected by adding weights to the tractor. Too much weight can shorten the productive life of the tractor. *See* SOIL.

Thomas S. Colvin

Harvesting

The practice of severing and reaping the plant or any of its parts is called harvesting.

Crops harvested for grain. The process of gathering such crops as corn, sorghums, wheat, oats, barley, rye, buckwheat, and rice is called grain harvesting. Ear corn is harvested by means of a corn picker (**Fig. 5**). The ears are snapped off by specially designed rollers which pass around the standing stalks. The husks are removed by a husking bed consisting of rolls of various types, over which the ears are passed.

Fig. 5. Corn picker. (*New Idea Farm Equipment Co.*)

Fig. 6. Forage harvester chopping haylage for livestock feed. (*Sperry New Holland, Division of Sperry Rand Corp.*)

Fig. 7. Mowing machine. (*Massey Ferguson Co.*)

Shelled corn is harvested by a picker-sheller or by the combine harvester. The picker-sheller snaps the ears from the stalks in the same manner as the picker, but the husking bed is replaced by a shelling unit. A trailing shelling unit attached to the picker can also be used.

Corn is also harvested by a self-propelled combine harvester. The header can be removed and replaced with a snapping unit, or the header can remain in place with the whole plant passing through the machine. Grain sorghums and cereals are harvested largely with the combine harvester, a machine that severs the standing crop, shells the grain, separates grain from straw, and removes chaff and trash in one operation. Sometimes these crops are severed and windrowed, allowed to dry, and threshed later.

Grain crops harvested for ensilage. This operation is used for corn, sweet sorghums, and cereals such as oats, wheat, barley, and rye.

Row crops, such as corn and some sorghums, are harvested with a forage harvester equipped with a row-crop attachment. High-moisture corn may be shelled by a picker-sheller and stored as ensilage.

Drilled crops and some row crops are harvested with the forage harvester equipped with a sickle-bar attachment (**Fig. 6**). Rotary-type harvesters are also used. The plants are severed at or near the soil surface and cut or shredded into short lengths.

Crops for silage, soilage, and hay. This type of harvesting is used for legumes (other than edible-podded legumes) and grasses (excluding corn, sorghums, and cereals). The sickle-bar or rotary-type forage harvester is used. It severs the crop near the ground surface and chops it into desired lengths for silage or soilage. It also chops hay from the windrow. The crop may be wilted slightly in the windrow to reduce moisture for silage preservation. The conventional mower, or the mower-crusher designed to speed up drying, is used to harvest crops for hay (**Fig. 7**).

Legumes and grasses for seed. Legumes and grasses are harvested largely by the combine harvester, either by direct or windrow methods (**Fig. 8**). Windrowing becomes necessary when the crop fails to ripen evenly. Because some seeds are lighter than cereal grains, machine adjustments differ widely. To increase overall efficiency, two combines may be hooked together in tandem. All straw and chaff from the lead combine passes through the rear one.

Podded legumes which are harvested include soybeans, dry edible beans, and peas. Soybeans are harvested exclusively by the combine-harvester direct method. Peas and beans may be harvested by the combine harvester or by bean threshers with multiple shelling cylinders (**Fig. 9**). In many cases, beans or peas may be removed or severed from the soil and windrowed prior to threshing. To prevent the cracking of seeds, cylinder speeds are reduced and concave clearance increased. Rubber-covered rolls, placed ahead of the cylinder, may be used to squeeze beans from pods.

Fig. 8. Combine-harvester. (*International Harvester Co.*)

Fig. 9. Bean thresher. (*C. B. Hay Co.*)

Root crops. Harvested root crops include sugarbeets, potatoes, and peanuts. Sugarbeets are gathered by special harvesters. One type tops the beets in place, after which the beets are lifted by specially designed blades or fingers. Another type lifts the beets by gripping the tops or by impaling the beets on a revolving spiked wheel (**Fig. 10**). The beets are then topped in the machine. An elevator conveys the beets to trucks for bulk handling.

Potatoes are harvested by several methods. They may be (1) dug with a one- or two-row digger and picked up by hand; (2) dug, sorted, and placed into containers by machine (vines, trash, and clods are removed mechanically and by workers riding the machine); (3) harvested by a fully mechanized procedure which includes digging, sorting, removal of vines, trash, and clods, and loading into bulk trucks (**Fig. 11**); and (4) dug with a standard digger, windrowed for drying, and picked up later by an indirect harvester. Sweet potatoes are harvested largely by the first method.

Peanuts are harvested by the following methods: (1) the pole-stack method, in which the peanuts are dug and hand-shaken, stacked around poles, and later picked; (2) a method in which they are dug with a one- or two-row digger, windrowed, and harvested later with a peanut picker (**Fig. 12**); and (3) the once-over method, in which all operations are accomplished with the same machine.

Crops harvested for fiber. Cotton is harvested by two methods: (1) pulling lint from the bolls by means of the cotton picker, a method which requires several pickings and is accomplished by broached spindles revolving rearward on which the lint is wound; and (2) pulling the entire boll from plants by a cotton stripper (**Fig. 13**), a once-over method accomplished by rolls of various types (steel, rubber paddle, or brush) which strip the bolls from the plants.

Special crops. There are several crops requiring special methods to be harvested. These include tobacco, castor beans, and sugarcane. Tobacco is harvested by two general methods. (1) Leaves may be primed from the stalks as they mature, in several primings starting with the lower, more mature, leaves; or (2) the stalks may be severed near the base (**Fig. 14**), upended, and speared, after which laths are inserted to handle the plants and support them in the curing barn. Machines have been developed

Fig. 10. Detail of spiked-wheel harvester gathering sugarbeets. (*Blackwelder Manufacturing Co.*)

Fig. 11. Potato harvester. (*USDA Agricultural Research Service, Red River Valley Potato Research Center*)

Fig. 12. Peanut digger-shaker-windrower. (*Department of Agricultural Engineering, North Carolina State College*)

Fig. 13. Cotton stripper. (*Allis-Chalmers*)

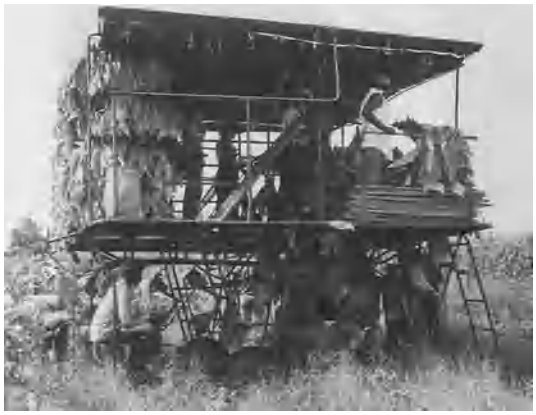


Fig. 14. Tobacco harvester. (*Department of Agricultural Engineering, North Carolina State College*)

Fig. 15. Sugarcane harvester. (*Agricultural Engineering Department, Louisiana State University*)

to speed up the priming process; workers ride rather than walk.

Castor beans are harvested by special machines that straddle the rows, with revolving beaters that strip the beans from the standing stalks.

Sugarcane is harvested by self-propelled harvesters which sever the cane at or slightly below the soil surface (**Fig. 15**). Additional knives top the cane. Tops and trash are removed by fans and distributed over the soil. Conveyors move the cane into heap rows or directly into trucks or wagons. Some machines cut the cane into short lengths for easier handling and processing. Edward A. Silver; R. B. Musgrave

Drying and Storage

Farm crops may be harvested at the most desirable stage of maturity and stored for weeks or months if properly dried or preserved. Field drying is an inexpensive procedure. However, in cool, humid areas a fine crop may deteriorate to a low-value feedstuff if it is damaged by rain during field drying; loss of quality may also occur as a result of mold or spontaneous heating in storage. To reduce such losses in forage crops, many farmers partially cure grasses or legumes in the field and then finish drying them in hay mows, bins, special wagons, or drying buildings by passing heated or unheated air through the forage with a power-driven fan unit attached to an air-duct system (**Fig. 16**).

Forage. Forage containing 70–80% moisture at harvest time is field-dried to 30–40% and finish-dried to 22–25% for safe storage as chopped or long hay, or to 20% as baled hay. Hay processed in this manner is superior to field-dried material in color, carotene content, and leafiness.

Because hay quality is to some extent a function of the rapidity of drying, heated air usually produces the best product. Very rapid drying can be accomplished with dehydrating equipment which can dry material from 75 to 10% moisture in 20 min or less. The quality of a dehydrated product is high; however, costs of labor and fuel are also high. Alfalfa to be used in mixed feed is the most frequently dehydrated crop (**Fig. 17**).

Field drying can be accelerated by the use of crushing machines which crack or shred the freshly cut forage as it passes through one or more pairs of crushing rollers (**Fig. 18**). Overall drying time is shortened because the stems, if crushed, will dry almost as fast as the leaves which, if the hay was improperly dried, often shatter and drop off.

Small grains and shelled corn. Small grains and shelled corn are dried to 12% moisture or less in either continuous or batch driers and stored in bins. Ear corn stored in an open crib must be dried to 16% moisture if mold growth in storage is to be prevented. Generally, temperatures of drying should not exceed 100°F (37°C) for seed and malting grains, 130°F (54°C) for milling corn, and 200°F (93°C) for feedstuffs (**Fig. 19**). Frequently rice drying is carried on in two stages to prevent cracking.

Ensiling. The anaerobic fermentation process of ensiling is used to preserve immature green corn, legumes, grasses, and grain plants. The crop is chopped and packed while at about 70–80% moisture and put into cyclindrical tower-type silos, horizontal trenchlike structures, or other containers to exclude the air. This tightly packed, juicy material is preserved by proper bacterial fermentation. Desirable microorganisms in grass and legume silage can be encouraged by field-wilting the crop to 70% moisture, or by adding chemical preservatives, sugar, and starch materials, such as ground corn or small grain.

Fig. 16. Special wagon rack attached to portable oil-burning crop drier for finish-drying chopped forage.

Fig. 17. Portable alfalfa-dehydrating equipment.

Fig. 18. Forage crusher for cracking or shredding stems to accelerate field drying rate of forage crops.

Fig. 19. Perforated metal bin for drying small grain and shelled corn with heated air from portable crop dryer.

Shelled or chopped ear corn is occasionally stored under similar anaerobic conditions. Hjalmar D. Bruhn

Processing Crops

Crop processing involves such operations as shelling, cleaning, separating, sorting, washing, treating, scarifying, testing, grinding, and ginning.

Shelling. The separation of corn kernels from the cob or the removal of the shell from nuts such as peanuts, walnuts, and hickory nuts is called shelling. This can be done with two types of machines, the spring type in which the kernels are rubbed from the ears by a notched metal bar called a rag iron, and the power-driven cylinder-type machine into which the ears are fed between a revolving shelling cylinder and stationary bars called concaves (**Fig. 20**). The kernels are rubbed off and separated before the cobs are removed. Proper shelling is obtained by control of the rate of feeding, tension of the shelling bar, choice of shelling concave, cleaning-air control, and cob-outlet control. The practice of picking high-moisture corn in the field and shelling it with a picker-sheller or combine-sheller is increasing.

Cleaning and separation. These procedures include the removal of foreign material, such as weed seeds, chaff, dead insects, and broken stems. The fanning mill, consisting of two vibrating screens and an air blast, is used on the farm for cleaning (**Fig. 21**). The most common methods of cleaning are by size, using a screen; by length, using a cylinder or disk with indented pockets to accept only short, small grain; by specific gravity, using a vibrating screen or inclined deck through which air is blown to remove the light material from the top; and by brine solutions of such density as to float light material and permit heavy material to settle. Seeds which become sticky when wet are separated from other seeds (for example, buckhorn seed from cloverseed) by moistening of the seed surfaces with wet iron filings or sawdust. The wetted seeds and iron filings or sawdust stick together, forming large clumps which are then removed by screening. Buckhorn seeds are removed in this manner from clover seeds. Seed shape can also be used as a means of separation because the round seeds roll and the flat seeds slide. Smooth and rough materials may be separated by using velvet- or velveteen-fabric-covered rolls, or by air in which the materials with irregular and rough surfaces are removed from those with smooth surfaces. Fruits may be cleaned by brushing, wet or dry, and by flailing.

Sorting. The separation of products into prescribed standards is called sorting. Grading is sorting to meet state and federal agency regulations. Grading, particularly of fruits and vegetables, is done by machinery to obtain the proper size or weight. Size grading is practiced to obtain desired diameter by

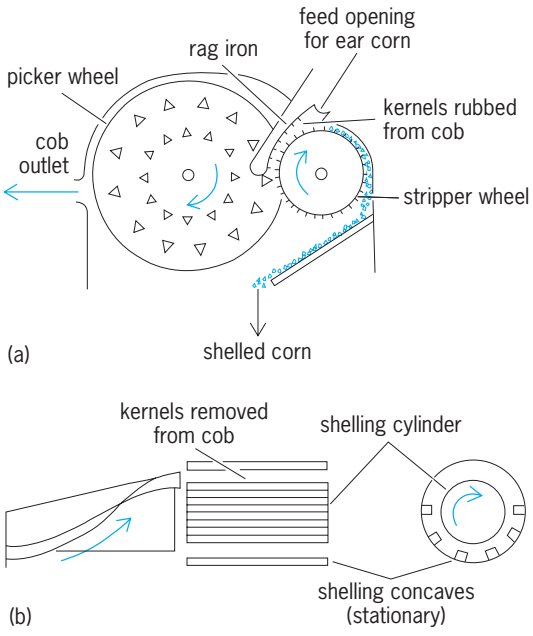


Fig. 20. Operation of two types of corn shellers. (*a*) Spring. (*b*) Cylinder.

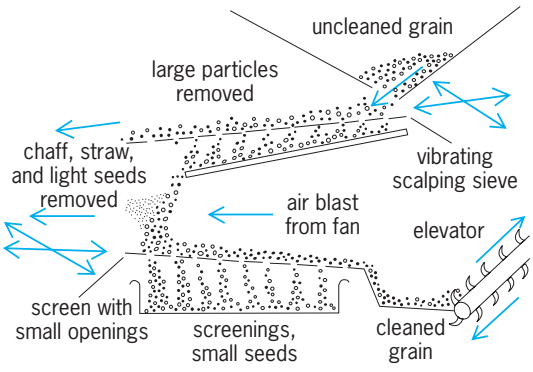


Fig. 21. Fanning mill operation.

using a belt with holes, a link chain, a diversion belt, diversion rollers, or spool units. Apples, peaches, potatoes, onions, lemons, and oranges are commonly sorted by size. The weight method is sometimes used to place products in appropriate grades. For beans, coffee, and lemons, color sorting is accomplished by scanning with an electric-eye device which rejects the materials that possess undesirable light-reflecting properties.

Washing. Foreign materials are removed from fruits, seeds, nuts, and vegetables by washing, often with a detergent and chlorine solution. Washing is done by soaking in tanks, spraying with high-pressure nozzles, or moving the materials through a washing solution in a rotary cylinder.

Treating. Seeds are treated for disease or growth control. Treating was formerly performed by heating the seeds in water, but now it is usually done with a chemical. The treating material may be applied as a chemical slurry, dust, liquid, or vapor. Mercury compounds are often used, but seed treated with mercury cannot be used for animal or human feed. Legume seeds are often inoculated with bacteria which convert the nitrogen of the soil air into a form which is suitable for use by the plant.

Scarifying. This process, usually preceded by hulling, is an operation in which hard seed, such as that of legumes, is scratched or scarred to facilitate water absorption and to speed germination.

Testing. Usually testing is performed on batches of seeds to determine the purity, moisture content, weight, and germination potential. The major problem in testing is to obtain a representative sample. The moisture content of grains and seeds indicates the keeping quality in storage and is a basis for estimating their commercial value. Weight per bushel is an important commercial criterion used for grain and is known as the test weight.

Grinding. Reduction in size of the material to improve utilization by animals and to make handling easier is called grinding. Measurement of the fineness of the ground material is designated by fineness modulus numbers from 1 (fine) to 7 (coarse). Units for size reduction include the hammer mill, a rotating high-strength beater which crushes the material until it will pass through a screen mounted above, below, or around the rotating hammer; the burr or attrition mill, in which two ribbed plates or disks rub or crush the material between them; and the roller mill, in which grain passes between pairs of smooth or corrugated rolls (**Fig. 22**). The last is used extensively for flour manufacture. The crimper-roller is used on farms to crush grains. The roller-crusher is also used to reduce the size of ear corn before it is fed into the hammer mill or burr mill.

Ginning. Separation of lint from cottonseed is called ginning. The cotton gin cleans the cotton and seed in addition to separating them. The saw gin, which consists of about 100 saws per gin stand mounted about $^5/_8$ in. (1.5 cm) apart on a shaft, is the most commonly used. The 12-in.-diameter (30-cm) saws are made up of spikes on wheels which

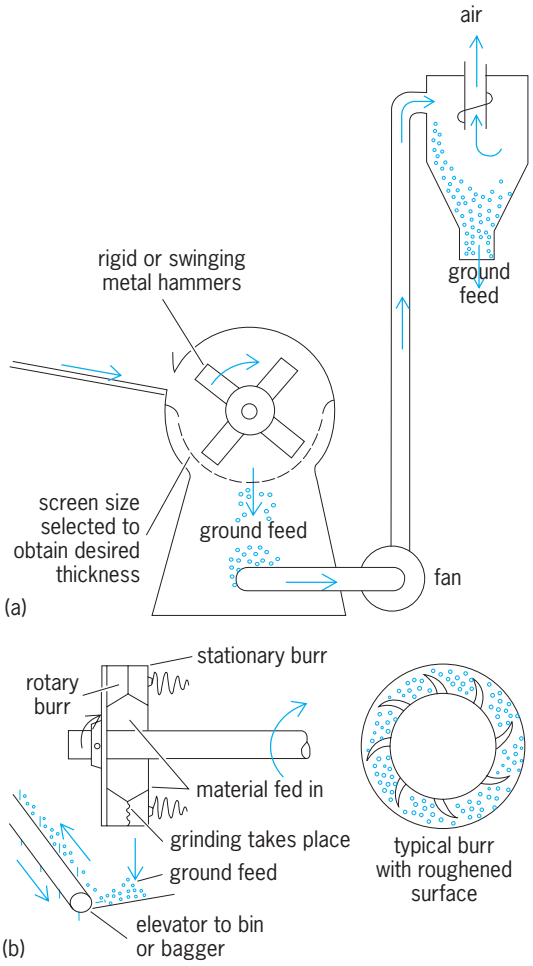


Fig. 22. Operation of two machines for grinding. (*a*) Hammer mill. (*b*) Burr mill.

pull the cotton through huller ribs, but the seed cotton (cotton containing seed) is not pulled through. The lint is removed from the saw blades with air or with a high-speed brush. *See* AGRICULTURAL ENGINEERING; AGRICULTURAL MACHINERY; AGRICULTURAL SCIENCE (PLANT); COTTON; FOOD MANUFACTURING.

Carl W. Hall

Sustainable Agriculture

Sustainable agriculture represents an integration of traditional, conservation-minded farming techniques with modern scientific advances. This system maximizes use of on-farm renewable resources instead of imported and nonrenewable resources, while earning a return that is large enough for the farmer to continue farming in an ecologically harmless, regenerative way. Agriculture in the United States is mostly conventional, and thus heavily dependent on the use of fossil-fuel-based inputs— inorganic fertilizers, pesticides, and labor-saving but energy-intensive farm machinery.

Conventional farming systems have greatly increased crop production and labor efficiency, but serious questions are being raised about their energy-intensive nature and their adverse effects on soil productivity, environmental quality, farm profitability, and human and animal health. This concern has

led to an increasing interest in sustainable farming systems, because they have the potential to reduce some of the negative effects of conventional agriculture.

Sustainable agriculture emphasizes use of natural, on-farm, nonchemical approaches to pest control that do not pose health threats to humans or decimate populations of earthworms and other beneficial life forms. Compared to conventional agriculture, sustainable agriculture improves wildlife habitat by reducing the possibility of pesticide poisoning and by increasing crop diversity. The use of crop rotations and green-manure legume crops in sustainable agriculture on sloping, erodible soils generally produces much less erosion than conventional agriculture.

Sustainable agriculture also emphasizes use of farm profits to finance the ownership and operation of the farm, rather than depending heavily on borrowed capital. It increases net income by reducing inputs, rather than by increasing output. Conventional farmers wanting to become sustainable farmers do not merely terminate pesticide and fertilizer use; they replace petrochemical inputs with a variety of practices, such as crop rotations, growing legumes or cover crops, use of green manures, and livestock enterprises.

More sustainable agriculture technologies are being developed and adopted, featuring reduced dependence on chemical technologies as well as making greater use of soil-conserving methods of growing crops. Farmers who consider their farming methods to represent organic, regenerative, reduced-input, biological, or other sustainable approaches are employing some of these increasingly popular farming technologies, in addition to a number of less conventional methods. Many of these farmers report that their costs of production are lower than those of their neighbors who use conventional methods. Some farmers report lower yields than their neighbors, but the sacrifice in yield is often more than offset by cost reductions.

J. P. Reganold; R. I. Papendic; J. F. Parr

Solarization

Soil solarization is a hydrothermal process involving solar radiation that disinfests soil of plant pathogens and weed pests and also effects other changes in the physical and chemical properties of soil that lead to improved growth and development of plants.

Solarization of soil is accomplished by placing a sheet of plastic on moist soil for 4–6 weeks during the warm summer months. The plastic sheeting must be laid close to the surface of the soil to reduce the insulating effect of an air layer under the plastic; the edges of the plastic are buried or anchored in the soil. Solar radiation then heats the soil, and the resulting heat and the moisture in the soil are conserved by the plastic cover. Over time, the heat and moisture bring about changes in soil microbial populations, release of soil nutrients, and changes in the soil structure that are beneficial to plants subsequently established there.

Technology. Soil solarization is a mulching process, and it may seem to resemble the practice of using plastic mulches to warm soils to increase seed germination and seedling growth in early spring months. However, soil solarization differs from such use of plastic in that killing soil temperatures (approximately 97°F or 37°C or higher) are attained for many soil-borne organisms, seeds and youg plants or seedlings.

Soil solarization has been especially useful in regions with high solar radiation where daily temperatures in summer months often reach 90°F (32°C) or more; it is a nonchemical, easily applied method that is of little or no danger to workers; it embodies the objectives of integrated pest management and reduces the need for toxic chemicals for plant disease and pest control.

Soil preparation. The area or field to be solarized should be disked, rototilled, or worked by hand to provide a smooth surface so that the plastic sheeting will lie flat. For crops grown on-the-flat or for solarization of areas such as lawns, the edges of the plastic sheeting are anchored or buried about 8 in. (20 cm) deep. For bedded-up soil, the plastic sheeting is laid over the length of the bed, and the edges are buried in the furrows. For large-scale use of solarization on flat or level soil, after the first sheet is laid and anchored in the soil, additional sheets are glued together on one side and anchored in the soil on the other side. The ends of the plastic sheeting are also anchored in the soil.

Soil moisture. Soil solarization is best accomplished when the soil is moist (at least 70% of field capacity), a condition that increases the thermal sensitivity of soil-borne pathogens and pests and also enhances heat conduction within the soil. When conditions permit, plastic sheeting may be laid on dry soil, and then irrigation water can be run under the plastic in order to wet the soil to a depth of at least 24 in. (60 cm). If the field or area is level, shallow furrows made by tractor wheels during the laying of plastic sheeting provide convenient courses for the water. If the area or field is not level, the soil should be preirrigated and the plastic sheeting then laid within a few days, or as soon as the soil will support machinery without compaction. When soils are bedded up before solarization, the plastic is laid over the bed and anchored on each side in the furrows. Irrigation water is then run in the furrows, and the moisture subs into the raised beds (moves by capillarity).

Plastic film. Plastic sheeting used in soil solarization provides a barrier to losses of moisture and heat from soil due to evaporation and heat convection. Clear or transparent polyethylene plastic sheeting is more effective for heating soil than black plastic, since the transparent plastic allows greater transmission of solar energy; it gives maximum transmittancy of radiation from wavelengths of 0.4 to 36 micrometers. Other plastics are also effective, and in some instances superior to polyethylene. Colored transparent plastics tend to reduce the formation of water droplets on the underside or soil side of the plastic sheeting that may act as reflectors of radiation.

Polyvinyl chloride films also have been used effectively for soil solarization in greenhouses.

The thickness of the polyethylene sheeting also affects the heating of the soil; film 25-μm thick results in temperature increases of soil several degrees higher than thicker films that reflect more solar energy. Polyethylene sheeting used for soil solarization requires the presence of an ultraviolet inhibitor to prevent its deterioration by sunlight.

Soil temperature. Day length, air temperature, cloud cover, and the darkness of soils all have marked effects on the heating of soil during solarization. The ideal time is during the summer months, when the days are longest and air temperatures are highest. Maximum temperatures in the upper 6–12 in. (15–30 cm) of soil are reached 3–4 days after solarization begins. Depending on soil depth, maximum temperatures of solarized soil are commonly between 108 and 131°F (42 and 55°C) at 2 in. (5 cm) depth and range from 90 to 97°F (32 to 36°C) at 18 in. (46 cm) depth. Dark soils reach higher temperatures during solarization than lighter-colored soils because of their greater absorption of solar radiation.

Modes of action. Soil solarization encompasses changes in the biological, chemical, and physical properties of soil, changes that are often interdependent, dynamic, and evident for two or more growing seasons. Reduction in plant diseases and pests, coupled with better soil tilth and increases of mineral nutrients in solarized soil, contributes to the improved growth response of plants and increases in the quantity and quality of crop yields.

Biological effects. Soil microorganisms whose growth and development are inhibited at temperatures of about 86°F (30°C) or higher include most plant pathogens and some pests. The death rates of these organisms in moist soil at temperatures of about 97°F (36°C) or higher increase linearly with temperature and time. In contrast, other microorganisms in soil are thermotolerant, and their populations decrease less during solarization. Thermotolerant organisms include Actinomycetes and *Bacillus* species and certain mycorrhizal fungi; especially in the absence of soil-borne pathogens and pests, they are often highly beneficial to plant growth. In some soils, solarization causes biological changes that suppress the reestablishment of pathogens.

Chemical effects. Among the chemical changes that occur in soils during solarization are significant increases in soluble nutrients, such as nitrogen, calcium, and magnesium, which are associated with the increased growth response of plants grown in solarized soils.

Physical effects. The physical effects of soil solarization are largely unknown, but marked improvement in soil tilth occurs. During solarization, soil moisture tends to rise to a cooler soil surface at night, whereas during the day the moisture tends to move toward a cooler temperature zone deeper in the soil. This diurnal cycling of soil moisture is believed to leach sodic layers and thus accounts for the reduction of salinity caused by solarization in some soils.

Applications. Soil solarization is a safe, nonchemical, and effective method for plant disease and pest control. When used in conjunction with agricultural chemicals and biological control agents, soil solarization can be extended to regions where environmental conditions are less suitable for it. Use of solarization is increasing in orchard, field, and truck crop farming as well as in greenhouse and nursery operations; it is also finding increased use for garden and landscape improvement. *See* IRRIGATION (AGRICULTURE); MYCORRHIZAE; SOIL; SOIL ECOLOGY; SOIL MICROBIOLOGY; SOLAR RADIATION. James E. DeVay

Bibliography. D. Baize, *Soil Science Analysis: A Guide to Current Use,* 1993; Deere and Company Staff (eds.), *Soil Management: FBM Farm Business Management,*1993; J. E. DeVay and J. J. Stapleton (eds.), *Soil Solarization,*1991; R. E. Phillips and S. A. Phillips, *No-Tillage Agriculture: Principles and Practices,*1984; K. A. Smith, *Soil Analysis: Physical Methods,*1990; M. A. Sprague and G. B. Triplett, *No-Tillage and Surface-Tillage Agriculture,*1986.

Agriculture

The art and science of crop and livestock production. In its broadest sense, agriculture comprises the entire range of technologies associated with the production of useful products from plants and animals, including soil cultivation, crop and livestock management, and the activities of processing and marketing. The term agribusiness has been coined to include all the technologies that mesh in the total inputs and outputs of the farming sector. In this light, agriculture encompasses the whole range of economic activities involved in manufacturing and distributing the industrial inputs used in farming; the farm production of crops, animals, and animal products; the processing of these materials into finished products; and the provision of products at a time and place demanded by consumers. The proportion of the United States economy dedicated to the production of food and fiber products is about one-sixth of the total national product.

Many different factors influence the kind of agriculture practiced in a particular area. Among these factors are climate, soil, topography, nearness to markets, transportation facilities, land costs, and general economic level. Climate, soil, water availability, and topography vary widely throughout the world. This variation brings about a wide range in agricultural production enterprises. Certain areas tend toward a specialized agriculture, whereas other areas engage in a more diversified agriculture. As new technology is introduced and adopted, environmental factors are less important in influencing agricultural production patterns. Continued growth in the world's population makes critical the continuing ability of agriculture to provide needed food and fiber.

Agriculture in the United States

Agriculture is the means by which the resources of land, water, and sunlight are converted into useful

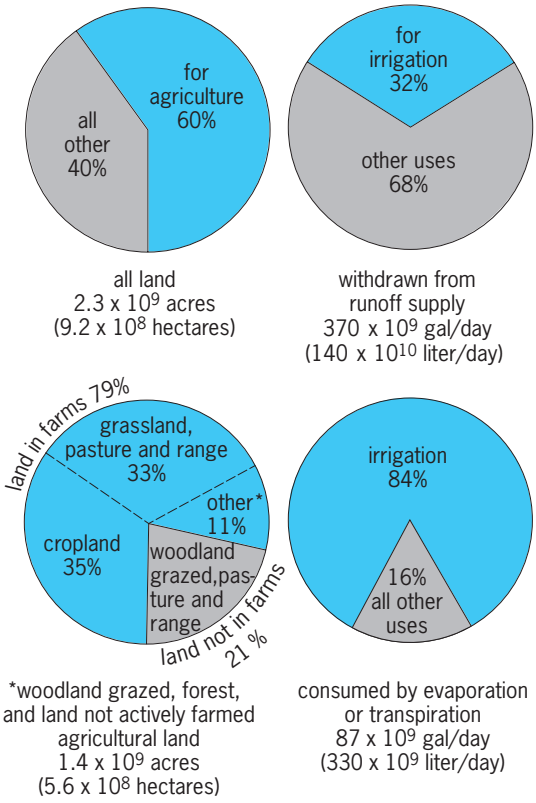


Fig. 1. Use of land and water resources in the United States. (*USDA*)

products. Land resources must be carefully managed to maintain and enhance the productivity of areas suitable for crops and for pasture and grazing, as well as to ensure effective use of rainfall so that the water needs of the nation will be met (**Fig. 1**). Of the 2.3×10^9 acres of land in the United States, about 1.3×10^9 acres (5.3×10^8 hectares) is used in agriculture, of which about 27% is cropland, 36% pasture and rangeland, and 25% forest. However, land-use categories are not static; pasture and rangeland may be shifted to forestry use, and as crop yields increase there is a net loss from cropland to less intense agricultural use. A significant portion of the land in pasture, range, and forestry can be considered as potential cropland, depending on agricultural demand and prices.

The United States has developed the most efficient agricultural system in the world, and this system continues to improve through scientific discoveries. Agriculture is increasingly a high-technology industry. There is now a tremendous array of powerful tools of modern plant science for the improvement of traditional crops through genetic manipulation and for finding alternative means of controlling pests (which could replace or greatly reduce pesticides that have environmental negatives).

The nation has evolved from a predominantly agricultural to a highly industrialized society (see **table**). The farm population grew in actual numbers up to 1940, but the percentage of the total population engaged in agricultural production steadily decreased as American farmers became more productive. A large population is engaged in supporting agriculture by providing agriculture inputs: tractors and trucks, machinery and equipment, petroleum products, fertilizers, pesticides and other agricultural chemicals, and other essential materials. A still larger group of people is engaged in agricultural industries that store, process, distribute, and sell farm products to the consumer. The number of people employed by these supporting industries is much greater than the number working on farms. The United States continues to depend on its lands and waters for the well-being of its people. In addition, the United States has become the world's greatest producer of foodstuffs for export to other nations.

Farm output has more than kept pace with the population growth of the nation. The output per worker-hour of farm workers has continued to increase because of the greater use of power and machinery, more effective uses of fertilizers to enhance crop productivity, use of higher-yielding crop varieties, more efficient feeding and management of livestock, and similar advances.

The output of farm products has increased more rapidly than inputs owing to the application of new knowledge and skills in farm operations. Labor costs have been reduced by the use of power and machinery; there have been great improvements in the use of fertilizers and other items such as chemical pesticides to control insects, diseases, and weeds, and in the use of livestock feed supplements. Total farm output and the output per unit of the various products that constitute inputs are closely related. Increases in American agricultural productivity are due to increased farmer efficiency and skill in applying agricultural science and technology. *See* AGRICULTURAL MACHINERY; AGRICULTURAL SCIENCE (ANIMAL); AGRICULTURAL SCIENCE (PLANT); FERTILIZER; FOOD MANUFACTURING; PESTICIDE.

American agriculture has provided a great diversity of food and fiber stuffs at lower costs than are enjoyed by most other nations.

Characteristics and trends. The skills and knowledge of the successful agricultural producer have changed drastically with the advent of tractors and trucks to replace horses, the development of more effective machines to replace much of the previous hand labor, the breeding of more productive

Chronology of American agriculture

Year	Total population, 10^6	Farm population, 10^6	Percent of total
1790	3.9	3.5	90+
1820	9.6	6.9	72
1850	23.2	11.7	50
1880	50.1	24.5	49
1910	92.0	28.5	31
1940	131.8	30.8	23.4
1950	151.0	25.0	16.5
1960	180.0	15.6	8.7
1970	204.0	9.7	4.8
1980	225.6	7.2	3.2
1990	247.8	4.6	1.9

crops and livestock, the development of chemical fertilizers to supplement native soil fertility, and the advent of chemical pesticides to control insects, diseases, and weeds. Increased processing has saved food products from spoilage and losses, and converted them into goods more readily utilized in meeting human needs. These agricultural industries now constitute a large segment of the national system for channeling agricultural products to consumers.

Nature of agricultural products. The primary agricultural products consist of crop plants for human food and animal feed and livestock products. The crop plants can be divided into 10 categories: grain crops (wheat, for flour to make bread, many bakery products, and breakfast cereals; rice, for food; maize, for livestock feed, syrup, meal, and oil; sorghum grain, for livestock feed; and oats, barley, and rye, for food and livestock feed); food grain legumes (beans, peas, lima beans, and cowpeas, for food; and peanuts, for food and oil); oil seed crops (soybeans, for oil and high-protein meal; and linseed, for oil and high-protein meal); root and tuber crops (principally potatoes and sweet potatoes); sugar crops (sugarbeets and sugarcane); fiber crops (principally cotton, for fiber to make textiles and for seed to produce oil and high-protein meal); tree and small fruits (apples, peaches, oranges, lemons, prunes, plums, cherries, grapes, and strawberries); nut crops (walnuts, almonds, and pecans); vegetables (melons, sweet corn, cabbage, cauliflower, lettuce, celery, tomatoes, eggplant, peppers, onions, and many others); and forages (for support of livestock pastures and range grazing lands and for hay and silage crops). The forages are dominated by a wide range of grasses and legumes, suited to different conditions of soil and climate.

Forest products include wood for construction material and panel products, fiber for pulp and paper, and extractives such as turpentine. Forests also serve the environment as a source of clean water and air, support of wildlife, and recreational and pleasant surroundings. Forestry, the culture of forestlands for human purposes, is often considered separately from agriculture, although the distinction between them is decreasing. It is clear that forest ecology interacts with all human condition and activities. *See* FOREST TIMBER RESOURCES.

Livestock products include cattle, for beef, tallow, and hides; dairy cattle, for milk, butter, cheese, ice cream, and other products; sheep, for mutton (lamb) and wool; pigs, for pork and lard; poultry (chiefly chickens but also turkeys and ducks) for meat and eggs; and horses, primarily for recreation.

In the United States, 65–90% of all feed units for cattle, sheep, and goats is provided by forage crops (grazing lands, hay, and silage), and the remainder from feed grains and other concentrated feeds. In most of the less-developed countries, such livestock subsist solely on forages.

Livestock play a vital role in using lands not suited for crops, to produce animal products that are indispensable to humans. However, commercially produced pigs and poultry are dependent on supplies of grains and concentrates. *See* BEEF CATTLE PRODUCTION; DAIRY CATTLE PRODUCTION; GOAT PRODUCTION; POULTRY PRODUCTION; SHEEP; SWINE PRODUCTION.

Exports. In the 1970s the world market for bulk commodities (maize, soybeans, and wheat) increased dramatically. In the 1980s the growth in the world market was for value-added products consisting of alternative and processed products, while the demand for bulk commodities, less than 10% of the exports from the United States, was flat. The United States is the world's largest exporter of agricultural products, principally grains and feed, oil seed and products, fruits and vegetables, cotton, and tobacco.

In addition to commercial sales of agricultural products, the United States has had a major role in providing foods and other supplies for relief and rehabilitation in developing countries. Such supplies have been sent as outright grants or as long-term loans without interest and with long grace periods. However, the reduction of grain surpluses in the United States reduced the nation's capacity to participate in large-scale food aid. Food give-away programs have a disastrous impact on the agriculture of developing countries and are, therefore, deemphasized except for famine conditions.

World Agricultural Conditions

The developing countries have had a mixed record of increasing their total food production in line with their increases in population growth. In South America increases in food production have kept up with population increases, while in Asia food production per capita increased as a result of the green revolution, the increased agricultural technology spearheaded by the introduction of improved wheat and rice. India, China, and Indonesia have made remarkable progress. Food production in Africa, however, has not increased, and per-capita production has actually declined in some countries, leading to grave problems. The application of science and technology to improve agriculture in Africa has great potential, but to be successful it requires changes in population control and political stability. The acquired immune deficiency syndrome (AIDS) epidemic in Africa is a complication that has affected long-term development there. In contrast, the developed countries in North America, western Europe, and Australia have continued to produce food surpluses.

In addition to the United Nations Development Program and the Food and Agricultural Organization (FAO) of the United Nations, there are various international agricultural research centers that are dedicated to solving the urgent agricultural problems in the tropics and subtropics. Agricultural development programs are also supported by the World Bank and by regional banks in Latin America and Asia. The bilateral assistance programs of the U.S. Agency for International Development is paralleled by bilateral programs of western European countries, Canada, and Japan. These programs deal not only with agricultural production and marketing but also with the

health, education, and institutional development of developing countries.

Environmental Concerns

A number of serious concerns impinged upon agriculture in the latter half of the twentieth century: rapid population increases, environmental deterioration, and energy price destabilization. These problems were shown to be part of a single issue—the distribution of the Earth's renewable and nonrenewable resources. Fear of emerging world food shortages caused by an unchecked population increase in the developing world did not lead to the dire consequences predicted earlier, largely because agricultural production increased in most, but not all, countries of the world. Populations in the developed world have reached a steady state.

Concerns over environmental safety as well as the impact of synthetic chemicals on human health dampened enthusiasm for chemical solutions to agricultural problems. This increasing concern resulted in enormous expense in the registration of new chemical materials because the agrochemical industry is forced to prove that substances are safe, a much more difficult task than to prove that a new substance is not harmful. Agricultural science has responded to worldwide concern by developing an integrated pest management (IPM) system which seeks a more rational basis for pest control through a combination of biological pest control strategies and a minimum use of pesticides, restricting use to those which are environmentally safe. Reducing agriculture's dependency on chemical solutions created new opportunities for genetic changes through both traditional and nontraditional techniques. However, the new genetics, known as genetic engineering, came under attack because of the unknown effects of foreign genes. Transfer of genes to crop species from other plant species has been accepted, but there is still considerable reluctance to accept the transfer of genes from animals or bacteria to crop plants.

Energy Costs

The United States food system uses about one-sixth of the energy in the nation, with food production accounting for one-fifth of this total. Thus, primary food production accounts for only about 3% of the nation's total energy demand. As shown in **Fig. 2**, it takes almost three times as much energy in home food preparation as in primary farm production. Notwithstanding, increasing concern over energy has generated various input/output accounting to calculate energy efficiency in agriculture. This has resulted in equating the energy in coal or oil with the energy in human labor or of food products—the results of which are clearly absurd.

The dramatic increases in energy costs will have large effects on the economics of various agricultural practices. A third of the energy used in actual agricultural production is related to the use of pesticides and fertilizers; other uses include field machinery, 23%; transportation, 18%; and irrigation, 13%

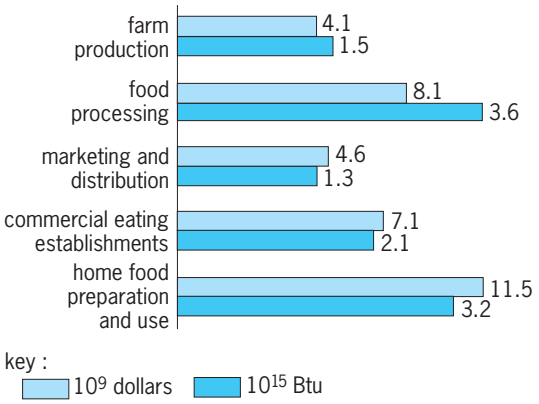


Fig. 2. Energy use in United States food system. (*USDA*)

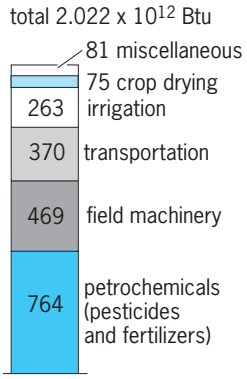


Fig. 3. Energy use in United States agricultural production. (*USDA*)

(**Fig. 3**). The cost/benefit ratio of many energy-dependent practices will change if energy costs increase out of proportion to other costs. This is clearly seen in the greenhouse industry, producer of flowers and some food crops. Greenhouse growers throughout the world have had to resort to various strategems to conserve energy, but present production has not declined, and the higher energy costs so far have been passed on to the consumer.

Jules Janick

Subsistence Agriculture

More than 500 million smallholders in the tropics farm under rain-fed conditions in diverse and risk-prone environments. In a constant struggle to survive, farm communities have developed innumerable ways of obtaining food and fiber from plants and animals. A wide range of different farming systems have evolved, each adapted to the local ecological conditions and inextricably entwined with the local culture. These traditional subsistence farming systems changed during the 1970s, 1980s, and 1990s primarily as a result of the expression of local creativity.

Subsistence farming. Subsistence farms comprise a patchwork of small fields containing mixtures of crops. Small farmers employ intricate farming systems to adjust to seasonal environmental changes,

marketing conditions, and the availability of family labor. More than 10% of the production is consumed directly on the farm's premises; there is little selling or trading. Subsistence farmers typically produce a great variety of crops and animals. For example, in Asian monsoon areas with more than 60 in. (1500 mm) of annual rainfall, it is not unusual to find as many as 20–30 tree crops, 30–40 annual crops, and 5–6 animal species on a single farm. Such multiple cropping methods are estimated to provide as much as 15–20% of the world's food supply. Throughout Latin America, farmers grow 70–90% of their beans with maize, potatoes, and other crops. Maize is intercropped on 60% of the region's maize-growing area.

Subsistence farming is characterized by diverse labor requirements. Having evolved to produce food year-round, the system provides continuous employment for unskilled labor to tend crops and livestock. *See* MULTIPLE CROPPING.

Typically, the subsistence farmer plants some commercial crops that entail relatively high risks. The farmer hedges against this risk by growing several less valued but more reliable crops, such as cassava or beans. In monsoon climates, with pronounced alternations of wet and dry seasons, the subsistence farmer ensures a stable production with long-duration root crops, tree crops, and animals. Although subsistence farming lacks the potential for producing a marketable surplus and thus supporting a higher standard of living for the farm family, it can often be adapted to increase productivity on more highly developed farms that are attempting to overcome resource limitations.

Uncultivated areas. Not only is diversity maintained within the area cultivated, but many peasants maintain uncultivated areas (such as forests, lakes, grasslands, streams, and swamps) in or adjacent to their fields, thus providing valuable products (including food, construction materials, medicines, organic fertilizers, fuels, and religious items). In humid, tropical conditions, collecting resources from primary and secondary forests is very intensive. In the Uxpanapa region of Veracruz, Mexico, peasants utilize about 435 wild plant and animal species, of which 229 are eaten. In many semiarid areas, gathering enables peasant and tribal groups to maintain their nutritional standards even during periods of drought.

Smallholder farming systems. Conventional agricultural development projects have demonstrated that although external inputs such as pesticides, fertilizers, and irrigation water can increase the productivity of peasant agriculture, such inputs can be maintained only at high financial and environmental cost. Moreover, they depend on the uninterrupted availability of commercial inputs—which is simply not viable given the level of impoverishment in most rural areas of the Third World. An agricultural strategy based on traditional cropping systems can bring moderate to high levels of productivity; and depending on prevalent socioeconomic and agroecological conditions, such diversified systems are sustainable at a much lower cost and for a longer period of time. *See* AGROECOSYSTEM.

Miguel A. Altieri

Remote Sensing

Information requirements of those who seek to manage the world's agricultural resources fall into five categories: the type of crop growing in each field; the state of health or vigor of the crop; the location and identity of crop-damaging agents or organisms; the crop yield likely to be obtained per unit of field area; and the total area of each field. Timely information in such areas would permit better crop management and increased production.

Agriculturists have used remote sensing to a significant extent since the mid-1930s. The only type of remote-sensing data ordinarily available until the late 1940s, however, consisted of small-scale, black-and-white aerial photos, most of which had been recently obtained for mapping purposes. Such photos were used by agriculturists during that period primarily in regional surveys for the annual identification of crops, the locating of fields, and the accurate delineation of field boundaries so that field acreages could be more accurately determined. In the 1950s it was found that early-warning evidence of certain crop diseases could be obtained through the use of a photographic film sensitive to radiation in the near infrared. Most of these applications continue to be made both in the United States and in many other parts of the world. *See* AERIAL PHOTOGRAPHY; PHOTOGRAMMETRY.

Space-acquired data. As operational space vehicles emerged, it became apparent that photos taken from a spacecraft in orbit at an altitude of 100 mi (160 km) or more above the Earth's surface had agricultural applications. The Johnson Manned Spacecraft Center conducted some highly successful photographic experiments from various crewed spacecraft, culminating in the *Apollo 9* flight in 1969. Furthermore, such vehicles offered the possibility of cost-effective monitoring of crop growth through the availability of repeated coverage of vast agricultural areas several times during each growing season.

Parallel developments in computer technology prompted consideration of computer-assisted analyses of the vast numbers of remote-sensing data. Computer applications required that some attribute of crops as imaged on a space photograph satisfy two conditions: the attribute must be of great diagnostic value in relation to the identification and monitoring of agricultural crops, and it must lend itself to being readily and unambiguously digitized.

Many photo image attributes had proved to be useful in the analysis by humans of agricultural crops on conventional aerial photographs. These included size, shape, shadow, pattern, tone, topographic site, and texture. However, only tone satisfied these two criteria. Furthermore, tone was the least degraded when remote sensing of agricultural crops was accomplished from a very high flying spacecraft rather than from a relatively low-flying aircraft.

During the 1970s progress was made in the development of a special kind of "camera" known as

a line scanner. As seen in **Fig. 4**, this device employs a very narrow-angle rotating prism which, as the remote-sensing vehicle advances, permits scene brightness to be observed and recorded for each small spot in an agricultural field. This digital record of tone can be telemetered directly and instantaneously to suitably located receiving stations on the ground and recorded there, pixel by pixel, on magnetic tape. Later, when the record is played back, it can be used to construct a complete swath of photolike imagery of the entire scene that was overflown by the spacecraft, as shown in **Fig. 5**. Alternatively, the data can be fed directly into a computer which has been programmed to identify, pixel-by-pixel and field-by-field, the crop types and crop condition classes.

For this computer-assisted analysis to be successfully employed, there must be a unique tone or scene brightness associated with each crop type and condition class that is to be identified. A typical line scanner recognizes and correspondingly digitizes as many as 256 gray levels, each indicative of a different scene brightness so that even very subtle but unique differences in scene brightness can be utilized. Many crops have been found to exhibit similar digitized signatures, at least throughout the vegetative stage of their growth. This problem is remedied in part by use of a line scanner system known as a multispectral scanner, which simultaneously acquires the digital records of scene brightness in each of several wavelength bands. More crop types and condition classes are computer-identifiable from the resulting digital multiband signature than from any single band. Through use of the proper combination of bands, an object's subatomic, atomic, molecular, and macromolecular properties can all be sensed.

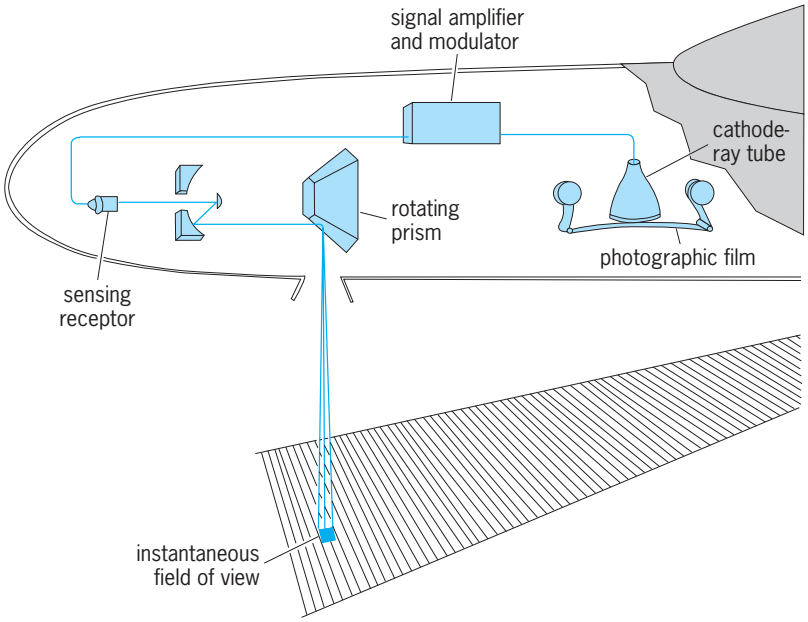
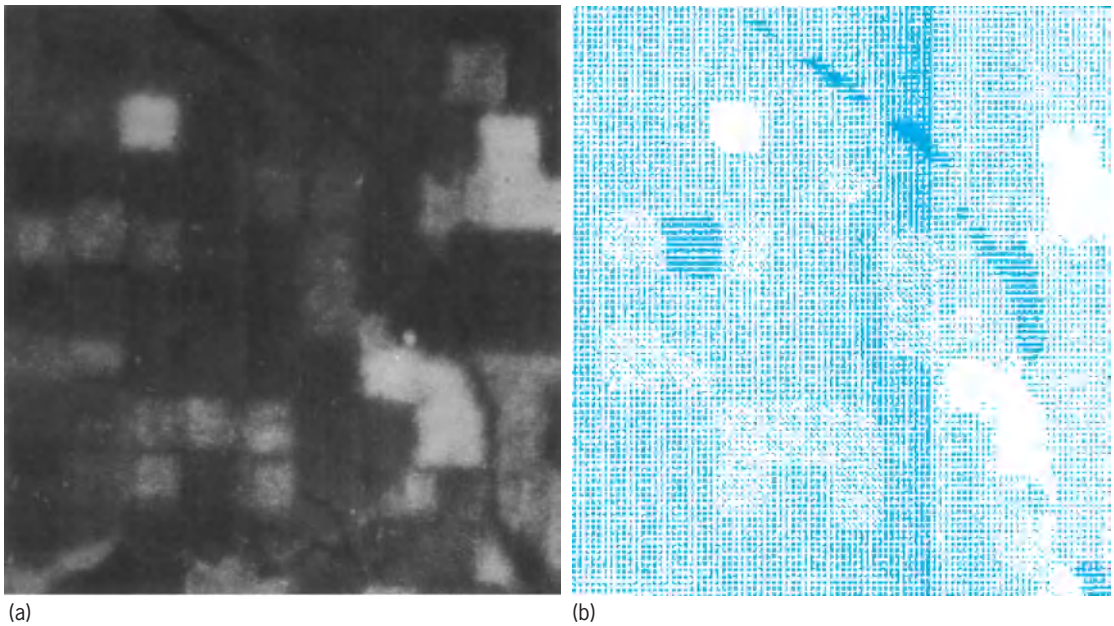


Fig. 4. Diagram of the operation of a multispectral scanning system.

In an effort to realize all of the benefits that might come from space-acquired remote-sensing data, NASA developed a crewless spacecraft for the specific purpose of repeatedly acquiring multispectral data about agricultural crops and other features on the surface of the Earth. This spacecraft (*Landsat*) was so programmed that, from its flight altitude of approximately 570 mi (912 km), all of the Earth's surface between about 80°N and 80°S latitudes would be systematically covered, once every 18 days (absence of cloud cover permitting) by its multispectral scanner. Still further progress toward the automation of agricultural remote sensing data analysis was made



(a) (b)

Fig. 5. Field-by-field images of an agricultural area, acquired as in Fig. 4. (*a*) Image multispectral lines obtained with a *Landsat* multispectral scanner from an altitude of 570 mi (912 km) [enlarged 20 times]. (*b*) Computer printout of the same area made from the line scanner's digital data as telemetered to Earth.

possible through the use of multidate records, typically acquired at 9- or 18-day intervals (cloud cover permitting) throughout the entire growing season from Sun-synchronous crewless satellites that are in the *Landsat* series.

The accuracy and timeliness of remote sensing–derived information about agricultural crops could be improved through the use of pertinent weather data as acquired by meteorological satellites with all-weather and day-and-night capabilities. Such capability steadily improves as a more extensive historical database develops that correlates crop production to the multidate weather data as acquired at several times throughout the growing season.

AgRISTARS. Undoubtedly the most substantial research and development activities relative to agricultural applications of remote sensing was the multiagency program of the United States government known as AgRISTARS (Agricultural Resource Inventories Through Aerospace Remote Sensing). This was a follow-on to the completed program, under similar sponsorship, known as LACIE (Large Area Crop Inventory Experiment). The overall goal of this multiagency effort was to determine the usefulness, costs, and extent to which aerospace remote sensing data can be integrated into existing or future USDA information systems to improve the objectivity, reliability, timeliness, and adequacy of information required to carry out USDA missions.

The phase known as early-warning/crop condition assessment provides a capability for the USDA to respond in a timely manner to factors which affect the quality and production of agricultural crops. Among the potentially adverse factors for which a remote sensing–based early-warning capability is needed are diseases, insects, untimely freezes, winds, floods, and droughts. Another phase of AgRISTARS known as foreign commodity production forecasting was to develop and evaluate remote sensing–based technology for making improved forecasts of agricultural production in foreign areas and to determine the suitability of that technology for integration into USDA crop information systems.

The *Landsat*-based multispectral scanner system, and certain derivatives of it, served as the primary data source for evaluation in relation to the AgRISTARS objectives. Listed below are some of the primary attributes of the *Landsat* vehicle and its multispectral scanner system that have value in relation to achieving those objectives. (At present, no other vehicle-sensor systems provide this important combination of characteristics.)

1. With multispectral capability, these scanner systems sense for the optimum wavelength bands needed to take inventory and monitor most types of crops and related agricultural resources, and provide high spectral fidelity within each such band.
2. With multitemporal capability, the systems provide multiple "looks" for monitoring the seasonal development of agricultural crops.
3. The constant repetitive observation point facilitates change detection by the matching of multitemporal images.
4. Sun synchronicity (nearly constant Sun angle) ensures nearly uniform lighting and uniform image tone or color characteristics for use in identification of crops and related features from their multiband tone signatures.
5. A narrow angular field characterizes the sensors. Because of the 570-mi (912-km) altitude and only 115-mi (184-km) swath width of a typical *Landsat* system, tone or color fall-off at the edges of a swath is minimized, thereby increasing still further the uniformity of image tone or color characteristics.
6. Systematic coverage provides a nearly uniform look at all agricultural areas of the entire Earth.
7. Computer-compatible products are provided directly, facilitating automatic data processing.
8. Potential minimum delay in data availability to users permits real-time analysis and thus facilitates the making of timely agricultural inventories, whether on a local, regional, national, or global basis.
9. Spatial resolution is optimum for the first-stage look and, in relation to concerns about military security, is politically palatable, both domestically and internationally.
10. The data can be routinely placed in the public domain for benefit of all peoples.

In addition to these attributes of *Landsat*, there is a capability for the satellite to receive crop-related data as broadcast from ground-based data platforms (such as periodic data on air, soil temperature, soil moisture, wind direction, and wind velocity), process the data with on-board computers, and relay the resulting information back to ground-based receiving stations for integration with data acquired directly from *Landsat*.

Livestock management. Most of the agricultural crops for which management can be improved and production increased through proper inventory and monitoring are of value, primarily, in providing the food and fiber products that are used directly by humans. Other plants, however, are valuable because they provide feed for livestock, leading to the production of meat, hides, and other animal products. Among plants in the latter category are those occupying the pasturelands and rangelands that are grazed by cattle, sheep, and a few other kinds of domesticated animals. Also included are various feed crops such as certain varieties of corn, oats, barley, rye, sorghum, and alfalfa.

On rangelands, there are likely to be wide variations from one year to another in temperature, precipitation, and other climatic factors that control the growth rate of plants and require monitoring by remote sensing from aircraft and spacecraft.

These climatic variations lead to variations not only in the time of range readiness for grazing but also in both the amount and quality of forage produced per unit area—the two primary factors that govern the animal-carrying capacity. Consequently, in any given year, the range manager, equipped with timely, accurate multidate remote sensing–derived information, is better able to determine how many grazing animals, and what kind, should be permitted to graze those lands, and at what time during the grazing season.

Despite the limited spatial resolution of aerospace photography, attempts have been made to take remote sensing–based inventories on rangelands as to the numbers of livestock present, area by area, by livestock type and also by vigor class. For such tasks to be performed, the remote-sensing data must exhibit a sufficiently high degree of spatial resolution. The quality of *Landsat* imagery and other space-acquired photography has not been sufficiently high to permit the making of such determinations.

Robert N. Colwell

Bibliography. M. A. Altieri, *Agroecology: The Scientific Basis of Alternative Agriculture,* 1987; M. A. Altieri and S. B. Hecht, *Agroecology and Small Farm Development,* 1991; E. C. Barrett and L. F. Curtis, *Introduction to Environmental Remote Sensing of Soils and Vegetation,* 1990; P. J. Curran, K. Kandratyev, and V. Kozogerov, *Remote Sensing of Soils and Vegetation,* 1990; C. Reinjtes, B. Haverkort, and A. Winters-Bayer, *Farming for the Future,* 1992.

Agroecosystem

A model for the functionings of an agricultural system, with all inputs and outputs. An ecosystem may be as small as a set of microbial interactions that take place on the surface of roots, or as large as the globe. An agroecosystem may be at the level of the individual plant–soil–microorganism system, at the level of crops or herds of domesticated animals, at the level of farms or agricultural landscapes, or at the level of entire agricultural economies.

Agroecosystems differ from natural ecosystems in several fundamental ways. First, the energy that drives all autotrophic ecosystems, including agroecosystems, is either directly or indirectly derived from solar energy. However, the energy input to agroecosystems includes not only natural energy (sunlight) but also processed energy (fossil fuels) as well as human and animal labor. Second, biodiversity in agroecosystems is generally reduced by human management in order to channel as much energy and nutrient flow as possible into a few domesticated species. Finally, evolution is largely, but not entirely, through artificial selection where commercially desirable phenotypic traits are increased through breeding programs and genetic engineering.

Energy flux. Agroecosystems are usually examined from a range of perspectives including energy flux, exchange of materials, nutrient budgets, and population and community dynamics. Each is a valid approach, but some approaches may be more appropriate in a particular case than others. For example, questions regarding production efficiency would likely require analyses of energy flux and nutrient budgets, while questions regarding crop–pest dynamics would likely necessitate a population or community approach.

Solar energy influences agroecosystem productivity directly by providing the energy for photosynthesis and indirectly through heat energy that influences respiration, rates of water loss, and the heat balance of plants and animals. At the level of individual plants, plant breeders have tried unsuccessfully to improve the effeciency of photosynthesis and reduce respiration losses. These basic metabolic pathways seem resistant to artificial selection. Breeders have successfully altered plant architecture, such as by changing leaf angles in wheat, so that the plant intercepts more sunlight. Similarly, animal breeders have selected livestock strains that are better able to withstand temperature extremes. *See* BIOLOGICAL PRODUCTIVITY; PHOTOSYNTHESIS.

The percentage of solar energy that is used by fields of crops also depends on the types of crops and spacing patterns. Densely planted crops such as wheat intercept more sunlight than more sparsely planted crops such as many vegetables. Comparisons of yields of different crops in similar fields show that net primary production at crop levels is a function of species-specific physiology and architecture and planting practices. For example, between 1930 and 1985 the average yield for maize in the United States increased from approximately 38.2 to 118 bushels/acre (3325 to 10,271 liters/hectare), and soya bean yields increased from 21.7 to 34.1 bu/acre (1915 to 2959 liters/ha). These yield differences reflected inherent species differences, increased planting densities, improved and more intense production inputs, and selection for new varieties. *See* ECOLOGICAL ENERGETICS.

Nutrient resources. Nutrient uptake from soil by crop plants or weeds is primarily mediated by microbial processes. Some soil bacteria fix atmospheric nitrogen into forms that plants can assimilate. Other organisms influence soil structure and the exchange of nutrients, and still other microorganisms may excrete ammonia and other metabolic by-products that are useful plant nutrients. There are many complex ways that microorganisms influence nutrient cycling and uptake by plants. Some microorganisms are plant pathogens that reduce nutrient uptake in diseased plants. Larger organisms may influence nutrient uptake indirectly by modifying soil structure or directly by damaging plants. When farms contain both domestic livestock and crops, animal manure may provide additional crop nutrients. *See* NITROGEN CYCLE; SOIL MICROBIOLOGY.

In most commerical farm operations, crop nutrients are greatly supplemented by fertilizer application to enhance yields. At the crop level, the long-term addition of fertilizer to soils may change soil pH, increase mineral salt concentration, and alter the

normal microbe–soil–plant interactions. *See* FERTILIZER.

Population and community interactions. Although agroecosystems may be greatly simplified compared to natural ecosystems, they can still foster a rich array of population and community processes such as herbivory, predation, parasitization, competition, and mutualism. Crop plants may compete among themselves or with weeds for sunlight, soil nutrients, or water. Cattle overstocked in a pasture may compete for forage and thereby change competitive interactions among pasture plants, resulting in selection for unpalatable or even toxic plants. Indeed, one important goal of farming is to find the optimal densities for crops and livestock. *See* HERBIVORY; POPULATION ECOLOGY.

However, other factors may influence a farmer's decision to plant at densities above or below this optimum. Risk of some mortality due to density-independent factors (such as from local variation in soil moisture) may lead to planting densities above the optimum. If the farmer is concerned about density-dependent mortality (for example, from a fungal pathogen whose spores are spread from plant to plant), crops may be planted more sparsely. Other factors of production, such as the cost of fertilizer, may also influence planting density. Optimal planting densities for a particular crop are determined with respect to a given set of production inputs, such as fertilizer levels, pesticides, irrigation, and tillage. Thus, at any given locale and planting year, a particular crop may have a range of optimal planting densities that depends on the amount and type of production factors used, and on the environmental risk factors that may concern the farmer.

The presence of weeds among crops can reduce crop yields through interspecfic competition. However, under many conditions the presence of weeds may not affect crop yields, and in certain circumstances weeds may actually enhance long-term production. Whether or not weeds will affect crop yields depends on the particular weed-crop species. Many weeds have little effect on yields and can be largely ignored, while other weeds, such as johnsongrass, can severely affect crop yields. At the appropriate densities, some weeds can improve crop production, particularly of orchard and vine crops, by providing ground cover, providing a habitat for beneficial insects, and improving soil nutrients. Weeds are also important for restoring soil nutrients and for protection against erosion during fallow periods. *See* WEEDS.

Biotechnology. Human influence on agroecosystems is increasing as advances are made in research geared toward artifically manipulating agricultural populations. New developments in pesticides and genetic engineering have had particularly strong impacts.

Pesticides. Since World War II, widespread use of synthetic chemical pesticides has bolstered farm production worldwide, primarily by reducing or eliminating herbivorous insect pests. Traditional broad-spectrum pesticides such as DDT, however, can have far-ranging impacts on agroecosystems. For instance, secondary pest outbreaks associated with the use of many traditional pesticides are not uncommon due to the elimination of natural enemies or resistance of pests to chemical control. The Colorado potato beetle, a common pest of many crops, quickly developed resistance to scores of pesticides; it examplifies the way in which long-term pesticide use may alter agroecosystem relationships. In addition, broad-spectrum pesticides may alter the soil community in an agroecosystem or kill pollinators and other beneficial organisms.

In 1996 the U.S. Congress passed the Food Quality Protection Act, which mandates a more restrictive means of measuring pesticide residues on crop plants. This legislation, borne primarily out of environmental and public health concerns, has resulted in the severe reduction and in some cases the banning of many broad-spectrum, traditional pesticides. As a result, growers and pesticide developers in temperate regions have begun to focus on alternative means of control. Pesticide developers have begun producing selective pesticides, which are designed to target only pest species and to spare natural enemies, leaving the rest of the agroecosystem community intact. Many growers are now implementing integrated pest management programs that incorporate the new breed of biorational chemicals with cultural and other types of controls. *See* FOREST PEST CONTROL; INSECT CONTROL, BIOLOGICAL; PESTICIDES.

Genetic engineering. The last few decades have seen tremendous advances in molecular approaches to engineering desirable phenotypic traits in crop plants. Traits such as resistance to insect attack, drought, frost, pathogens, and herbicides are among many that have been genetically spliced into numerous crop plants and made commercially available in the global market. Although artificially modifying crop plants is nothing new, the techniques used in genetic engineering allow developers to generate new varieties an order-of-magnitude faster than traditional plant breeding. In addition, genetic engineering differs from traditional breeding in that the transfer of traits is no longer limited to same-species organisms. Many of these genetically modified organisms are being designed with an eye toward increased marketability (for example, better-tasting crops and consistent ripening) as well as better environmental protection (for example, reduced reliance on pesticide sprays). At the same time, scientists are still assessing the effects that the widespread deployment of these traits may have on agroecosystems and natural ecosystems. There is some concern, for instance, that engineered traits may escape, via genes in pollen transferred by pollinators, and become established in weedy populations of plants in natural ecosystems, in some cases creating conservation management problems and new breeds of superweeds. As with pesticides, there is evidence

Maize production practices in the United States and Guatemala

Production factors	Guatemala	United States
Cost	Personal labor accounts for nearly all factors of production.	Fertilizers and pesticides amount to more than one-third of total production costs.
Planting technique	Maize is planted with other crops (intercropping) such as beans and squash. Total combined yield is greater than crops grown as monocultures in separate fields.	Maize is planted as a monoculture; chemical inputs set planting densities; sometimes maize is rotated with other crops based on market demand.
Distribution	Maize is used for human consumption; small livestock is fed on surplus grain and crop residue. Livestock is allowed to forage in fields after harvest.	Maize is used mostly for animal feeds or as additives (oils and meal) to prepared foods for human consumption.
Yield	Relatively low.	High.
Environmental factors	Land pressures force steep-slope cultivation, causing significant erosion. The chemical control of weeds can lead to pollution.	High environmental costs include soil erosion and chemical contamination of soil and water.

that insects are already becoming resistant to some more widespread traits used in transgenic plants, such as the antiherbivore toxin produced by the bacterium *Bacillus thuringiensis*. *See* BIOTECHNOLOGY; GENETIC ENGINEERING.

Comparative agroecosystems. A contrast between extreme forms of farm production illustrates the different principles of agroecosystem structure and function (see **table**). Commercial maize production in the United States corn belt can be contrasted with traditional maize cultivation in the highlands of Guatemala. In both regions, maize is a major part of the farm economy.

In highly commercial farms, the agroecosystem is more open in that soil loss from erosion and nutrient and chemical leaching are high; and in that yields are sold off the farm. Because fertilizer, pesticides, and other inputs are readily available, the weaknesses of the system are more than compensated by the large amounts of inputs. In the more traditional system, the agroecosystem is more closed with a greater cycling of nutrients taking place within the farm. Chemical inputs are low because the complexity of the fields offers some protection against pests and because the combination of maize and legumes may provide a mutualistic benefit. Bacteria associated with the legume roots fix atmospheric nitrogen, possibly providing the maize with a source of nitrogen. The vine legumes benefit from using the maize as a trellis.

The specter of incorporating biotechnological advances into commercial farms has increased hopes of augmenting on farm productivity while also reducing the use of environmentally damaging chemical inputs. *See* AGRICULTURAL SOIL AND CROP PRACTICES; AGRICULTURE; ECOSYSTEM; SOIL CONSERVATION.

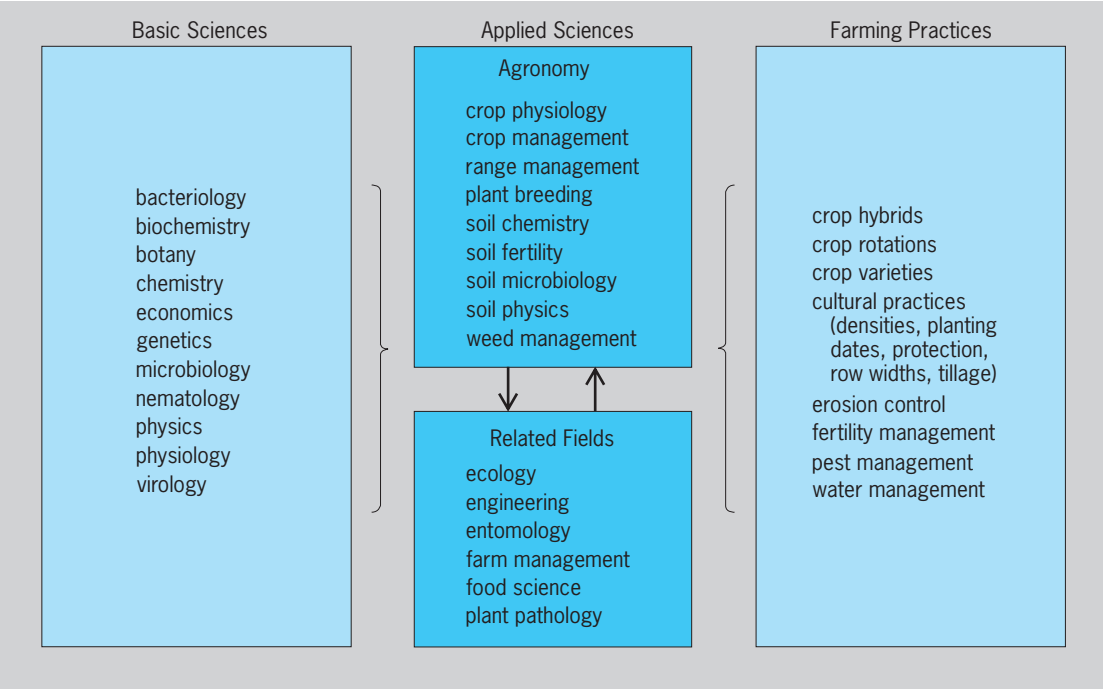
C. Ronald Carroll; Carol A. Hoffman; John E. Banks

Bibliography. C. R. Carroll, J. H. Vandermeer, and P. M. Rosset (eds.), *Agroecology*, 1990; S. R. Gliessman, *Agroecology: Ecological Processes in Sustainable Agriculture*, 1998; S. R. Gliessman, *Agroecology: Researching the Ecological Basis for Sustainable Agriculture*, 1990; S. R. Gliessman, *Agroecosystem Sustainability: Developing Practical Strategies*, CRC Press, 2001; O. Koul and G. S. Dhaliwal (eds.), *Predators and Parasitoids* (Advances in Biopesticide Research Series), Taylor & Francis, 2003; S. Krimsky and R. Wrubel, *Agricultural Biotechnology and the Environment*, 1996; National Research Council, *Alternative Agriculture*, 1989; National Research Council, *Environmental Effects of Transgenic Plants: The Scope and Adequacy of Regulation*, National Academy Press, 2002.

Agronomy

The science and study of crops and soils. Agronomy is the umbrella term for a number of technical research and teaching activities (see **illus.**): crop physiology and management, soil science, plant breeding, and weed management frequently are included in agronomy; soil science may be treated separately; and vegetable and fruit crops generally are not included. Thus, agronomy refers to extensive field cultivation of plant species for human food, livestock and poultry feed, fibers, oils, and certain industrial products. *See* AGRICULTURE.

Agronomic studies include some basic research, but the specialists in this field concentrate on applying information from the more basic disciplines (see illus.), among them botany, chemistry, genetics, mathematics, microbiology, and physiology. Agronomists also interact closely with specialists in other applied areas such as ecology, entomology, plant pathology, and weed science. The findings of these collaborative efforts are tested and recommended to farmers through agricultural extension agents or commercial channels to bring this knowledge into practice. Agronomy is an integrative science that brings together principles and practices from other fields into combinations that are useful to farmers. This critical area is now focused on the efficiency of resource use, profitability of management practices, and minimization of the impact of farming on the immediate and the off-farm environment.



Relationship of agronomy to basic sciences, related fields, and practical farming methods.

History of cropping. Agronomy has its roots in the prerecorded history of human settlements. The first agronomists were no doubt women who nurtured small plants in garbage areas near campsites and then began to consciously bring back plants and seeds and tend them to useful maturity in these primitive gardens. The most vigorous and desirable plants were used, and seeds were selected for the next planting. Such trial-and-error planting led to the development of more permanent settlements based on agriculture to complement gathering, fishing, and hunting. The more organized farming systems with crops in rows and intensive management are barely more than one to two centuries old. This so-called agricultural revolution and the subsequent application of machines during the industrial revolution brought increased efficiency to farming and reduced much physical drudgery. Agronomy now brings practical science and the basic results of research to bear on increased productivity and efficiency. *See* AGRICULTURAL MACHINERY.

Science has made major contributions to agriculture through agronomy in the latter half of the twentieth century. Hybrid corn and sorghum, new soybean and wheat varieties, manufactured or enhanced chemical fertilizers, and a range of pesticides contributed greatly to labor productivity and crop yield during this time. In some developing countries in the 1960s and 1970s, a so-called green revolution based on new plant varieties, fertilizer and pesticide usage, and improved water control brought new hope to areas facing severe population and food crises. The green revolution was effective primarily on fertile, mechanized farms where infrastructure was well developed, inputs and markets were available, and substantial capital was invested in agriculture. *See* BREEDING (PLANT); FERTILIZER; PESTICIDE.

Technical specialists. Agronomists promote the use of improved technology through research, teaching, and extension activities. In the United States, much of this is in the state and federal systems. Commercial employment includes soil testing and recommendations, scouting for insect, weed, and disease problems, and consulting on fertilizer, pesticide, and irrigation water use. Field representatives and salespeople for chemical and fertilizer companies, as well as technical seed production specialists, have training in agronomy. Many farmers and consultants for large commercial farms have formal college or technical school courses in the specific details of science as related to agriculture.

Future directions. The positive and negative impacts of current farming practices are becoming more apparent. High-productivity-per-unit labor has been achieved through large investments of fossil-fuel-derived inputs, such as fertilizers, pesticides, and diesel fuel for drying crops and pumping water for irrigation. In addition, the amounts and off-farm effects of nitrates and pesticide residues in surface waters and underground aquifers can be measured with greater precision. Agriculture in some areas has been identified as a major nonpoint source of pollution of the environment. Agronomists have been called upon to provide solutions in these areas. *See* IRRIGATION (AGRICULTURE).

Principles of plant ecology and the biological processes and interactions in natural communities are being studied to gain insight into the design of future cropping systems. Crop rotations, reduced tillage, and contour strip cropping can help prevent the loss

of soil and nutrients from sloping lands or reduce the need for added nitrogen before the planting of each cereal crop. Rotations break the reproductive cycles of weeds and insects and reduce pesticide requirements. Plant breeders can develop new hybrids and varieties that are more tolerant to diseases and insects. Commercial applications of biotechnology will result in products that are biologically derived, are used in ultralow volume, or have a more specific effect on the target species. The result is a less negative impact on desirable species and on the total environment. Agronomists in the public and private sectors integrate these new technologies with current farm practices to create a more resource-efficient, profitable, and sustainable agriculture. *See* AGRICULTURAL SOIL AND CROP PRACTICES; AGROECOSYSTEM; CONSERVATION OF RESOURCES; SOIL CONSERVATION.

Charles A. Francis

Bibliography. J. A. Barden, R. G. Halfacre, and D. J. Parrish, *Plant Science*, 1987; W. R. Fehr, *Principles of Cultivar Development*, 1987; M. S. Kipps, *Production of Field Crops: A Textbook of Agronomy*, 6th ed., 1970; J. H. Martin, W. H. Leonard, and D. L. Stamp, *Principles of Field Crop Production*, 3d ed., 1989; N. C. Stoskopf, *Cereal Grain Crops*, 1985; S. L. Tisdale and W. L. Nelson, *Soil Fertility and Fertilizers*, 6th ed., 1998.

Aharonov-Bohm effect

The predicted effect of an electromagnetic vector or scalar potential in electron interference phenomena, in the absence of electric or magnetic fields on the electrons.

The fundamental equations of motion for a charged object are usually expressed in terms of the magnetic field $\vec{B}$ and the electric field $\vec{E}$. The force $\vec{F}$ on a charged particle can be conveniently written as in Eqs. (1), where q is the particle's charge, $\vec{v}$ is

$$\vec{F} = q\vec{E} \qquad \vec{F} = q\vec{v} \times \vec{B} \tag{1}$$

its velocity, and the symbol $\times$ represents the vector product. Associated with $\vec{E}$ is a scalar potential V defined at any point as the work W necessary to move a charge from minus infinity to that point, $V = W/q$. Generally, only the difference in potentials between two points matters in classical physics, and this potential difference can be used in computing the electric field according to Eq. (2), where d is the distance

$$\vec{E} = \frac{V_1 - V_2}{\vec{d}} \tag{2}$$

between the two points of interest. Similarly, associated with $\vec{B}$ is a vector potential $\vec{A}$, a convenient mathematical aid for calculating the magnetic field; the magnetic field is the curl of the vector potential, Eq. (3).

$$\vec{B} = \nabla \times \vec{A} \tag{3}$$

See CALCULUS OF VECTORS; ELECTRIC FIELD; POTENTIALS.

In quantum mechanics, however, the basic equations that describe the motion of all objects contain $\vec{A}$ and V directly, and they cannot be simply eliminated. Nonetheless, it was initially believed that these potentials had no independent significance. In 1959, Y. Aharonov and D. Bohm discovered that both the scalar and vector potentials should play a major role in quantum mechanics. They proposed two electron interference experiments in which some of the electron properties would be sensitive to changes of $\vec{A}$ or V, even when there were no electric or magnetic fields present on the charged particles. The absence of $\vec{E}$ and $\vec{B}$ means that classically there are no forces acting on the particles, but quantum-mechanically it is still possible to change the properties of the electron. These counterintuitive predictions are known as the Aharonov-Bohm effect.

Effect for vector potentials. According to quantum mechanics, any object with mass can behave as both a particle and a wave. Classical interference phenomena, such as the interference pattern produced when a point source of light is imaged by two slits, are still observed when this light source is replaced by an electron beam, demonstrating the electron's wavelike nature. The phase δ of the quantum-mechanical wave function for the electron can be written as in Eq. (4), where δ_0 is an arbitary phase factor, e is

$$\delta = \delta_0 + \frac{2\pi e}{h}\int \vec{A}\cdot\vec{dl} + \frac{2\pi e}{h}\int V(t)dt \tag{4}$$

the charge of the electron, h is Planck's constant, the integral of the dot product of the vector potential $\vec{A}$ with $\vec{dl}$ is taken over the electron path, and

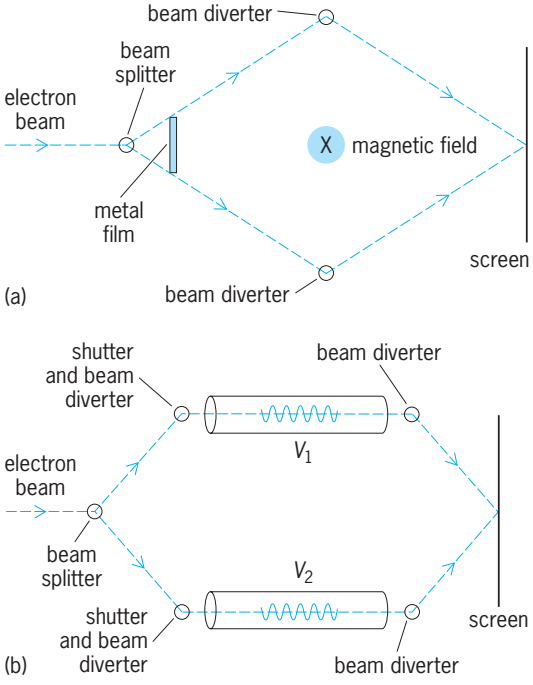


Fig. 1. Aharonov and Bohm's proposed experiments to demonstrate the effects of electromagnetic potentials in electron interference phenomena. (*a*) Experiment to verify the effect of a vector potential. (*b*) Experiment to demonstrate the effect of a scalar potential.

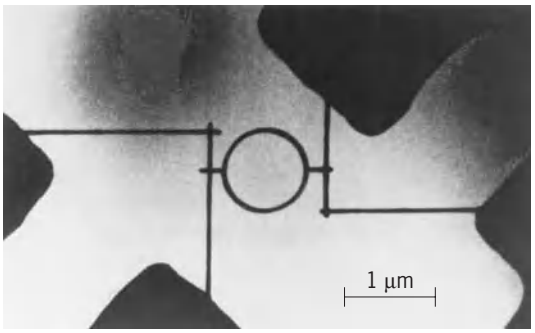


Fig. 2. Very small ring fabricated from a gold film 38 nm thick. LInes forming the ring are 40 nm wide, and the average diameter is 0.82 μm. The electrical resistance is measured by applying a constant current through two of the leads connecting the sample and measuring the voltage difference between the two other leads. (*From R. A. Webb et al., Observation of h/e Aharanov-Bohm oscillations in normal-metal rings, Phys. Rev. Lett., 54:2696–2699, 1985*)

the scalar potential $v(t)$ can vary with time t. *See* ELECTRON DIFFRACTION; INTERFERENCE OF WAVES; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

In Aharonov and Bohm's first proposed experiment (**Fig. 1***a*), a beam of electrons traveling in a vacuum is split such that half of the beam travels clockwise around a region containing a magnetic field and the other half counterclockwise. The integral of the vector potential over the electron path given in Eq. (4) is identically equal to the magnetic flux Φ (the magnetic field times the area over which the field is nonzero) enclosed by the two beams. When the beams are recombined, the intensity of the resulting beam will oscillate periodically as the enclosed magnetic field is changed. This oscillation, with a period in Φ of $\Phi_0 = h/e$, represents patterns of constructive and destructive interference occurring between

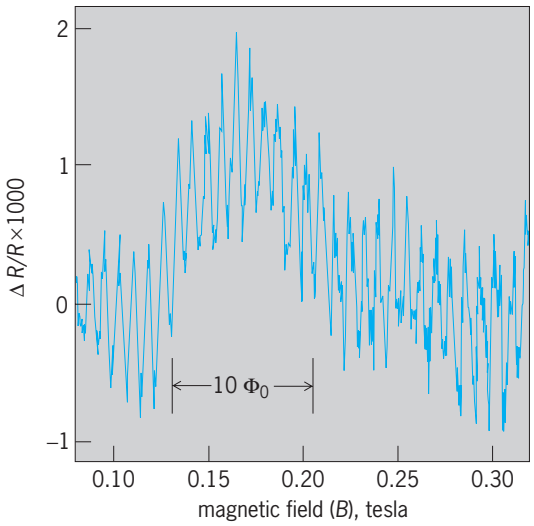


Fig. 3. Resistance oscillations observed in the ring of Fig. 1 at temperature of 0.01 K over a very small magnetic field range. Vertical scale measures deviation ΔR of resistance from its mean value R. The period of the h/e Aharonov-Bohm oscillations in magnetic field is 0.0076 tesla (76 gauss). The arrows show the field scale for 10 flux quanta (10 Φ_0) in the average area of the ring.

the two beams. If the magnetic field (produced, for example, by an infinite solenoid) is confined to the interior of the region within the electron paths, the recombined beam will still oscillate periodically as the field is varied because $\vec{A}$ is not zero in the region where the electrons travel; only $\vec{B}$ is.

Within a year of this prediction, evidence for the Aharonov-Bohm effect and the physical reality of the vector potential was obtained in electron-beam experiments, but only in 1986 was a conclusive demonstration obtained.

Effect for scalar potentials. Aharonov and Bohm's second interference experiment (Fig. 1*b*) is designed to demonstrate that a scalar potential can also change the phase of the electron wave function. An electron beam is again split in half; each half passes through a long, cylindrical, high-conductivity metal tube; and the beam is recombined. Two electrical shutters chop the beam into segments that are much shorter than the tube length but much longer than the de Broglie wavelength. If a different scalar potential (V_1 and V_2 in Fig. 1*b*) is applied to each metal tube after the chopped beam enters the tube and is turned off before the beam leaves the tube, the intensity of the recombined beam will be altered and will oscillate periodically as the potential difference between the tubes ($V_1 - V_2$) is changed. Since a potential applied to a highly conducting metal tube produces no electric field inside the tube, no classical forces act on these electrons during the experiment. While this prediction has not been conclusively demonstrated, evidence for its validity has been obtained in electron-beam experiments and in electrical measurements on normal metals. *See* ELECTROSTATICS.

Effect for normal metals. Surprisingly, the Aharonov-Bohm effect plays an important role in understanding the properties of electrical circuits whose wires or transistors are smaller than a few micrometers. The electrical resistance in a wire loop oscillates periodically as the magnetic flux threading the loop is increased, with a period of h/e, the normal-metal flux quantum. In single wires, the electrical resistance fluctuates randomly as a function of magnetic flux. Both these observations reflect an Aharonov-Bohm effect. They have opened up a new field of condensed-matter physics because they are a signature that the electrical properties are dominated by quantum-mechanical behavior of the electrons, and that the rules of classical physics are no longer operative.

These experiments were made possible by advances in the technology for fabricating small samples. The metallic ring shown in **Fig. 2** is about 100 atoms wide and 100 atoms thick, and contains only about 10^8 atoms. As the magnetic field is varied, the electrical resistance of the ring oscillates with a characteristic period of $B = 0.0076$ tesla (**Fig. 3**). From measurements of the average area S of the ring, the period is equal to the normal-metal flux quantum $\Phi_0 = h/e = BS$. These data demonstrate that the Aharonov-Bohm effect exists in resistive samples.

Richard A. Webb

Bibliography. M. Peshkin and A. Tonomura, *The Aharonov-Bohm Effect*, 1989.

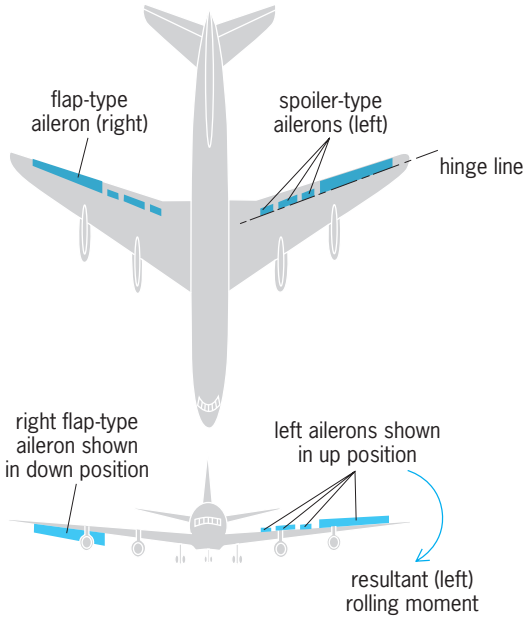
Aileron

The hinged rear portion of an aircraft wing, moved differentially on each side of the aircraft to obtain lateral or roll control moments. The angular settings of the ailerons are controlled by the human or automatic pilot through the flight control system. Typical flap- and spoiler-type ailerons are shown in the **illustration**. *See* FLIGHT CONTROLS.

Operating principles. The operating principles of ailerons are the same as for all trailing-edge hinged control devices. Deflection of an aileron changes the effective camber, or airfoil curvature relative to the wing chord, of the entire wing forward of the aileron. With the trailing edge deflected upward, reduced low flow velocities are produced on the upper wing surface, and increased local flow velocities are produced on the lower wing surface. By Bernoulli's law, this results in a reduction of lift over the portion of the wing forward of the aileron, and on the aileron itself. Conversely, trailing-edge down deflection of a flap-type aileron increases the lift in the same areas. Ailerons are located as close as possible to the wing tips, to maximize rolling moment by increasing the moment arm of the force due to the change in wing lift. In the case of flap-type ailerons, when the trailing edge is raised on one wing, say the left, the trailing edge of the aileron on the opposite or right wing is lowered by about the same amount. The decrease in lift on the left wing is accompanied by a lift increase on the right wing. While the net wing lift remains about the same, a rolling moment or torque about the aircraft's fore-and-aft axis develops in a left, or counterclockwise, direction as seen by the pilot. *See* BERNOULLI'S THEOREM.

Aileron effectiveness is reduced at supersonic speeds, since trailing-edge devices cannot affect pressures on the wing surface forward of themselves when the flow velocity exceeds the speed of pressure propagation (speed of sound). Lift changes thus occur only on the ailerons themselves at supersonic speeds, except for end effects. *See* SUPERSONIC FLIGHT.

Flight maneuvers. Ailerons are used to perform rolling or banking maneuvers. Banked turns, in which one wing is lower than the other, are used to change aircraft heading. Ailerons are deflected to establish a rolling moment and hence a rolling acceleration. The rolling acceleration becomes a steady rolling velocity usually less than 2 s after the ailerons have been deflected, as a result of the heavy damping in roll (rolling moment opposing rolling velocity) of the wings. Neutralizing the ailerons stops the rolling velocity, thus allowing the rolling motion to be stopped at the desired bank angle. Aileron control effectiveness is evaluated relative to given levels of rolling velocity, or alternatively to banking a given amount, such as 30 degrees, in a specified time. Other functions of the ailerons are the maintenance of lateral balance under conditions of aerodynamic, inertial, or thrust asymmetries, crosswind takeoffs and landings, deliberate sideslips, and lateral gusts. Ailerons have become the primary spin-recovery control for aircraft with low rolling inertia relative to pitching inertia; this characteristic is typical of modern fighter aircraft with thin wings of short span.



Flap and spoiler-type ailerons on jet transport airplane.

Spoiler ailerons. Flap-type ailerons are replaced or supplemented by spoiler-type ailerons for a variety of reasons. Spoiler ailerons are usually installed forward of the landing flaps on commercial jet transports, in order to supplement aileron effectiveness during landing approaches, when the landing flaps are extended. Greatly reduced takeoff and landing speeds can be obtained by devoting the trailing edge of the entire wing to high-lift flaps. This is made possible by substituting spoilers for flap-type ailerons.

Spoilers have the additional advantages of reduced wing twisting moments due to chordwise aerodynamic loading and of improved effectiveness at transonic speeds. Spoilers designed for lateral control at transonic or supersonic speeds are often located on the inboard portion of the wings. This further reduces wing twisting due to chordwise aerodynamic loading.

A major disadvantage of spoiler ailerons is that a practical means of self-balancing the operating torques with other aerodynamic torques has not been found. All current spoilers are hinged at the leading edge and must be forced open against large positive aerodynamic pressures acting on their outer surfaces. *See* AERODYNAMICS; WING.

Malcolm J. Abzug

Bibliography. J. D. Anderson, *Introduction to Flight*, 4th ed., 1999; E. H. Pallett, *Automatic Flight Control*, 4th ed., 1994.

Air

A predominantly mechanical mixture of a variety of individual gases enveloping the terrestrial globe to form the Earth's atmosphere. In this sense air is one of the three basic components, air, water, and land (atmosphere, hydrosphere, and lithosphere), that form the life zone at the face of the Earth. *See* AIR POLLUTION; AIR PRESSURE; AIR TEMPERATURE; ATMOSPHERE; ATMOSPHERIC CHEMISTRY; ATMOSPHERIC GENERAL CIRCULATION; METEOROLOGY.

Radio, radar, rockets, satellites, and interplanetary probes are stimulating investigation of the upper atmosphere and the transition zone to outer space. *See* IONOSPHERE.

Some aspects of air important in engineering are well known. For example, pneumatic equipment commonly transmits force and energy in a variety of mining machines and drills, in air brakes, automotive air suspension, and other devices. The study of forces and moments exerted on aircraft in motion through the air is a special field of aerodynamics. *See* AERODYNAMIC FORCE; AERODYNAMICS; COMPRESSOR; WIND STRESS.

Charles V. Crittenden

Bibliography. M. Allaby, *Air*, 1992.

Air armament

That category of weapons which are typically delivered on target by fixed or rotary-wing aircraft, with the exception of nuclear weapons. Specifically included are guns and ammunition, rockets, free-fall bombs, cluster weapons that consist of a dispenser and submunitions, air-to-air and air-to-surface guided weapons, mines, and antiradiation missiles. Nuclear weapons such as the air-launched cruise missile and nuclear bombs, which are delivered by aircraft, are considered strategic weapons, a specific class of air armament. Related to, but not included as, armament are carrying racks and ejector mechanisms, fire-control computers, interface electronics, and head-up displays, which are critical to delivering the unguided weapons accurately. *See* ATOMIC BOMB; HYDROGEN BOMB.

Aircraft guns and cannon. Since World War II, all major nations have practiced optimization, or fitting the weapon to the job, in attempting to obtain maximum effectiveness against anticipated targets with minimum weight and volume systems. The most widely used aircraft guns outside the former Communist countries are the French DEFA and British Aden 30-mm guns, followed by the United States 20-mm M61 Vulcan.

Since World War II, the Russians have employed a variety of calibers. Their dominant weapons are all designed to fire the same 23-mm round, although .50-caliber guns are still seen on some helicopters and 30 mm on some fixed-wing aircraft.

The most modern aircraft guns are the 27-mm Mauser developed for multirole use on the German, Italian, and British multirole combat aircraft (MRCA) Tornado, and the large 30-mm GAU-8 and GAU-13 antiarmor guns which carry high-explosive incendiary (HEI) and armor-piercing incendiary (API) ammunition. The GAU-8 is carried inside the United States A-10 aircraft, for which it was developed. The GAU-13, developed for the A-7/F-4, can be mounted in the GPU-5A gunpod, outside most modern tactical fighters.

A 20-mm aircraft gun system for air-to-air combat has been developed that greatly expands the gun engagement envelope by employing telescoped ammunition. This gun was tailored for the F-22 and Joint Strike Fighter (JSF) aircraft. With this type of ammunition, the muzzle velocity of the round is increased to over 5000 ft/s (1500 m/s), compared with 3300 ft/s (1000 m/s) for conventional ammunition. This feature, combined with a lead-computing, optical gunsight, gives the pilot an all-aspect engagement capability.

Aircraft rockets. Free rockets, or unguided rockets, developed during World War II, are more accurate than bombs but less accurate than guns. Their warheads are generally larger than shells that can be fired from aircraft guns but smaller than most bombs. Most rocket usage is air-to-surface, where reasonable accuracy is attainable. Although rocket systems are often considered obsolete, virtually all major powers maintain one or more in their arsenal. Most current systems are fin-stabilized (as opposed to spin-stabilized), and vary from about 2.4 in. (60 mm) to over 8 in. (200 mm) in diameter. They are designed to be employed at ranges from 3000 to 30,000 ft (1 to 10 km). *See* ROCKET.

Bombs. Aircraft-delivered bombs were considered to be the ultimate weapon until the end of World War II. Conventional bombs employing TNT or TNT-based mixtures have been made in many types and sizes. Armor-piercing bombs were designed for use against concrete fortifications and submarine pens. Demolition bombs have thin steel cases and maximum explosive loads. General-purpose bombs are a compromise between armor-piercing and demolition. Bombs have been made to deliver virtually anything against any type of target. Some examples are incendiary bombs to ignite wooden buildings; napalm or jellied gasoline for use against anything that can be damaged by flame; underwater bombs for use against hydroelectric dams; cluster bombs, which separate and disperse to saturate an area; and leaflet bombs to deliver propaganda.

General-purpose, demolition, and armor-piercing bombs are made in a variety of sizes up to about 2000 lb (900 kg). The advent of nuclear weapons made larger bombs obsolete. Smaller, more accurate, conventional bombs are preferred to destroy targets while minimizing collateral damage. Examples are the large number of smart weapons employed during Operation Desert Storm in 1991 and subsequent air strikes against Iraq to enforce the United Nations sanctions.

The majority of bombs (and other stores) are carried on ejector racks, which forcibly eject the bomb through aerodynamic interference to ensure that it

leaves the aircraft. Pylons, racks, and ejectors have become a special field of aircraft armament. *See* AERODYNAMICS.

Cluster weapon. An entirely new class of aircraft armament known as cluster weapons has evolved since the 1950s. It is partly an outgrowth of cluster bombs from World War II and partly a result of the realization that many targets, notably personnel and light vehicles, could be destroyed more efficiently with several small bombs than with one large bomb. The primary advantage of cluster weapons over unitaries is their large footprint or area covered, which compensates for delivery errors and target uncertainty errors incurred by unguided weapons.

Both captive dispensers, which stay with the aircraft, and droppable stores, which sometimes fly preprogrammed maneuvers, are used. The means of dispersing the munitions are varied. The munitions may be shot out of their recesses, ejected by air or gas pressure, separated after ejection by various aerodynamic means, or even separated by centrifugal force from a spinning canister. The trend is toward dispensers that deploy a wide pattern of munitions from aircraft at treetop height.

The 950-lb (430-kg) combined effects munition (CBU-87/B, where CBU stands for cluster bomb unit) can be delivered at supersonic speeds and is used to attack armor, vehicles, and personnel over a wide area. Each free-fall dispenser carries 202 submunitions, weighing 3.4 lb (1.5 kg) apiece, which consist of a steel fragmentation case, a shaped charge that can defeat armored vehicles, and fire-starting materials. Fusing can be selected as a function of time for low-level dispensing or as a function of ground proximity when the weapon is released at higher altitudes. Adjustable dispenser-spin and submunition-dispersion characteristics allow the weapon to be tailored to the situation, such as vehicles parked in an area or strung out along a moving column.

The CBU-89/B or Gator mine system carries a combination of 72 antitank/antivehicle mines and 22 antipersonnel mines. This 710-lb (320-kg) free-fall cluster weapon is the first and only scatterable mine system in the U.S. Air Force inventory, and can be used against tanks, trucks, personnel, and armored vehicles. After release from an aircraft, the dispenser is opened by either a timed fuse or a ground-proximity sensor, depending on the aircraft altitude. This altitude determines how widely the mines are scattered. The mines detonate when either a preprogrammed self-destruct time-out is reached, the battery voltage in the mine falls below a certain level, or a target is detected. The antitank/antivehicle mines sense the proximity of the target with a magnetic sensor. The antipersonnel variety explodes if any one of four triplines is disturbed or someone tries to move it.

An even more sophisticated cluster weapon, the sensor-fused weapon, CBU-97/B, is designed to carry 10 submunition delivery vehicles (SDVs), each containing four warheads. After dispersion, the SDVs orient, stabilize, and drop by parachute to the optimal altitude as determined by the on-board altimeter. The 40 sensor-fused warheads are then dispersed, each searching for targets by using individual infrared sensors. When a target is sensed, the warhead fires a high-velocity, self-forging projectile at the target vehicle.

The cluster bomb units, CBU-87/B, CBU-89/B, and CBU-97/B, greatly improved in accuracy with the addition of a fin control kit containing an inertial guidance system, which achieved initial operational capability in 1999. With the added kit, the cluster weapons are known as wind-corrected munitions dispensers, CBU-103/B, CBU-104/B, and CBU-105/B. The need for such kits became apparent during the high-altitude bombing of Iraq during Desert Storm, where wind effects made the basic cluster weapons ineffective.

Fire control. Although not normally included as part of the field of aircraft armament, fire control is the term that covers the sighting, aiming, and computation which enables the pilot or aircrew to hit the target. Basically, it examines the conditions of the engagement and indicates when to release the armament in order to obtain hits.

The most primitive system, the fixed sight, is a mechanical or optical device that is aligned to coincide with the gun-bore line some fixed distance in front of the aircraft. To use such a sight requires estimating range, lead, deflection, and a number of lesser variables. *See* GUNSIGHTS.

The first degree of sophistication, the lead computing optical sight, uses gyroscopic sensors to measure own-vehicle motion and, based on the assumption that both aircraft are flying similar maneuvers, predicts the future position of the target and lags the reticle, or pipper, so that when the reticle is on the target the correct lead has been generated. This system is also known as a disturbed-reticle sight because the reticle is in constant motion in a maneuvering encounter.

The ultimate fire-control system, the director, measures all conditions, own-vehicle motion, velocity, air density, angle of attack, and other deterministic variables; tracks the target with radar or electrooptical sensors; measures the range to target; computes in real time the ballistics of the round; predicts where it will pass the target at target range; and displays a reticle on the head-up display in the pilot's field of view. If the pilot pulls the trigger when this reticle is on the target, hits will likely result.

Guided weapons. Guided weapons is a generic term which applies to any of the previously described ballistic systems when they are deliberately perturbed from their ballistic path after launch in order to increase the probability of hitting a target. There are three fundamental problems in guided weapons: determining where the target is or will be, determining where the weapon is, and correcting the weapon's location to coincide with the target's location at the time of closest encounter. The first two problems are called guidance; the third, control. Control is the most straightforward problem and is usually accomplished by aerodynamic-control-surface deflection, although it may also be achieved

by rocket-thrust-vector control, gas thrusters, explosive squib control, or any other device to alter the velocity vector of the weapon. *See* AUTOPILOT; FLIGHT CONTROLS; GUIDANCE SYSTEMS.

There are five fundamental concepts of guidance, namely inertial guidance, command guidance, active guidance, semiactive guidance, and passive guidance, and many different implementations of each.

Inertial guidance. This requires that the initial conditions at launch be precisely known, that is, distance and heading to target, and velocity and orientation of the launch vehicle. On-board sensors (gyroscopes and accelerometers) measure all changes in initial conditions and, using computers, direct the guidance elements (autopilot) to maneuver the weapon to the target. This guidance concept is not very accurate because it is open-loop and is generally used with more precise terminal guidance which is closed-loop.

Aided inertial guidance using an inertial navigation system and the constellation of Global Positioning System (GPS) satellites has become an affordable way to achieve accurate terminal guidance where the targets are not moving and the locations of the targets in the GPS coordinate system are known. GPS provides accurate position and velocity information to the inertial system, thereby bounding the inertial system's errors, while the inertial system provides short-term filtering of the guidance information between GPS updates. The GPS system errors [typically, 15 m (50 ft) spherical error probable] and the target location uncertainty limit the accuracy achievable. *See* CONTROL SYSTEMS; INERTIAL GUIDANCE SYSTEM; SATELLITE NAVIGATION SYSTEMS.

Command guidance. In command guidance, the weapon system operator or on-board sensors observe the relative location of weapon and target and direct trajectory corrections. In the simplest application, the operator may visually observe the weapon and target and remotely maneuver the weapon. A second degree of refinement is a television camera in the nose of the missile, through which the operator observes the target and guides the missile. In the most refined state, sensors on the launch aircraft track both the target and missile, and command the flight trajectory of the weapon without human intervention.

Active guidance. In this type of system, electromagnetic emissions, for example, radar, microwave, or laser, are transmitted from the weapon to the target, and the return energy reflections are measured to determine range and angle to the target. This return is continuously processed to provide control signals. Depending upon the sophistication of the weapon's seeker, the reflected energy may be processed in a manner that will produce imagery of the target to increase the likelihood of hitting a particular aim point. *See* LIDAR; RADAR.

Semiactive guidance. This resembles active guidance except that the illumination of the target is provided by a designator not located on the weapon. It may be on the launch aircraft, another aircraft, or the ground. When illumination is provided by a source other than the launch platform, the term bistatic guidance is used.

Passive guidance. This uses the natural emissions radiating from targets (such as infrared, visible, radio-frequency, or acoustic), which are usually referred to as the target's characteristic signature, to uniquely acquire a target and subsequently guide the weapon to the target.

Smart weapons. The ultimate objective of a weapon system is to be completely autonomous. An example of an autonomous system is one combining inertial and passive or active terminal guidance, and the appropriate algorithms to acquire the target after launch without operator intervention. In this case, the weapon is launched into an area where targets are known to exist, and, upon reaching the area, the weapon searches and finds its own target, homes in on it, and destroys it. The trend toward weapons that can autonomously acquire targets allows weapons to be built that have a substantial standoff capability which increases the survivability of the launch platform, improves accuracy, increases proficiency, and reduces the logistical burden.

There are two classes of smart weapons. The first class consists of those guided weapons that possess some form of terminal guidance and home in on the target. Weapon systems in this class include laser-guided bombs such as Paveway II and Paveway III, imaging systems such as the GBU-15, AGM-130, and Maverick where television or infrared imagery is used, and hotspot tracking systems such as the air-to-air AIM-9S Sidewinder missile and the high-speed antiradiation missile (HARM).

Laser-guided bombs. The Paveway III (GBU-24, GBU-27) low-level laser-guided bomb can alter its course from midpoint to impact by homing in on the reflections off the target from a laser illuminator. The aircraft using the weapon flies low enough to avoid falling victim to air-defense weapons. The weapon can be delivered in any of three modes—level, lofted, or diving—adapting automatically to the mode employed. Target illumination is either continuous or delayed until after the weapon is well along in its flight path, and provided by the releasing aircraft, another aircraft, or supporting ground forces.

As a result of the buried command-and-control bunkers encountered during the Desert Storm conflict, a 4700-lb (2130-kg) variant of the 2000-lb (900-kg) GBU-24 was produced in 28 days. Coincidentally, the GBU-28 was designed to defeat just such buried command-and-control sites, and successfully accomplished this, ending the Desert Storm conflict one day after being deployed in theater and used effectively only one time.

Imaging systems. The Maverick, AGM-65D, an air-to-surface missile, uses infrared imagery to develop the terminal guidance information, and must be locked on to the target prior to launch. Subsequent guidance is completely autonomous until target impact.

The GBU-15 air-to-surface guided glide weapon is similar in some respects to the Maverick and uses either television (visible) or imaging infrared guidance and a data link. Information supplied by the data link

is used to acquire the target and perform midcourse maneuvers. Once the target is acquired, the missile's seeker can be locked on and the aimpoint selected. Terminal guidance can then be performed on-board the weapon automatically. A longer-range version of the GBU-15 is the AGM-130, where a 600-lb (270-kg) rocket motor has been added to the GBU-15 airframe to increase standoff. Also, as a consequence of experience gained during Desert Storm, the AGM-130 is fitted with a GPS system to reduce the workload of the weapons officer during the midcourse phase of the weapon's flight.

Hot-spot tracking systems. In the case of the AIM-9S Sidewinder air-to-air missile, lock-on to the target must occur before launch. Terminal guidance is performed on-board by the missile's electronics and seeker following launch.

Brilliant weapons. The second class of smart weapons includes those that autonomously acquire the target after launch, and are usually termed lock-on-after-launch, fire-and-forget, or brilliant weapons. Among such weapons are the high-speed antiradiation missile, the advanced medium-range air-to-air missile (AMRAAM), and developmental weapons like the Autonomous Guided Conventional Weapon (AGCW), the Joint Direct Attack Munition (JDAM), the Joint Standoff Weapon (JSOW), and the Joint Air-to-Surface Standoff Missile (JASSM) for air-to-surface application.

The AGM-88 high-speed antiradiation missile has the mission of suppressing air defense threats. It can be employed in a prebriefed, self-protect, or target-of-opportunity mode. After launch, using a selfcontained, broadband, radar-emission homing seeker and a sophisticated signal processor, the missile autonomously searches for and acquires the target, and terminally guides itself to strike the radiating source.

The AGM-120 advanced medium-range air-to-air missile operates in a similar manner against aircraft targets beyond visual range. Its guidance consists of an active radar seeker in conjunction with an inertial reference unit, allowing the missile to be almost independent of the aircraft's fire control system. Once the missile closes in on the target, its active radar seeker autonomously locks on and guides it to intercept. This allows the pilot to simultaneously aim and fire several missiles at multiple targets, and then disengage from the attack and turn away before they hit.

The reduction of nuclear arms in the arsenals of the United States and Russia has resulted in the conversion of a number of cruise missiles from their nuclear role to that of conventional weapons. The Conventional Air Launched Cruise Missile (CALCM), AGM-86C, was developed to provide B52H bombers with a precision strike capability just in time for Desert Storm. The cruise missile was modified to use a GPS-aided inertial navigation system and carry a 3000-lb (1350-kg) blast fragmentation warhead.

Advanced weapons. Developmental weapons are all categorized as brilliant weapons and do not rely on a human in the loop, provide pinpoint accuracy, and use warheads tailored to maximize the probability of defeating the target with minimal collateral damage.

An improvement to the Paveway class of laser-guided weapons is the Autonomous Guided Conventional Weapon, which is designed to use infrared imagery to autonomously acquire high-value targets. The weapon is first loaded with navigation algorithms, target data, and the location of the intended launch point. After launch, the weapon is designed to autonomously detect, acquire, and attack specific targets by using an imaging infrared seeker, inertial

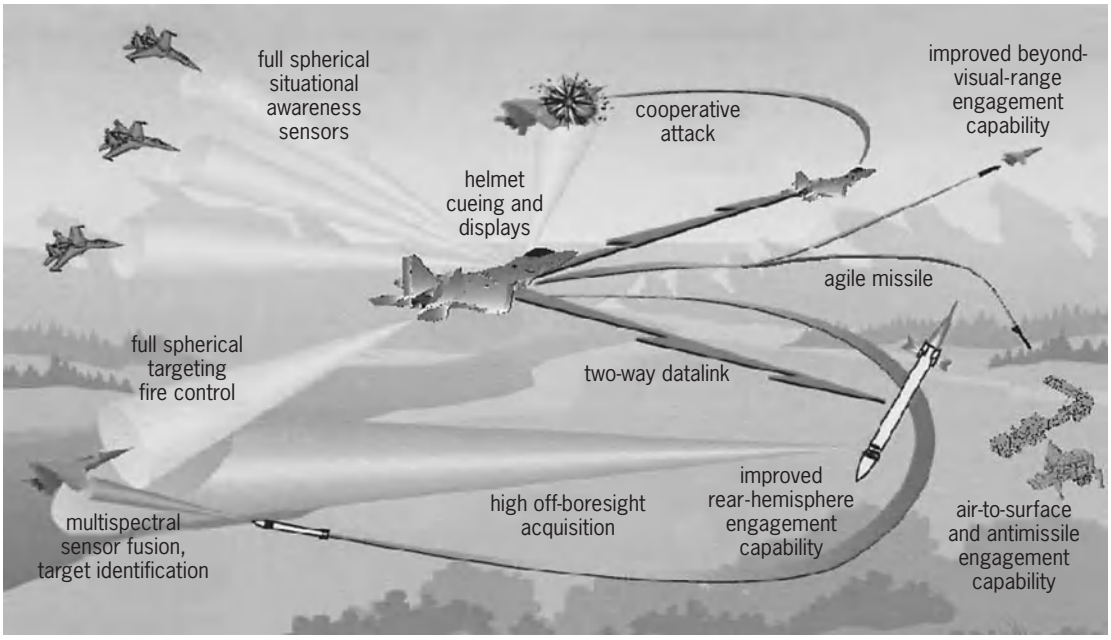
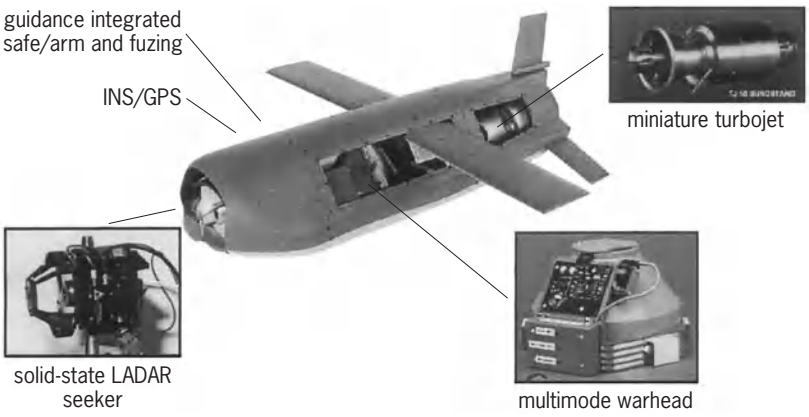


Fig. 1. Air Superiority Missile Concept.



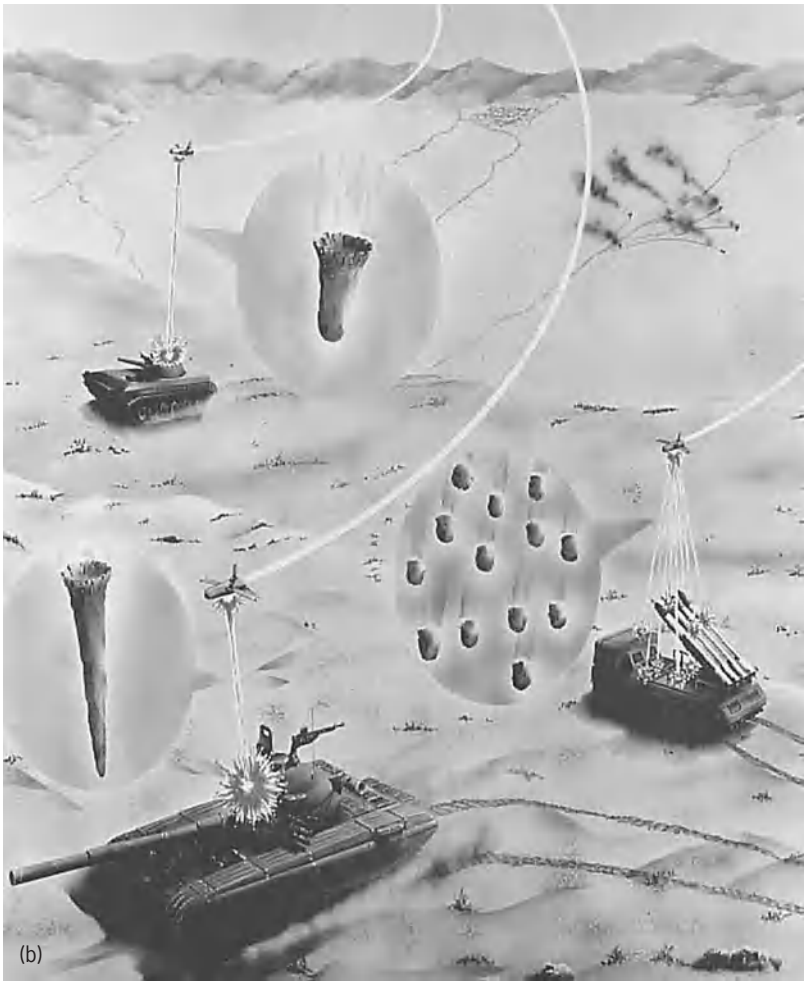


Fig. 2. Low-Cost Autonomous Attack System (LOCAAS; formerly Low-Cost Antiarmor Submunition). (*a*) Autonomous submunition, with length of 30 in. (0.75 m), wingspan of 40 in. (1.0 m), and weight of 90–100 lb (40–45 kg). (*b*) Various warhead modes.

navigation, and signal processing for target acquisition and aimpoint selection. Although this weapon appears to resemble a Paveway III, it is truly autonomous and does not require a human in the loop.

The Joint Direct Attack Munition is an autonomous precision strike weapon with an all-weather capability and consists of a conventional MK-84 general-purpose bomb and an inexpensive tail kit. The tail kit uses a GPS-aided inertial navigation system to derive guidance information, and controls the weapon's trajectory until target impact.

For air targets, the AIM-9X missile is being developed to replace the AIM-9S. It has increased maneuverability and a high degree of countermeasure resistance since it makes use of an imaging seeker. This missile is a stepping stone to the Air Superiority Missile Concept (**Fig. 1**), which will incorporate an advanced seeker and be extremely agile. Moreover, it will be capable of engaging targets anywhere in the vicinity of the launching aircraft and will rely on a helmet-mounted sight, or cueing system, in which the weapon or other sensors are controlled to automatically look in the same direction that the pilot looks.

Weapon systems for defeating mobile targets are evolving around the developmental Low-Cost Autonomous Attack System (LOCAAS, formerly the Low-Cost Anti Armor Submunition). This is an autonomous submunition (one of several similar units which are packaged together to form the munition or weapon) with a laser radar (LADAR) seeker capable of autonomously searching for targets, acquiring the target, performing target classification, and guiding the weapon to the target. The LADAR seeker develops an image of the target using high-resolution range measurements of the area in front of the submunition's flight path. Classification of the target is performed to enhance the probability of kill of the warhead by matching the mode of the warhead to the type of target encountered (**Fig. 2**).

The Joint Air-to-Surface Standoff Missile, in development, is designed to attack hard, soft, relocatable, and area-type targets at long standoff ranges comparable to those typical of a cruise missile scenario, that is, well beyond area defenses, so as not to risk launch platforms and crews. *See* ARMY ARMAMENT; GUIDED MISSILE; NAVAL ARMAMENT.

Samuel Lambert

Bibliography. W. Armstrong, Jr., Replacing the Sidewinder, *Armed Forces J. Int.*, pp. 30–33, January, 1999; B. Gunston, *The Illustrated Encyclopedia of Aircraft Armament*, Orion Books, New York, 1988; E. Ulsamer, "Brilliant" weapons gather momentum, *Air Force Mag.*, pp. 74–80, August, 1988.

Air brake

A friction type of energy-conversion mechanism used to retard, stop, or hold a vehicle or other moving element. The activating force is applied by a difference in air pressure. With an air brake, a slight effort by the operator can quickly apply full braking force. *See* FRICTION.

The air brake, operated by compressed air, is used in buses; heavy-duty trucks, tractors, and trailers; and off-road equipment. The air brake is required by law on locomotives and railroad cars. The wheel-brake mechanism is usually either a drum or a disk brake. The choice of an air brake instead of a mechanical, hydraulic, or electrical brake depends partly on the

availability of an air supply and the method of brake control.

For example, in a bus in which compressed air actuates doors, air may also actuate the brakes. On railroad cars, compressed air actuates the brakes so that, in the event of a disconnect between cars, air valves can automatically apply the brakes on all cars. Air brakes can also be applied and held on mechanically, even if no compressed air is available and the vehicle is not in use. Regulations require alternate and fail-safe methods of applying air brakes.

The force with which air brakes are applied depends on the area of an air-driven diaphragm as well as on air pressure. As a result, a large activating force is developed from moderate pressure. On passenger trains, trucks, and buses, air brakes operate at about 110 psi (760 kilopascals).

System operation. In a motor vehicle, the air-brake system consists of three subsystems: the air-supply, air-delivery, and parking/emergency systems.

Air-supply system. This system includes the compressor, reservoirs, governor, pressure gage, low-pressure indicator, and safety valve. The engine-driven compressor takes in air and compresses it for use by the brakes and other air-operated components. The compressor is controlled by a governor that maintains air compression within a preselected range. The compressed air is stored in reservoirs. If the pressure in an air reservoir becomes too high, a safety valve allows air to escape. A pressure gage indicates the pressure within the system. A low-pressure indicator turns on a warning lamp or sounds a buzzer when the air pressure drops below a safe minimum for normal operation. *See* SAFETY VALVE.

Air-delivery system. This system includes a foot-operated brake valve, one or more relay valves, the quick-release valve, and the brake chambers. The system delivers compressed air from the air reservoirs to the brake chambers, while controlling the pressure of the air. The amount of braking is thereby regulated.

The brake valve meters the delivery of air pressure to the front brake chambers and the rear relay valve, relative to the distance that the brake-valve foot pedal is depressed. The rear relay valve is constantly supplied with full-system pressure from the primary reservoir. As the brake valve meters the pressure to the relay valve, the relay valve controls and delivers the air pressure to the rear brake chambers.

In the brake chambers, the air pressure is converted into a mechanical force to apply the brakes. As the pressure increases in each brake chamber, movement of the diaphragm pushrod forces the friction element against the rotating surface to provide braking. When the driver releases the brake valve, the quick-release valve and the relay valve release the compressed air from the brake chambers.

Parking/emergency system. This system includes a parking-brake control valve and spring brake chambers. These chambers contain a strong spring to mechanically apply the brakes (if the brakes are properly adjusted) when air pressure is not available. During normal vehicle operation, the spring is held compressed by system air pressure acting on a diaphragm.

For emergency stopping, the air-brake system is split into a front brake system and a rear brake system. If air pressure is lost in the front brake system, the rear brake system will continue to operate. However, the supply air will be depleted after several brake applications. Loss of air pressure in the rear brake system makes the front brake system responsible for stopping the vehicle, until the supply air is depleted.

In a semitrailer (a vehicle consisting of a tractor pulling a trailer), a tractor protection valve, or breakaway valve, automatically closes to apply the trailer brakes and preserve the tractor air supply if the trailer should break away and snap the air lines. The valve also closes if severe air leakage develops in the tractor or trailer.

Combined brake systems. A straight air-brake system, described above, may be combined with other brake systems. In an air-assisted hydraulic brake system, the hydraulic system is actuated by an air-hydraulic cylinder or power unit. The driver's effort is combined in the power unit with force from the air-brake cylinder piston or chamber diaphragm to displace fluid under pressure for actuation of the brakes. Other combined brake systems include the use of an air-powered hydraulic master cylinder actuated by an air-brake cylinder or chamber.

An antilock-braking system may also be used with air brakes to improve braking performance, reduce stopping distance, and help prevent trailer jackknifing. Thus, the vehicle can retain its directional stability and steerability even during severe braking on slippery surfaces. Another benefit is longer tire life resulting from reduced wear caused by tire skidding.

Railroad air brake. On railroad trains, a pressurized pipe runs the length of the train. Release of air pressure in the pipe reduces the pressure to the triple valve, which connects the compressor, auxiliary reservoir, and brake cylinder. The triple valve then admits compressed air from each car's auxiliary reservoir to the car's brake cylinders, which actuate the brake rods, or pushrods, to apply the brakes.

Vacuum brake. This alternate form of air brake maintains a vacuum in the actuating chamber or cylinder. By opening one side of the cylinder to the atmosphere, atmospheric pressure moves the diaphragm to apply the brakes.

Aircraft air brake. A different kind of air brake is used on aircraft. Energy of momentum is transferred to the air as heat by an air brake that consists of a flap or other device for producing drag. When needed, the braking device is extended from an aircraft wing or fuselage into the airstream. *See* AUTOMOTIVE BRAKE; BRAKE.

Donald L. Anglin

Bibliography. A. K. Baker, *Vehicle Braking*, 1986; R. Limpert, *Brake Design and Safety*, 2d ed., 1999; *Ford Truck Shop Manual*, annually; E. J. Schulz, *Diesel Equipment I*, 1982.

Air conditioning

The control of certain environmental conditions including air temperature, air motion, moisture level, radiant heat energy level, dust, various pollutants, and microorganisms.

Comfort air conditioning refers to control of spaces to promote the comfort, health, or productivity of the inhabitants. Spaces in which air is conditioned for comfort include residences, offices, institutions, sports arenas, hotels, factory work areas, and motor vehicles. Process air-conditioning systems are designed to facilitate the functioning of a production, manufacturing, or operational activity. There are many examples. Heat-producing electronic equipment in an airplane cockpit must be kept cool to function properly, while the occupants of the cockpit are maintained in comfortable conditions. A multicolor printing press requires accurate registration of the colors so the environment must be maintained at a constant relative humidity to avoid paper expansion or shrinkage, and press heat and ink mists must be removed as potential health hazards for personnel. Maintenance of conditions within surgical suites of hospitals and in "clean" or "white" rooms of manufacturing plants, where an atmosphere almost free of microorganisms and dust must be maintained, is a specialized subdivision of process air conditioning. *See* AUTOMOTIVE CLIMATE CONTROL.

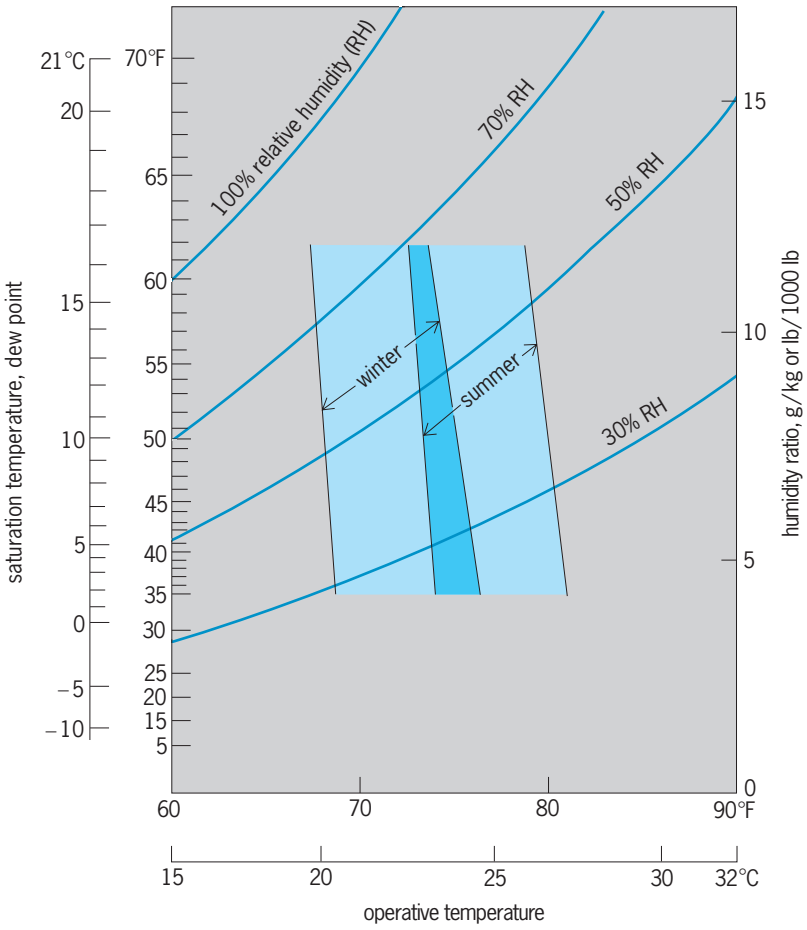


Fig. 1. Ranges of operative temperature and relative humidity acceptable to at least 80% of sedentary persons in typical summer or winter clothing and doing light activities. In the overlap area, people with summer clothing might feel slightly cool, while those in winter clothing might feel slightly warm.

Physiological principles. A comfort air-conditioning system is designed to help maintain body temperature at its normal level without undue stress and to provide an atmosphere which is healthy to breathe.

The human body produces heat at various rates, depending upon the person's weight and degree of activity. Normally more heat is generated than is required to maintain body temperature at a healthful level. Hence, proper air environment is required to permit normal cooling, even in winter. Heating the room air does not heat people; rather it affects the rate at which they lose heat and thereby affects their comfort and health. Since all people do not have the same metabolic rate or wear comparable clothing, some people in a room may feel comfortable while others do not. The acceptable ranges of operative temperatures and relative humidities shown in shaded areas for winter and summer in **Fig. 1** were developed by the American Society of Heating, Refrigerating and Air-Conditioning Engineers.

The comfort chart (Fig. 1) describes average responses to given environmental conditions. Preferences vary considerably from person to person. For instance, in office buildings women may complain of chill or drafts under conditions where men, in different attire, are comfortable. Metabolism rates vary with individuals. Whenever a building budget permits, an engineer designs an air-conditioning system that is flexible enough to allow adjustment by individual occupants of both temperature and air motion in the air-conditioned space. Conditions that would satisfy all persons within the same space have yet to be achieved. *See* PSYCHROMETRICS.

Control of body temperature is accomplished by control of the emission of energy from the body by radiation, by convection to air currents that impinge on the skin or clothing, by conduction of clothing and objects that are contacted, and by evaporation of moisture in the lungs and of sweat from the skin. Radiant emission from the body is a function of the amount of clothing worn (or blankets used) and the temperature of the surrounding air and objects, so that a body may experience a net loss by radiation to its surroundings or a net gain. Consequently, air-conditioning system operation is affected substantially by the basic construction of the building thermal envelope; cold outer walls induce more body heat loss for the same average room temperature than do warmer walls. Evaporation and convection heat losses are functions of air temperature and velocity. Evaporation is a function, in addition, of relative humidity.

When the amount and type of clothing and the temperature, velocity, and humidity of the air are such that the heat produced by the body is not dissipated at an equal rate, blood temperature begins to rise or fall and discomfort is experienced in the form of fever or chill, in proportion to the departure of body temperature from the normal 98.6°F

(37°C) [for individuals this standard may be slightly low or high]. Hence, space conditions to maintain comfort depend upon the degree of human activity in the space, the amount and type of clothing worn, and, to a certain extent, the physical condition of the occupants, because old age or sickness can impair the body's heat-producing and heat-regulating mechanisms.

The heat-dissipating factors of temperature, humidity, air motion, and radiant heat flow must be considered simultaneously. Within limits, the same amount of comfort (or, more objectively, of heat-dissipating ability) is the result of a combination of these factors in an enclosure. Conditions for constant comfort are related to the operative temperature. The perception of comfort is related to one's metabolic heat production, the transfer of this heat to the environment, and the resulting physiological adjustments and body temperature. The heat transfer is influenced by the environmental air temperature, thermal radiation, air movement, relative humidity, and the personal effects of activity and clothing. The effect of clothing is evaluated by its thermal resistance, or clo value; clo = 0.155 m^2 K/W (0.88 ft^2 h °F/Btu).

For some years, the effective temperature was used to indicate the uniform temperature of a radiantly black enclosure at 50% relative humidity in which an occupant would feel the same comfort, physiological strain, and heat exchange as in the actual environment with the same air motion. The preferred indicator is the operative temperature, the uniform temperature of a radiantly black enclosure in which an occupant would exchange the same amount of heat by radiation plus convection as in the actual nonuniform environment. Operative temperature is the average of the air and mean radiant temperatures weighted by the heat transfer coefficients for convection and radiation. At mean radiant temperatures less than 120°F (50°C) and air speeds of 80 ft/m (0.4 m/s) or less, the operative temperature is approximately the simple average of the air and the mean radiant temperature, and is equal to the adjusted dry bulb temperature.

In practice, most air-conditioning systems for offices, schools, and light-work areas are designed to maintain temperature in the range 68–75°F (20–24°C). In hot weather, in order to minimize the effects of sudden temperature differences when people leave temperature-controlled spaces to go outdoors, higher temperatures are maintained than in cold weather. Lower cold-weather temperatures are used when energy conservation is a factor. Depending on climatic characteristics, in hot weather the relative humidity is reduced to below 50%, while in cold weather it is increased above the 10–15% relative humidity that may occur when no humidification is provided. However, great care must be used in establishing relative humidity limits for each building design and occupancy so that condensation of moisture on and within building components does not create deteriorating conditions, many of which may not be detectable until performance problems occur.

Calculation of loads. Engineering of an air-conditioning system starts with selection of design conditions; air temperature and relative humidity are principal factors. Next, loads on the system are calculated. Finally, equipment is selected and sized to perform the indicated functions and to carry the estimated loads.

Design conditions are selected on the bases discussed above. Each space is analyzed separately. A cooling load will exist when the sum of heat released within the space and transmitted to the space is greater than the loss of heat from the space. A heating load occurs when the heat generated within the space is less than the loss of heat from it. Similar considerations apply to moisture.

Heat generated within the space consists of body heat, heat from all electrical appliances and lights, and heat from other sources such as cooking stoves and industrial ovens. Heat is transmitted through all parts of the space envelope, which includes walls, floor, slab on ground, ceiling, doors, and windows. Whether heat enters or leaves the space depends upon whether the outside surfaces are warmer or cooler than the inside surfaces. The rate at which heat is conducted through the building envelope is a function of the temperature difference across the envelope and the thermal resistance of the envelope (R-value). Overall R-values depend on materials of construction and their thickness along the path of heat flow, and air spaces with or without reflectances and emittances, and are evaluated for walls and roofs exposed to outdoors, and basements or slab exposed to earth. In some cases, thermal insulations may be added to increase the R-value of the envelope.

Solar heat loads are an especially important part of load calculation because they represent a large percentage of heat gain through walls, windows, and roofs, but are very difficult to estimate because solar irradiation is constantly changing. Intensity of radiation varies with the seasons [it rises to 457 Btu/h ft^2 (2881 W/m^2) in midwinter and drops to 428 Btu/h ft^2 (2698 W/m^2) in midsummer]. Intensity of solar irradiation also varies with surface orientation. For example, the half-day total for a horizontal surface at 40° north latitude on January 21 is 353 Btu/h ft^2 (2225 W/m^2) and on June 21 it is 1121 Btu/h ft^2 (7067 W/m^2), whereas for a south wall on the same dates comparable data are 815 Btu/h ft^2 (5137 W/m^2) and 311 Btu/h ft^2 (1961 W/m^2), a sharp decrease in summer. Intensity also varies with time of day and cloud cover and other atmospheric phenomena. *See* SOLAR RADIATION.

The way in which solar radiation affects the space load depends also upon whether the rays are transmitted instantly through glass or impinge on opaque walls. If through glass, the effect begins immediately but does not reach maximum intensity until the interior irradiated surfaces have warmed sufficiently to reradiate into the space, warming the air. In the case of irradiated walls and roofs, the effect is as if the outside air temperature were higher than it is. This apparent temperature is called the sol-air temperature, tables of which are available.

In calculating all these heating effects, the object is proper sizing and intelligent selection of equipment; hence, a design value is sought which will accommodate maximums. However, when dealing with climatic data, which are statistical historical summaries, record maximums are rarely used. For instance, if in a particular locality the recorded maximum outside temperature was 100°F (38°C), but 95°F (35°C) was exceeded only four times in the past 20 years, 95°F may be chosen as the design summer outdoor temperature for calculation of heat transfer through walls. In practice, engineers use tables of design winter and summer outdoor temperatures which list winter temperatures exceeded more than 99% and 97.5% of the time during the coldest winter months, and summer temperatures not exceeded 1%, 2.5%, and 5% of the warmest months. The designer will select that value which represents the conservatism required for the particular type of occupancy. If the space contains vital functions where impairment by virtue of occasional departures from design space conditions cannot be tolerated, the more severe design outdoor conditions will be selected.

In the case of solar load through glass, but even more so in the case of heat transfer through walls and roof, because outside climate conditions are so variable, there may be a considerable thermal lag. It may take hours before the effect of extreme high or low temperatures on the outside of a thick masonry wall is felt on the interior surfaces and space. In some cases the effect is never felt on the inside, but in all cases the lag exists, exerting a leveling effect on the peaks and valleys of heating and cooling demand; hence, it tends to reduce maximums and can be taken advantage of in reducing design loads.

Humidity as a load on an air-conditioning system is treated by the engineer in terms of its latent heat, that is, the heat required to condense or evaporate the moisture, approximately 1000 Btu/lb (2324 kilojoules/kg) of moisture. People at rest or at light work generate about 200 Btu/h (586 W). Steaming from kitchen activities and moisture generated as a product of combustion of gas flames, or from all drying processes, must be calculated. As with heat, moisture travels through the space envelope, and its rate of transfer is calculated as a function of the difference in vapor pressure across the space envelope and the permeance of the envelope construction.

Years ago engineers and scientists contributed their experiences with moisture migration rates into walls of ordinary houses in ordinary climates; this led to defining a limiting rate of moisture flow that was designated a perm (from permeance). Permeance is the rate of moisture flow through a material or construction as used, whereas permeability is a property of a material that expresses moisture flow rate for a unit thickness. A unit perm was considered the highest rate of water vapor flow into a wall (or roof) that would differentiate the probability of moisture problems from the probability of no moisture problems. As stated above, excessive moisture adversely affects the thermal performance of constructions, and thereby the condition of air for comfort and health of people; in addition, it affects energy waste that could be avoided.

While moisture migrates by diffusion, building designs and the air-conditioning designs must also consider the effects of air infiltration and exfiltration, especially from wind in a particular location.

Another load-reducing factor to be calculated is the diversity among the various spaces within a building or building complex served by a single system. Spaces with east-facing walls experience maximum solar loads when west-facing walls have no solar load. In cold weather, rooms facing south may experience a net heat gain due to a preponderant solar load while north-facing rooms require heat. An interior space, separated from adjoining spaces by partitions, floor, and ceiling across which there is no temperature gradient, experiences only a net heat gain, typically from people and lights. Given a system that can transfer this heat to other spaces requiring heat, the net heating load may be zero, even on cold winter days.

Air-conditioning systems. A complete air-conditioning system is capable of adding and removing heat and moisture and of filtering airborne substitutes, such as dust and odorants, from the space or spaces it serves. Systems that heat, humidify, and filter only, for control of comfort in winter, are called winter air-conditioning systems; those that cool, dehumidify, and filter only are called summer air-conditioning systems, provided they are fitted with proper controls to maintain design levels of temperature, relative humidity, and air purity.

Design conditions may be maintained by multiple independent subsystems tied together by a single control system. Such arrangements, called split systems, might consist, for example, of hot-water baseboard heating convectors around a perimeter wall to offset window and wall heat losses when required, plus a central cold-air distribution system to pick up heat and moisture gains as required and to provide filtration for dust and odor. *See* HOT-WATER HEATING SYSTEM.

Air-conditioning systems are either unitary or built-up. The window or through-the-wall air conditioner (**Fig. 2**) is an example of a unitary summer air-conditioning system; the entire system is housed in a single package which contains heat removal, dehumidification, and filtration capabilities. When an electric heater is built into it with suitable controls, it functions as a year-round air-conditioning system. Unitary air conditioners are manufactured in capacities as high as 100 tons (1 ton of air conditioning equals 12,000 Btu/h or 76,000 W/m^2) and are designed to be mounted conveniently on roofs, on the ground, or other convenient location, where they can be connected by ductwork to the conditioned space.

Built-up or field-erected systems are composed of factory-built subassemblies interconnected by means such as piping, wiring, and ducting during final assembly on the building site. Their capacities range up to thousands of tons of refrigeration and millions of Btu per hour of heating. Most large buildings are so conditioned.

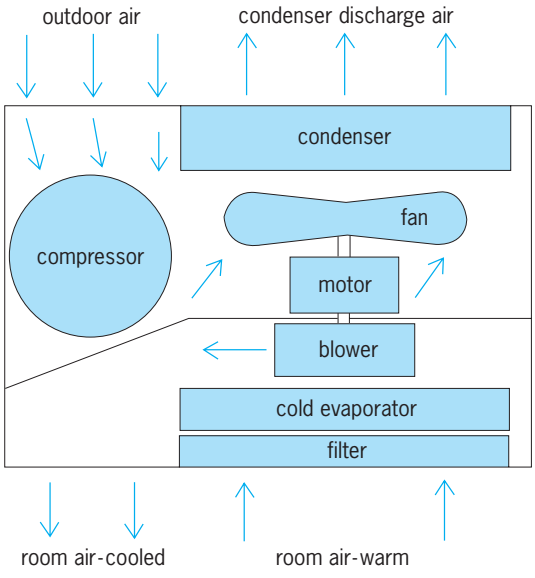


Fig. 2. Schematic of room air conditioner. (*After American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc., Guide and Data Book, 1967*)

Another important and somewhat parallel distinction can be made between incremental and central systems. An incremental system serves a single space; each space to be conditioned has its own, self-contained heating-cooling-dehumidifying-filtering unit. Central systems serve many or all of the conditioned spaces in a building. They range from small, unitary packaged systems serving single-family residences to large, built-up or field-erected systems serving large buildings. *See* CENTRAL HEATING AND COOLING.

When many buildings, each with its own air-conditioning system which is complete except for a refrigeration and a heating source, are tied to a central plant that distributes chilled water and hot water or steam, the interconnection is referred to as a district heating and cooling system. This system is especially useful for campuses, medical complexes, and office complexes under a single management. *See* DISTRICT HEATING.

Conditioning of spaces. Air temperature in a space can be controlled by radiant panels in floor, walls, or ceiling to emit or absorb energy, depending on panel temperature. Such is the radiant panel system. However, to control humidity and air purity, and in most systems for controlling air temperature, a portion of the air in the space is withdrawn, processed, and returned to the space to mix with the remaining air. In the language of the engineer, a portion of the room air is returned (to an air-handling unit) and, after being conditioned, is supplied to the space. A portion of the return air is spilled (exhausted to the outdoors) while an equal quantity (of outdoor air) is brought into the system and mixed with the remaining return air before entering the air handler. *See* PANEL HEATING AND COOLING.

Typically, the air-handling unit contains a filter, a cooling coil, a heating coil, and a fan in a suitable casing (**Fig. 3**). The filter removes dust from both return and outside air. The cooling coil, either containing recirculating chilled water or boiling refrigerant, lowers air temperature sufficiently to dehumidify it to the required degree. The heating coil, in winter, serves a straightforward heating function, but when the cooling coil is functioning, it serves to raise the temperature of the dehumidified air (to reheat it) to the exact temperature required to perform its cooling function. The air handler may perform its function, in microcosm, in room units in each space as part of a self-contained, unitary air conditioner, or it may be a huge unit handling return air from an entire building. *See* AIR COOLING; AIR FILTER; HUMIDITY CONTROL.

There are three principal types of central air-conditioning systems: all-air, all-water, and air-processed in a central air-handling apparatus. In one type of all-air system, called dual-duct, warm air and chilled air are supplied to a blending or mixing unit in each space. In a single-duct all-air system, air is supplied at a temperature for the space requiring the coldest air, then reheated by steam or electric or hot-water coils in each space.

In the all-water system the principal thermal load is carried by chilled and hot water generated in a central facility and piped to coils in each space; room air then passes over the coils. A small, central air system supplements the all-water system to provide dehumidification and air filtration. The radiant panel system, previously described, may also be in the form of an all-water system.

In an air-water system, both treated air and hot or chilled water are supplied to units in each space. In winter, hot water is supplied, accompanied by cooled, dehumidified air. In summer, chilled water is supplied with warmer (but dehumidified) air. One supply reheats the other.

All-air systems preceded the others. Primary motivation for all-water and air-water systems is their capacity for carrying large quantities of heat energy in small pipes, rather than in larger air ducts. To accomplish the same purpose, big-building all-air

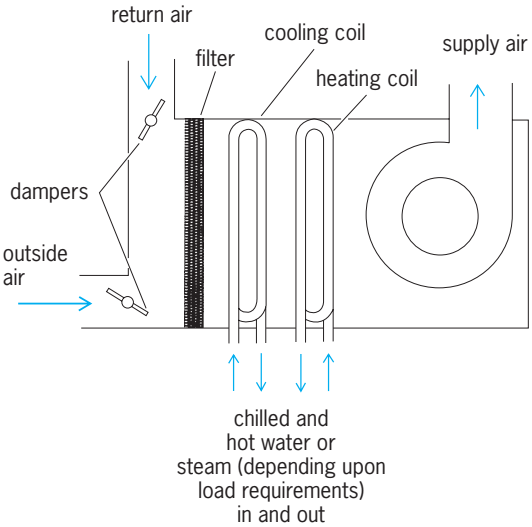


Fig. 3. Schematic of central air-handling unit.

systems are designed for high velocities and pressures, requiring much smaller ducts.

Richard L. Koral; E. C. Shuman

Bibliography. B. C. Langley, *Fundamentals of Air Conditioning Systems*, 2d ed., 2000; C. McPhee (ed.), *ASHRAE Pocket Guide for Air-Conditioning, Heating, Ventilation, and Refrigeration*, 2005; F. C. McQuiston, J. D. Parker, and J. D. Spitler, *Heating, Ventilating and Air Conditioning Analysis and Design*, 2004.

Air cooling

Lowering of air temperature for comfort, process control, or food preservation. Air and water vapor occur together in the atmosphere. The mixture is commonly cooled by direct convective heat transfer of its internal energy (sensible heat) to a surface or medium at lower temperature. In the most compact arrangement, transfer is through a finned (extended surface) coil, metallic and thin, inside of which is circulating either chilled water, antifreeze solution, brine, or boiling refrigerant. The fluid acts as the heat receiver. Heat transfer can also be directly to a wetted surface, such as waterdroplets in an air washer or a wet pad in an evaporative cooler. *See* COMFORT TEMPERATURES; HEAT TRANSFER.

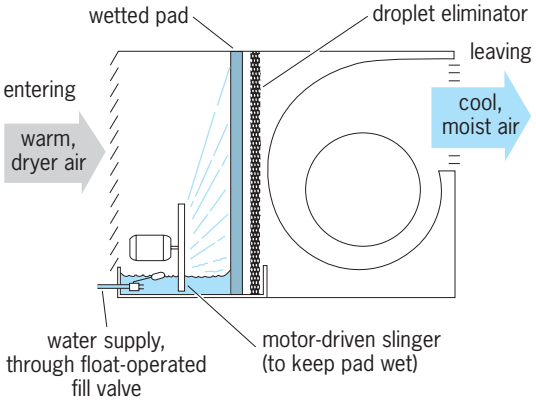


Fig. 1. Schematic view of simple evaporative air cooler.

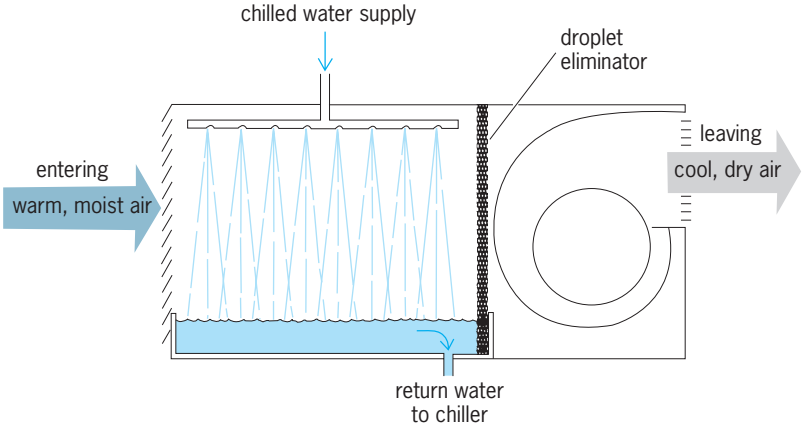


Fig. 2. Schematic of air washer.

Evaporative cooling. For evaporative cooling, nonsaturated air is mixed with water. Some of the sensible heat transfers from the air to the evaporating water. The heat then returns to the airstream as latent heat of water vapor. The exchange is thermally isolated (adiabatic) and continues until the air is saturated and air and water temperatures are equal. With suitable apparatus, air temperature approaches within a few degrees of the theoretical limit, the wet-bulb temperature. Evaporative cooling is frequently carried out by blowing relatively dry air through a wet mat (**Fig. 1**). The technique is employed for air cooling of machines where higher humidities can be tolerated; for cooling of industrial areas where high humidities are required, as in textile mills; and for comfort cooling in hot, dry climates, where partial saturation results in cool air at relatively low humidity. *See* HUMIDITY.

Air washer. In the evaporative cooler the air is constantly changed and the water is recirculated, except for that portion which has evaporated and which must be made up. Water temperature remains at the adiabatic saturation (wet-bulb) temperature. If water temperature is controlled, as by refrigeration, the leaving air temperature can be controlled within wide limits. Entering warm, moist air can be cooled below its dew point so that, although it leaves close to saturation, it leaves with less moisture per unit of air than when it entered. An apparatus to accomplish this is called an air washer (**Fig. 2**). It is used in many industrial and comfort air-conditioning systems, and performs the added functions of cleansing the airstream of dust and of gases that dissolve in water, and in winter, through the addition of heat to the water, of warming and humidifying the air.

Air-cooling coils. The most important form of air cooling is by finned coils, inside of which circulates a cold fluid or cold, boiling refrigerant (**Fig. 3**). The latter is called a direct-expansion (DX) coil. In most applications the finned surfaces become wet as condensation occurs simultaneously with sensible cooling. Usually, the required amount of dehumidification determines the temperature at which the surface is maintained and, where this results in air that is colder than required, the air is reheated to the proper temperature. Droplets of condensate are entrained in the airstream, removed by a suitable filter (eliminator), collected in a drain pan, and wasted.

In the majority of cases, where chilled water or boiling halocarbon refrigerants are used, aluminum fins on copper coils are employed. Chief advantages of finned coils for air cooling are (1) complete separation of cooling fluid from airstream, (2) high velocity of airstream limited only by the need to separate condensate that is entrained in the airstream, (3) adaptability of coil configuration to requirements of different apparatus, and (4) compact heat-exchange surface.

Defrosting. Wherever air cooling and dehumidification occur simultaneously through finned coils, the coil surface must be maintained above 32°F (0°C) to prevent accumulation of ice on the coil. For this reason, about 35°F (1.7°C) is the lowest-temperature

air that can be provided by coils (or air washers) without ice accumulation. In cold rooms, where air temperature is maintained below 32°F (0°C), provision is made to deice the cooling coils. Ice buildup is sensed automatically; the flow of cold refrigerant to the coil is stopped and replaced, briefly, by a hot fluid which melts the accumulated frost. In direct-expansion coils, defrosting is easily accomplished by bypassing hot refrigerant gas from the compressor directly to the coil until defrosting is complete.

Cooling coil sizing. Transfer of heat from warm air to cold fluid through coils encounters three resistances: air film, metal tube wall, and inside fluid film. Overall conductance of the coil, U, is shown in the equation below, where K_o is film conductance of the out-

$$\frac{1}{U} = \frac{1}{K_o} + r_m + \frac{R}{K_i}$$

side (air-side) surface in Btu/(h)(ft^2)(F); r_m is metal resistance in (h)(ft^2)(F)/Btu, where area is that of the outside surface; K_i is film conductance of the inside surface (water, steam, brine, or refrigerant side) in Btu/(h)(ft^2)(F); U is overall conductance of transfer surface in Btu/(h)(ft^2)(F), where area again refers to the outside surface; and R is the ratio of outside surface to inside surface.

Values of K_o are a function of air velocity and typically range from about 4 Btu/(h)(ft^2)(F) at 100 feet per minute (fpm) to 12 at 600 fpm. If condensation takes place, latent heat released by the condensate is in addition to the sensible heat transfer. Then total (sensible plus latent) K_o increases by the ratio of total to sensible heat to be transferred, provided the coil is well drained.

Values of r_m range from 0.005 to 0.030 (h)(ft^2)(F)/Btu, depending somewhat on type of metal but primarily on metal thickness.

Typical values for K_i range from 250 to 500 Btu/(h)(ft^2)(F) for boiling refrigerant. In 40°F (4.4°C) chilled water, values range from about 230 Btu/(h)(ft^2)(F) when water velocity is 1 foot per second (fps) to 1250 when water velocity is 8 fps.

Use of well water. Well water is available for air cooling in much of the world. Temperature of water from wells 30 to 60 ft (9 to 18 m) deep is approximately the average year-round air temperature in the locality of the well, although in some regions overuse of available supplies for cooling purposes and recharge of ground aquifers with higher-temperature water has raised well water temperature several degrees above the local normal. When well water is not cold enough to dehumidify air to the required extent, an economical procedure is to use it for sensible cooling only, and to pass the cool, moist air through an auxiliary process to dehumidify it. Usually, well water below 50°F (10°C) will dehumidify air sufficiently for comfort cooling. Well water at these temperatures is generally available in the northern third of the United States, with the exception of the Pacific Coast areas.

Ice as heat sink. For installations that operate only occasionally, such as some churches and meeting halls, water recirculated and cooled over ice offers an economical means for space cooling (**Fig. 4**). Cold water is pumped from an ice bunker through an extended-surface coil. In the coil the water absorbs heat from the air, which is blown across the coil. The warmed water then returns to the bunker, where its temperature is again reduced by the latent heat of fusion (144 Btu/lb) to 32°F (0°C). Although initial cost of such an installation is low, operating costs are usually high.

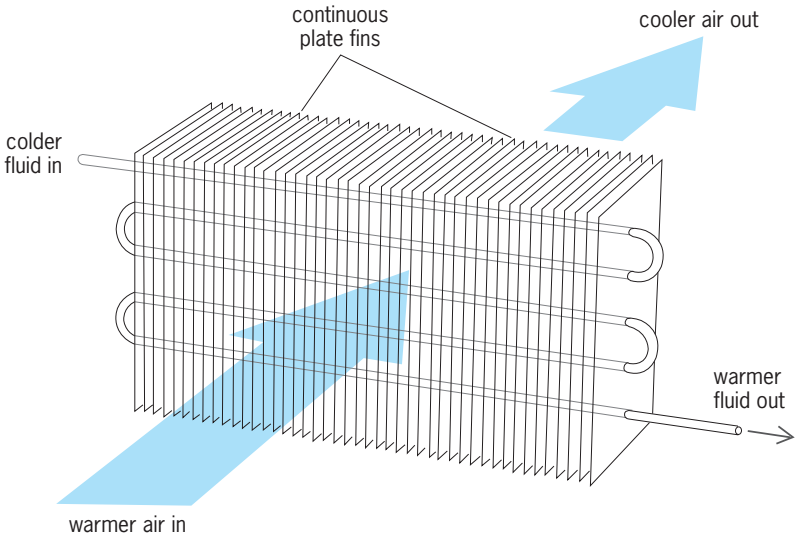


Fig. 3. Typical extended-surface air-cooling coil.

Refrigeration heat sink. Where electric power is readily available, the cooling function of the ice, as described above, is performed by a mechanical refrigerator. If the building complex includes a steam plant, a steam-jet vacuum pump can be used to cause the water to evaporate, thereby lowering its temperature by the latent heat of evaporation (about 1060 Btu/lb, depending on temperature and pressure). High-pressure steam, in passing through a primary ejector, aspirates water vapor from the evaporator, thereby maintaining the required low pressure that causes the water to evaporate and thus to cool itself (**Fig. 5**). *See* REFRIGERATION.

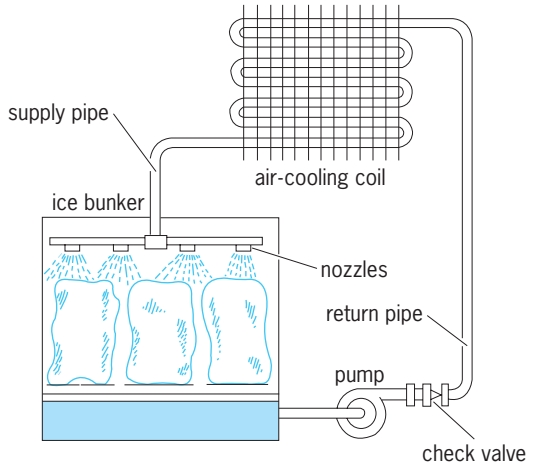


Fig. 4. Air cooling by circulating ice-cooled water.

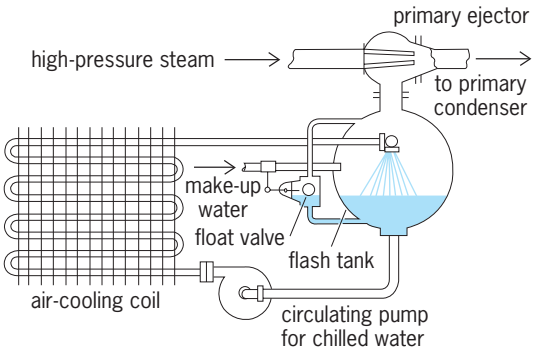


Fig. 5. Air cooling by circulating water that is cooled, in turn, by evaporation in flash tank.

Where electric power is costly compared to low-temperature heat, such as by gas, absorption refrigeration may be used. Two fluids are used: an absorbent and a refrigerant. The absorbent is chosen for its affinity for the refrigerant when in vapor form; for example, water is used as the absorber with ammonia as the refrigerant. Concentrated ammonia water is pumped to a high pressure and then heated to release the ammonia. The high-pressure ammonia then passes through a condenser, an expansion valve, and an evaporator, as in a mechanical system, and is reabsorbed by the water. The cycle cools air circulated over the evaporator. *See* AIR CONDITIONING.

Richard L. Koral

Bibliography. Air Conditioning and Refrigeration Institute, *Refrigeration and Air Conditioning*, 3d ed., 1998; A. Althouse et al., *Modern Refrigeration and Air Conditioning,* 18th ed., 2003; American Society of Heating and Air Conditioning Engineers, *Guide and Data Book,* annual; L. Jeffus, *Refrigeration and Air Conditioning: An Introduction to HVAC*, 4th ed., 2003.

Air-cushion vehicle

A transportation vehicle, also called a hovercraft, that rides slightly above the Earth's surface on a cushion of air. The air is continuously forced under the vehicle by a fan, generating the cushion that greatly reduces friction between the moving vehicle and the surface. The air is usually delivered through ducts and injected at the periphery of the vehicle in a downward and inward direction (**Fig. 1**). The design of the vehicle's underside, combined with seals or skirts attached below the hull around the perimeter, restrains the air, creating the cushion. Because the vehicle is not in contact with the surface, it has six dimensions of motion. *See* DEGREE OF FREEDOM (MECHANICS); DUCTED FAN.

Generally, an air-cushion vehicle is an amphibious aerostatic craft capable of slight vertical lift regardless of forward speed. This type of air-cushion vehicle can operate equally well over ice, water, marsh, or relatively level land. A variation is the surface-effect ship (SES), which has rigid side hulls that ride in the water like a catamaran, containing the air cushion and reducing air loss. Flexible air seals across the forward and aft ends of the surface-effect ship reduce drag, while the side hulls and the use of screw propellers or waterjets increase drag. These also limit operation to water and make the surface-effect ship nonamphibious. *See* AERODYNAMIC FORCE.

Design and construction. The air-cushion vehicle is basically a load-carrying hull fitted with a seal or skirt system to contain the air cushion, a lift system, a propulsion system, and a control system. One or more engines supply power to provide lift and propulsion. Some air-cushion vehicles have gas turbine engines which drive free or ducted air propellers through gearboxes. Depending on size, other air-cushion vehicles have diesel engines, automotive engines, or combination power sources such as a lawnmower engine for lift and a snowmobile engine for propulsion. *See* GAS TURBINE.

Many air-cushion vehicles have a length-to-beam ratio of approximately 2:1. Speed generally is in the 30–100-mi/h (50–160-km/h) range. Cushion pressure usually is in the range of 30–50 lb/ft^2 (1.50–2.40 kilopascals), although some air-cushion vehicles have used a pressure of approximately 100 lb/ft^2 (4.80 kPa). Cushion depth varies from about 8 in. (20 cm) on recreational vehicles carrying two or three passengers to 9 ft (3 m) on large vehicles capable of carrying hundreds of passengers. A greater cushion depth usually allows an increase in the maximum wave height over which the vehicle can operate.

Applications. Commercial uses of air-cushion vehicles include transportation, supply, and pipeline and cable inspection and maintenance. In scheduled service, large air-cushion vehicles have been used to ferry cars and passengers across the English Channel (**Fig. 2**). These craft displace 336 short tons (305 metric tons), and have a maximum speed of 65 knots (120 km/h), with propulsion provided by swivel-mounted air propellers. Military missions of aircushion vehicles, which can be lightly armed, include patrolling, and transporting troops and equipment from ship to shore and across the beach. Small air-cushion vehicles are available in single-seat, two-seat, and four-seat versions. Most are personal sport craft for recreational use on land or over calm water. Air-cushion vehicles of various sizes and maximum speeds are employed by government agencies in law enforcement, fire fighting, and disaster relief. *See* NAVAL SURFACE SHIP.

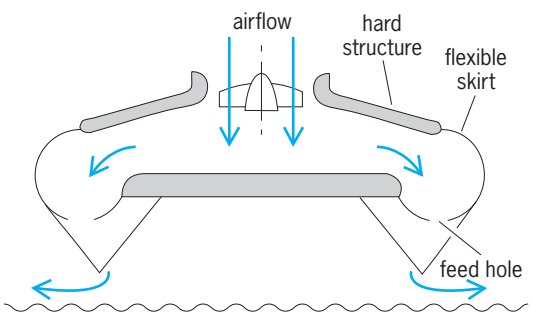


Fig. 1. Basic construction of an aerostatic air-cushion vehicle.

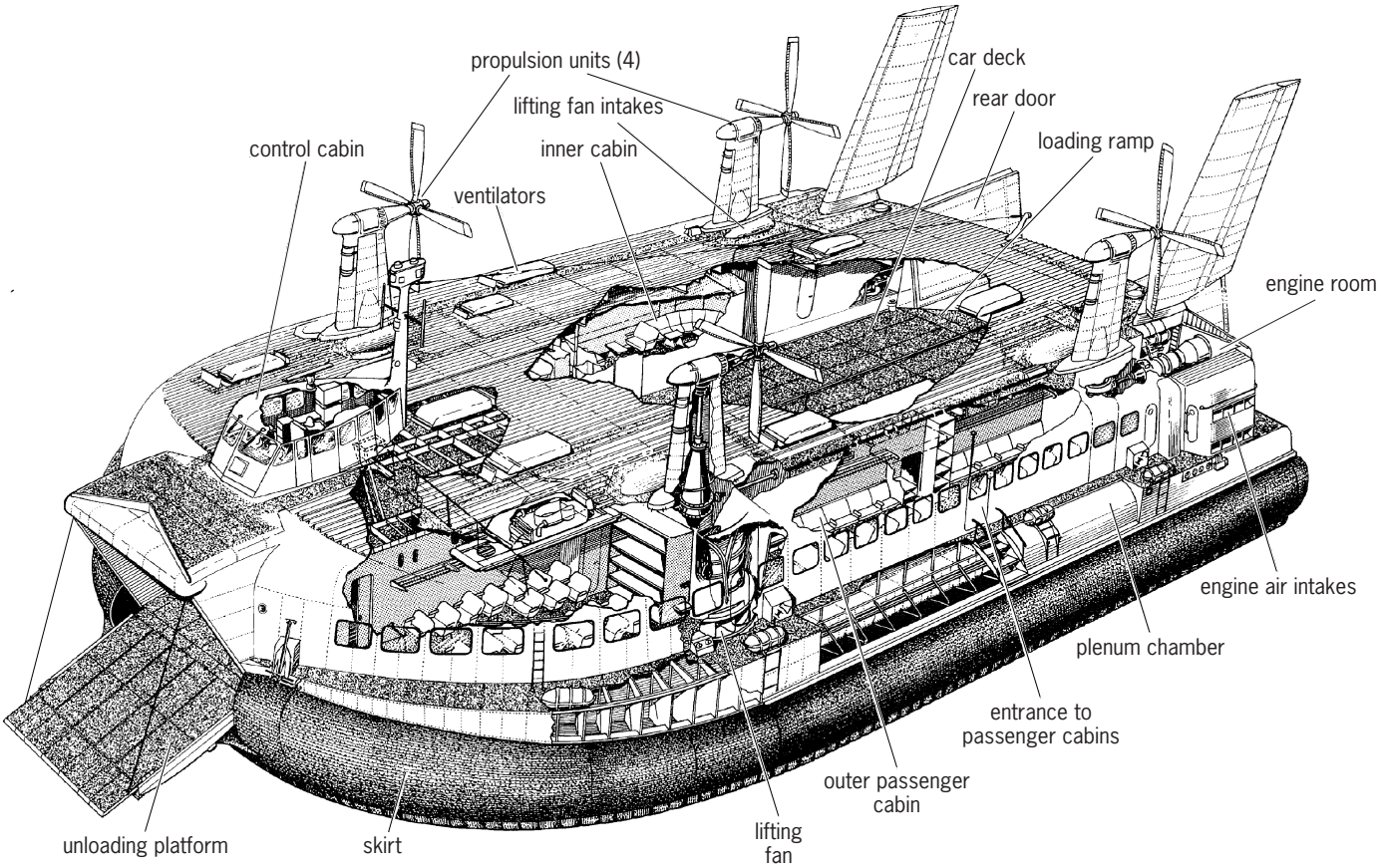


Fig. 2. Air-cushion vehicle used to transport cars and passengers across the English Channel. (*British Hovercraft Corp.*)

Wing-in-ground-effect craft. Another type of vehicle relying on a cushion of air for support is the wing-in-ground-effect (WIG) craft, also called the wing-in-surface-effect (WISE) craft. Similar in design to an aircraft, the ground-effect craft is an aerodynamic vehicle that generally travels only over water and requires forward speed to provide lift. It is designed to take advantage of two characteristics of an airfoil moving through air while close to the surface: (1) Air pressure builds up under the wing, producing greater lift than if the wing were in conventional flight; and (2) the induced drag of the wing is less than in conventional flight because wingtip vortices are weaker and downwash angle decreases. The result is an increase in the ratio of lift to drag when a craft is flying in ground effect. *See* AERODYNAMIC FORCE; AERODYNAMICS; AIRCRAFT; AIRCRAFT DESIGN; AIRFOIL; AIRPLANE.

A ground-effect vehicle can achieve greater fuel economy and range than an aircraft in conventional flight, since flying at a height of about 10% of wingspan reduces induced drag by about 50%. However, the ceiling and maneuverability of the ground-effect vehicle are limited, although some vehicles can fly out of ground effect but only with great loss in efficiency. Cruising speed of most ground-effect vehicles has been in the range of 70–170 mi/h (110–270 km/h), with some models cruising at 310–340 mi/h (500–550 km/h).

Beginning in the 1960s, the Russians built and flew a variety of wing-in-ground-effect craft, called ekranoplans, which were developed primarily for military applications. One ekranoplan displaced up to 595 short tons (540 metric tons), making it probably the largest vehicle to ever lift itself from the water and into the air under its own power. The Russians and other designers of ground-effect craft have used ducted fans, deflected exhaust, and separate turbofan engines to blow air or exhaust under the wings to augment takeoff power. *See* JET PROPULSION; TURBOFAN.

Several types and sizes of wing-in-ground-effect vehicles are available (**Fig. 3**). Some designs utilize

Fig. 3. Wing-in-ground-effect craft used over water as a commuter vehicle. (*FlareCraft Corp.*)

composite construction and are comparable in performance with hydrofoils. The first wing-in-ground-effect vehicle in commercial service served as a water taxi. *See* HYDROFOIL CRAFT. Donald L. Anglin

Bibliography. *Jane's High Speed Marine Transportation*, Jane's Information Group Ltd., London, 1999; Y. Liang and A. Biaut, *Theory and Design of Air Cushion Craft*, Wiley, New York, 1999; Special edition: Modern ships and craft, *Nav. Eng. J.*, 97(2):1–336, February 1985.

Air filter

A component of most systems in which air is used for industrial processes, for ventilation, or for comfort air conditioning. The function of an air filter is to reduce the concentration of solid particles in the airstream to a level that can be tolerated by the process or space occupancy purpose. Degrees of cleanliness required and economics of the situation (life cycle costs) influence the selection of equipment. *See* AIR; AIR CONDITIONING; VENTILATION.

Solid particles in the airstream range in size from 0.01 micrometer (the smallest particle visible to the naked eye is estimated to be 20 micrometers in diameter) up to things that can be caught by ordinary fly screens, such as lint, feathers, and insects. The particles generally include soot, ash, soil, lint, and smoke, but may include almost any organic or inorganic material, even bacteria and mold spores. This wide variety of airborne contaminants, added to the diversity of systems in which air filters are used, makes it impossible to have one type that is best for all applications.

Three basic types of air filters are in common use today: viscous impingement, dry, and electronic. The principles employed by these filters in removing airborne solids are viscous impingement, interception, impaction, diffusion, and electrostatic precipitation. Some filters utilize only one of these principles; others employ combinations. A fourth method, inertial separation, is finding increasing use as a result of the construction boom throughout most of the Middle East.

Viscous impingement filters. The viscous impingement filter is made up of a relatively loosely arranged medium, usually consisting of spun glass fibers, metal screens, or layers of crimped expanded metal. The surfaces of the medium are coated with a tacky oil, generally referred to as an adhesive. The arrangement of the filter medium is such that the airstream is forced to change direction frequently as it passes through the filter. Solid particles, because of their momentum, are thrown against, and adhere to, the viscous coated surfaces. Larger airborne particles, having greater mass, are filtered in this manner, whereas small particles tend to follow the path of the airstream and escape entrapment.

Operating characteristics of viscous impingement filters include media velocities ranging from 300 to 500 ft/min (1.5–3 m/s), with resistance to airflow about 0.10 in. (2.5 mm) of water gage, or w.g., when clean and 0.50 in. (13 mm) w.g. when dirty. A glass-fiber filter is thrown away when the final resistance is reached; the metal type is washed, dried, recoated with adhesive, and reused. Automatic types, employing rolls of glass-fiber material in blanket form, or an endless belt of metal plates which pass through an oil-filled reservoir to remove the accumulated dirt, are often chosen to minimize maintenance labor.

Dry filters. Dry-air filters are the broadest category of air filters in terms of the variety of designs, sizes, and shapes in which they are manufactured. The most common filter medium is glass fiber. Other materials used are cellulose paper, cotton, and polyurethane and other synthetics. Glass fiber is used extensively because of its relatively low cost and the unique ability to control the diameter of the fiber in manufacture (in general, the finer the fiber diameter, the higher will be the air-cleaning efficiency of the medium). *See* MANUFACTURED FIBER; NATURAL FIBER.

Dry filters employ the principles of interception, in which particles too large to pass through the filter openings are literally strained from the airstream; impaction, in which particles strike and stick to the surfaces of the glass fibers because of natural adhesive forces, even though the fibers are not coated with a filter adhesive; and diffusion, in which molecules of air moving in a random pattern collide with very fine particles of airborne solids, causing the particles to have random movement as they enter the filter media. It is this random movement which enhances the likelihood of the particles coming in contact with the fibers of filter media as the air passes through the filter. Through the process of diffusion a filter is able to separate from the airstream particles much smaller than the openings in the medium itself. *See* MOLECULAR ADHESION.

Most dry filters are of the extended surface type, with the ratio of filtered area to face area (normal to the direction of air travel) varying from 7.5:1 to as much as 50:1. The higher the filter efficiency, the greater the ratio of areas, in order to hold down the air friction loss. Clean resistance through dry filters can be as low as 0.10 in. (2.5 mm) w.g. or, for very-high-efficiency filters, as much as 1 in. (25 mm) w.g. These filters are generally allowed to increase 0.5–1 in. (13–25 mm) in pressure drop before being changed. Face velocities from 400 ft/min (2 m/s) to 600 ft/min (3 m/s) are common for this class of filter, although newer energy-conscious design favors the lower portion of this range.

Electronic air cleaners. Limited primarily to applications requiring high air-cleaning efficiency, these devices operate on the principle of passing the airstream through an ionization field where a 12,000-V potential imposes a positive charge on all airborne particles. The ionized particles are then passed between aluminum plates, alternately grounded and connected to a 6000-V source, and are precipitated onto the grounded plates.

The original design of the electronic air cleaner utilizes a water-soluble adhesive coating on the plates, which holds the dirt deposits until the plates require cleaning. The filter is then deenergized, and the dirt and adhesive film are washed off the plates. Fresh adhesive is applied before the power is turned on again. Other versions of electronic air cleaners are designed so that the plates serve as agglomerators; the agglomerates of smaller particles are allowed to slough off the plates and to be trapped by viscous impingement or dry-type filters downstream of the electronic unit.

Designs of most electronic air cleaners are based on 500 ft/min (2.5 m/s) face velocity, with pressure losses at 0.20 in. (5 mm) w.g. for the washable type and up to 1 in. (25 mm) w.g. for the agglomerator type using dry-type after-filters. Power consumption is low, despite the 12,000-volt ionizer potential, because current flow is measured in milliamperes. *See* ELECTRIC CHARGE.

Inertial separators. These are devices for utilizing the principle of momentum to separate larger particulate, primarily that above 10 μm in diameter, from a moving airstream. The larger particles tend to keep going in the same direction in which they entered the device while the light-density air changes direction. Commonly offered in manifolded multiples of V-shaped pockets with louvered sides or in small-diameter cyclones, the inertial separators are used most frequently in hot, arid, desertlike regions where winds generate significant airborne dust and sand particles. A secondary bleed-air fan is used to discharge the separated particulate to the outside of the building.

Testing and rating. ASHRAE Test Standard 52 is becoming universally accepted as the optimum procedure for the testing and rating of all types of air filters and for comparing the performance of the products of competing manufacturers. An arrestance test, which measures the ability of a filter to remove the weight component of airborne particulate, is used for measuring the effectiveness of inertial separators, all viscous impingement filters, and some of the less effective dry filters. An efficiency test (dust spot efficiency using atmospheric air) provides ratings for dry filters and for electronic air cleaners. Enforcement of the certification procedures, under which manufacturers may claim certified performance, is the responsibility of the Air Conditioning and Refrigeration Institute (ARI) under their Standard 680.

The ASHRAE arrestance test measures the ability of a filter to remove a specified dust sample from the airstream which passes through the test filter. Constituents of the dust sample are carefully controlled to ensure reproducibility of test results. In the dust spot efficiency test, samples of unfiltered atmospheric air and filtered air (downstream of the test filter) are drawn off through chemical filter paper. A photoelectric cell scans the blackness of the dust spots that appear on the two pieces of filter paper; the times required to achieve equal discoloration of both pieces of filter paper, one in the unfiltered air and the other in the filtered air, are translated into a mathematical ratio, which becomes the efficiency of the test filter. *See* GALVANOMETER.

The development of ultrahigh-efficiency (HEPA) filters for special purposes (such as protection from radioactive and bacterial contaminants) led to the development of the dioctylphthalate (DOP) test. Of all the methods used to evaluate air filters, the DOP test procedure is the most sophisticated; it is used only for testing filters to be employed in critical air-cleaning applications. In a homogeneous smoke of controlled particle size (0.3-μm particles are used in this test), the DOP vapor passes through the filter. A light-scattering technique counts the number of particles entering the test filter, and a similar device counts those that emerge from the filter. Certain applications call for filters of such high efficiency that only 0.001% of incident 0.3-μm particles is permitted to emerge.

Morton A. Bell

Air heater

A component of a steam-generating unit that absorbs heat from the products of combustion after they have passed through the steam-generating and superheating sections. Heat recovered from the gas is recycled to the furnace by the combustion air and is absorbed in the steam-generating unit, with a resultant gain in overall thermal efficiency. Use of preheated combustion air also accelerates ignition and promotes rapid burning of the fuel.

Air heaters frequently are used in conjunction with economizers, because the temperature of the inlet air is less than that of the feedwater to the economizer, and in this way it is possible to reduce further the temperature of flue gas before it is discharged to the stack. *See* BOILER ECONOMIZER.

Air heaters are usually classed as recuperative or regenerative types. Both types depend upon convection transfer of heat from the gas stream to a metal or other solid surface and upon convection transfer of heat from this surface to the air. In the recuperative type, exemplified by tubular- or plate-type heaters, the metal parts are stationary and form a separating boundary between the heating and cooling fluids, and heat passes by conduction through the metal wall (**Fig. 1**). In rotary regenerative air heaters

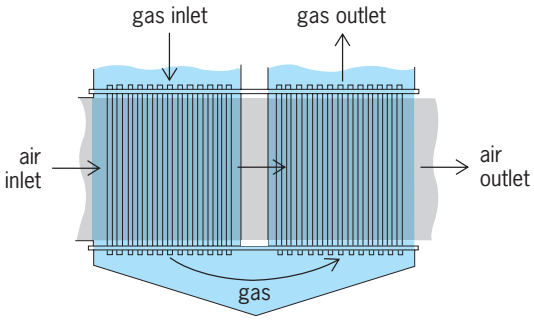


Fig. 1. Tubular air heater, two-gas single-air pass.

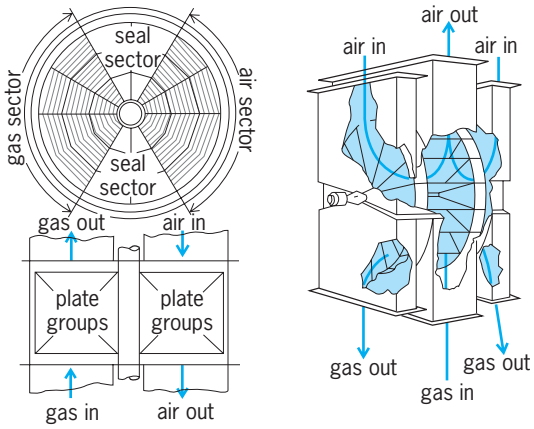


Fig. 2. Two types of rotary regenerative air heaters.

(**Fig. 2**) heat-transferring members are moved alternately through the gas and air streams, thus undergoing repetitive heating and cooling cycles; heat is transferred to or from the thermal storage capacity of the members. Other forms of regenerative-type air heaters, which seldom are used with steam-generating units, have stationary elements, and the alternate flow of gas and air is controlled by dampers, as in the refractory stoves of blast furnaces; or they may employ, as in the pebble heaters used in the petroleum industry for high-temperature heat exchange, a flow of solid particles which are alternately heated and cooled. *See* CONVECTION (HEAT).

In convection heat-transfer equipment, higher heat-transfer rates and better utilization of the heat-absorbing surface are obtained with a counterflow of gases through small flow channels. The rotary regenerative air heater readily lends itself to the application of these two principles and offers high performance in small space. However, leakage of air into the gas stream necessitates frequent maintenance of seals between the moving and stationary members, and fly ash often is transported into the combustion air system. These problems are not experienced with recuperative air heaters of the tubular type. *See* STEAM-GENERATING UNIT.

George W. Kessler

Bibliography. American Society of Heating, *Refrigerating and Air Conditioning Engineers, 2004 ASHRAE Handbook: Heating, Ventilating, and Air-Conditioning: Systems and Equipment*, 2004; F. C. McQuiston, J. D. Parker, and J. D. Spitler, *Heating, Ventilating and Air Conditioning Analysis and Design*, 6th ed., 2004.

Air mass

In meteorology, an extensive body of the atmosphere which is relatively homogeneous horizontally. An air mass may be followed on the weather map as an entity in its day-to-day movement in the general circulation of the atmosphere. The expressions air mass analysis and frontal analysis are applied to the analysis of weather maps in terms of the prevailing air masses and of the zones of transition and interaction (fronts) which separate them.

The relative horizontal homogeneity of an air mass stands in contrast to sharper horizontal changes in a frontal zone. The horizontal extent of important air masses is reckoned in millions of square miles. In the vertical dimension an air mass extends at most to the top of the troposphere, and frequently is restricted to the lower half or less of the troposphere. *See* FRONT; METEOROLOGY; WEATHER MAP.

Development of concept. Practical application of the concept to the air mass and frontal analysis of daily weather maps for prognostic purposes was a product of World War I. A contribution of the Norwegian school of meteorology headed by V. Bjerknes, this development originated in the substitution of close scrutiny of weather map data from a dense local network of observing stations for the usual far-flung international network. The advantage of air-mass analysis for practical forecasting became so evident that during the three decades following World War I the technique was applied in more or less modified form by nearly every progressive weather service in the world. However, the rapid increase of observational weather data from higher levels of the atmosphere during and since World War II has resulted in a progressive tendency to drop the careful application of air-mass analysis techniques in favor of those involving the kinematical or dynamic analysis of upper-level air flow, usually involving numerical methods supported by large computers.

Origin. The occurrence of air masses as they appear on the daily weather maps depends upon the existence of air-mass source regions, areas of the Earth's surface which are sufficiently uniform so that the overlying atmosphere acquires similar characteristics throughout the region. *See* ATMOSPHERIC GENERAL CIRCULATION.

Weather significance. The thermodynamic properties of air mass determine not only the general character of the weather in the extensive area that it covers, but also to some extent the severity of the weather activity in the frontal zone of interaction between air masses. Those properties which determine the primary weather characteristics of an air mass are defined by the vertical distribution of water vapor and heat (temperature). On the vertical distribution of water vapor depend the presence or absence of condensation forms and, if present, the elevation and thickness of fog or cloud layers. On the vertical distribution of temperature depend the relative warmth or coldness of the air mass and, more importantly, the vertical gradient of temperature, known as the lapse rate. The lapse rate determines the stability or instability of the air mass for thermal convection and consequently, the stratiform or convective cellular structure of the cloud forms and precipitation. The most unstable moist air mass, in which the vertical lapse rate may approach 1°F/170 ft (1°C/100 m), is characterized by severe turbulence and heavy showers or thundershowers. In the most stable air mass there is observed an actual increase (inversion) of

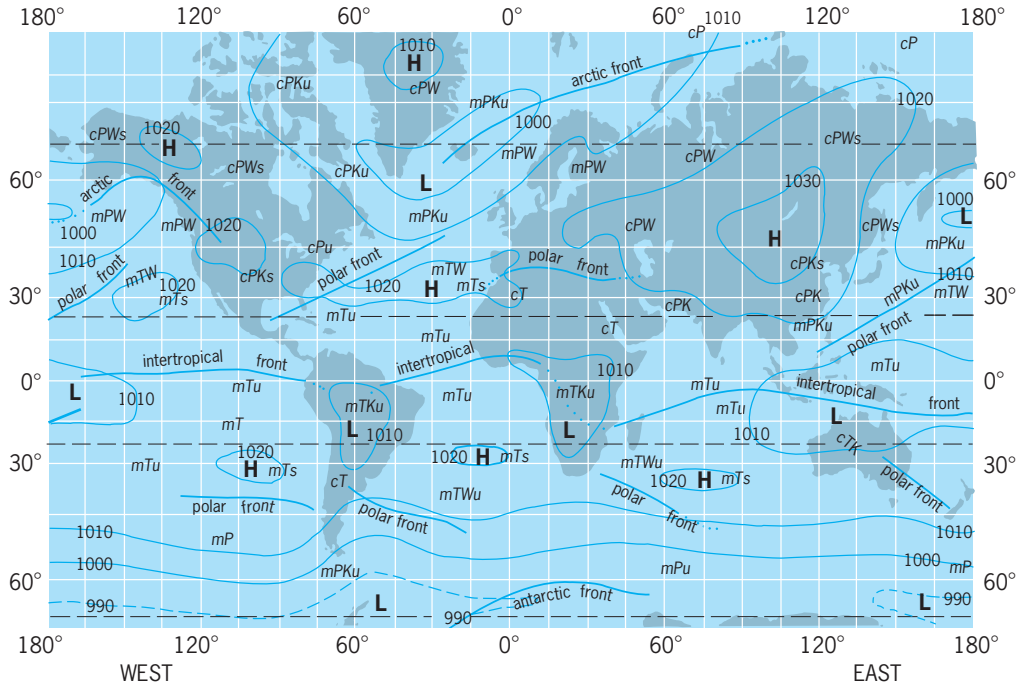


Fig. 1. Air-mass source regions, January. High- and low-atmospheric-pressure centers are designated *H* and *L* within average pressure lines numbered in millibars (such as 1010); 1 millibar = 10^2 Pa. Major frontal zones are labeled along heavy lines. (*After H. C. Willett and F. Sanders, Descriptive Meteorology, 2d ed., Academic Press, 1959*)

temperature with increase of height at low elevations. With this condition there is little turbulence, and if the air is moist there is fog or low stratus cloudiness and possible drizzle, but if the air is dry there will be low dust or industrial smoke haze. *See* TEMPERATURE INVERSION.

Classification. A wide variety of systems of classification and designation of air masses was developed by different weather services around the world. Most systems of air-mass classification are based on a designation of the character of the source region and the subsequent modifying influences to which the

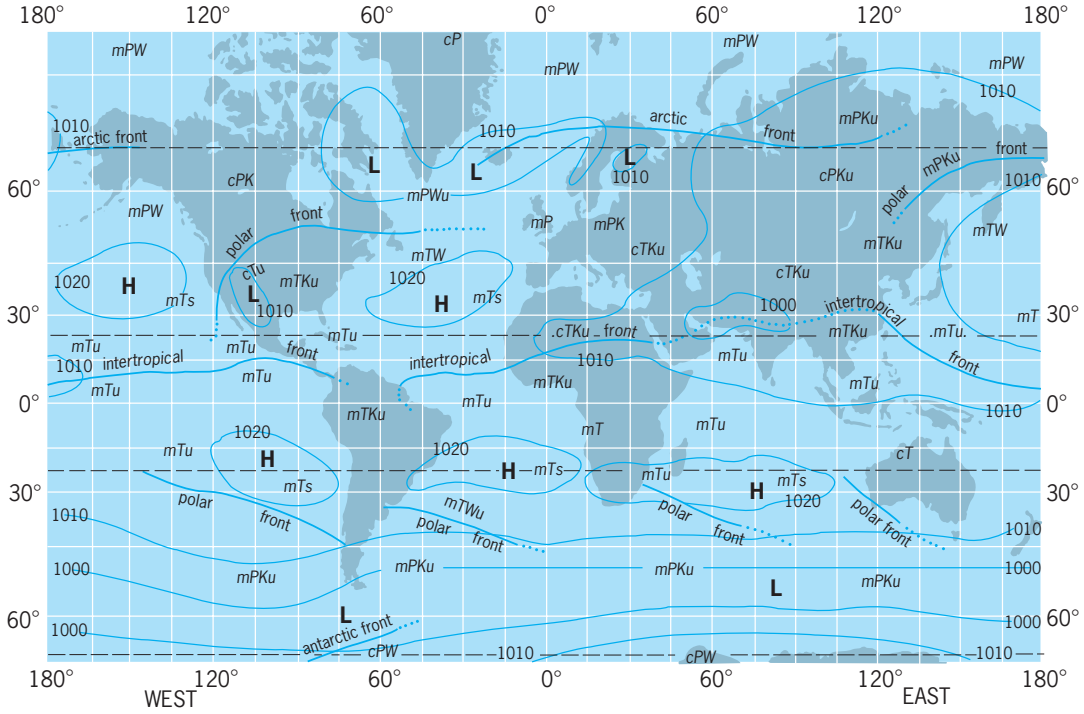


Fig. 2. Air-mass source regions, July. The symbols which are used in this figure are the same as those for Fig. 1. (*After H. C. Willett and F. Sanders, Descriptive Meteorology, 2d ed., Academic Press, 1959*)

air mass is exposed. Probably the most effective and widely applied system of classification is a modification of the original Norwegian system that is based on the following four designations.

Polar versus tropical origin. All primary air-mass source regions lie in polar (*P* in **Figs. 1** and **2**) or in tropical (*T*) latitudes. In middle latitudes there occur the modification and interaction of air masses initially of polar or tropical origin. This difference of origin establishes the air mass as cold or warm in character.

Maritime versus continental origin. To be homogeneous, an air-mass source region must be exclusively maritime or exclusively continental in character. On this difference depends the presence or absence of the moisture necessary for extensive condensation forms. However, a long trajectory over open sea transforms a continental to a maritime air mass, just as a long land trajectory, particularly across major mountain barriers, transforms a maritime to a continental air mass. On Figs. 1 and 2, *m* and *c* are used with *P* and *T* (*mP*, *cP*, *mT*, and *cT*) to indicate maritime and continental character, respectively.

Heating versus cooling by ground. This influence determines whether the air mass is vertically unstable or stable in its lower strata. In a moist air mass it makes the difference between convective cumulus clouds with good visibility on the one hand and fog or low stratus clouds on the other. Symbols *W* (warm) and *K* (cold) are used on maps—thus, *mPK* or *mPW*.

Convergence versus divergence. Horizontal convergence at low levels is associated with lifting and horizontal divergence at low levels with sinking. Which condition prevails is dependent in a complex manner upon the large-scale flow pattern of the air mass. Horizontal convergence produces vertical instability of the air mass in its upper strata (*u* on maps), and horizontal divergence produces vertical stability (*s* on maps). On this difference depends the possibility or impossibility of occurrence of heavy air-mass showers or thundershowers or of heavy frontal precipitation.

Examples of the designation of these tendencies and the intermediate conditions for maritime polar air masses are *mPWs*, *mPW*, *mPWu*, *mPs*, *mPu*, *mPKs*, *mPK*, and *MPKu*. Hurd C. Willett; Edwin Kessler

Bibliography. S. Ackerman and J. A. Knox, *Meteorology: Understanding the Atmosphere*, 2002; R. Anthes, *Meteorology*, 7th ed., 1996; L. J. Battan, *Fundamentals of Meteorology*, 2d ed., 1984; J. M. Moran and M. D. Morgan, *Meteorology*: *The Atmosphere and the Science of Weather*, 5th ed., 1996; R. B. Stull, *Meteorology for Scientists and Engineers*, 2d ed., 1999.

Air navigation

A discipline that combines the means to know the location and heading of an aircraft with respect to some map or coordinate system, with guidance which means steering the aircraft along some desired path toward a destination. The construction of the desired path is done with respect to the same a map or coordinate system used for location. In civil aviation, the coordinate system that is becoming a de facto standard is the World Geodetic System of 1984 (WGS-84). This is largely due to the fact that positioning in the Global Positioning System (GPS) is based on this mapping and coordinate system.

In the aircraft, the process of planning and directing its progress between selected geographic points or over a selected route is the main navigation task. The primary tasks are planning the flight path or route, guiding the aircraft safely along the desired route, conforming with the rules of flight and with special procedures such as noise abatement, maximizing fuel efficiency, departure takeoff, and landing. Landing may require precision navigation and guidance if flight instruments are used.

Modern air navigation makes use of four dimensions: latitude, longitude, altitude, and time. Time becomes a factor whenever time constraints have to be met. Examples are arrival at specific waypoints or at the destination at a precisely specified time.

The simplest form of air navigation is pilotage, in which the pilot directs the aircraft by visual reference to objects on the ground. Pilotage is the most basic skill, and despite the accuracy of modern air navigation instruments, it can be drawn upon in the case of an emergency where navigational aids experience failure. More complex methods rely upon navigational aids external to the aircraft or upon self-contained, independent systems. *See* PILOTAGE.

Dead reckoning is the process of estimating one's current position by measuring the change in position since the last accurate position fix. The simplest form of dead reckoning requires an airspeed indicator, outside-air-temperature gage, clock, compass, and chart (map). Using these tools and a few simple computations, the pilot can follow a compass heading from a point of departure for a specified time and arrive at a dead-reckoned position. Dead reckoning is periodically validated by position information from a source external to the aircraft, such as a ground sighting, radio bearing, celestial observation, or a direct position solution.

All navigation instruments that propagate the-break position based on position increments, velocity measurement, or acceleration measurement are dead-reckoning systems. Examples are the Doppler velocity systems (DVS) and inertial navigation systems (INS). *See* DEAD RECKONING.

Navigation systems that derive a position fix from range measurements or from range and angle measurements are direct position solutions. These navigation instruments measure range or angles between the aircraft and known points. Examples are distance measurement equipment (DME) and very high frequency (VHF) omnidirectional radio range (VOR), respectively. Depending on the geometry of the range and angle measurements, the position solution may not be unique. Examples of nonunique position

solutions are DME-DME; examples of unique position solutions are GPS position fixes. Whenever available, unique position solutions are regarded as the last accurate position fix.

Flight Planning

Air navigation begins with a flight plan. For a flight in visual meteorological conditions (VMC), a simple plan may start with drawing a course line on a specially designed aeronautical chart between the point of departure and the destination. From the chart and the course line, the pilot, using a protractor, determines the true course (in degrees clockwise from true north), magnetic variation, and distance to be flown, usually in nautical miles. The pilot also checks the chart for obstacles or hazards that may be on or adjacent to the planned route.

The three minimum considerations that must be accounted for in the flight plan are:

1. The pilot must properly calculate the course and fuel consumption.
2. The pilot must be able to accurately fly the aircraft along the intended route without hazard.
3. The pilot must obtain the weather conditions, particularly the winds aloft, from the weather service.

The first two considerations are under the pilots control and can be planned. The third is not but it establishes the VMC.

During flight planning, the pilot will apply corrections to the true course for magnetic variation (the angular difference between true north and magnetic north), deviation (the angular difference between the nominal magnetic heading and the actual aircraft compass heading), and computed wind drift.

A preflight briefing is customary and may be obtained from a flight service station (FSS). It will contain the expected weather conditions for the flight and pertinent NOTAMS (Notices to Airmen), and can be obtained in person, by telephone, or by a direct user access terminal (DUAT) system, requiring a personal computer and modem. Following the preflight weather briefing, the pilot has the necessary facts to prepare a simple flight plan: wind direction and velocity, which were included in the briefing; true course, which was determined from the course line on the chart; and true airspeed, which is calculated by applying the corrections for temperature and altitude (about +2% for each 1000 ft or 300 m above sea level) to the indicated airspeed. With the wind, true course, and true airspeed, the pilot can construct a wind triangle to determine the effect of the wind on speed and heading. The wind triangle can be constructed manually by using a protractor and ruler, or it may be set up on an electronic calculator or a manually operated navigational computer. The wind triangle computation provides groundspeed (speed over the surface) and wind drift angle, in degrees left or right of the course (**Fig. 1**). During flight, a correction counter to the drift angle (called wind correction angle or CRAB) will be applied to the aircraft heading to prevent drifting off course. With the planned groundspeed and the distance to be flown

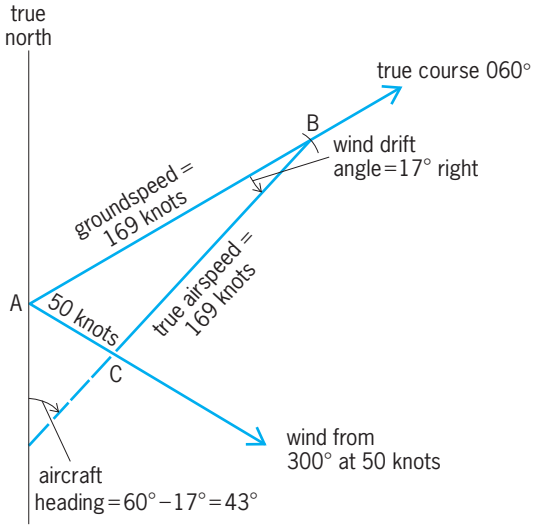


Fig. 1. Typical wind triangle, drawn to scale. 1 knot = 1 nmi/h = 0.514 m/s.

in miles, the pilot can compute the flight time and the fuel required. Extra fuel is added to assure an adequate reserve upon landing.

The above description applies to the formation of a simple, fair-weather flight plan. The process assumes greater complexity for a higher-performance aircraft, for a longer flight, or for a flight in adverse weather.

Weather forecasts for the departure airport, for the route of flight, and for the destination airport must be considered in formulating the plan. Severe storms or turbulence can sometimes be avoided by taking a different route or flying at a different altitude. More favorable winds may also be found by selecting a different route or altitude. *See* AERONAUTICAL METEOROLOGY.

Air navigation is three-dimensional, and selection of a flight altitude is an important part of the planning process. Light, piston-engine aircraft are usually operated at altitudes of 10,000 ft (3000 m) or less, while jet aircraft commonly operate at much higher altitudes, where they may take advantage of the jet's increased fuel economy. During flight planning for large turboprop and jet aircraft, the time en route and the fuel used at various altitudes or over different routes are often compared to select the optimum altitude and route.

Aircraft flying in the lower altitudes must select altitudes that will safely clear the terrain, including artificial obstacles such as television towers and tall buildings. Instrument flights are required to fly at least 1000 ft (300 m) above the highest obstacle in nonmountainous areas or 2000 ft (600 m) above the highest obstacle in designated mountainous areas.

The selection of a flight altitude is further constrained by requirements for separation from other aircraft. Generally, flights proceeding in a northerly or easterly direction fly at odd altitudes (3000, 5000, 7000 ft, and so forth, equivalent to 900, 1500, 2100 m), while flights proceeding in a southerly or westerly direction fly at even altitudes (2000, 4000, 6000 ft, and so forth, equivalent to 600, 1200,

1800 m). Flights operating under visual flight rules (VFR) add 500 ft (150 m) to the altitudes above. Above 29,000 ft (8800 m), somewhat different rules apply because of increased vertical separation required between aircraft.

Because of the cost of fuel and the large quantities used, fuel conservation is important. Flying at higher altitudes and reducing a jet's speed slightly will usually produce significant fuel savings with little increase in time; consequently, many operators regularly fly at the optimal, most fuel-efficient speed. *See* AIRCRAFT FUEL.

When the destination weather is expected to be poor, an alternative airport must be designated and additional fuel carried to fly there should a diversion be necessary.

Navigation Charts

All navigation requires special maps or charts such as topographic maps intended for pilotage or radio navigation charts intended to be used with ground radio aids. Aeronautical maps such as sectional charts show terrain contours and the elevation above mean sea level of significant terrain and other obstacles, such as mountain peaks, buildings, and television towers.

With a few exceptions, most charts for aeronautics use the Lambert conformal projection. Significant advantages of the Lambert projection are that a straight line drawn between two points is a great circle, which is the shortest distance between those points, and distances are uniformly scaled across the entire chart.

Mercator projection charts draw lines of longitude and latitude in straight lines. A straight line on a Mercator chart is called a rhumb line. It defines a constant heading path between departure and destination points and is not in general a great circle path. Distances are not uniformly scaled across the chart from north to south. *See* MAP PROJECTIONS.

Some special tasks, such as polar navigation, utilize a grid chart whose meridians and parallels are equally spaced, resembling lines on graph paper. *See* POLAR NAVIGATION.

Operational Air Navigation

The operational phase of air navigation commences with the preflight cockpit check of clocks, radios, compasses, flight management systems, and other flight and navigation equipment. Some equipment such as inertial navigation requires certain initialization data to be inserted before the units can operate properly. An air-traffic control clearance for the planned flight is obtained from the controlling authorities during preflight.

The flight management systems (FMS) can automate a very significant portion of the cockpit workload, allowing the flight crew to manage the aircraft systems instead of interacting with the various systems. However, the FMS must be set up properly, including entry of the flight path data bases, vertical navigation performance parameters and policies, weather information (if not done so automatically), and any other pertinent data. The FMS can ensure optimal fuel burn and arrival at the destination on time while providing horizontal and vertical guidance to the flight director or autopilot, achieving true four-dimensional navigation. *See* AUTOPILOT.

Departure. During taxi to the takeoff runway, the pilot follows a detailed airport chart and instructions received by radio from the airport ground controller. After a takeoff in visual flight conditions, the pilot turns to a heading to intercept the planned course and climbs to the proper altitude for the direction of flight.

The departing pilot may be guided by a standard instrument departure (SID) chart for the airport or may be radar-guided by instructions from the radar departure controller. SIDs are used to automate and completely specify the departure route. Aircraft that can follow SID procedures have the advantage of quick and efficient dispatch. However, SIDs are complex in general and are typically executed by an FMS.

En route. The pilot's primary navigational task is to assure that the aircraft stays on course. If navigation is by pilotage, the pilot frequently checks the aircraft's position by reference to topographic features on the aeronautical chart being used, and the aircraft's heading is corrected as necessary to follow the course. Ground speed is determined from the time required to fly the distance from the departure airport to the first and subsequent fixes. Fuel is checked to verify that the quantity used thus far agrees with the planned use of fuel.

The pilot estimates the time that the aircraft is expected to be over the next fix, based on the speed thus far made good and the speed anticipated over the next segment. This process will continue throughout the flight, with the pilot repeatedly assessing speed and fuel used versus the flight plan and making alterations as needed.

An FMS automates this task. It collects the positioning information from the various navigation sensors such as GPS, INS, VOR, and DME, and produces the best probable position fix. It may employ sophisticated techniques such as Kálmán filtering and hierarchical logic to produce the position solution. Simultaneously, the FMS will check flight policies, check altitude and noise restrictions, and provide guidance to the autopilot, and in addition it can monitor fuel consumption.

VOR navigation. The most common form of en route navigation over landmasses in the Western world uses a network of VOR stations. VOR transmitters are spaced at strategic intervals, averaging about 100 mi (160 km) between stations within the United States but sometimes much farther apart. Designated routes between VOR stations are referred to as airways.

VOR transmitters emit a special signal that, when received, indicates heading toward a VOR station. The pilot can then determine the direction from the ground station to the aircraft, measured in degrees magnetic and referred to as a radial. Distance to the

station is displayed when a DME transmitter is colocated with the VOR. Such an installation is referred to as a VOR/DME. When the DME function is provided by a colocated Tacan transmitter, the installation is referred to as a VORTAC. *See* DISTANCE-MEASURING EQUIPMENT; TACAN.

An airway is defined by a series of VOR stations along an airway route. Navigation charts show the location of the VORs, their radio frequencies, and the radials (courses) between them which make up the airways. The pilot refers to a chart for the frequency of the VOR and the radial to be followed, which are then set on the appropriate cockpit equipment. After following the outbound radial to the midpoint or designated changeover point between stations, the next station on the airway is tuned in and the inbound radial to that station is followed. When the aircraft is equipped with an FMS, the FMS performs the tuning function automatically, leaving the pilot to monitor the flight and make necessary adjustments as succeeding stations are selected along the route.

It is not necessary to follow charted airways in order to use VOR stations for navigation. Flights are often cleared by air-traffic control to proceed "direct" from one VOR to another in order to detour around adverse weather or to change routes. *See* RHO-THETA SYSTEM; VOR (VHF OMNIDIRECTIONAL RANGE).

Arrival. Arrival routes and procedures at major airports are often so complex that it is necessary to detail them on standard terminal arrival route (STAR) charts. Details include navigational aids, routes, required altitudes, and communications frequencies.

A high-altitude jet starts descending miles from the landing airport. A jet flying at 39,000 ft (11,900 m) might start to descend more than 100 nautical miles (185 km), or 15–20 min, from the airport. Unpressurized aircraft carrying passengers fly much lower, but their descent must be more gradual to minimize passenger discomfort in adjusting to the change in atmospheric pressure.

Regardless of the aircraft type, the pilot estimates as accurately as possible the time and point to start descending. This is particularly important in jet-type aircraft since an early descent causes a significant increase in fuel consumption while a late descent introduces other problems.

Every instrument approach for landing has its own chart, whether it uses an instrument landing system (ILS), the Space-Based Augmentation System (SBAS), VOR, nondirectional beacon (NDB), radar, or some other aid. The approach chart contains such things as the radio frequency and identification of the aids to be used, altitudes to be observed at various points during the approach, heading of the final approach course, minimum weather conditions required for the approach, missed approach instructions, and airport altitude.

The preferred aid for instrument landing use is ILS. It consists of two separate radio beams: one called the localizer is aligned very precisely with the runway centerline, and the other called the glideslope is projected upward at a slight angle from the landing touchdown point on the runway. By following the two beams very closely on the cockpit instruments, the pilot can descend safely for landing, regardless of weather.

ILS accuracy permits aircraft to safely descend to lower weather minimums than are possible with other aids such as VOR or NDB. Descent for landing is commonly authorized to 200 ft (60 m) above ground, with 0.5-mi (0.8-km) visibility (Category I or Cat-I). Most ILS approaches in small aircraft are controlled manually, but many large aircraft are equipped with automatic pilots that can be coupled to follow the ILS. Aircraft with more sophisticated approach and landing capability may be permitted to land on specially designated runways when weather is less than Category I. Category III is the most stringent weather condition, and properly equipped aircraft may be authorized to make autopilot-coupled ILS approaches and automatic landings on those runways when ceiling and visibility are zero or near zero. *See* AUTOMATIC LANDING SYSTEM; INSTRUMENT LANDING SYSTEM (ILS).

SBAS approaches have now been introduced and are capable of providing an ILS CAT-I equivalent service. Approach procedures based on SBAS are called localizer with precision vertical (LPV) approaches. Such approaches require no airport ground aides, only an SBAS/GPS avionics receiver. The LPV approach charts are similar to ILS charts, the major difference being that an LPV approach requires an SBAS area navigation (RNAV) capability, typically with addition of an FMS. It is expected that SBAS will replace ILS CAT-I approaches.

After landing, the pilot taxis the aircraft to the parking area following the airport chart and instructions received by radio from the ground controller. Some major airports are equipped with special radar for following taxiing aircraft during very low visibility. *See* SURVEILLANCE RADAR.

Altimetry

Aircraft altitude may be obtained from a barometric altimeter, a radio altimeter, or GPS. Although all of these instruments produce an altitude measurement, they are not the same kind of altitude measurement and they cannot be used interchangeably. The reason is that their reference systems are different. In general, baro-altimeters produce mean sea level altitude (for corrected low altitudes) or an altitude with respect to a standard atmospheric altitude; GPS receivers produce altitude with respect to the WGS-84 coordinate system; and radar altitude is simply the distance between the aircraft and the closest terrain beneath it (including mountains).

Mean sea level (MSL) altitude. The MSL is an average sea level surface that defines the zero altitude and is used to define the altitude reference in all navigation and mapping charts. The MSL zero altitude is determined by measuring the Earth's gravity with satellites. The zero altitude surface cannot be directly observed due to tidal actions, changes in the Earth's

weather, melting of glaciers, and other effects. *See* GEODESY.

Barometric altitude. The primary altitude instrument in the aircraft is the baro-altimeter. The baro-altimeter measures air pressure and coverts it to a displayed altitude in meters or feet. *See* ALTIMETER.

Barometric altimeters employ three different operating modes: by setting altitude above mean sea level, as reported by a nearby weather station (QNH); by using a standard setting (QNE) of 29.92 in. of mercury (1013.2 millibars) as the sea level or zero MSL altitude; and by setting the altimeter to read zero when on the airport surface (QFE), thereby showing height above the airfield when in flight. A small number of airlines use QFE when approaching the destination airport, but most use QNH.

At and above a transition altitude, all aircraft use the standard QNE altimeter setting, and altitudes are termed flight levels (FL). For example, 29,000 ft becomes FL290. The transition altitude in United States airspace is 18,000 ft (FL180) or 5500 m. Elsewhere, transition altitudes are determined by the controlling governmental agency. The QNE barometric altitude has the same definition (and conversion from air pressure to altitude) for all aircraft.

Radar altitude. All radar altimeters measure the smallest range between the terrain immediately below the aircraft and the aircraft. The radar altimeter is intended for instrument approaches that require accurate height above the landing runway threshold. Most radar altimeters operate only below 2500 ft (750 m) above ground.

GPS altitude. GPS receivers provide accurate WGS-84 altitude information. GPS altitude may be converted to MSL altitude using an MSL model such as the Earth Gravitational Model of 1996 (EGM96), developed by the National Imagery and Mapping Agency, the NASA Goddard Space Flight Center, and Ohio State University.

GPS altitude is unaffected by the errors found in barometric altimetry. It has an accuracy of approximately 35 ft (10 m). The accuracy of standard GPS positioning is much improved when augmented by the Space-Based Augmentation System.

SBAS altitude. SBAS receivers provide improved GPS altitude that can be used for performing LPV landing approaches equivalent to ILS CAT-I approaches with 200-ft (60-m) decision height. LPV approach plates (single-page aeronautical charts that provide the pilot with the detailed information needed to perform a precision approach) reference MSL; however, the final approach segment (FAS) is specified in WGS-84 coordinates.

Reduced Vertical Separation Minima (RVSM)

East-west traffic has resulted in a heavily congested airspace. The vertical separation required of aircraft between FL290 and FL410 has been 2000 ft (600 m). Special RVSM rules were brought in to allow the reduction of this vertical separation to 1000 feet (300 m). The result is an additional six flight levels between FL290 and FL410, allowing more aircraft to fly efficient time-fuel profiles. RVSM-qualified aircraft must have higher-accuracy baro-altimeters (operating in QNE mode) and other surveillance equipment such as TCAS-II, and must meet altitude-keeping performance standards. *See* AIRCRAFT COLLISION AVOIDANCE SYSTEM.

There are several airspaces to which RVSM applies. In the United States, such airspace is referred to as Domestic RVSM or DRVSM. When flying in an RVSM airspace, the pilot must abide by special rules governing that particular airspace.

Additional Navigation Aids and Systems

The **table** lists the electronic systems discussed, gives their principal frequency bands, and indicates which systems are internationally standardized.

Flight management system (FMS). An FMS integrates aircraft communication, navigation, flight guidance and control, and aircraft performance management functions. The navigation function may be as simple as the area navigation described below, but today it is more likely to utilize inputs from several sensors: VOR/DME, GPS/SBAS, Loran-C, and INS. Typically, the FMS will select the navigation sensors that produce the best possible position, favoring SBAS and then proceeding down to less accurate sources of navigation. Navigation sources may be integrated through a Kálmán filter algorithm. The Kálmán filter models and estimates the behavior of dynamical systems such as an INS, and it maintains the dynamical system error (the variance). In air navigation, Kálmán filters are typically used to integrate the inertial reference systems (IRS) with GPS/SBAS, taking advantage of the self-contained short-term navigation performance of IRS and combining it with the long-term accuracy and integrity provided by SBAS. *See* ESTIMATION THEORY.

Extensive aircraft performance and regularly updated regional or worldwide navigation databases are stored in and utilized by the FMS. Commonly used routes can be stored in the FMS and called up when needed. In addition to lateral navigation (LNAV), a flight management computer can manage vertical

Electronic navigation systems

System	Frequency band
VOR*	108-118 MHz
DME*, Tacan	960-1215 MHz
ILS localizer*	108-112 MHz
ILS glideslope*	329-335 MHz
Doppler radar	10-20 GHz
Loran-C	90-110 kHz
MLS	5.0-5.25 GHz
GPS	1227, 1575 MHz
SBAS	1575 MHz
ADF/NDB*	190-1600 kHz
Weather radar	5, 9 GHz
Radar altimeter	4.2 GHz
Doppler velocity radar	13.325 GHz

*Internationally standardized systems.

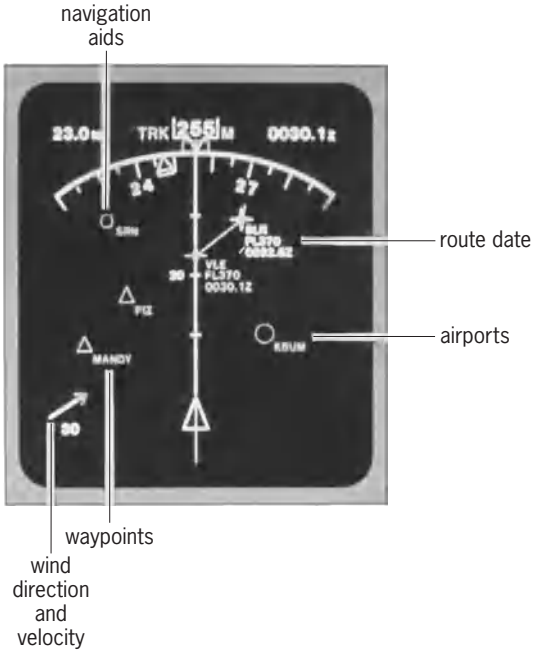


Fig. 2. Electronic map display developed by a flight management computer on the pilot's instrument panel of a Boeing 767 airplane. (*United Airlines*)

navigation (VNAV), that is, climbs and descents on the proper profiles.

The FMS can also control aircraft speed to achieve required time of arrival at specific waypoints along the intended path of flight. The time management function of the FMS can be used in conjunction with air-traffic control management to achieve precise arrival control of the aircraft for improved airspace and runway utilization. The combination of lateral and vertical navigation with management of speed and time is referred to as 4-D RNAV. Besides coupling with the autopilot and autothrottle, in some installations the flight management computer is coupled with an electronic map display which can be used for primary navigation guidance (**Fig. 2**). *See* AIRCRAFT INSTRUMENTATION.

Current FMS also computes a cost performance index (CPI). The CPI is an indicator of the aircraft flight performance with respect to an optimal path. The optimal path is developed by an operator, or airline, according to its policies. These policies may require a strict time of arrival, a minimum fuel burn, or any combination of policies and requirements. Variable factors such as winds, air-traffic control directives, and aircraft weight have an impact on the actual CPI.

Global Navigation Satellite System (GNSS). The Global Positioning System, sometimes called NAVSTAR, is a part of the Global Navigation Satellite System, and is operated by the U.S. Department of Defense. GPS is a space-based positioning, radionavigation, and time-distribution system designed for worldwide military use. (A comparable system called GLONASS developed by the former Soviet Union is still in operation.) GPS has been made available for worldwide civil and military use.

The Wide Area Augmentation System (WAAS) was designed in the United States by the Federal Aviation Administration to provide GPS-independent real-time GPS integrity, differential corrections, and ranging signals. The WAAS complements GPS and extends GPS capability as a primary-means-of-navigation system and as a landing approach system.

The WAAS has been adopted internationally by the International Civil Aviation Organization (ICAO) and has become the Space Based Augmentation System (SBAS). Several nations are developing SBAS systems: in Europe as the European Geostationary Navigation Overlay System (EGNOS), in India as GAGAN, in China as BEIDOU, and in Japan as MSAS. All of these systems are interoperable and conform to the ICAO SBAS standards. SBAS provides GPS integrity, real-time differential corrections, and ranging signals similar to GPS. SBAS will eventually cover the Earth except for latitudes exceeding 75° due to the fact that SBAS satellites are parked in geosynchronous orbits. All continents with the exception of Antarctica will be covered by SBAS.

Future GNSS development includes the European Galileo satellite navigation system, and GPS and SBAS modernization. Galileo is similar to GPS in function; however, its signal transmission formats differ from GPS. Equipment that can receive GPS, Galileo, and GLONASS signals is being developed. Future GPS and SBAS will provide a second civil frequency called L5, centered at 1176 MHz, as well as adding a civil signal to the L2 frequency at 1227 MHz. Furthermore, the civil C/A codes on the L1 band at 1575 MHz will be augmented by new civil codes with less cross-correlation that support CAT-IIIb functionality. The resulting benefits are projected to support CAT-II, and possibly CAT-III, with GPS, SBAS, and Galileo.

Airborne GPS/SBAS navigation receivers continuously process GPS and SBAS satellite data to update the real-time position, latitude, longitude, and altitude in WGS-84 coordinates. The position may be displayed as a digital readout of present position, sent as an input to the course deviation indicator on the pilot's instrument panel, or sent as a sensor input to a navigation computer such as the flight management system or inertial navigation system.

All certified, airborne GPS/SBAS receivers are distinguished by providing integrity with the position solution. Integrity is an estimate of the position fix error, assuming that a satellite signal is erroneous or misleading. The integrity algorithms are referred to as either the Receiver Autonomous Integrity Monitor (RAIM) or the Fault Detection and Exclusion (FDE) algorithms. In one case, the SBAS network determines the integrity of each GPS satellite in real time, and relays it to the SBAS receivers and to the avionics displays and enunciators within 6 seconds of the total elapsed time since detection.

Terrestrial enhancements of GNSS. Two terrestrial enhancements of GNSS have been proposed. The first is the Local Area Augmentation System (LAAS) that has been internationally adopted by ICAO as the

Ground Based Augmentation System (GBAS). It is based on the same principles and techniques used in differential GPS (DGPS) satellite surveying to determine the satellite positioning error. The identified errors and corrections are valid only within the local area.

GBAS requires precisely located ground reference stations which receive satellite signals, compute a position based upon that data, and compare it with the reference stations' known position to compute corrections to the satellite signals. The resultant corrections are transmitted to all GBAS receivers via a standard VHF data broadcast (VDB) link. The defined operational radius of the VDB link is 23 nautical miles (43 km).

GBAS is intended to provide the required accuracy for precision instrument approaches, including Category III instrument landings. Due to the effects of spatial decorrelation, cross-correlation effects of the GPS C/A positioning code, and the lack of an additional radio-frequency channel, achieving a certified CAT-II/III capability has been elusive. It is believed that with a dual-frequency GPS constellation, and perhaps with Galileo, GBAS will achieve certified CAT-II/III operational capability. *See* SATELLITE NAVIGATION SYSTEMS.

The second proposed terrestrial enhancement is the Ground Regional Augmentation System (GRAS), developed by Australia. It intends to provide SBAS functionality, but by using specially modified GBAS ground stations. The VDB link remains with a 23-nautical-mile radius.

Area navigation (RNAV). An area navigation system provides the pilot with navigational guidance along random routes which are not restricted to airways or along direct routes between ground-based navigation aids such as VOR or NDB.

VORTAC-based RNAV relies upon precise DME distance from two or more VORTACs or VOR/DMEs (rho/rho mode) to continuously fix aircraft position. SBAS-based RNAV uses the position fix capability provided by an SBAS avionics receiver.

To plot an RNAV course, a direct course line is drawn on a navigation chart between the departure and destination points. The pilot then locates several VOR/DMEs or VORTACs along the proposed route and draws lines from those facilities to intersect the direct course. Those intersections will serve as waypoints. This task is automated by an FMS. SBAS position is always available and requires no planning such as with VOR/DME.

Prior to departure or en route, the pilot enters the waypoints into the RNAV course-line computer (CLC), which will compute direct navigational guidance to the waypoints. The cockpit equipment displays the course to be followed, distance remaining, and time-to-go to the next waypoint. The course to be followed may be coupled to an autopilot or the course deviation indicator (CDI).

RNAV is particularly useful in locating otherwise hard-to-find airports which are not served by a navigational facility. More advanced area navigation (random route) capability is available by using Loran-C, an inertial navigation system, or SBAS.

Inertial navigation system (INS). An INS is a self-contained system. It relies on highly accurate gyroscopes and accelerometers to produce an aircraft position solution, true heading, groundspeed, track, and attitude. The INS does this by continuously integrating accelerometer and angular rate sensor data into its position and attitude solution. The INS is a dead-reckoning sensor, and its position and attitude solution is subject to drift and error growth as well as to initialization error. An INS must be initialized before it is used for navigation.

Some INS installations have provisions for automatically updating (correcting) position information with input from GPS/SBAS, Loran-C, or DME stations via a Kálmán filter. Lacking such updates or data augmentation, INS position accuracy degrades in a predictable way with operating time. The INS is not considered sufficiently accurate to be used as an approach aid; however, an INS may be required for executing missed approaches, depending on terrain.

Prior to the introduction of FMS, INS provided steering information used by the autopilot to maintain the computed great circle route of flight. An INS was commonly used on long, overwater flights and over large, undeveloped land areas, and frequently used for point-to-point direct navigation in more developed regions. In modern installations, inertial reference systems (IRSs) are used in combination with FMSs and GPS/SBAS. The IRS is a sensor that measures orientation and acceleration, whereas an INS includes the computations required to convert these measurements into a navigation solution. *See* INERTIAL GUIDANCE SYSTEM.

Loran-C. Loran-C is a low-frequency, hyperbolic navigation system which operates by measuring the difference in arrival times of synchronized radio signals from transmitters located hundreds of miles apart. Inexpensive Loran-C receiver technology, system reliability, accuracy, and coverage have greatly increased the system's availability and acceptance at all levels of civil aviation. Stations transmitting on a common frequency make up a so-called chain which is composed of a master and two or more secondary transmitters. **Figure 3** shows a chain of stations which serves the southeastern United States and is typical of Loran-C installations.

Chain coverage is determined by the transmitted power from each station, their geometry including distance between stations, and their orientation. Within the nominal coverage range, 1000 nautical miles (1850 km), propagation is affected by the physical condition of the Earth's surface and the atmosphere.

In the United States, Loran-C is authorized by the Federal Aviation Administration (FAA) for use as a supplemental area navigation system over much of the 48 conterminous states. Because of the popularity of Loran-C as a highly accurate (to within 0.25 nautical mile or 0.5 km) navigation system,

the FAA has designed the National Aerospace System (NAS) to include Loran-C and nonprecision approaches at selected airports with adequate Loran-C coverage. Despite the extensive availability of Loran-C, however, it is not feasible to provide worldwide coverage because of range limits on the signals and the requirement that transmitters be land-based. Loran-C is not typically found on commercial air transport aircraft but on general aviation aircraft. *See* LORAN.

Microwave landing system (MLS). Like the ILS, MLS transmits two signals: one for runway alignment (localizer) and one for glidepath guidance. Unlike ILS, MLS allows multiple paths, including curved approaches to the runway, and variable glideslope angle selection. Those qualities make it particularly suited for airports located in mountain valleys or in noise-sensitive areas. It allows reduced intervals between aircraft to increase runway acceptance rates, and facilitates short-field operations for short and vertical takeoff and landing (STOL and VTOL) aircraft. MLS has not been accepted or deployed for civil aviation; its main use has been in military aircraft. *See* MICROWAVE LANDING SYSTEM (MLS); SHORT TAKEOFF AND LANDING (STOL); VERTICAL TAKEOFF AND LANDING (VTOL).

Nondirectional beacon (NDB). Low-frequency, nondirectional beacons are used in many parts of the world and are primarily associated with instrument approaches to airport runways. In less developed regions, they are also used as en route radio aids, which have a longer range and are less expensive to install and maintain than VORs but are less accurate. The NDB relies upon the directional properties of the aircraft's automatic direction finder (ADF) loop antenna, but bearings taken on an NDB are relatively unreliable when the station is more than a short distance away. This is especially true in the presence of heavy buildup of static electricity, which is common in the vicinity of thunderstorms. *See* DIRECTION-FINDING EQUIPMENT.

Radar. Navigation of civil aircraft by radar is largely limited to instructions provided by ground-based radar controllers. A common application is the use of radar vector headings to guide departing and arriving aircraft. Using radar, air-traffic controllers monitor airways, following all aircraft operated under instrument flight rules, and call deviations to the pilot's attention. Radar controllers also simplify long-range navigation across the United States by providing radar-guided direct routes to distant airports. *See* AIR-TRAFFIC CONTROL.

Surveillance radar may be used to provide instructions for a nonprecision approach to a specific runway. Using a radar map overlay and an identified return from the aircraft being controlled, the controller issues headings and recommended altitude instructions to the pilot so as to place the aircraft in a position from which a landing can be completed by visual reference to the ground.

Airborne weather radar has limited ground mapping capability which can be used as a backup to other navigation systems, particularly over coastal areas. Approach capabilities using airborne weather radar in the ground mapping mode have been established for helicopter operations at offshore oil-drilling platforms. *See* AIRBORNE RADAR.

Fig. 3. Loran-C chain GRI 7980 in the southeastern United States, comprising master station M and slave stations W, X, Y, and Z. A position fix is shown at point A, where the lines of position 7980-Y-45000 and 7980-Z-61000 intersect.

Tacan. TACAN (Tactical Air Navigation) is a primary aid to navigation for the U.S. Department of Defense, NATO, and a small number of civil users. It is designated for en route and terminal use, including naval shipboard, and operates on radio frequencies in the ultrahigh-frequency (UHF) band. It provides azimuth and distance information with reference to the station. The instrument indications are virtually the same as those provided by VOR; distance information is the same as that provided by a VORTAC. TACAN stations for en route use are usually colocated with VORs so that military aircraft which are unable to receive VORs can navigate on the same airway structure as civil aircraft.

Doppler velocity system (DVS) radar. Doppler velocity radar navigation is a self-contained dead-reckoning navigation system. It has been largely replaced on civil airplanes by more modern systems such as INS and GPS, but is still in used on military aircraft. The DVS can produce excellent velocity measurements, but it does track moving water currents. While this may be desirable in helicopter sea rescue, it is also the cause of navigation system error. *See* DOPPLER RADAR; ELECTRONIC NAVIGATION SYSTEMS; NAVIGATION.

John Studenny

Bibliography. D. J. Clausing, *Aviator's Guide to Flight Planning*, 1989; D. J. Clausing, *Aviator's Guide to Navigation*, 3d ed., 1997; A. Helfrick, *Principles of Avionics*, 3d ed., 2004; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., 1997;

C. F. Lin, *Modern Navigation, Guidance, and Control Processing*, 1991; M. S. Nolan, *Fundamentals of Air Traffic Control*, 4th ed., 2003; Radio Technical Commission for Aeronautics, *Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment*, RTCA/DO-229D, 2006; J. Sanderson, *Private Pilot Manual*, 1993; A. T. Wells, *Flight Safety*, 1992.

Air pollution

The presence in the atmosphere of natural and artificial substances that affect human health or well-being, or the well-being of any other specific organism. Air pollution also applies to situations where contaminants affect structures and artifacts or esthetic sensibilities (such as visibility or smell). Most artificial impurities are injected into the atmosphere at or near the Earth's surface. The lower atmosphere (troposphere) cleanses itself of some of these pollutants in a few hours or days as the larger particles settle to the surface and soluble gases and particles encounter precipitation or are removed through contact with surface objects. Removal of some particles (such as sulfates and nitrates) by precipitation and dry deposition results in acid deposition, which may cause serious environmental damage. Also, mixing of the pollutants into the upper atmosphere may dilute the concentrations near the Earth's surface, but can cause long-term changes in the chemistry of the upper atmosphere, including the ozone layer. *See* ATMOSPHERE; TROPOSPHERE.

Characteristics

All particulate matter and contaminant gases exist in the atmosphere in variable amounts. Typical natural sources of particles are salt particles from the oceans, volcanoes, and soil. Natural sources of gases are from active volcanoes, vegetation, and soils. Typical artificial contaminants are waste smokes and gases formed by industrial, municipal, transportation, and residential processes. Pollens, spores, and rusts are natural aerosols augmented artificially by human land-use practices. *See* ATMOSPHERIC CHEMISTRY; SMOKE.

There are a number of air pollutants that are regulated by the U.S. Environmental Protection Agency (EPA). The Clean Air Act, which was last comprehensively amended in 1990, requires the EPA to set National Ambient Air Quality Standards (NAAQS) for pollutants considered harmful to public health and the environment. The act established two types of national air quality standards. Primary standards set limits to protect public health, including the health of "sensitive" populations such as asthmatics, children, and the elderly. Secondary standards set limits to protect public welfare against decreased visibility and damage to animals, crops, vegetation, and buildings.

The EPA set National Ambient Air Quality Standards for six principal pollutants, which are called criteria pollutants (**Table 1**). Units of measure for the standards are parts per million (ppm) by volume, milligrams per cubic meter of air (mg/m^3), and micrograms per cubic meter of air ($\mu g/m^3$) at 25°C (77°F).

In addition to the criteria pollutants, certain fluorocarbon compounds (gases) significantly affect the ozone layer in the stratosphere. Also, trace gases and metals (such as dioxin, toxaphene, and mercury) that are emitted as by-products of industrial operations or refuse burning are suspected of being toxic at low concentrations. And some substances, such as benzene, are emitted by burning gasoline.

There are some gases that do not directly affect human health at ambient concentrations but pose a threat due to their capacity to absorb radiation. They

TABLE 1. National Ambient Air Quality Standards

Pollutant	Standard value	Standard type
Carbon monoxide (CO)		
8-h average	9 ppm (10 mg/m^3)	Primary
1-h average	35 ppm (40 mg/m^3)	Primary
Nitrogen dioxide (NO_2)		
Annual arithmetic mean	0.053 ppm (100 $\mu g/m^3$)	Primary and secondary
Ozone (O_3)*		
8-h average	0.08 ppm (157 $\mu g/m^3$)	Primary and secondary
Lead (Pb)		
Quarterly average	1.5 $\mu g/m^3$	Primary and secondary
Particulate <10 μm (PM_{10})		
Annual arithmetic mean	50 $\mu g/m^3$	Primary and secondary
24-h average	150 $\mu g/m^3$	Primary and secondary
Particulate <2.5 μm ($PM_{2.5}$)*		
Annual arithmetic mean	15 $\mu g/m^3$	Primary and secondary
24-h average	65 $\mu g/m^3$	Primary and secondary
Sulfur dioxide (SO_2)		
Annual arithmetic mean	0.03 ppm (80 $\mu g/m^3$)	Primary
24-h average	0.14 ppm (365 $\mu g/m^3$)	Primary
3-h average	0.50 ppm (1300 $\mu g/m^3$)	Secondary

*The 8-h average standard for ozone, and the standards for PM 2.5, have been remanded by the D.C. Circuit Count of Appeals for reconsideration by the agency.
†Parenthetical value is an approximately equivalent concentration.

TABLE 2. Global emissions for several important pollutants*

Pollutant	Anthropogenic sources	Natural sources	Annual emissions, 10^{12} oz/yr (10^{12} g/yr)	
			Anthropogenic	Natural
SO_2	Fossil fuel combustion	Volcanoes, biogenic processes	7.48 (212)	0.71 (20)
CO	Auto exhaust, general combustion	Forest fires, photochemical reactions	25 (700)	74 (2100)
NO, NO_2	Combustion	Biogenic processes in soil, lightning	2.7 (75)	6.3 (180)
CO_2	Combustion	Biological processes	776 (22,000)	10^4 (10^6)
Nonmethane hydrocarbons	Combustion	Biogenic processes in soil and vegetation	1.4 (40)	705 (20,000)

*Based on 1978 emissions figures.

are known as greenhouse gases and include carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), as well as many of the fluorocarbon compounds. Greenhouse gases absorb radiation emitted from the Earth's surface and, in increasing concentrations, are capable of increasing the atmosphere's temperature.

History. Air pollution has been a significant problem since the industrial revolution, when the intensive burning of coal and oil in centralized locations began. The problem was compounded because the population of the Earth had also been rapidly growing, and the growth in use of motor vehicles in the twentieth century added to the problem.

Air pollution regulations were enacted after the occurrence of three notorious episodes associated with light winds and reduced vertical mixing that persisted for several days. Many deaths were recorded in 1930 in the Meuse Valley in Belgium, in 1948 in Donora, Pennsylvania, and in 1952 in London. Smoke and sulfur dioxide (SO_2) concentrations measured during the episode in London were on the order of 10 times the current air quality standards, resulting in the passage of the British Clean Air Act. Through elimination of some sources and controls on others, the smoke concentration in London was cut by more than one-half in the decade following the Clean Air Act.

Types of sources. Sources may be characterized in a number of ways. A distinction may be made between natural and anthropogenic sources. Another frequent classification is in terms of stationary (power plants, incinerators, industrial operations, and space heating) and moving (motor vehicles, ships, aircraft, and rockets) sources. Another classification describes sources as point (a single stack), line (a line of sources), or area (city).

Different types of pollution are specified in various ways: gaseous, such as carbon monoxide, or particulate, such as smoke, pesticides, and aerosol sprays; inorganic, such as hydrogen fluoride, or organic, such as mercaptans; oxidizing substances, such as ozone, or reducing substances, such as oxides of sulfur and oxides of nitrogen; radioactive substances, such as iodine-131; inert substances, such as pollen or fly ash; or thermal pollution, such as the heat produced by nuclear power plants.

Air contaminants are produced in many ways and come from many sources. It is difficult to identify all the various producers. For some pollutants, such as carbon dioxide and methane, natural emissions sometimes far exceed anthropogenic emissions (**Table 2**).

Both anthropogenic and natural emissions are variable from year to year, depending on fuel usage, industrial development, climate, and volcanic activity. In some countries where pollution control regulations have been implemented, emissions have been significantly reduced. In dry regions, natural emissions of nitrogen oxides (NO_x), carbon dioxide (CO_2), and hydrocarbons can be greatly increased during a season with high rainfall and above-average vegetation growth. *See* NITROGEN OXIDES.

Most estimates of the methane budget put the anthropogenic component at about two-thirds. Ruminant production and emissions from rice paddies are regarded as anthropogenic because they result from human agricultural activities. Increases in carbon dioxide since the industrial revolution are also principally the result of human activities. These emissions have not yet equilibrated with the rest of the carbon cycle and so have had a profound effect on atmospheric levels, even though emissions from fossil fuel combustion are dwarfed by natural emissions.

An overlooked source of air pollution is the interior of homes, offices, and industrial buildings. Indoor air pollution has received increased attention since the discovery that concentrations of radon gas (naturally emitted from soils and structures) and household chemicals in indoor air can reach 5–10 times the levels in the dirtiest outside air. These high concentrations are caused by inadequate ventilation. As an example of controlling indoor air pollution, bans on cigarette smoking have proliferated since the late 1980s. *See* RADON.

The possibility of accidental release of toxic gases into the atmosphere was studied a great deal following the 1984 industrial accident in Bhopal, India, where over 2000 people died. Since then, the types of chemicals and source scenarios that might occur have been better defined.

Effects. In order to design regulations for air pollution, it is first necessary to estimate the effects on

humans and animals, vegetation, materials, and the atmospheric system. This assessment is not easy at low air pollution levels, where effects are subtle, occur gradually over long periods of time, or are the result of several confounding factors occurring simultaneously.

Humans and animals. The major concern with air pollution relates to its effects on humans. Since most people spend most of their time indoors, there has been increased interest in air pollution concentrations in homes, workplaces, and shopping areas. Much of the early information on health effects came from occupational health studies completed prior to the implementation of general air-quality standards.

Air pollution principally injures the respiratory system, and health effects can be studied through three approaches: clinical, epidemiological, and toxicological. Clinical studies use human subjects in controlled laboratory conditions, epidemiological studies assess human subjects (health records) in real-world conditions, and toxicological studies are conducted on animals or simple cellular systems. Epidemiological studies are the most closely related to actual conditions, but they are the most difficult to interpret because of the lack of control and the subsequent problems with statistical analysis. Another difficulty arises because of differences in response among different people. *See* EPIDEMIOLOGY.

The response of humans and animals to air pollution also depends on the exposure time (**Fig. 1**). It is well known that survival at much higher concentrations is possible if exposure time is only a few seconds. These short averaging times are of interest for accidental release of toxic chemicals. Conversely, low levels of air pollution can be harmful if there is continual exposure to them over a period of a year or more. *See* BIOMETEOROLOGY.

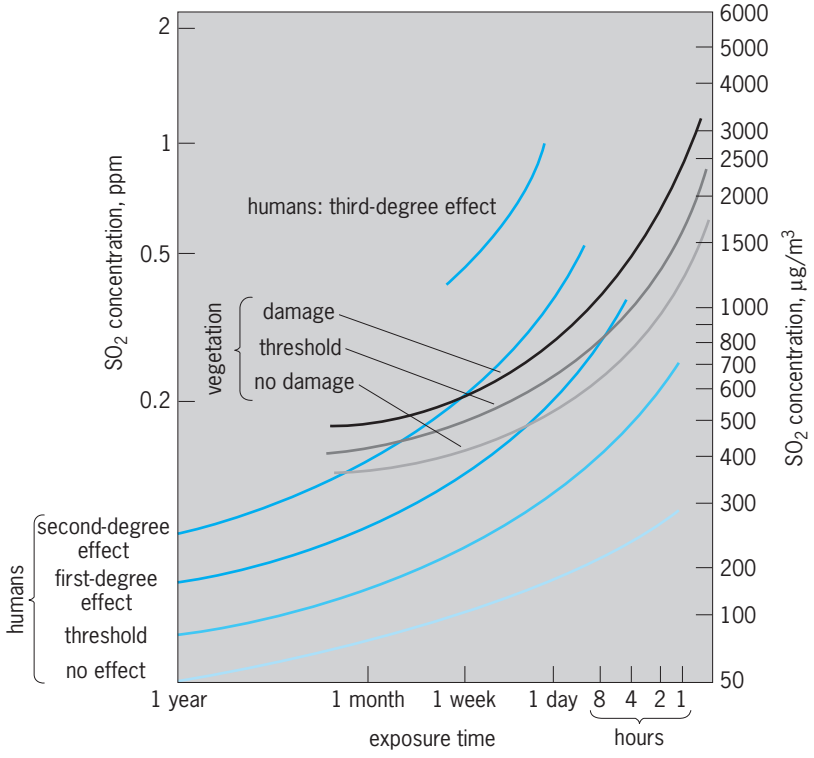


Fig. 1. Effects of sulfur dioxide on humans and vegetation. (*After L. J. Brasser et al., Sulphur Dioxide: To What Level Is It Acceptable?, Res. Inst. Pub. Health Eng. [Delft, Netherlands] Rep. G300, 1967*)

Vegetation. Damage to vegetation by air pollution is of many kinds. Sulfur dioxide may damage field crops such as alfalfa and trees such as pines, especially during the growing season (Fig. 1). Both hydrogen fluoride (HF) and nitrogen dioxide (NO_2) in high concentrations have been shown to be harmful to citrus trees and ornamental plants, which are of economic importance in central Florida. In addition, ozone and ethylene can cause damage to certain kinds of vegetation.

Since the early 1970s, the long-term effect of acid rain (rain with low pH caused by regional sulfate and nitrate pollution) on vegetation and fish has been recognized. Injury to conifer trees in Scandinavia and eastern North America has been documented, as well as fish kills in lakes in those regions. However, in most cases the damage can be blamed on other causes, such as climate change.

Materials. Corrosion of materials by atmospheric pollution is a major problem. Damage occurs to ferrous metals; to nonferrous metals, such as aluminum, copper, silver, nickel, and zinc; to building materials; and to paint, leather, paper, textiles, dyes, rubber, and ceramics. *See* CORROSION.

Atmospheric system. Gases emitted into the atmosphere can affect the dynamics of the atmosphere through changes in longwave and shortwave radiation processes. Particles can absorb or reflect incoming shortwave solar radiation, reducing the amount reaching the Earth's surface during the day. Greenhouse gases can absorb longwave radiation emitted by the Earth's surface and atmosphere.

Carbon dioxide, methane, fluorocarbons, nitrous oxides, ozone, and water vapor are important greenhouse gases that selectively absorb longwave radiation. This effect (the greenhouse effect) warms the temperature of the Earth's atmosphere and surface higher than would be found in the absence of an atmosphere. Because the amount of greenhouse gases in the atmosphere is rising, it is likely that the temperature of the atmosphere will gradually rise, possibly resulting in a general warming of the global climate over time. The carbon dioxide increase is primarily due to fossil fuel burning. It is very difficult to model the magnitude and distribution of the possible warming accurately because of complicated atmospheric responses and unknown natural sources and sinks, but it is estimated that the global temperature may rise as much as 5°C (8°F) by 2040. Because the warming would melt sea and land ice, the polar regions are believed to be warmed the most. As ice melts, the exposed land surface will be able to absorb more solar radiation than when it was ice covered, further exacerbating warming in that region. The warming would result in significant environmental effects, such as a rise in mean sea level of a few meters due to the

melting of ice sheets and a shift in vegetation patterns. *See* CLIMATE HISTORY; GREENHOUSE EFFECT.

Researchers are also concerned with pollution of the stratosphere (10–50 km or 6–30 mi above the Earth's surface) by aircraft and by broad surface sources. The stratosphere is important, because it contains the ozone layer, which absorbs part of the Sun's shortwave radiation and keeps it from reaching the surface. If the ozone layer is significantly depleted, an increase in skin cancer in humans is expected. Each 1% loss of ozone is estimated to increase the skin cancer rate 3–6%. *See* STRATOSPHERE.

Visibility is reduced as concentrations of aerosols or particles increase. The particles do not just affect visibility by themselves but also act as condensation nuclei for cloud or haze formation. In each of the three serious air pollution episodes discussed above, smog (smoke and fog) was present with greatly reduced visibility. Before emissions controls were imposed in the United States, visibility in many urban areas was often less than 1 km (0.6 mi). Even with the reductions in emissions in urban areas, a study by the U.S. Department of the Interior found that visibility in the eastern United States has been cut more than 50% in 40 years. Visibility is also an issue in pristine areas where the air historically has been very clean, with visibilities normally greater than 100 km (60 mi). If a power plant plume impacts such an area, the plume can be visible over the entire region. Likewise, the broad plumes from urban areas can significantly reduce visibilities in pristine regions, such as the Grand Canyon National Park, many hundreds of kilometers downwind. *See* SMOG.

Meteorology. The dynamics of the atmosphere will affect the concentration of air pollutants. It is often important to understand the physical processes leading to an observed concentration at a given point, or to model the expected concentration at the given point based on future emission scenarios. In both cases, it is necessary to estimate the fraction of the total emissions at the source that arrives at the receptor location in question. The physical processes relating to air quality are described below.

Wind transport. For distance scales up to regional scales (about 1000 km or 600 mi), the key question in understanding air quality variations in the transport wind is in what direction and how fast is the wind blowing. In general, for near-surface emissions, the surface air pollution concentrations will be greatest at low wind speeds. With elevated stacks, there is a trade-off between the diluting influence of plume rise (which is greater with lower wind speeds) and the diluting influence of greater ventilation (which is greater with higher wind speeds).

To estimate a source's contribution to a receptor, it is necessary to measure or estimate the wind flow patterns at and downwind of the source region. It is best if winds are measured at several locations between the sources and the receptor, since the wind field is often highly variable over distances of 10 km (6 mi) or more. But the wind field can also be extrapolated or interpolated from limited measurements, or can be modeled by using fundamental meteorological laws.

Trajectories are sometimes used to simulate the motion of the plume over multiple hours or days. Backward trajectories are used to identify the probable sources of observed pollutants; forward trajectories estimate the likely path of released pollutants. *See* WIND.

Stability. The vertical dispersion or spread of air pollution near the ground can be very rapid on sunny afternoons (unstable conditions), or can be nearly nonexistent on calm clear nights (stable conditions). In the first case the temperature decreases with height at a rate slightly greater than 1°C/100 m (0.55°F/100 ft), and in the second case the temperature can decrease with height at a slower rate than 1°C/100 m or even may increase with height (inversion). Neutral stability occurs during high-wind, cloudy conditions, when the temperature decreases with height at a rate of 0.98°C/100 m (0.53°F/100 ft). Specific patterns of plume dispersion are associated with each condition (**Fig. 2**). The terms stable and unstable refer to whether a parcel of air displaced upward adiabatically (with no exchange of heat with its environment) will tend to return to its original level or will be accelerated upward.

If the wind is blowing, the air near the ground will be turbulent even at night, but the turbulence will drop off rapidly with height so that the atmosphere has little turbulence at heights above 100 m (330 ft). During the day, the well-mixed turbulent layer adjacent to the ground is usually capped by an abrupt inversion at a height of above 1000 m (3300 ft), marking the tops of the thermals driven by surface heating. The atmosphere above this height shows little diurnal variability and is, in general, slightly stable. The depth of the turbulent air adjacent to the ground is commonly called the mixing depth, and is important because it marks the maximum vertical extent of diffusion in the mixed layer.

Dispersion. Dispersion is defined as the spread of pollutants caused by atmospheric turbulence, or

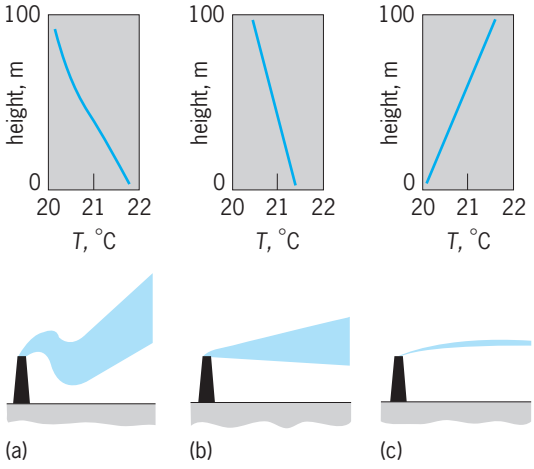


Fig. 2. Typical temperature variation with height and plume behavior for (*a*) unstable, (*b*) neutral, and (*c*) stable conditions. 1 m = 3.3 ft. °F = (°C × 1.8) + 32.

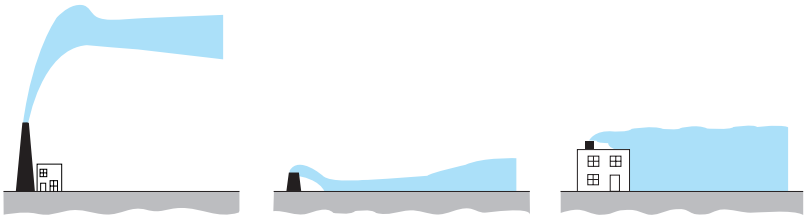


Fig. 3. Examples of the behavior of (*a*) a buoyant plume, (*b*) a dense gas plume, and (*c*) a plume being drawn into the aerodynamic wake of a building.

random fluctuations in wind velocity. Over the course of an hour during the daytime, the fluctuations in wind speed are about 20% of the average speed of the wind. These fluctuations are caused by random whorls or eddies of air that include scales of motion from about 0.1 mm to a few kilometers for vertical eddies and up to hundreds of kilometers for horizontal eddies. The rate of dispersion is faster when the size of the plume is smaller than the size of the eddies affecting the plume.

In the United States, the EPA offers a series of regulatory models for the purpose of estimating the impact of specific sources or groups of sources on downwind receptors. These models assume an averaging time of an hour or longer, and that the dispersion within the hour is random so the plume can be described as a straight line with concentration decreasing from the centerline. The rate of decrease of concentration laterally and vertically from the centerline of the plume is described by a statistically "normal" (Gaussian) distribution with standard deviations that increase with downwind distance at a rate dependent on the atmospheric stability.

Turbulence and dispersion are enhanced over rough surfaces, such as forests, hills, or urban areas, because of their larger surface friction. For a given wind speed, dispersion is about four times as great over a rough urban surface as over a smooth water surface. Dispersion is even more strongly influenced by stability, and vertical plume spread can increase by a factor of 10 or more from night to day.

There has been much research to improve dispersion estimates in situations where maximum concentrations are observed for pollutants emitted from tall stacks. For example, highest short-term (1 h average) ground-level concentrations in flat terrain around tall stacks are observed during light-wind, sunny, daytime conditions when vigorous eddies frequently bring the plume to the ground near the stack. Another serious condition occurs in mountainous terrain where plumes during stable conditions may affect the terrain directly, leading to high concentrations in a narrow belt on the mountainside.

In the case around tall stacks during light-wind, sunny, daytime conditions, it has been observed that, on the average, the plume will slowly be moved toward the ground as the areas of downward-moving air are larger than the areas of more rapidly upward-moving thermals. Depending on the height of the tall stack, plumes will affect the surface with a greater probability than is predicted by using the Gaussian plume simulation techniques.

Air pollution variability. The atmosphere is characterized by intense turbulence over a wide range of space or time scales. Just as instantaneous measurements of the wind speed will vary randomly about their long-term mean, so will air pollution measurements. This variability is greatest for instantaneous measurements and slowly decreases as averaging time increases. For a given hourly average wind speed and stability, the hourly average concentration from a source could vary by a factor of 2 from one day to the next because of turbulence. This natural variability imposes a limit on the accuracy of any transport and dispersion model.

Source effects. Before atmospheric dispersion has a chance to act on a plume, there are sometimes significant source effects due to differences between the density and velocity of the effluent and the ambient air. Most stack plumes are much warmer and hence less dense than the ambient air, and also usually have a significant upward velocity (on the order of 10 m/s or 33 ft/s) out of the stack. In this case, the plume can rise as much as several hundred feet above the stack before leveling off (**Fig. 3***a*). For constant source conditions, plume rise decreases as ambient stability and wind speed increase.

Sometimes the effluent is denser than the ambient air (for example, cold plumes or plumes containing gases that are denser than air), in which case the plume sinks toward the ground. Other dense gas sources may occur as a result of accidental rupturing of gas tanks. For example, a dense gas plume hugs the ground until it becomes sufficiently dilute to be dispersed by ambient turbulence (Fig. 3*b*).

At many short industrial stacks, the possibility exists that the plume may be drawn into the aerodynamic wake caused by the obstruction of the airflow by a nearby building. This may happen if the stack height is less than about 2.5 times the height of the structure (Fig. 3*c*). The ground-level concentration of air pollutants will increase because the plume is "downwashed" to the ground, but the average concentration in the plume is less because of the increased ambient turbulence.

Chemistry. Air pollution can be divided into primary and secondary compounds, where primary pollutants are emitted directly from sources (for example, carbon monoxide and sulfur dioxide) and secondary pollutants are produced by chemical reactions between other pollutants and atmospheric gases and particles (for example, sulfates and ozone). Most of the chemical transformations are best described as oxidation processes. In many cases, these secondary pollutants can have significant environmental effects, such as acid rain and smog.

Photochemistry. Los Angeles smog is a well-known example of secondary pollutants formed by photochemical processes, as a result of primary emissions of nitric oxide (NO) and reactive hydrocarbons from anthropogenic sources such as transportation and industry as well as natural sources. Energy from the

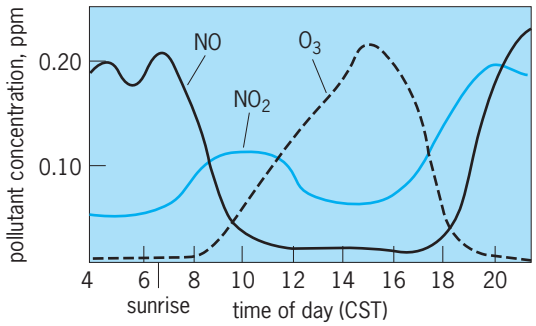


Fig. 4. The strong dependence of pollutant concentration on the intensity of sunlight. Note that ozone (O_3) concentration is a maximum in the afternoon, and zero at night. (*After A. C. Stern et al., Fundamentals of Air Pollution, 2d ed., Academic, 1984*)

Sun causes the formation of nitrogen dioxide, ozone (O_3), and peroxyacetalnitrate, which cause eye irritation and plant damage (**Fig. 4**).

Acid rain. The study of acid rain grew dramatically in the 1980s; prompted by the discovery of damage to sensitive lakes in areas where acidity of rain was the strongest. It has been shown that when emissions of sulfur dioxide and nitrogen oxide from tall power plant and other industrial stacks are carried over great distances and combined with emissions from other areas, acidic compounds can be formed by complex chemical reactions. In the absence of anthropogenic pollution sources, the average pH of rain is around 5.6 (slightly acidic). In the eastern United States, acid rain with a pH less than 5.0 has been measured and consists of about 65% dilute sulfuric acid, 30% dilute nitric acid, and 5% other acids (**Fig. 5**).

The chemistry of acid rain involves phase changes, catalysts, and aerosol formation, and is greatly influenced by cloud processes. A major component of sulfate deposition onto the surface is due to dry deposition, which occurs during all periods when precipitation is absent. Dry deposition is caused by impaction of particles or absorption of gases by the ground, water, or vegetation surface. It is possible that the chemical reactions are such that sulfate deposition is linearly related to emissions, implying that a 50% cut in industrial sulfur emissions would result in a 50% decrease in sulfate deposition. *See* ACID RAIN.

Particles. Particles in the atmosphere are characterized by size. The EPA has designated particles as supercoarse, coarse, fine, and ultrafine (**Table 3**). These distinctions generally represent differing sources and varying potentials for the particle to behave in the atmosphere and to affect human health. Supercoarse particles generally come from natural sources, including windblown sands and soil, volcanic activity, sea salt, and pollen. They are considered less of a human health threat since they have an appreciable mass and tend to fall out of the atmosphere quickly upon release. Moreover, the nose is reasonably efficient in removing these particles before they can reach the lungs.

Particulate matter less than approximately 10 micrometers (PM_{10}) can penetrate the lower respiratory tract. Course particles are a mix of natural and anthropogenic sources.PM_{10} is regulated as a specific

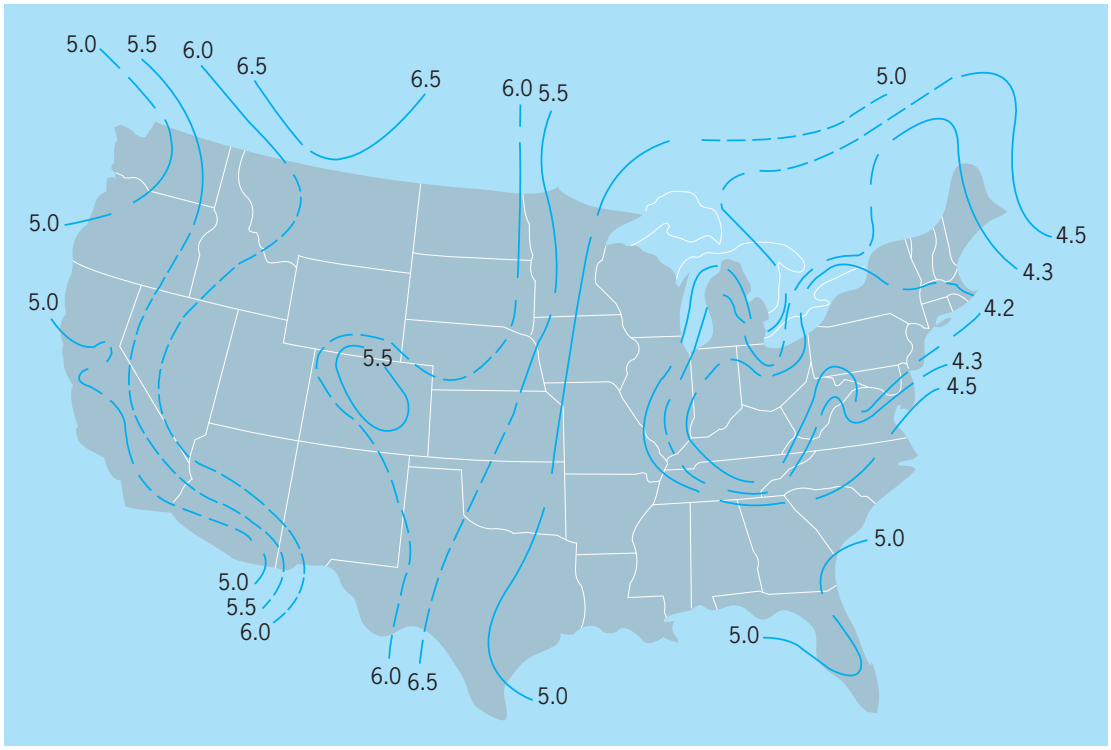


Fig. 5. Average pH of rain in the United States, showing the broad area of low pH in the northeastern United States, downwind of the large power plants in the Ohio Valley. A change in pH of 1.0 indicates a factor-of-10 change in acidity. (*After J. Wisniewski and E. L. Keitz, Acid rain deposition patterns in the continental United States, Air Soil Pollut., 19:327–339, 1983*)

TABLE 3. Terminology for particle sizes

Description	Particle size
Supercoarse	$D_p > 10\ \mu m$
Coarse	$2.5\ \mu m < D_p < 10\ \mu m$
Fine	$0.1\ \mu m < D_p < 2.5\ \mu m$
Ultrafine	$D_p < 0.1\ \mu m$

type of pollutant because this size range is considered respirable. The particle-size range between 0.1 and 10 μm is especially important in air pollution studies. A major fraction of the particulate matter generated in some industrial sources is in this size range.

Fine particles tend to be of anthropogenic origin and the result of particle growth from the ultrafine particles. Particles of this size, also referred to as the accumulation mode, grow through a cycle that can include condensation of flue gases, coagulation of smaller particles, and processes involved in the formation and evaporation of cloud droplets. Some of these particles have chemical characteristics that attract water vapor (hygroscopic) and will form "cloud condensation nuclei"—the basis for cloud droplet growth. Chemical reactions inside the cloud droplets and the collision and coalescence of the droplets with subsequent evaporation are believed to speed the formation of these size particles.

EPA chose 2.5 μm as the partition between fine and coarse particulate matter. Particles less than approximately 2.5 μm are regulated as $PM_{2.5}$. Air emission testing and air pollution control methods for $PM_{2.5}$ particles are different from those for coarse and supercoarse particles.

$PM_{2.5}$ particles settle quite slowly in the atmosphere, relative to coarse and supercoarse particles. Normal weather patterns can keep $PM_{2.5}$ particles airborne for several hours to several days and enable these particles to cover hundreds of miles. $PM_{2.5}$ particles can cause health problems due to their potentially long airborne retention time and the inability of the human respiratory system to defend itself against particles of this size.

In addition, the chemical makeup of $PM_{2.5}$ particles is quite different from coarse and supercoarse particles. EPA data indicate that $PM_{2.5}$ particles are composed primarily of sulfates, nitrates, organic compounds, and ammonium compounds. The EPA also determined that $PM_{2.5}$ particles often contain acidic materials, metals, and other contaminants believed to be associated with adverse health effects.

Particles less than 1 μm in diameter are called ultrafine (or submicrometer) particles and can be the most difficult size to collect. Particles in the range 0.2–0.5 μm are common in many types of combustion, waste incineration, and metallurgical sources. Particles in the range 0.1–1.0 μm are important because they can represent a significant fraction of the particulate emissions from some types of industrial sources and because they are relatively hard to collect.

Particles can be much smaller than 0.1 μm. In fact, particles composed of as little as 20 to 50 molecules clustered together can exist in a stable form. Some industrial processes, such as combustion and metallurgical sources, generate particles in the range 0.01–0.1 μm. These sizes are approaching the size of individual gas molecules, which are in the range 0.0002–0.001 μm. However, particles in the size range 0.01–0.1 μm tend to agglomerate rapidly to yield particles in the >0.1-μm range. Accordingly, very little of the particulate matter entering an air pollution control device or leaving the stack remains in the size range 0.01–0.1 μm.

Visibility reduction. Particles in the atmosphere are distributed over a wide range of sizes and shapes. They can be divided into Aiken particles (less than 0.2 μm), fine particles (0.2–2 μm), and large particles (greater than 2 μm). The Aiken particles are, in general, the result of condensation of hot gases within plumes. They often coagulate onto the fine particles through Brownian motion. The fine particles are too small to be removed from the atmosphere through sedimentation, and this size becomes the point of accumulation for particles in the atmosphere. Sources for the large particles can be either anthropogenic or natural, but the ratio of natural particles is highest for this size category.

The large particles have a size range that overlaps with that of the wavelengths of visible light and as such are very efficient at scattering light. Consequently, large particles play a significant role in reducing visibility. Research has shown the existence of large regions of low-visibility air that can be tracked from day to day as it moves around the eastern United States in the summertime. These regions are thought to be largely the result of the chemical production of sulfate particles from sulfur dioxide that is exacerbated by the presence of water vapor.

In the western United States, where the visibility is generally better than in the eastern half, individual plumes have been shown to degrade visibility in specific National Parks on specific days. However, the problem of plume impaction by aggregates of plumes from urban areas in the summertime in National Parks such as the Grand Canyon and Yosemite is significant and more difficult to regulate.

Air Quality Index. In an effort to create a more accessible measure of air quality, the EPA developed the Air Quality Index (AQI). The AQI indicates how clean or polluted the air is, as well as the associated health effects. The AQI focuses on health effects experienced within a few hours or days after breathing polluted air. The EPA calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide. For each of these pollutants, the EPA has established national air quality standards to protect public health (Table 1).

The AQI can be considered a yardstick that runs from zero to 500. The higher the AQI value, the greater the level of air pollution and the greater the

health concern. For example, an AQI value of 50 represents good air quality with little potential to affect public health, while an AQI value over 300 represents hazardous air quality.

An AQI value of 100 generally corresponds to the national air quality standard for the pollutant, which is the level EPA has set to protect public health. AQI values below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is considered to be unhealthy—at first, for certain sensitive groups of people, then for everyone as AQI values get higher.

Stratospheric ozone layer. The stratospheric ozone layer absorbs much of the Sun's ultraviolet radiation, and any depletion of this ozone could have serious effects on animals and vegetation. Several types of air pollutants may attack the ozone layer, including fluorocarbons and nitrogen oxides. Photochemical reactions with hydrocarbons, nitrogen oxides, and ozone may lead to a new chemical balance in the stratosphere. The possibility of these changes was thought to have diminished because of a reduction in fluorocarbon use during the 1970s. However, in the mid-1980s observations of ozone at great elevations in the atmosphere suggested that there was a steady depletion, with major losses over Antarctica. *See* STRATOSPHERIC OZONE.

Accidental chemical releases. The chemical thermodynamics of some accidental releases are quite complicated. For example, if liquid chlorine is stored in a tank at high pressure and a nozzle ruptures, the releases will consist of two phases: liquid chlorine and gaseous chlorine. The evaporation process can lead to extreme cooling of the cloud close to the source.

Steven R. Hanna; Perry J. Samson

Controls

Air pollution controls have been designed for both mobile sources (mainly motor vehicles) and stationary sources.

Mobile sources. Air pollution control technologies for mobile sources involve vehicles fueled by gasoline and by diesel fuel. In addition, the fuel quality is an important factor.

Gasoline-fueled vehicles. A wide variety of emissions-reduction technologies has been developed for gasoline-fueled vehicles. Before controls were required, engine crankcases were vented directly to the atmosphere. Crankcase emissions controls, which basically consist of closing the crankcase vent port, were included in designs for automotive vehicles manufactured in the United States beginning in the early 1960s, with the result that control of these emissions is no longer considered a serious technical concern.

Exhaust emissions of hydrocarbons, carbon monoxide, and nitrogen oxides are related to the air–fuel mixture inducted, the peak temperatures and pressures in each cylinder, and other engine design parameters. Variations in these parameters are, therefore, capable of causing significant increases or decreases in emissions. *See* GASOLINE.

Dilution of the incoming charge has been shown to reduce peak cycle temperature by slowing flame speed and absorbing some heat of combustion. Recirculating a portion of the exhaust gas back into the incoming air–fuel mixture, thereby lowering peak cycle temperature, is used to lower the concentration of nitrogen oxides in the emissions. Improvements in mixture preparation, air-fuel intake systems, and ignition systems can increase dilution tolerance. It can also be increased by increasing the burn rate or flame speed of the air-fuel charge.

Both electronic control and exhaust aftertreatment devices are utilized to reduce air pollution.

1. *Electronic controls.* With so many interrelated engine design and operating variables playing an increasingly important role in the modern automotive engine, the control system has taken on increased importance. Modifications in spark timing must be closely coordinated with air–fuel ratio changes and degrees of exhaust gas recirculation to prevent significant decreases in fuel economy or performance which may result from emissions reductions, or to prevent an increase of NO_x emissions as carbon monoxide (CO) emissions decrease. In addition, controls that are selective in responding to engine load or speed have been found beneficial in preventing widespread adverse impacts resulting from emissions of pollutants. Therefore, electronic controls have replaced more traditional mechanical controls. For example, electronic control of ignition timing has demonstrated an ability to optimize timing under all engine conditions, and has the added advantage of reduced maintenance requirements and improved durability compared with mechanical systems. When it is coupled with electronic control of exhaust gas recirculation, NO_x emissions can be reduced without affecting fuel economy, in some cases resulting in improvements. *See* IGNITION SYSTEM.

2. *Exhaust aftertreatment devices.* When stringent exhaust emissions standards (especially for hydrocarbons or nitrogen oxides) are mandated, exhaust aftertreatment devices such as catalytic converters tend to be used to supplement engine modifications. An oxidation catalyst is a device that is placed on the tailpipe of a car and that, if the chemistry and thermodynamics are correct, will oxidize almost all the hydrocarbons and carbon monoxide in the exhaust stream. Three-way catalysts (which can lower levels of hydrocarbons, carbon monoxide, and nitrogen oxides simultaneously) were introduced in the United States in 1977, and subsequently became widely used when the United States standard for nitrogen oxides was lowered to 1.0 g per mile. To work effectively, these catalysts require precise control of air–fuel mixtures. As a result, three-way systems have indirectly fostered improved air-fuel management systems, such as throttle body fuel injection systems as well as electronic controls. One unique advantage of catalysts is their ability to selectively eliminate some of the more hazardous compounds in vehicle exhaust such as aldehydes, reactive hydrocarbons, and polynuclear

hydrocarbons. *See* AUTOMOBILE; CATALYSIS; CATALYTIC CONVERTER; FUEL INJECTION; INTERNAL COMBUSTION ENGINE.

Diesel-fueled vehicles. While the major technical problems associated with reducing emissions from gasoline-fueled cars have been solved, it is apparent that reductions from these vehicles alone are not sufficient to solve the air pollution problems in many areas. Therefore, diesel trucks and buses have received attention as significant sources of particulates and nitrogen oxides. Degradation of emissions is very low for diesel-fueled vehicles compared to gasoline-fueled ones; and diesel vehicles have inherently low hydrocarbon emissions, although these hydrocarbons are higher in molecular weight and thus of a different character from those emitted by gasoline engines. Further, uncontrolled diesel engines emit objectionable exhaust odors, which are a frequent source of complaints from the public. In the United States, standards have been adopted for these vehicles, which foster technological developments similar to those that have already been achieved for gasoline-fueled vehicles.

Smoke emissions from diesel engines are composed primarily of unburned carbon particles from the fuel. This type of emission usually results when there is an excess amount of fuel available for combustion. This condition is most likely to occur under high engine load conditions, such as acceleration and engine lugging, when the engine needs additional fuel for power. Failure to clean or replace a dirty air cleaner, a common maintenance error, may produce a large quantity of smoke emissions; a dirty air cleaner can choke off available air to the engine, resulting in a lower-than-optimum air–fuel mixture. The manner in which the vehicle is operated can also be important, since smoke emissions from diesel engines are minimized by selection of the proper transmission gear to keep the engine operating at the most efficient speeds. Moderate accelerations and fewer changes in highway cruising speed as well as reduced speed for hill climbing also minimize smoke emissions.

Basic approaches to diesel engine emission control fall into three major categories: (1) engine modifications, including combustion chamber configuration and design, fuel-injection timing and pattern, turbocharging, and exhaust gas recirculation; (2) exhaust aftertreatment, including traps, trap oxidizers, and catalysts; and (3) fuel modifications, including control of fuel properties, fuel additives, and alternative fuels. *See* COMBUSTION CHAMBER; TURBOCHARGER.

Design strategies for controlling emissions of nitrogen oxides in diesel-fueled vehicles include variable injection timing and pressure, charge cooling (cooling of the incoming air to make it more dense), and exhaust gas recirculation. Retarding injection timing, while a well-known method of reducing formation of nitrogen oxides, can lead to increases in fuel consumption and emissions of particulates and hydrocarbon. These problems can be mitigated by varying the injection timing with engine load or speed. Also, high-pressure injection can reduce these problems. If this injection is coupled with electronic controls, emissions of nitrogen oxides can be reduced significantly with a simultaneous improvement in fuel economy (although not as great as could occur if electronics were added without any emission requirements).

With relatively lenient particulate standards, engine modifications are generally sufficient to lower engine-out emission levels (emissions out of the engine but before they would go into some type of aftertreatment device; in this case, synonymous with tailpipe emission levels). The necessary modifications include changes in combustion chamber design, fuel injection timing and spray pattern, turbocharging, and the use of exhaust gas recirculation. Further particulate controls are possible through greater use of electronically controlled fuel injection. With such a system, signals proportional to fuel rate and piston advance position are measured by sensors and are electronically processed by the electronic control system to determine the optimum fuel rate and timing.

One of the important factors in diesel control design is that emissions of nitrogen oxides, particulates, and hydrocarbons are closely interdependent. For example, retarded fuel timing within certain ranges decreases emission of nitrogen oxides and particulates but can lead to increases in emission of hydrocarbons. As technology has advanced, these potential trade-offs have diminished. For example, engine designs such as modifications in the combustion chamber and electronically controlled fuel injection have resulted in simultaneous reduction in emissions of hydrocarbons, particulates, and nitrogen oxides.

Exhaust aftertreatment methods include traps, trap oxidizers, and catalysts. Prototypes of trap oxidizer systems have achieved 70–90% reductions from engine-out particulate emissions rates and, with proper regeneration, have demonstrated the ability to achieve these rates for high milage. Basically, all these designs rely on trapping a major portion of the engine-out particles and consuming them before they accumulate sufficiently to saturate the filter and cause problems in fuel economy and performance or other difficulties. *See* DIESEL ENGINE.

Improved fuel quality. Improvements in fuel quality are a significant factor in control of air pollution from mobile sources. Modifications to diesel fuel composition are a quick and cost-effective means of reducing emissions from existing vehicles. The two modifications that show the most promise are reductions in sulfur content and in the fraction of aromatic hydrocarbons in the fuel. In the United States, the EPA has reduced sulfur content in diesel fuel to a maximum of 0.05% by weight.

Lowering the sulfur content of diesel fuel not only causes a direct reduction in emissions of sulfur dioxide (SO_2) and sulfate particles but also reduces the indirect formation of sulfate particles from SO_2 in the atmosphere. In Los Angeles, it is estimated that each pound of SO_2 emitted results in roughly 2.2 kg (1 lb) of fine particulate matter in the atmosphere.

In this case, therefore, the indirect particulate emissions due to SO_2 from diesel vehicles are roughly equal to their direct particulate emissions. Conversion of SO_2 to particulate matter is highly dependent on local meteorological conditions, so the effects could be greater or less in other cities.

A reduction in the aromatic hydrocarbon content of diesel fuel may also help to reduce emissions, especially where fuel aromatic levels are high. For existing diesel engines, a reduction in aromatic hydrocarbons from 35% to 20% by volume would be expected to reduce transient particulate emissions by 10–15% and emissions of nitrogen oxides by 5–10%. Hydrocarbon emissions, and possibly the mutagenic activity of the particulate soluble organic fraction, would also be reduced. Aromatic reductions of this magnitude can often be obtained through alterations in methods of diesel fuel production and blending strategy without a need for major new investments in additional processing capacity.

Reduced aromatic content in diesel fuels would have other environmental and economic benefits. The reduced aromatic content would improve the fuel's ignition quality, facilitating cold starting and idling performance and reducing engine noise. The reduction in the use of catalytically cracked blending stocks should also have a beneficial effect on deposit-forming tendencies in the fuel injectors, and would reduce maintenance costs. On the negative side, the reduced aromatics might result in some impairment of cold flow properties, due to the increased paraffin content of the fuel.

Detergent additives in diesel fuel have the ability to prevent and remove injector tip deposits, thus reducing smoke levels. The reduced smoke probably results in reduced particulate emissions as well, but this has not been demonstrated as clearly because of the great expense of particulate emissions tests on in-use vehicles. Cetane-improving additives can also produce some reduction in emissions of hydrocarbons and particulates in marginal fuels. *See* CETANE NUMBER.

Alternative fuels. The possibility of substituting cleaner-burning fuels for diesel fuel has drawn attention since the mid-1980s. The advantages provided by this substitution include conservation of oil products and the security of readily available energy sources, as well as the reduction or elimination of particulate emissions and visible smoke. The most promising of the alternative fuels are natural gas, methanol made from alcohol, and, in limited applications, liquefied petroleum gas (LPG).

1. *Natural gas.* Natural gas has many desirable qualities as an alternative to diesel fuel in heavy-duty vehicles. Clean burning, cheap, and abundant in many parts of the world, it already plays a significant role as an energy source for vehicles in a number of countries. The major disadvantage is its gaseous form at normal temperatures. Pipeline-quality natural gas is a mixture of several different gases, but the primary constituent is methane, which typically makes up 90–95% of the total volume. Methane is a nearly ideal fuel for Otto cycle (spark ignition) engines. As a gas under normal conditions, it mixes readily with air in any proportion, eliminating cold-start problems and the need for cold-start enrichment. It is flammable over a fairly wide range of air–fuel ratios. With a research octane number of 130 (the highest of any commonly used fuel), it can be used with engine compression ratios as high as 15:1 (compared to 8–9:1 for gasoline), thus giving greater efficiency and power output. The low lean flammability limit permits operation with extremely lean air-fuel ratios, having as much as 60% excess air. However, its high flame temperature tends to result in high emissions of nitrogen oxides, unless very lean mixtures are used.

Because of its gaseous form and poor self-ignition qualities, methane is a poor fuel for diesel engines. Since diesels are generally somewhat more efficient than Otto cycle engines, natural gas engines are likely to use somewhat more energy than the diesels they replace. The high compression ratios achievable with natural gas limit this efficiency penalty to about 10% of the diesel fuel consumption, however. *See* LIQUEFIED NATURAL GAS (LNG); METHANE; NATURAL GAS; OTTO CYCLE.

2. *Liquefied petroleum gas.* Liquefied petroleum gas is widely used as a vehicle fuel in the United States, Canada, the Netherlands, and elsewhere. As a fuel for spark ignition engines, it has many of the same advantages as natural gas, with the additional advantage of being easier to carry aboard the vehicle. Its major disadvantage is the limited supply, which would rule out any large-scale conversion to liquefied petroleum gas. *See* LIQUEFIED PETROLEUM GAS (LPG).

3. *Methanol.* Methanol has many desirable combustion and emissions characteristics, including good lean combustion characteristics, low flame temperature (leading to low emissions of nitrogen oxides), and low photochemical reactivity. As a liquid, methanol can either be burned in an Otto cycle engine or be injected into the cylinder in a diesel. With a fairly high octane number of 112 and excellent lean combustion properties, methanol is a good fuel for lean-burn Otto cycle engines. Its lean combustion limits are similar to those of natural gas, while its low energy density results in a low flame temperature compared to hydrocarbon fuels, and thus lower emissions of nitrogen oxides. Methanol burns with a sootless flame and contains no heavy hydrocarbons. As a result, particulate emissions from methanol engines are very low, consisting essentially of a small amount of unburned lubricating oil. Methanol's high octane number results in a very low cetane number, so that methanol cannot be used in a diesel engine without some supplemental ignition source. The use of ignition-improving additives, spark ignition, glow plug ignition, or dual injection with diesel fuel in converted heavy-duty diesel engines has been demonstrated. *See* ALCOHOL FUEL; METHANOL; OCTANE NUMBER.

Methanol combustion does not produce soot, so particulate emissions from methanol engines are limited to a small amount derived from lubricating oil. Methanol's flame temperature is also lower than that

for hydrocarbon fuels, resulting in emissions of nitrogen oxides that are typically 50% lower. Emissions of carbon monoxide are generally comparable to or somewhat greater than those from a diesel engine (except for stoichiometric Otto cycle engines, for which emissions of carbon monoxide may be much higher). These emissions can be controlled with a catalytic converter, however.

The major pollution problems associated with methanol engines come from emissions of unburned fuel and formaldehyde. Methanol (at least in moderate amounts) is relatively innocuous—it has low photochemical reactivity—and, while acutely toxic in large doses, displays no significant chronic toxicity effects. Formaldehyde, the first oxidation product of methanol, is much less benign. A powerful irritant and suspected carcinogen, it also displays very high photochemical reactivity. Although all internal combustion engines produce some formaldehyde, some early designs of methanol engines exhibited greatly increased formaldehyde emissions as compared to those from diesel engines. The potential for large increases in emissions of formaldehyde with the widespread use of methanol vehicles has raised considerable concern about what would otherwise be a very benign fuel from an environmental standpoint. Formaldehyde emissions can be reduced through changes in the design of the combustion chamber and the injection system; they are also readily controllable through the use of catalytic converters, at least under warmed-up conditions. *See* FORMALDEHYDE.

Michael P. Walsh

Stationary sources: volatile and hazardous pollutants. Volatile organic compounds and hazardous air pollutants come from many sources. In the United States, hazardous air pollutants are identified in Title III of the Clean Air Amendments Act of 1990, as pollutants that present, or may present, through inhalation or other routes of exposure, a threat of adverse human health effects. Volatile organic compounds react in the atmosphere to form photochemical oxidants (including ground-level ozone) that affect health, damage materials, and cause crop and forest losses. In the United States, the EPA has established a 120-parts-per-billion ambient ozone standard (1-h average, daily maximum) and an 80-parts-per-billion standard (8-h average). The 120-ppb standard applies only to areas that were designated nonattainment when the 80-ppb standard was adopted in July 1997. Millions of people live in areas where the ozone standards are routinely exceeded. Hazardous air pollutants can also cause detrimental effects on human health.

Controlling the emissions of volatile organic compounds and hazardous air pollutants is important but complex, because there are many sources that usually emit mixtures of compounds. Because organic compounds may undergo complex reactions in the atmosphere, there are a number of technologies designed to control emissions. Often, it is preferable to prevent the formation and emission of organic compounds rather than attempting to control them. Organic emissions can be reduced or prevented by using alternative chemicals, by modifying chemical processes, and by changing the products being used. In addition to preventing pollution, such measures can save energy and reduce costs.

Conventional control technologies. These include thermal incinerators, catalytic oxidizers, adsorbers, flares, boilers and industrial furnaces, and condensers.

1. *Thermal incinerators.* Thermal incinerators raise the temperature of emission streams so that organic compounds are destroyed; destruction efficiencies can exceed 99%. Thermal incinerators are typically applied to emission streams that contain dilute mixtures of volatile organic compounds and hazardous air pollutants. The effectiveness of a thermal incinerator is largely dependent on the temperature in the combustion chamber. Supplementary fuels (natural gas or fuel oil) must be used to maintain desired combustion temperatures. Thermal incinerators are available in many sizes.

2. *Catalytic oxidizers.* Catalytic oxidizers use a catalyst to speed up the reaction of organic compounds at temperatures lower than those required by thermal incinerators, thus reducing requirements for supplementary fuel. Destruction efficiencies are in the 95% range, but they can be higher if more catalyst and supplementary fuel are used. The catalyst may be fixed (fixed bed) or mobile (fluid bed). Catalytic oxidation is not as widely used as thermal incineration because some materials may damage the catalyst. Even with proper usage, the catalyst must be replaced periodically. Catalytic oxidizers are available in the same size range as thermal incinerators.

3. *Adsorbers.* Adsorbers employ a bed packed with an adsorbent material that captures volatile organic compounds and hazardous air pollutants. The most common adsorbent is carbon, but other materials, such as silica gel, alumina, and polymers, are also used. Removal efficiencies up to 99% can be obtained in adsorbers, and they provide the option of reclaiming the adsorbed material for reuse. There are many important organic compounds for which carbon adsorption is quite effective. Some examples are toluene, benzene, and methyl ethyl ketone. Carbon adsorbers are sensitive to excessive humidity, emission stream temperatures above 54°C (130°F), and inlet organic compound concentrations of 10,000 ppm and above. *See* ADSORPTION.

4. *Absorbers.* Absorbers are control devices that provide contact between the emission stream and a carefully selected solvent in which the volatile organic compounds and hazardous air pollutants readily dissolve. When the solvent has absorbed the pollutants, the solvent must be regenerated or disposed of. Regeneration of the solvent can be expensive, and disposal can be a problem because of the potential for surface-water or ground-water contamination. Thus, absorbers are often used in conjunction with other control devices such as thermal incinerators. *See* ABSORPTION.

5. *Flares.* Flares are emission control devices that destroy volatile organic compounds and hazardous air pollutants by combustion. Often, supplemental fuel is added to maintain the combustion process

and to destroy the pollutants. Flares are often chosen as emission control devices for volatile organic compounds and hazardous air pollutants when emission stream flow is uncertain or intermittent, as in process upsets and emergencies.

6. *Boilers and industrial furnaces.* In-place boilers and industrial furnaces can control emissions of volatile organic compounds and hazardous air pollutants. Not only can these units control an emission stream that is the main fuel supply, but they can also control a waste stream that has a small flow rate in comparison to the flow rate of the fuel–air mixture used to fire the unit. Since proper functioning of such systems is essential to the proper operation of a plant, their use as control devices for volatile organic compounds and hazardous air pollutants must be carefully monitored to avoid adverse impacts on their performance and reliability.

7. *Condensers.* Condensers remove volatile organic compounds and hazardous air pollutants from emission streams by lowering the temperature of the emission stream, thus condensing these substances. Condensers are widely used to reduce the concentrations of volatile organic compounds and hazardous air pollutants in emission streams before they enter other control devices such as thermal incinerators, catalytic oxidizers, and adsorbers.

Alternative control technologies. These include biofilters, separation systems, ultraviolet oxidizers, corona destruction, and methods involving plasmas.

1. *Biofilters.* Biofilters use microorganisms to destroy volatile organic compounds and hazardous air pollutants. In the biofilter, the emission stream comes in contact with microorganisms that feed on these materials. Biofilters may have cost advantages over other control technologies, but significant issues such as proper matching of microorganisms to specific emission streams, long-term operational stability, and waste disposal may also be factors.

2. *Separation systems.* Separation technologies have been used extensively to control emissions of volatile organic compounds and hazardous air pollutants. Absorption technology has utilized such improvements as high-efficiency packing and trays, highly selective solvents, and advanced process configurations. Hydrophobic molecular sieves and superactivated carbons have enhanced the efficiency and applicability of adsorption technology. Improved membranes to separate and recover organic compounds may also serve to control emissions. *See* ACTIVATED CARBON; CHEMICAL SEPARATION (CHEMICAL ENGINEERING); MEMBRANE SEPARATIONS; MOLECULAR SIEVE.

3. *Ultraviolet oxidizers.* Ultraviolet radiation has the potential to destroy some volatile organic compounds and hazardous air pollutants. When organic compounds absorb ultraviolet light energy, they are activated; and they may decompose (photodissociation) or react with other compounds (radical oxidation). This technology may be combined with the use of a catalyst to increase reaction rates.

4. *Corona destruction.* There are two principal types of corona destruction reactors: packed bed corona and pulsed corona. The packed bed reactor is packed with high dielectric material, and high voltage (15,000–20,000 V) is applied to electrodes at both ends of the bed. High-energy electric fields are created in the spaces between the packing material, and volatile organic compounds and the hazardous air pollutants in an emission stream passed through the bed are destroyed. The pulsed corona reactor has a wire electrode in the center, and the wall of the reactor serves as the other electrode. When high voltages are applied to the two electrodes at nanosecond intervals, high-energy fields are established, and pollutants passing through the reactor are destroyed. *See* CORONA DISCHARGE.

5. *Plasmas.* Processes involving plasma technology include mechanisms for creating very high temperatures. The volatile organic compounds and hazardous air pollutants in incoming waste streams are destroyed as a result of the high temperatures and the interaction with other chemical species that exist at these temperatures.

Wade H. Ponder

Stationary sources: particulates. Particulate control devices remove particulate matter from a gas stream prior to its discharge to the atmosphere. Generally these devices are in the form of mechanical collectors, wet scrubbers, electrostatic precipitators, and fabric filters. Occasionally two or more of these mechanisms are used in one device. Control of fine particles from large industrial sources predominantly employs electrostatic precipitators and fabric filters.

Mechanical collectors. These devices range from simple settling chambers, which make use of gravity, to rotary-fan collectors, to cyclone and multiclone inertial devices. In general, these devices can collect particles of about 5 μm in diameter with efficiencies of 90% or more. They are ineffective for the submicrometer particles whose control is required by modern air pollution regulations. Mechanical collectors are useful for applications in which the particulate matter consists of only large particles. A second use is to remove most of the particulate matter from a gas stream before it enters a high-efficiency control device in order to minimize material handling. *See* CENTRIFUGATION.

Wet scrubbers. This generic term covers many types of control devices for capturing particles within a gas stream by bringing them into contact with a liquid. The capture mechanisms can be inertial impaction, direct interception, or diffusion resulting from Brownian motion of very fine particles. Inertial impaction and direct interception depend on the number of waterdroplets and the relative velocities of the particles and droplets. This implies high energy to the scrubber supplied through either the gas or liquid streams or by a mechanical device such as an impeller. Diffusion capture of small particles is enhanced by adding energy to increase the quantity of small liquid droplets, which limits the length of the diffusion path required for capture. To meet modern air pollution requirements for control of fine particles, a high-energy input is needed, usually as a high-pressure drop across the device. In addition, liquid waste treatment is required for the scrubbing

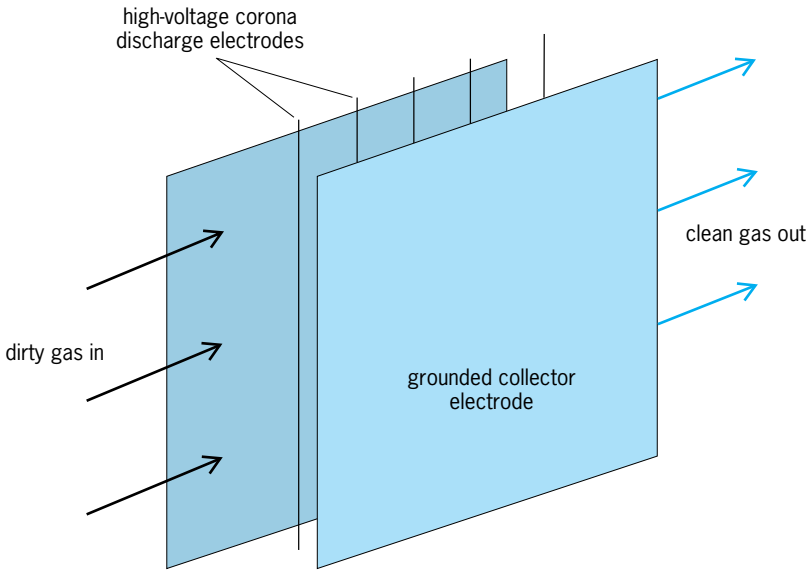


Fig. 6. Single lane of one section of an electrostatic precipitator.

process. However, a potential side benefit of scrubbing is the simultaneous ability to collect, by mass transfer, gaseous pollutants in the gas stream.

To help offset the high power requirement that results from the pressure drop across the scrubber, it is possible that electrostatic attractive forces can be introduced to improve capture. Energy needs can also be reduced by causing condensation upon the particles in the saturated atmosphere of the scrubber.

Electrostatic precipitators. These devices give particles an electrical charge and subject them to an electrical field that exerts a force to move them to the collecting electrode. In the basic process, particles are charged and collected simultaneously (**Fig. 6**). Incoming particles are charged by ions that are generated by the corona, which surrounds the high-voltage discharge electrodes. The high voltage also establishes the electric field to the grounded electrodes, which collect the particles. Periodically, the collected particles are knocked from the collector electrodes by mechanical rapping, and they fall into a hopper located beneath them.

An industrial-size electrostatic precipitator, such as at an electric utility, has collector electrodes 7.5–13.5 m (25–45 ft) in height and 1.5–4.5 m (5–15 ft) in length. A section might have 30 to 60 or more adjacent lanes with widths of 200–450 mm (8–18 in.). The unit might also have three to eight consecutive sections. Gas flow through the electrostatic precipitator could range from 250,000 to 5,000,000 m^3/h (150,000 to 3,000,000 ft^3/min) at temperatures from 150 to 343°C (300 to 650°F). The voltage applied to the corona discharge electrode ranges from 30 to more than 80 kV. The power required for operation is predominantly electrical. The pressure drop through the units is considerably less than for scrubbers.

Variations to the basic electrostatic precipitator process improve performance. Gas conditioning (addition of chemical additives to a gas stream) and pulse energization (intense pulses of direct-current high voltage supplied to the electrostatic precipitator) help overcome the detrimental effects of high-resistivity particle matter. Particles having a high electrical resistivity are more difficult to charge and collect than are particles of lower electrical resistivity. Electrostatic precipitators with two stages separate the charging and collecting functions, allowing both to be optimized for maximum efficiency. Continuous flushing away of the collected particles with water allows simultaneous collection by mass transfer of gaseous pollutants in the gas stream. However, operating the unit in the wet mode introduces a wastewater treatment requirement.

Electrostatic evaporators can be designed to specific efficiency requirements. Computer performance prediction models help in the engineering design of these devices. *See* ELECTROSTATIC PRECIPITATOR.

Fabric filtration. Filtration by collection of particles on a fabric makes use of inertial impaction, direct interception, and diffusion capture. The filtration fabric may be configured as bags, envelopes, or cartridges. For use with processes having high volumes of emissions, bag filters are the most important. The fabric bags are contained in a baghouse, fitted with automatic cleaning mechanisms to remove the collected particles, making it unnecessary to remove and replace dirty bags.

There are two basic methods for arranging and cleaning filter bags in baghouses. These are inside-to-outside flow (**Fig. 7***a*) in which the particulate matter is collected on the interior surface, and outside-to-inside flow (Fig. 7*b*) in which the particulate matter is collected on the outside of the bags. Inside-to-outside bags are cleaned by reversing the gas flow momentarily, shaking the bags, or a combination of both methods. Outside-to-inside flow bags are cleaned by directing a jet of air down the center of the bag to

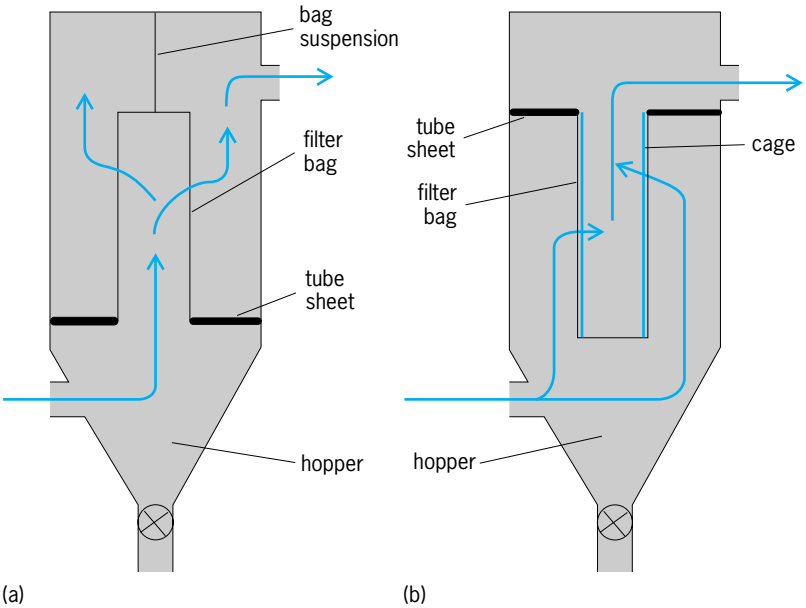


Fig. 7. Basic configurations of fabric filtration baghouses. (*a*) Inside-to-outside flow. (*b*) Outside-to-inside flow. Arrows indicate direction of gas flow.

dislodge the deposited dust. In both cases the particulate matter falls into a hopper, from which it is removed. The bags are attached to a tube sheet that separates the clean side of the baghouse from the dirty side. Outside-to-inside flow bags contain an internal cage to prevent their collapse during filtration. Inside-to-outside bags, in industrial applications, are as large as 10.8 m (36 ft) in length with a diameter of 12 in. (300 mm). Outside-to-inside bags range from 2.5 to 7.2 m (8 to 24 ft) in length with diameters from 100 to 200 mm (4 to 8 in.). Depending upon the size of the process to be controlled, a baghouse might contain from fewer than 100 bags to several thousand bags. Power required for operation is needed predominantly to move the gas through the device. The pressure drop for filtration is greater than for electrostatic precipitators but less than for scrubbers.

Inside-to-outside bags are usually made from woven fibers, whereas outside-to-inside bags use felted fibers. The temperature and chemical characteristics of the gas stream determine the type of fabric material used. Performance prediction models are utilized in the design; however, data from field testing are usually required for their use. For example, one focus has been on the use of electrostatics with fabric filtration to increase fine-particle capture and decrease pressure drop. *See* AIR FILTER; DUST AND MIST COLLECTION; FILTRATION.

Norman Plaks

Hybrid electrostatic precipitator–fabric filter concepts. At least three hybrid systems are being developed to reduce fine PM emissions: the Compact Hybrid Particulate Collector (COHPAC), developed by the Electric Power Research Institute; the Advanced Hybrid Particulate Collector (AHPC), developed by the Energy and Environmental Research Center of North Dakota under U.S. Department of Energy (DOE) sponsorship; and the Electrostatically Stimulated Fabric Filter (ESFF), being developed by Southern Research Institute under EPA sponsorship. The COHPAC I concept adds a small pulse-jet fabric filter baghouse immediately downstream of an electrostatic precipitator or replaces the latter stages within the precipitator structure (COHPAC II). Residual particle charge is reported to improve the fabric filter performance, so that very low emissions have been experienced.

The AHPC system combines reduced-size electrostatic precipitator and fabric filter compartments within a single vessel and introduces proprietary synergistic principles to achieve reduction of reentrainment and recollection of dust caused by close bag spacing, reduction of chemical attack on bagfilters, and ultrahigh collection efficiencies.

The ESFF system consists of dust precharging at the baghouse inlet and vertical charging wires placed at the centerline of each four-bag array that induce an electrostatic field between the wire and the grounded cage supporting each bagfilter. Modeling based on small-scale research indicates up to an order of magnitude reduction of submicrometer particles and two orders of magnitude supermicrometer particle reduction at a pressure loss of only 20–30% compared to that of conventional fabric filtration.

Stationary sources: nitrogen and sulfur oxides. The major gaseous pollutants emitted from stationary emission sources in the United States include sulfur oxides (SO_x) and nitrogen oxides (NO_x) from fuel combustion, waste combustion, metallurgical processes, and ore reduction. Sulfur dioxide (SO_2) and sulfur trioxide (SO_3) are emitted from fuel combustion. Sulfur dioxide is the dominant pollutant in most applications, but there is also a concern with the emission of acid gases, notably hydrochloric acid (HCl) from waste combustion. Adverse health effects from exposure to both SO_x and NO_x have been documented and have been the focus of most major air pollution legislation, along with particulate matter, in the United States. For example, the Clean Air Act of 1970 established National Ambient Air Quality Standards for emissions of SO_2 and NO_x from electric power plants. The Clean Air Act Amendments of 1990 identify SO_2 and NO_x as the major contributors to acid rain, and specify major reductions in emissions from the largest emitting sources—coal-fired electric power plants. *See* ACID RAIN.

Sulfur oxides are formed during fuel combustion by the oxidation of naturally occurring sulfur compounds in both mineral and organic portions of fossil fuels. Fuel nitrogen oxides are formed by oxidation of nitrogen compounds in fuel, and thermal nitrogen oxides are formed by the reaction of nitrogen and oxygen in air within a high-temperature flame. When wastes are burned, chlorinated compounds in synthetic materials are converted to HCl gas, while sulfur- and nitrogen-bearing wastes are oxidized to SO_x and NO_x. Many metallurgical processes use ores that occur naturally as sulfides: copper, zinc, iron, lead, and mercury are examples. Roasting these ores liberates fairly high concentrations of SO_x compared to fossil fuel combustion.

Control of SO_x, NO_x, and HCl may occur prior to combustion by pretreatment of fuels or wastes (precombustion), within the combustion zone furnace, or may take place after combustion. For solid fuels, the precombustion technologies are generally lower in cost and lower in efficiency, and the postcombustion technologies are higher in cost and higher in efficiency. For liquid and gaseous fuels, fuel pretreatment is generally more effective and cheaper than postcombustion technology. The strategies for control vary with the type and size of the parent process and the stringency of regulations.

Conventional technologies. These include precombustion controls, combustion controls, and postcombustion controls.

1. *Precombustion controls.* Fuel or waste pretreatment can reduce the sulfur, nitrogen, or chlorine content and subsequent emissions from the combustion process. Examples are gas sweetening, hydrotreating of liquid fuels, physical coal cleaning plants, and sorting of wastes prior to incineration.

• Gas sweetening. Natural gas and gases derived from oil, oil shale, bitumen, and coal may contain significant amounts of hydrogen sulfide, which when combusted forms SO_x. Absorbers that use regenerable sorbent liquids are used to remove and

concentrate the hydrogen sulfide gas into a feed stream for a sulfur plant, which converts gaseous hydrogen sulfide into elemental sulfur liquid.

• Oil desulfurization. Petroleum liquids and crude oil may be treated by injection of hydrogen gas, vacuum distillation, and partial oxidation to liberate and remove sulfur and nitrogen impurities. When burned, the cleaned fuels emit significantly lower concentrations of SO_x and NO_x than untreated fuels.

• Coal cleaning. A major industry within coal mining and distribution is that of physically cleaning coal to upgrade the coal heating value. Undesirable ash is physically removed by a number of washing processes. In the case of coals with higher sulfur content that are found in the eastern and midwestern United States, a significant amount of sulfur is also removed in the coal cleaning process, typically 20–30% of the total sulfur. This directly results in comparable reduction of SO_x when coal is burned.

2. *Combustion controls.* Within the combustion process, a number of techniques have been developed to minimize the formation of NO_x. Attempts to premix chemical sorbents with fuel or inject sorbents into the combustor for control of SO_x have not been as successful. Preferred controls for combustion NO_x vary according to fuel and combustor type. Major combustion sources of NO_x include steam boilers, gas-fired turbines, internal combustion engines, and waste combustors.

For steam boilers, minimizing the combustion air (low excess air) is commonly practiced. Minimizing the air-to-fuel ratio in the combustor reduces NO_x by 5–10% over uncontrolled combustion. Staging burners are devices that either lower oxygen in the flame to render it fuel-rich or lower peak flame temperature to render it air-rich; either strategy lowers NO_x formation. Staging the air or fuel to lower the peak temperature and availability of oxygen for NO_x formation can reduce emissions of NO_x by 10–45%. Special staging burners (low NO_x burners) are used in large utility boilers to reduce emissions of NO_x up to 45%. Supplemental, low-nitrogen fuels, such as natural gas or distillate oil, can be added via a special fuel-staging system, known as fuel reburning, which reduces NO_x by up to 70% in large coal-fired boilers.

Techniques for reducing flame temperature, and hence NO_x, in other NO_x sources include water injection in turbines and special ignition systems for diesel-fuel and natural-gas-fuel engines.

3. *Postcombustion controls.* These controls reduce sulfur oxide, nitrogen oxide, and hydrogen chloride.

• Sulfur oxide. For control of SO_x, the dominant system is the wet flue-gas desulfurization system, or wet scrubber, which contacts flue gas in an absorber vessel with a slurry or solution spray of water and alkaline sorbent (**Fig. 8**). Depending on the source of the SO_x, the economic choice of sorbent may be lime, limestone, or sodium carbonate. The SO_x reacts to form a salt in the scrubber liquid, which exits the bottom of the scrubber. A water treatment system separates the solids from water, and the wet solids are disposed of in ponds or landfills. In the case of large coal-fired utility boilers, the limestone scrubber is dominant, and the resulting solid salts or sludge create a considerable disposal problem. Use of a special oxidation step can convert the sludge to

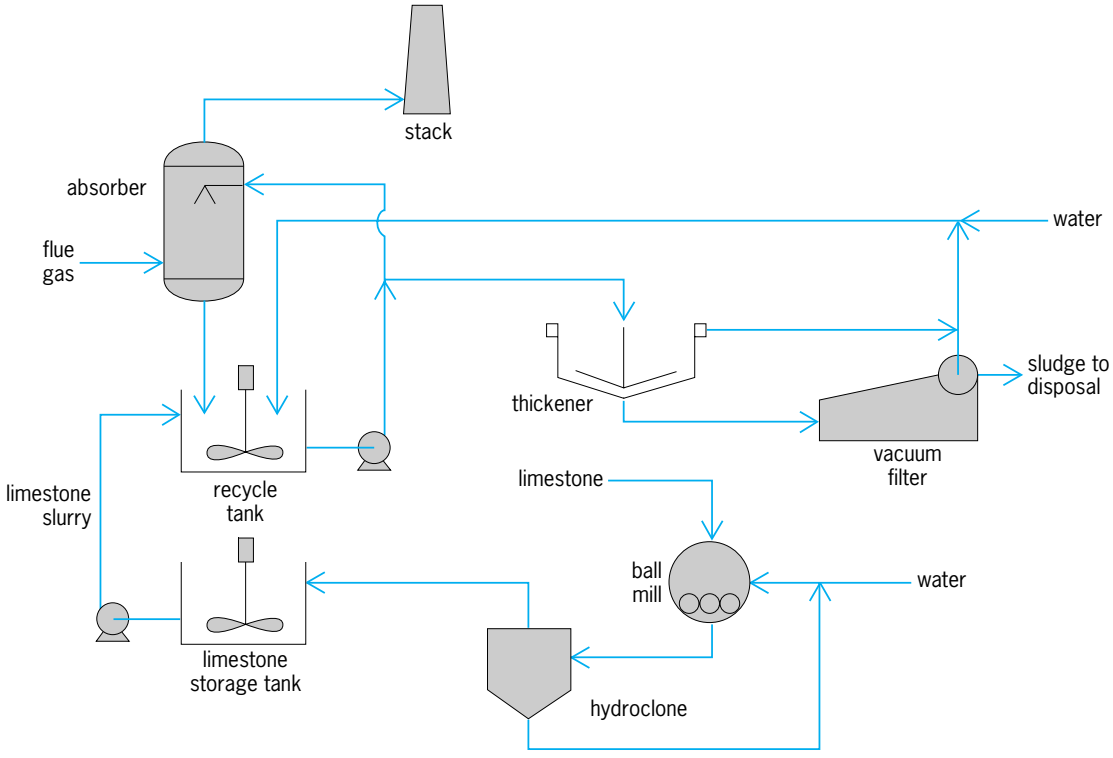


Fig. 8. Wet scrubber SO_x control system.

gypsum, which has commercial value in construction. A special subset of scrubbing includes the use of reduced water scrubbers (spray dryer absorbers). The process is sometimes known as dry scrubbing (which is a misnomer since the scrubbing is wet but the product is collected dry). Spray dryers are common in smaller industrial boilers, cogeneration boilers, and waste combustors. The advantages of lime spray drying are reduced water usage, no water–solids separation, and lower capital cost. Whereas halogenated acids such as hydrochloric acid (HCl) adversely affect wet scrubber operations in mass transfer, solids separation, and materials, the presence of chlorides in spray dryers enhances mass transfer and collection of fine particles. Sulfuric acid mist is also more readily collected in the spray dryer–dust collector system. Disadvantages are a dry waste of fly ash and spent sorbent which is difficult to market, usually ending up in a landfill; and limited application to low-to-moderate sulfur fuels (95% SO_2 removal on <2% sulfur fuel is the upper limit for most applications).

Another type of gas absorber is the circulating-bed scrubber using hydrated lime to absorb acid gases. The system operates upstream of an existing dust collector, much like the spray dryer, but is essentially a vertical pipe with internals that encourage solids–gas mixing and flash drying of moisture, within a fraction of the gas residence time of its competitors. With no lime slurrying or spray nozzles, many of the operating problems of dry scrubbers are avoided. With high solids recirculation, 98% removal of sulfur oxides has been achieved in European applications, and the system has been applied to higher-sulfur (4% sulfur) fuels. The disadvantages of this system are dry solids disposal, upper size limits on single-module operation, and larger solids loadings on the dust collector.

In ore roasting, the concentrations of SO_x are generally an order of magnitude higher than for fuel burning. These high concentrations of SO_x make recovery of SO_x by conversion to sulfuric acid somewhat profitable. Therefore, many primary smelters generate by-product sulfuric acid from SO_x in conventional acid plants. *See* SULFURIC ACID.

• Nitrogen oxide. Postcombustion control of NO_x consists of two techniques that use ammonia-based liquids injected into the flue gas. With selective noncatalytic reduction, the ammonia liquid is injected in the cooler regimes of the combustor to reduce NO_x to nitrogen and oxygen, achieving reductions of NO_x in the range 30–60%. Injecting ammonia in front of a catalyst system designed for control of NO_x can remove up to 90% of NO_x. This selective catalytic reduction control system is applicable to practically all sources of NO_x, and it is generally the highest-capital-cost system for such control.

Oxidation–absorption of NO_x is another alternative where a gas absorber is already installed or simultaneous SO_x/NO_x control is desirable. In this technique, an oxidant is added to the flue gas, such as ozone or chloric acid, or absorber liquor (sodium chlorate, for example) that causes NO_x to be oxidized to water-soluble nitric acid or nitrous acid vapor and absorbed in the scrubber liquor.

• Hydrogen chloride. Control of HCl has been achieved by adapting postcombustion SO_x control systems to combined HCl/SO_x removal systems. Wet scrubbers, spray dryers, and dry sorbent injection are applicable to waste combustors. In the United States, spray dryer scrubbers are a popular system for controlling HCl, typically removing 90–95% HCl and 70–90% SO_x.

Toxic gases and vapors (mercury). In the December 1997 report on mercury to Congress, coal-fired boilers represented over 50% of the total anthropogenic mercury emissions in the United States, estimated at 144 megagrams per year. The control of mercury is complicated by several factors, the most prominent being the very low concentrations in flue gas and the species present. Mercury vapor occurs in coal combustor flue gas at concentrations ranging from 3 to 11 $\mu g/nm^3$ as elemental (Hg^0) form or ionic mercuric chloride ($HgCl_2$) or mercuric oxide (HgO) form. Elemental mercury is insoluble in water and therefore not easily captured by wet or semidry scrubbing technology. $HgCl_2$ is water-soluble and readily absorbed in alkaline scrubbing liquor. Little is known about the behavior of HgO in flue gases.

Activated carbon has been used in the United States to reduce mercury emissions from municipal waste incinerators, but has not been commercially applied to coal-fired boilers. Substantial differences in flue gas species and concentration make transfer of technology impractical. Wet and dry flue gas desulfurization (FGD) units have been shown to be effective in removing oxidized mercury species.

Sulfuric acid mist. Since 1998, fossil fuel–fired power plants are required to report annual release inventories of nearly 600 toxic chemicals covered under the national Toxic Release Inventory, established under Section 313 of the Emergency Planning and Community Right-to-Know Act of 1986. This generally applies to any release exceeding 11,340 kg (25,000 lb) per year. Unwittingly power plants have become manufacturers of several toxic agents by virtue of their formation in the combustion of fuel and subsequent cooling and treating of exhaust flue gas. One recent study estimates that of a potential 17 metal species, 12 organic compounds, and 13 other acids and oxidants, a 650-MWe power plant burning a coal of 2.5 lb/MBtu sulfur content can be expected to report on 10 of these, with major components being hydrochloric acid (controls discussed above) and sulfuric acid (condensed sulfur trioxide). Primary particle collectors operating above the acid dew point can remove solid sulfates that have reacted with alkaline compounds such as lime, sodium, or ammonia; therefore any injection of these compounds for other reasons can directly reduce sulfuric acid (H_2SO_4) emissions. Some control is likely to occur by sorption or condensation of acids on collectible fly ash particles. Upstream NO_x controls may also influence H_2SO_4 emissions—catalytic systems may oxidize SO_2 to SO_3 and increase potential emissions, while ammonia slip from SCR and SNCR will likely react with SO_3

to form ammonium sulfate and bisulfate particles, effectively reducing H_2SO_4 emissions at the stack. The wet scrubber for SO_2 control ironically exacerbates SO_3 problems because (1) the particle collector is upstream and sees SO_3 as a vapor; (2) the scrubber cools the gas, converting SO_3 vapor to H_2SO_4 aerosol; and (3) the scrubber is designed for gas absorption only, allowing the submicrometer H_2SO_4 particles to exit, largely uncontrolled. One study indicates that only 15% of the SO_3 entering the FGD system is removed by the scrubbing process. Conversely, the so-called dry systems that scrub acid gases upstream of the particle collector seem particularly adept at converting SO_3 to solids that are effectively collected in the dust collector as particles. In severe cases of wet FGD-induced H_2SO_4 stack plumes, a wet electrostatic precipitator has been installed between the FGD system and stack, or the mist eliminator inside the FGD has been replaced by a wet electrostatic precipitator, to reduce plume opacity to acceptable levels.

Multipollutant controls. The concept of multipollutant controls is to develop a compact system that removes many pollutants in a minimum of process steps, and is upgradable to meet future emission codes. One approach is to convert all pollutant vapors and gases to particles upstream of the particle collector, thereby collecting all gaseous pollutants as condensed or sorbed particles. The class of circulating bed absorbers offers such an opportunity for multiple-pollutant control. Special sorbents, such as activated carbon, zeolites, or calcium silicates for sequestering toxic metals and organics, may be easily added to the base lime or sodium stream, and the large effective solids residence time allows for better utilization of sorbent. Oxidants for converting NO_x into more reactive acid vapors, which would cause potential wet chemistry problems in wet scrubbers, would not present such problems in dry solids absorption; hence simultaneous removal of SO_x, NO_x, and toxic vapors by a multipollutant sorbent or mix of sorbents introduced into a rapidly cooling gas stream, typical of that in a circulating bed, followed by an efficient particle collector, would be possible.

Alternative technologies. In the United States, the focus was upon clean energy in the late 1980s and early 1990s. The U.S. Clean Coal Technology Program administered by the Department of Energy emphasized the concept of repowering as opposed to construction of conventional coal-fired boilers. The concept of repowering includes various configurations of advanced power cycles that use coal gasification, fluidized bed combustion, gas turbines, and heat recovery boilers to achieve higher efficiencies in conversion of coal to electric power. These configurations avoid conditions favorable for the formation of NO_x normally associated with coal combustion. The amount of sulfur oxides is also much reduced, since coal gasification converts fuel sulfur (sulfur oxides formed by the oxidation of sulfur compounds in fuel) to hydrogen sulfide, 99% of which is typically removed by absorption and conversion to liquid sulfur.

Charles B. Sedman

Legal issues

Industry often has little economic incentive to voluntarily install pollution controls, and private litigation has often not addressed what is normally a communitywide problem. Thus, many countries have established schemes of legislative and administrative regulation. In the United States, the Clean Air Act directs the EPA to establish national ambient air quality standards that limit the permissible concentration of air pollutants at any outdoor point to which the public has access. There are two types of ambient air quality standards: primary and secondary (Table 1). Primary standards must protect the public health with an adequate margin of safety. Secondary standards must safeguard public welfare from known or anticipated adverse effects. (Generally, as is shown in Table I, primary and secondary standards have been set at the same levels.) The EPA may not take compliance costs into account in setting ambient air quality standards. In 1997, the EPA tightened the ambient air quality standards for particulate matter and ozone. New standards have been set to control particulate matter with an effective aerodynamic diameter of 2.5 μm or less ($PM_{2.5}$). The ozone standard was revised to regulate concentrations over an 8-h averaging time to reflect scientific findings that health effects are associated with exposures over an extended time.

Each state is required to submit for EPA approval a State Implementation Plan (SIP). This must consist of measures, such as limitations on emissions from individual sources of pollution, sufficient to demonstrate that the state will attain and maintain the primary standards by specific deadlines and the secondary standards as expeditiously as practicable. Each state has a great deal of discretion in allocating the reduction burden among sources as long as the state shows that its SIP plan will result in timely attainment and maintenance of the standards. In addition, each state's plan must be adequate to prevent the transport of air pollution across state lines that would significantly interfere with another state's ability to meet the standards. The EPA has become more aggressive in implementing this prohibition.

Each SIP must include programs for the review of the proposed construction or modification of major sources of air pollution, both in areas in which the ambient air quality standards are violated (nonattainment areas) and in clean-air areas. Under both programs, a source may receive a permit to construct or modify only if it uses the best pollution controls available. The program in nonattainment areas requires that the permit applicant furnish offsets—decreases in air pollution that go beyond what is required by the SIP—to compensate for the new or modified source's effect on air quality. The program in clean-air areas, known as prevention of significant deterioration, establishes increments that, in effect, cap the growth of concentrations of sulfur dioxide,

particulate matter, and nitrogen oxide in clean areas. These increments are especially stringent in national parks and wilderness areas. There is much controversy about which changes at existing sources constitute modifications that are subject to the programs.

Some kinds of sources are regulated directly by the federal government. For example, the federal government establishes emission limits for new cars and trucks and regulates the content of automotive and truck fuel. These regulations have virtually eliminated lead emissions in the United States. The federal government also regulates sources of hazardous air pollutants that pose a risk of cancer or other life-threatening disease. The 1990 Amendments to the Clean Air Act list 189 substances as air pollutants, subject to addition or deletion by the EPA. The EPA must establish emissions standards that require the use of maximum available control technology by major sources of such pollutants, and must set additional, stricter standards for a substance if the agency finds that, even with maximum available control technology, a major source poses a lifetime risk of more than one in a million to the most exposed individual. These additional standards must protect health with an ample margin of safety.

The 1990 Amendments established an innovative program to decrease emissions of sulfur dioxide, which turn into sulfates that degrade visibility, increase concentrations of health-threatening fine particles, and cause acid rain. The Amendments require a 10-million-ton reduction in annual emissions of sulfur dioxide and establish annual allowances for individual sources. These sources may trade allowances among themselves so that the reduction can be accomplished in the most cost-effective way. The EPA is encouraging states to use the same approach in its new Clean Air Implementation Rule. This rule requires 28 eastern and midwestern states to further reduce sulfur dioxide and nitrogen oxide emissions that cross state lines and interfere with the attainment of the new air quality standards for ozone and particulate matter.

Craig N. Oren

Bibliography. R. W. Boubel et al., *Fundamentals of Air Pollution*, 3d ed., 1994; W. T. Davis (ed.), *Air Pollution Engineering Manual*, 2d ed., 2000; R. P. Donovan, *Fabric Filtration for Combustion Sources*, 1985; M. Lippmann (ed.), *Environmental Toxicants: Human Exposures and Their Health Effects*, 2d ed., 2005; L. R. Paulson (ed.), *Human Exposure Assessment for Airborne Pollutants: Advances and Opportunities*, 1991; K. R. Parker (ed.), *Applied Electrostatic Precipitation*, 1996; K. C. Schifftner, *Air Pollution Control Equipment Selection Guide*, 2002; J. H. Seinfeld and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate*, 2d ed., 2006; J. D. Spengler et al., *Indoor Air Quality Handbook*, 2000; L. K. Wang et al. (eds.), *Air Pollution Control Engineering*, 2004; K. Wark et al., *Air Pollution: Its Origin and Control*, 3d ed., 1997; A. Wellburn, *Air Pollution and Climate Change*, 2d rev. ed., 1994; Y. Zhang, *Indoor Air Quality Engineering*, 2004.

Air pollution, indoor

The presence of gaseous and particulate contaminants in the indoor environment. Most pollution is due to human sources, although natural sources do exist, including plants, animals, and other living organisms and bodies of water that release various chemical aerosols. Contamination can occur from infiltration indoors of atmospheric pollutants generated outdoors, and thus the indoor environment is affected by meteorological conditions.

Contamination of indoor air is considered very important because of the amount of time that people spend indoors (65–95%), and the effects of indoor pollutants on humans, pets, and materials. It was not until the twentieth century that serious scientific measurements were taken and studies of air pollution done. Later developments included investigations of indoor sources and effects. Another concern has been the amount of human cancer caused by airborne toxins and radon from both indoor and outdoor sources. Research involving energy-saving measures, human activities, and consumer products has led to evaluation of the health effects of the indoor environment specifically, as well as of the total air pollution.

Sources. Natural sources of air pollution are soils and water that release radon decay products, volatile organic compounds, fungi, and so on, and living organisms that release allergens (for example, dander from pets). Human-produced (anthropogenic) chemical releases (emissions) originate from various types of appliances, combustion sources (including vehicle combustion engines), and building materials. There has been an increase in the home use of secondary combustion heating sources, such as wood-burning and gas and kerosene space heaters (often unvented) due to increased costs of regular heating, and they are very important contributors to high levels of indoor air pollutants. Carbon dioxide (CO_2) is generated both by combustion and by living organisms. Other pollutants include molds, microorganisms (such as bacteria and viruses), and pollen from outdoor and indoor plants (**Table 1**). Human activity has been noted to increase indoor air contamination by all of these. Some pollutants are generated by individual behavior, such as tobacco smoking (environmental tobacco smoke, or ETS), or through use of consumer products.

Classification and behavior of pollutants. The chemicals in pollutants can be divided into inorganic and organic, and can be further classified on the basis of their thermal and hydroscopic characteristics. Particles are characterized also by functional size (usually mass-median aerodynamic diameter), density, chemical identity, water content, and affinity for absorption and adsorption. The mass-median aerodynamic diameter is the measurement of size and density of particles (including viable and nonviable organic particles); it is a major characteristic of particles and determines their behavior (dispersal, deposition), especially their likelihood of inhalation and

TABLE 1. Major sources and types of indoor pollutants

Sources	Pollutants
Fossil fuel combustion (natural gas and kerosene appliances); wood combustion (wood stoves, fireplaces)	Particulates (PM); nitrogen oxides (NO_x); carbon monoxide and dioxide (CO, CO_2); lead and other trace metals; polynuclear aromatic hydrocarbons (PNAs) and other volatile organic compounds (VOCs)
Tobacco smoking	PM; CO; CO_2; NO_x; PNA; VOCs; radon progeny
Building and furnishing materials	PNAs (especially aldehydes) and other VOCs; PM; radon progeny; also act as reservoirs for molds and other allergens
Water reservoirs (fixtures; air conditioning/cleaning/treating)	Mold; *Bacillus* sp. and other bacteria
Consumer products	PNAs and other VOCs; trace metals
Other product examples	Cleaners, solvents, disinfectants, cosmetics, air purifiers, etc.
Animals (pets and opportunistic dwellers) and plants	Allergens; CO_2
Infiltration	PM; NO_x; sulfur oxides (SO_x); pollen; mold

deposition in different parts of the respiratory system. For instance, environmental tobacco smoke produces very small (fine) particles, and intense combustion can produce ultrafine particles, both of which are more deeply inhaled than larger particles. Sometimes, gaseous pollutants are further divided into vapor and nonvapor phase, and degree of volatility. *See* PARTICULATES.

Nitrogen oxide (NO_x) can be produced as a gas [nitric oxide (NO) or nitrogen dioxide (NO_2)], as particles [nitrate (NO_3^-)], or in aerosol form [such as nitrous acid (HNO_2)]. Radon decay products are more important when adsorbed on particles, as are aldehydes, other volatile organic compounds, and compounds derived from environmental tobacco smoke; all of these are normally gaseous, but all except radon progeny can occur in the form of particles. *See* NITROGEN OXIDES; RADON.

The generation and behavior of pollutants in enclosed environments are also affected by most meteorological factors. Indoor environments are often sealed to the extent that the weather is cold or hot. Infiltration of outdoor pollutants is affected by climate, temperature, humidity, barometric pressure, and wind speed and direction. The small particles and gases that infiltrate most—NO_2 and volatile organic compounds—are affected by outdoor concentrations, tightness of the indoor space (being sealed with little air exchange), operation of heating and cooling systems, convection currents, full growth (or lack thereof) of local trees and shrubs, and so forth. Indoor environments also possess individual characteristics of ventilation, dispersion, and deposition. Some gases, for example sulfur dioxide (SO_2) and ozone (O_3), infiltrating the indoor environment are absorbed readily if absorbent materials (such as carpets, drapes, cloth-covered furniture, clothes) are present. These gases thus exist in high concentrations only when the outdoor concentrations are very high or such materials are not present. *See* OZONE.

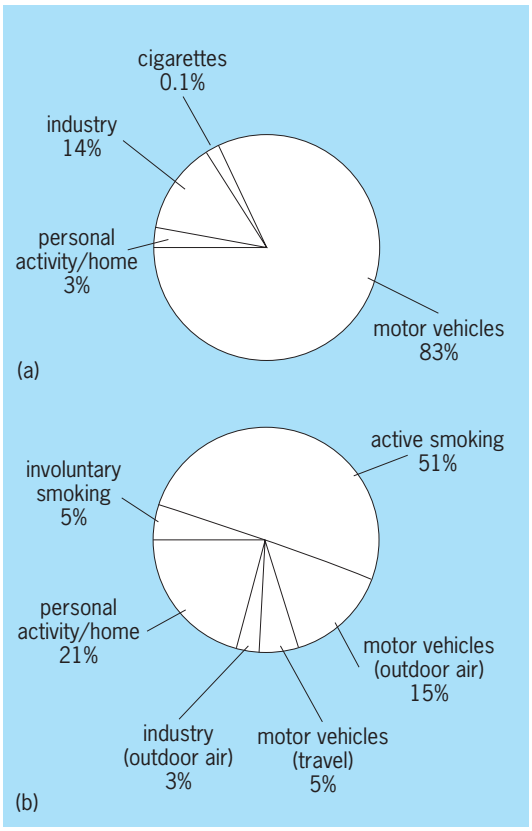
Exposure assessment. Chemical-physical mass-balance models are used to estimate pollutant concentrations, as are statistical models. Assessments of exposures provide information on the concentration and spatial-temporal distribution of each pollutant and are related to spatial-temporal activities of the people assessed. Personal exposure factors include time spent in different indoor environments (**Table 2**), activities and behavior, breathing rates, and so forth. These factors indicate that individuals experience pollution levels very different from those measured at a nearby fixed monitoring station. The pollutant benzene provides a good example (**Fig. 1**). Personal exposure must also take into account the major indoor sources, such as environmental tobacco smoke, which produces particulate matter leading to concentrations that frequently occur in excess of ambient standards (set by law in the United States). Lower socioeconomic status groups not only live in areas with higher ambient exposures, but also live in poorly ventilated homes and use more inappropriate combustion sources having higher smoking levels.

Biological markers of exposures to pollutants have been developed for a variety of toxic materials in the air. Lead is measured by well-established blood assays. These tests have indicated declines of lead in the body that parallel the declines in amount of lead in the atmosphere, the amount of lead in gasoline, and the amount of lead in paint and dust that were mandated by law in the United States. Cotinine, a nicotine metabolite, is the marker used for environmental tobacco smoke. Using metal biomarkers as estimates of exposures to elemental metals (air exposures mostly from smaller particles) in the National Human Exposure Assessment Survey (NHEXAS), researchers find that children have higher levels of

TABLE 2. Time activity patterns*

Location	Percentage of time spent by age group, years		
	<5	6–65	>65
Home (inside)	79	63	78
Work (inside)	—	11	1
Other (inside)	8	7	5
Home (outside)	6	5–6	8
Work (outside)	—	3	1
Other (outside)	3	4–5	3
In transit	4	6	4

*From Arizona National Human Exposure Assessment Survey.



Benzene emissions versus exposures. (*a*) Emissions. (*b*) Exposures. (*From L. Wallace, Enviromental Protection Agency, 1989; National Research Council, Human Exposure Assessment for Airborne Pollutants, 1991*)

cadmium, nickel, chromium, and lead if they have been exposed to environmental tobacco smoke. This survey found also that children had higher exposures than adults to pesticides.

Indoor monitoring has specific requirements related to the type of pollutants and enclosed spaces monitored. Monitoring protocols are developed with regard to the availability, practicality, and expense of continuous or integrated sampling methods to measure the pollutant over the periods of interest. These monitors are usually fixed or portable. Personal monitors have been developed for some pollutants. Good surrogate measurements are often utilized to represent more complex exposures (for example, total hydrocarbons for all organics).

Effects. Characterization of the effects of air pollution requires determination of concentration/exposure, exposure/dose (for humans and animals), and exposure/dose/response relationships. These studies have involved humans, animals, and materials.

Humans. Effects experienced by humans include acute and chronic symptoms (morbidity), increases in acute respiratory illness, declines in lung function and other physiological and immunological changes, and deaths, especially in those with chronic diseases. The risks of multiple exposures may contribute additively or synergistically to health consequences. There are several indoor pollutants that have similar, and possibly synergistic, effects: particulate matter, NO_2, and formaldehyde (HCHO) produce irritation of eyes and mucous membranes; various types of particulate matter, HCHO, and allergens produce allergic symptoms; NO_2, particulate matter, and HCHO produce changes in pulmonary function; carbon monoxide (CO), NO_2, and HCHO affect cognitive skills; CO, NO, NO_2, nitrates, and methylene chloride (CH_2Cl_2) change carboxyhemoglobin and methemoglobin levels and affect the heart and brain; and radon progeny, especially when associated with tobacco smoke, can promote lung cancer. Lead has been shown to have neurological effects at concentrations that occurred frequently in the past, especially in infants and children. Benzene has been shown to produce a form of leukemia, and both radon progeny and asbestos have been shown to produce lung cancer in humans. Other types of cancers may occur with exposures to carcinogens found in occupational settings. *See* ASBESTOS; BENZENE; MUTAGENS AND CARCINOGENS; RESPIRATORY SYSTEM DISORDERS.

Aeroallergens are also of great importance, as many people have allergic problems, including asthma. It is likely that aeroallergens and interactive exposures with other pollutants can aggravate allergies and produce respiratory diseases. Thus, a large amount of acute disability is related to indoor pollutant exposures. *See* ALLERGY; ASTHMA.

Animals. Sudies of pollutant effects on animals have shown effects similar to those found in humans. These studies used mortality, lung pathology, and changes in pulmonary physiology and immunology as criteria. Animal studies of cancer have provided the evidence to protect humans from those carcinogens.

Materials. Indoor air pollution produces destruction by corrosion. In addition, outdoor pollutants can infiltrate and interact with indoor pollutants to produce new compounds that are damaging to materials; an example is O_3 interacting with indoor volatile organic compounds to produce corrosive compounds affecting electronic equipment. Economically significant damage has occurred in various metals, ferrous (iron-based) and nonferrous (copper, silver, nickel, aluminum, zinc). Extensive damage to rubber, paint and dyes, leather and textiles, and ceramics has been demonstrated. *See* CORROSION.

Controls. For some pollutants, control measures used to protect health are based on regulatory law. Unfortunately, the experience gained in developing and implementing strategies for the outdoor environment is not very applicable to indoor environments, and new strategies are needed. Only a few countries have regulations for indoor pollutants, specifically radon progeny, formaldehyde, and CO_2. Controls should include, where necessary, practices, labeling requirements, building codes, guidelines, and recommendations.

ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) standards are used in the United States. Some European countries and Japan also have such standards, and World Health Organization (WHO) publications have recommendations for control as well. Primary control measures

include removal of sources (including water, plants, pets, and certain building materials) or reductions in source emissions through manufacturing processes, changes in building materials, prevention of water leakage, and source usage recommendations. Source removal is the first and most logical form of control. Reductions in source emissions have been achieved for several products through changes in manufacturing process or construction. Controls have included recommendations for source placement, use, and ventilation (including the amount of fresh air intake and filtration). Reduction of exposure also includes standard direct methods such as filtration and cleaning, as well as good building and source design. Appropriate building maintenance is necessary. Ventilation that produces adequate exfiltration of contaminated air is very efficient. Filtration is useful but can be expensive. Active charcoal can be used for filtering most gases, and HEPA filters can be used to filter most particle matter, but they need to be monitored and maintained appropriately.

Governmental agencies can control product manufacturing, promote ventilation in workplaces, use public health education to help people minimize risks and promote safe indoor environments, and develop regulations (such as banning tobacco smoking in workplaces and public buildings). Laws can be passed to prevent exposures in public locations. Finally, individuals can limit their exposures, especially in residences, by several measures, including avoidance. *See* AIR POLLUTION; ENVIRONMENTAL TOXICOLOGY.

Michael D. Lebowitz

Bibliography. American Society for Heating, Refrigeration and Air Conditioning Engineers (ASHRAE), Standards for heating, cooling and ventilation, ASHRAE Web site, National Human Exposure Assessment Survey (NHEXAS), *J. Expos. Anal. Environ. Epidemiol.*, vol. 9, 1999; R. Bertollini et al. (eds.), *Environmental Epidemiology: Exposure and Disease*, CRC/Lewis, London, 1996; M. Lippmann (ed.), *Environmental Toxicants*, 2d ed., Wiley, New York, 2000; National Research Council (NRC), *Human Exposure Assessment for Airborne Pollutants*, National Academy Press, Washington, DC, 1991; NRC, *Indoor Allergens*, National Academy Press, Washington, DC, 1993; NRC, *Policies and Procedures for Control of Indoor Air Quality*, National Academy Press, Washington, DC, 1987.

Air pressure

The force per unit area that the air exerts on any surface in contact with it, arising from the collisions of the air molecules with the surface. It is equal and opposite to the pressure of the surface against the air, which for atmospheric air in normal motion approximately balances the weight of the atmosphere above, about 15 lb/in.^{2} at sea level. It is the same in all directions and is the force that balances the weight of the column of mercury in the torricellian barometer, commonly used for its precise measurement. *See* BAROMETER.

Units. The units of pressure traditionally used in meteorology are based on the bar, defined as equal to 1,000,000 dynes /cm^2. One bar equals 1000 millibars or 100 centibars.

In the meter-kilogram-second or International System of Units (SI), the unit of force, the pascal (Pa), is equal to 1 newton/m^2. One millibar equals 100 pascals. The normal pressure at sea level is 1013.25 millibars or 101.325 kilopascals.

Also widely used in practice are units based on the height of the mercury barometer under standard conditions, expressed commonly in millimeters or in inches. The standard atmosphere (760 mmHg) is also used as a unit, mainly in engineering, where large pressures are encountered. The following equivalents show the conversions between the commonly used units of pressure, where $(\text{mmHg})_n$ and $(\text{in. Hg})_n$ denote the millimeter and inch of mercury, respectively, under standard (normal) conditions, and where $(\text{kg})_n$ and $(\text{lb})_n$ denote the weight of a standard kilogram and pound mass, respectively, under standard gravity.

$$\begin{aligned}
1\ \text{kPa} &= 10\ \text{millibars} = 1000\ \text{N/m}^2 \\
&= 7.50062\ (\text{mmHg})_n \\
&= 0.95300\ (\text{in. Hg})_n \\
1\ \text{millibar} &= 100\ \text{Pa} = 1000\ \text{dynes/cm}^2 \\
&= 0.750062\ (\text{mmHg})_n \\
&= 0.0295300\ (\text{in. Hg})_n \\
1\ \text{atm} &= 101.325\ \text{kPa} = 1013.25\ \text{millibars} \\
&= 760\ (\text{mmHg})_n = 29.9213\ (\text{in. Hg})_n \\
&= 14.6959\ (\text{lb})_n/\text{in.}^2 \\
&= 1.03323\ (\text{kg})_n/\text{cm}^2 \\
1\ (\text{mmHg})_n &= 1\ \text{torr} = 0.03937008\ (\text{in. Hg})_n \\
&= 1.333224\ \text{millibars} \\
&= 133.3224\ \text{Pa} \\
1\ (\text{in. Hg})_n &= 33.8639\ \text{millibars} \\
&= 25.4\ (\text{mmHg})_n \\
&= 3.38639\ \text{kPa}
\end{aligned}$$

Variation with height. Because of the almost exact balancing of the weight of the overlying atmosphere by the air pressure, the latter must decrease with height, according to the hydrostatic equation (1),

$$dP = -g\ \rho\ dZ \tag{1}$$

where P is air pressure, ρ is air density, g is acceleration of gravity, Z is altitude above mean sea level, dZ is infinitesimal vertical thickness of horizontal air layers, and dP is pressure change which corresponds to altitude change dZ. Integration of Eq. (1) yields Eq. (2), where P_1 is pressure at altitude Z_1, and P_2

$$P_1 - P_2 = \int_{Z_1}^{Z_2} \rho g\ dZ \tag{2}$$

is pressure at altitude Z_2. The expressions on the right-hand side of Eq. (2) represent the weight of the column of air between the two levels Z_1 and Z_2.

In the special case in which Z_2 refers to a level above the atmosphere where the air pressure is nil, one has $P_2 = 0$, and Eq. (2) yields an expression for air

pressure P_1 at a given altitude Z_1 for an atmosphere in hydrostatic equilibrium.

By substituting in Eq. (1) the expression for air density based on the well-known perfect gas law and by integrating, one obtains the hypsometric equations for dry air under the assumption of hydrostatic equilibrium, Eqs. (3), valid below about 54 mi (90 km),

$$\log_e\left(\frac{P_1}{P_2}\right) = \frac{M}{R}\int_{Z_1}^{Z_2}\frac{g}{T}dZ \qquad (3a)$$

$$Z_2 - Z_1 = \frac{R}{M}\int_{P_2}^{P_1}\frac{T}{g}\frac{dP}{P} \qquad (3b)$$

where g is the gravitational acceleration; M is the gram-molecular weight, 28.97 for dry air; R is the gas constant for 1 mole of ideal gas, or 8.315×10^7 erg/(mole)(K); and T is the air temperature in K.

Equation (3) may be used for the real moist atmosphere if the effect of the small amount of water vapor on the density of the air is allowed for by replacing T by T_v, the virtual temperature given by Eq. (4), in which e is partial pressure of water vapor

$$T_v = T\left[1 - \left(1 - \frac{M_w}{M}\right)\frac{e}{P}\right]^{-1} \qquad (4)$$

in the air, M_w is gram-molecular weight of water vapor (18.016 g/mole), and $(1 - M_w/M) = 0.3780$.

Equation (3*a*) is used in practice to calculate the vertical distribution of pressure with height above sea level. The temperature distribution in a standard atmosphere, based on mean values in middle latitudes, has been defined by international agreement. The use of the standard atmosphere permits the evaluation of the integrals of Eqs. (3*a*) and (3*b*) to give a definite relation between pressure and height. This relation is used in all altimeters which are basically barometers of the aneroid type. The difference between the height estimated from the pressure and the actual height is often considerable; but since the same standard relationship is used in all altimeters, the difference is the same for all altimeters at the same location, and so causes no difficulty in determining the relative position of aircraft. Mountains, however, have a fixed height, and accidents have been caused by the difference between the actual and standard atmosphere. *See* ALTIMETER.

Horizontal and time variations. In addition to the large variation with height discussed in the previous paragraph, atmospheric pressure varies in the horizontal and with time. The variations of air pressure at sea level, estimated in the case of observations over land by correcting for the height of the ground surface, are routinely plotted on a map and analyzed, resulting in the familiar "weather map" representation with its isobars showing highs and lows. The movement of the main features of the sea-level pressure distribution, typically from west to east in middle and high latitudes, and from east to west in the tropics, produces characteristic fluctuations of the pressure at a fixed point, which vary by a few percent within a few days.

Smaller-scale variations of sea-level pressure, some even too small to appear on the ordinary weather map, are also present. These are associated with various forms of atmospheric motion, including small-scale wave motion and turbulence. Relatively large variations in short distances are found in hurricanes and in intense winter storms, and in and near thunderstorms; the most intense is the low-pressure region in a tornado. The pressure drop within tornadoes and hurricanes can be extreme, about 10% of normal pressure. *See* ISOBAR (METEOROLOGY); WEATHER MAP.

It is a general rule that in middle latitudes at localities below 3280 ft (1000 m) in height above sea level, the air pressure on the continents tends to be slightly higher in winter than in other seasons; whereas at considerably greater heights on the continents and on the ocean surface, the reverse is true.

Various maps of climatic averages indicate certain regions where systems of high and low pressure predominate. Over the oceans there tend to be areas or bands of relatively high pressure, most marked during the summer, in zones centered near latitude 30°N and 30°S. The Asiatic landmass is dominated by a great high-pressure system in winter and a low-pressure system in summer. Deep low-pressure areas prevail during the winter over the Aleutian, the Icelandic-Greenland, and Antarctic regions. These and other centers of action produce offshoots which may travel for great distances before dissipating.

Thus during the winter, spring, and autumn in middle latitudes over the land areas, it is fairly common to experience the passage of a cycle of low- and high-pressure systems in alternating fashion over a period of about 6–9 days in the average, but sometimes in as little as 3–4 days, covering a pressure amplitude which ranges on the average from roughly

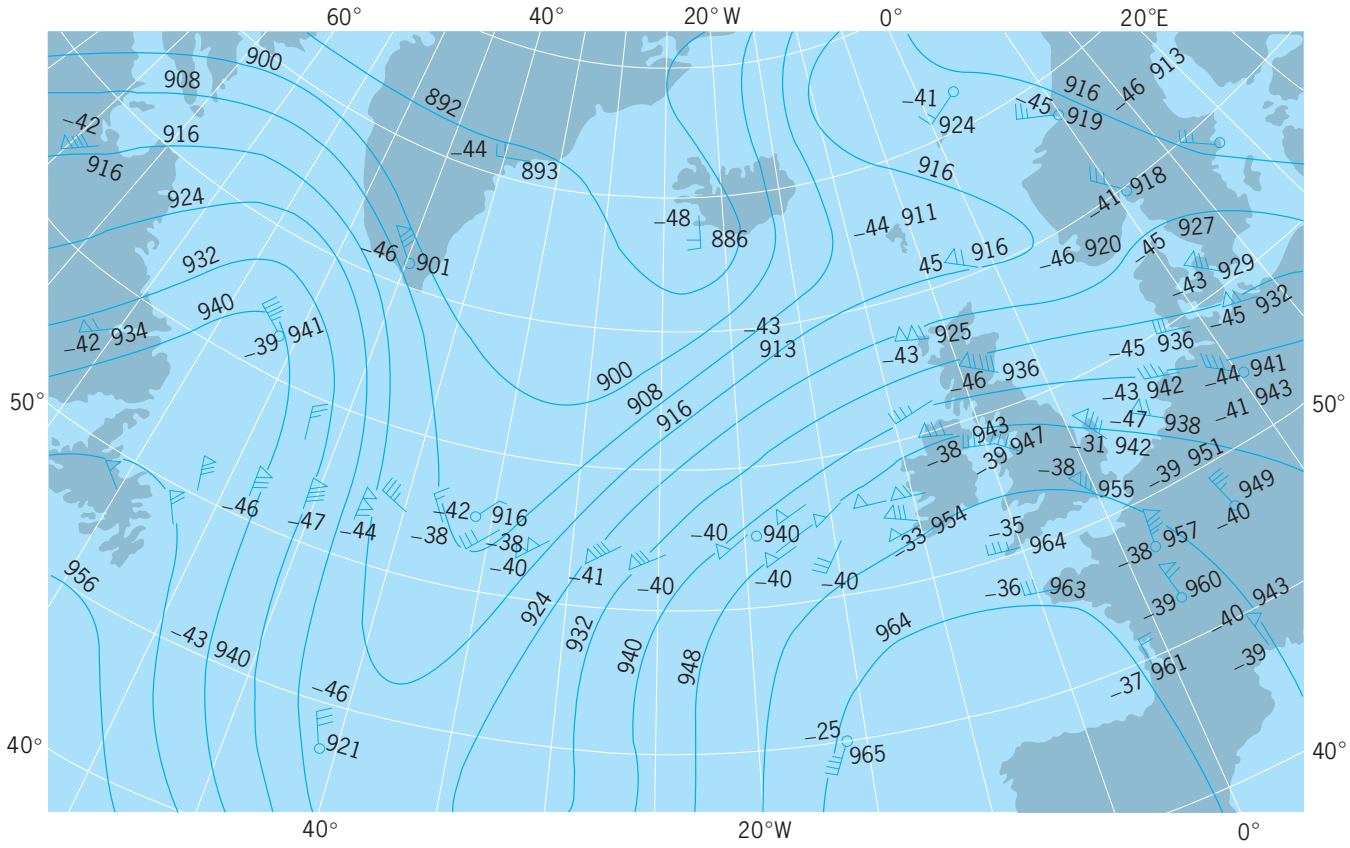
Mean atmospheric pressure and temperature in middle latitudes, for specified heights above sea level*

Altitude above sea level†			
Standard geopotential meters (m′)	Meters at latitude 45°32′40″	Air pressure, millibars	Assumed temperature, K‡
0	0	1.01325×10^3	288.15
11,000	11,019	2.2632×10^2	216.65
20,000	20,063	5.4747×10^1	216.65
32,000	32,162	8.6798×10^0	228.65
47,000	47,350	1.1090×10^0	270.65
52,000	52,429	5.8997×10^{-1}	270.65
61,000	61,591	1.8209×10^{-1}	252.65
79,000	79,994	1.0376×10^{-2}	180.65
88,743	90,000	1.6437×10^{-3}	180.65

*Approximate annual mean values based on radiosonde observations at Northern Hemisphere stations between latitudes 40 and 49° N for heights below 105,000 ft (32,000 m) and on observations made from rockets and instruments released from rockets. Some density data derived from searchlight observations were considered. Values shown above 105,000 ft (32,000 m) were calculated largely on the basis of observed distribution of air density with altitude. In correlating columns 1 and 2, the acceleration of gravity, G, is taken to be 98,066.5 cm^2/s^2 per standard geopotential meter (m′). Data on first three lines are used in calibration of aircraft altimeters. 1 m = 3.281 ft.

†Above 295,000 ft (90,000 m) there occurs an increase of temperature with altitude and a variation of composition of the air with height, resulting in a gradual decrease in molecular weight of air with altitude.

‡Temperature (°F) = 1.8 [temperature (K)] − 459.67.



Contours of 300-millibar surface, in tens of meters, with temperature in °C [°F = (°C × 1.8) + 32°] and measured winds at the same level. Winds are plotted with arrow pointing in direction of the wind, with each bar of the tail representing 10 m/s (33 ft/s or 19.4 knots). Triangle represents 50 m/s (165 ft/s or 97 knots).

15–25 millibars less than normal in the low-pressure center to roughly 15–20 millibars more than normal in the high-pressure center. During the summer in middle latitudes the period of the pressure changes is generally greater, and the amplitudes are less than in the cooler seasons (see **table**).

Within the tropics, where there are comparatively few passages of major high- and low-pressure systems during a season, the most notable feature revealed by the recording barometer (barograph) is the characteristic diurnal pressure variation. In this daily cycle of pressure at the ground there are, as a rule though with some exceptions, two maxima, at approximately 10 A.M. and 10 P.M., and two minima, at approximately 4 A.M. and 4 P.M., local time.

The total range of the diurnal pressure variation is a function of latitude as indicated by the following approximate averages (latitude N): 0°, 3 millibars; 30°, 2.5 millibars; 35°, 1.7 millibars; 45°, 1.2 millibars; 50°, 0.9 millibar; 60°, 0.4 millibar. These results are based on the statistical analysis of thousands of barograph records for many land stations. Local peculiarities appear in the diurnal variation because of the influences of physiographic features and climatic factors. Mountains, valleys, oceans, elevations, ground cover, temperature variation, and season exert local influences; while current atmospheric conditions also affect it, such as amount of cloudiness, precipitation, and sunshine. Mountainous regions in the western United States may have only a single maximum at about 8–10 A.M. and a single minimum at about 5–7 P.M., local time, but with a larger range than elsewhere at the same latitudes, especially during the warmer months (for instance, about 4 millibars difference between the daily maximum and minimum). *See* ATMOSPHERIC TIDES.

At higher levels in the atmosphere the variations of pressure are closely related to the variations of temperature, according to Eq. (3*a*). Because of the lower temperatures in higher latitudes in the lower 6 mi (10 km), the pressures at higher levels tend to decrease toward the poles. The **illustration** shows a typical pattern at approximately 6 mi (10 km) above sea level. As is customary in representing pressure patterns at upper levels, the variation of the height of a surface of constant pressure, in this case 300 millibars, is shown rather than the variation of pressure over a horizontal surface. *See* AIR TEMPERATURE.

Besides the latitudinal variation, the illustration also shows a wave pattern typical of the pressure field, and the midlatitude maximum in the wind field known as the jet stream, with its "waves in the westerlies." In the stratosphere the temperature variations are such as to reduce the pressure variations at higher levels, up to about 50 mi (80 km), except that in winter at high latitudes there are relatively large variations above 6 mi (10 km). At altitudes above 50 mi (80 km) the relative variability of the pressure increases again. Although the pressure and density at these very high levels are small, they are important

for rocket and satellite flights, so that their variability at high altitudes is likewise important.

Relations to wind and weather. The practical importance of air pressure lies in its relation to the wind and weather. It is because of these relationships that pressure is a basic parameter in weather forecasting, as is evident from its appearance on the ordinary weather map.

Horizontal variations of pressure imply a pressure force on the air, just as the vertical pressure variation implies a vertical force that supports the weight of the air, according to Eq. (1). This force, if unopposed, accelerates the air, causing the wind to blow from high to low pressure. The sea breeze is an example of such a wind. However, if the pressure variations are on a large scale and are changing relatively slowly with time, the rotation of the Earth gives rise to geostrophic or gradient balance such that the wind blows along the isobars. This situation occurs when the pressure variations are due to the slow-moving lows and highs that appear on the ordinary weather map, and to the upper air waves shown in the illustration, in which the relationship is well illustrated. *See* CORIOLIS ACCELERATION; GEOSTROPHIC WIND.

The wind near the ground, in the lowest few hundred meters of the atmosphere, is retarded by friction with the surface to a degree that depends on the smoothness or roughness of the surface. This upsets the balance mentioned in the previous paragraph, so that the wind blows somewhat across the isobars from high to low pressure.

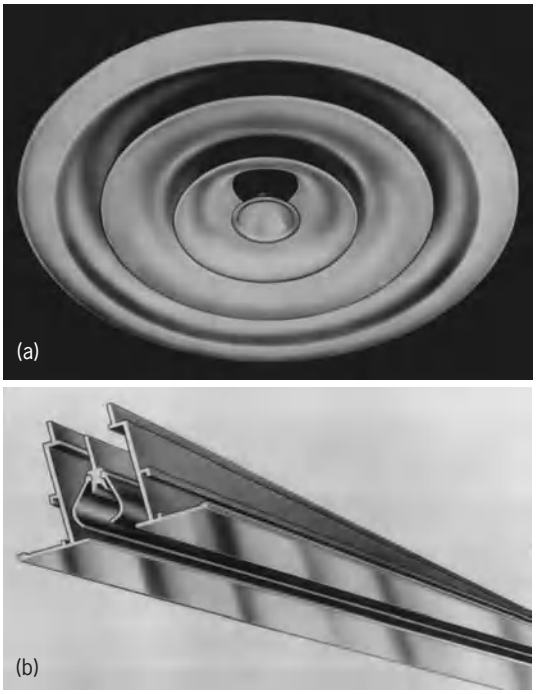
The large-scale variations of pressure at sea level shown on a weather map are associated with characteristic patterns of vertical motion of the air, which in turn affect the weather. Descent of air in a high heats the air and dries it by adiabatic compression, giving clear skies, while the ascent of air in a low cools it and causes it to condense and produce cloudy and rainy weather. These processes at low levels, accompanied by others at higher levels, usually combine to justify the clear-cloudy-rainy marking on the household barometer. Raymond J. Deland; Edwin Kessler

Bibliography. R. A. Anthes, *Meteorology*, 7th ed., 1996; R. G. Fleagle, *An Introduction to Atmospheric Physics*, 2d ed., 1980; D. D. Houghton (ed.), *Handbook of Applied Meteorology*, 1985; A. Miller and R. A. Anthes, *Meteorology*, 4th ed., 1980.

Air register

A device attached to an air-distributing duct for the purpose of discharging air into the space to be heated or cooled. These openings are referred to as registers, diffusers, supply outlets, or grills. By common acceptance, a register is an opening provided with means for discharging the air in a confined jet, whereas a diffuser is an outlet which discharges the air in a spreading jet (see **illus.**). Both registers and diffusers may be placed at a number of locations in a room, including the floor, baseboard, low or high on the sidewall, window sill,or ceiling.

For heating, the preferred location is in the floor, at the baseboard, or at the low sidewall of the outside wall, preferably under a window. For cooling, the preferred location is high on the inside wall or the ceiling. For year-round air conditioning in homes, a compromise location is the floor, baseboard, or low sidewall at the exposed wall, especially if adequate air velocity in an upward direction is provided at the supply outlet.



Some of the more common diffusers. (*a*) Round ceiling diffuser. (*b*) Single-slot ceiling diffuser.

A well-designed register effectively conceals the hole at the end of the duct, throws or projects the air in the direction and at the distance desired, limits the velocity usually to 1000 ft/min (300 m/min) or slower, and deflects the air away from walls and obstructions. The register also adjusts the direction of airflow to provide on-the-spot manipulation of the airstream, and adjusts the airflow rate to lesser amounts. It should accomplish these functions without producing dust streaks on nearby walls and ceilings, excessive air noise, or large pressure losses. Many registers, diffusers, slots, and air panels are commercially available and satisfy a majority of these qualifications. *See* OIL BURNER; WARM-AIR HEATING SYSTEM. Seichi Konzo

Bibliography. American Society of Heating, Refrigerating, and Air Conditioning Engineers, *Handbook and Product Directory*: *Equipment*, 1979; N. C. Harris, *Modern Air Conditioning Practice*, 3d ed., 1983.

Air separation

Separation of atmospheric air into its primary constituents. Nitrogen, oxygen, and argon are the primary constituents of air. Small quantities of neon, helium, krypton, and xenon are present at constant concentrations and can be separated as products. Varying quantities of water, carbon dioxide,

Composition of dry air

Component	Percent by volume	Component	Parts per million by volume
Nitrogen	78.084	Carbon dioxide	350–400
Oxygen	20.946	Neon	18.2
Argon	0.934	Helium	5.2
		Krypton	1.1
		Xenon	0.09
		Methane	1–15
		Acetylene	0–0.5
		Other hydrocarbons	0–5

hydrocarbons, hydrogen, carbon monoxide, and trace environmental impurities (sulfur and nitrogen oxides, chlorine) are present depending upon location and climate. Typical quantities are shown in the **table**. These impurities are removed during air separation to maximize efficiency and avoid hazardous operation. *See* AIR; ARGON; HELIUM; KRYPTON; NEON; XENON.

Three different technologies are used for the separation of air: cryogenic distillation, ambient temperature adsorption, and membrane separations. The latter two have evolved to full commercial status. Membrane technology is economical for the production of nitrogen and oxygen-enriched air (up to about 40% oxygen) at small scale. Adsorption technology produces nitrogen and medium-purity oxygen (85–95% oxygen) at flow rates up to 100 tons/day. The cryogenic process can generate oxygen or nitrogen at flows of 2500 tons/day from a single plant and make the full range of products.

Cryogenic air separation. The process of cryogenic distillation takes place at liquid-air temperature and has several steps: air compression; purification to remove water, carbon dioxide, and hydrocarbons; heat exchange between cooling air and warming products; distillation; and turbine expansion to provide refrigeration (**Fig. 1**). Products may be obtained as cryogenic liquids by incorporating additional refrigeration equipment. *See* CRYOGENICS; LIQUEFACTION OF GASES.

Air compression and purification. Air is compressed in a multistage centrifugal compressor to about 90 lb/in.^{2} absolute pressure (psia; 6.2 bars). The heat of compression is removed by water cooling. Condensed water is removed in a separator vessel. Water vapor, carbon dioxide, and hydrocarbons except for methane are then removed by adsorption onto solid adsorbent pellets of alumina or molecular sieves. These impurities would otherwise condense and accumulate in the cold process equipment. The adsorbent beds are regenerated at intervals by purging them with dry, hydrocarbon-free, waste gas from the distillation process.

Cooling and separation. The purified air is cooled to near liquefaction temperature in a heat exchanger, where cold product and waste streams are simultaneously rewarmed. The cold air enters the bottom of a multiple-column distillation system, where it separates into oxygen, nitrogen, and argon products and anitrogen-rich waste stream. The vapor in

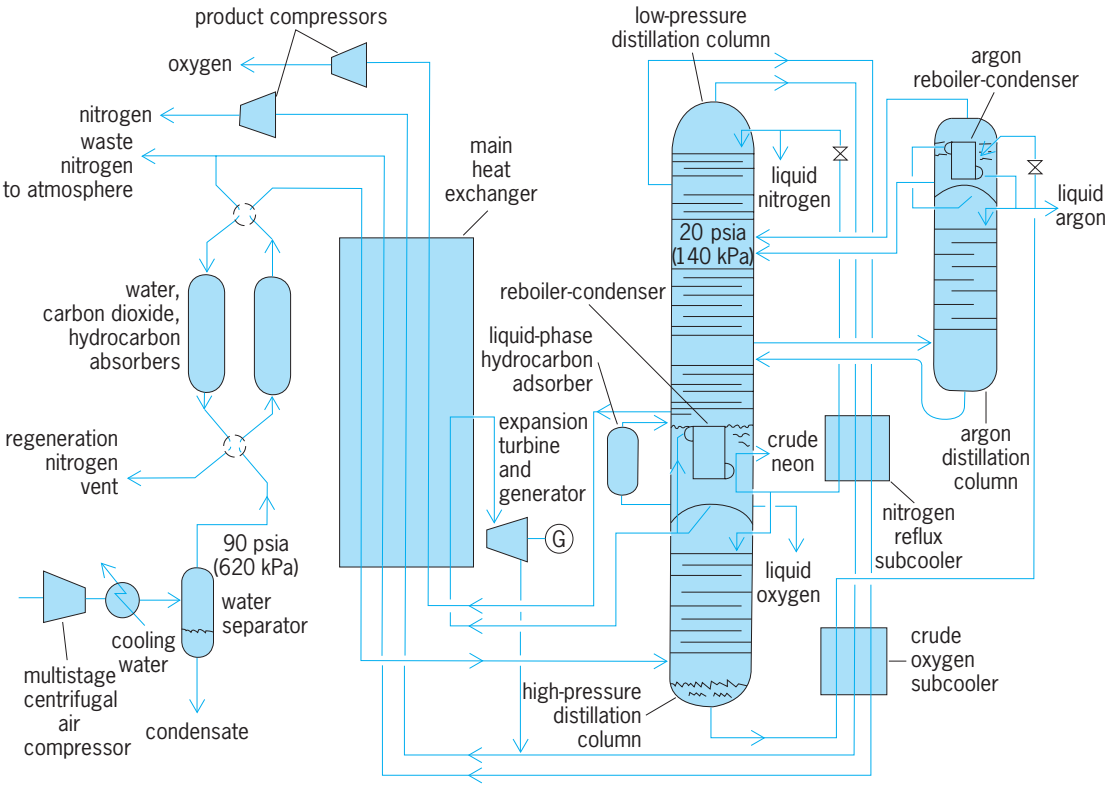


Fig. 1. Cryogenic air separation process. Air is compressed, purified, and cooled to liquefaction temperature; it is then distilled in a series of three columns to separate oxygen, nitrogen, and argon products.

equilibrium with a boiling liquid-air mixture contains an increased proportion of nitrogen. This principle is used to progressively enrich nitrogen in vapor and oxygen in liquid by successively contacting the streams in a series of stages. Thus as the air vapor rises through a sequence of liquid contact stages in the lower distillation column, it is depleted in oxygen until a pure nitrogen gas is obtained.

A small part of this nitrogen flows to the main heat exchanger, where it warms and expands through a turbine to provide process refrigeration. The rest of the nitrogen condenses to liquid in the reboiler-condenser, where oxygen is simultaneously boiled in the upper distillation column.

The condensed nitrogen divides into two streams. One stream passes to the upper column, and the other flows down the lower column and is enriched to an oxygen content of approximately 35%. This oxygen-rich stream is reduced in pressure and partly vaporized in a second reboiler-condenser that condenses argon column overhead vapor. It is then fed to the upper distillation column, which operates at a pressure of about 20 psia (1.4 bars). The difference in pressure between the upper and lower columns is determined by the temperature difference (about 2°F or 1°C) between the condensing nitrogen vapor and boiling oxygen liquid. In the upper column, the oxygen and nitrogen are separated to produce liquid oxygen at the bottom (purity > 99.5%) and nitrogen gas at the top.

The oxygen stream is boiled in the reboiler-condenser to provide both product gas and the reboil vapor for the upper column. A small stream of the liquid oxygen is circulated continuously through a silica gel adsorption bed to remove traces of acetylene, which might otherwise accumulate to hazardous levels in the oxygen reboiler. Nitrogen at the column top is purified by the reflux liquid nitrogen stream from the lower column. A second (waste) nitrogen stream containing a small percentage of oxygen is removed a few trays below. Argon has a boiling temperature between that of oxygen and nitrogen, and thus it accumulates in the middle of the distillation column. The argon-enriched stream (10% argon) is removed and is distilled in a third column to a purity of 96–99% argon. Further refining can provide high purity. While this stream has a small flow (<1% of feed air), it has a high value. Neon may be recovered from a noncondensed vent stream leaving the reboiler-condenser in the upper column; and a mixture of krypton and xenon from liquid oxygen at the bottom of the upper column.

The distillation columns have cylindrical shells containing numerous perforated metal plates that cause vapor to bubble through the liquid. An important innovation has been to replace the plates with high-efficiency structured packing. This modification reduces pressure loss in each stage and allows the use of more stages and lower operating pressure, which in turn reduces the energy requirement for air compression. There are many variations of the process arrangement to suit specific product flows, pressures, and purity. For example, to produce nitrogen alone, the upper columns are not required, and the enriched oxygen stream may be vaporized and expanded directly to provide refrigeration. The heat exchangers in the process have very large contact areas to provide high thermal efficiency. *See* DISTILLATION; HEAT EXCHANGER.

Adsorptive air separation. Adsorptive separation occurs when one component of a gas mixture preferentially accumulates onto a solid surface, thus depleting its concentration in the gas phase. Zeolite molecular sieves preferentially adsorb nitrogen over oxygen and argon at ambient temperature and pressure, and they are used in the vacuum-swing adsorption (VSA) or pressure-swing adsorption (PSA) process to produce oxygen (**Fig. 2**). Air is compressed by a centrifugal blower to a pressure of about 20 psia (1.4 bars) and then is cooled, and any water condensate is separated. The air passes through a bed of zeolite on which nitrogen is adsorbed. The oxygen product contains about 5% argon and up to 5% nitrogen. After a period (usually less than 1 min) the adsorbent bed is saturated with nitrogen and must be regenerated. At this time the airflow is switched to a second zeolite bed to maintain continuous oxygen production. *See* MOLECULAR SIEVE; ZEOLITE.

The saturated bed is evacuated to a pressure below about 5 psia (0.35 bar) to remove nitrogen. It is then repressurized to atmospheric pressure by using a part of the oxygen product from the on-line bed. This step ensures that the bed is clean before repeating the nitrogen adsorption step. Two or more beds may be used in the process with various pressure equalization steps to achieve maximum process efficiency. Oxygen may be obtained from this process at an energy consumption comparable to that required for cryogenic distillation (240 kWh/ton oxygen) but at a lower purity.

Nitrogen may also be produced by an adsorption process using carbon molecular sieves. Oxygen is adsorbed faster than nitrogen on such materials because of smaller molecular diameter. Air is compressed to about 125 psia (8.6 bars) and cooled, and water is separated. The dehydrated air then passes through the adsorbent bed, where oxygen is removed. The process again uses two or more beds operating in sequence to obtain a continuous nitrogen flow at purities to 0.1% oxygen. When high-purity nitrogen is required, residual oxygen may be reduced to part-per-million levels by catalytic reaction with added hydrogen. Once a bed is saturated with oxygen (<1 min), it is regenerated by depressurizing to atmospheric pressure. Process efficiency is enhanced by partial repressurization with some of the product nitrogen. In both adsorption processes, a computerized automatic control system is used to determine valve switching and optimize production. *See* ADSORPTION.

Membrane separation. Oxygen permeates faster than nitrogen through many organic polymers. This characteristic may be used for air separation. A typical membrane separator contains small hollow polymer fibers 100–500 micrometers in diameter and 1–3 m (3–10 ft) in length. These are assembled in

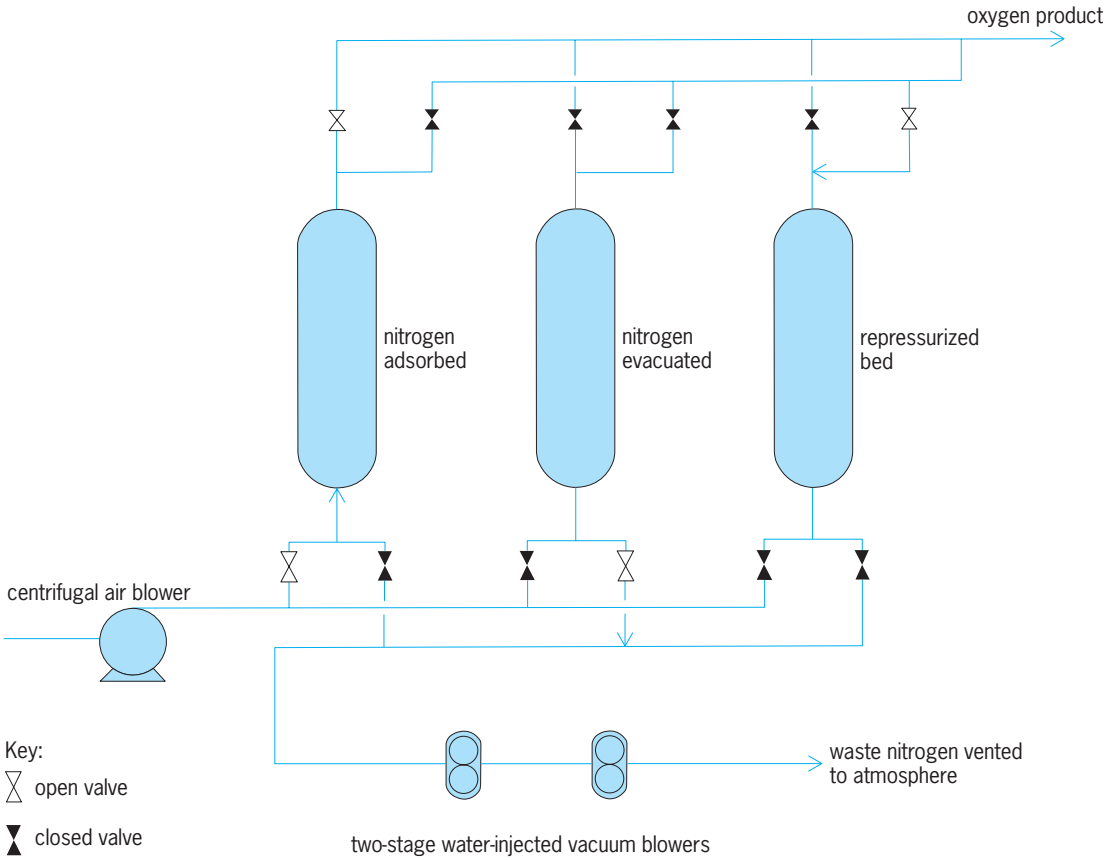


Fig. 2. Process for oxygen production by vacuum swing adsorption.

bundles of 0.1–0.25 m (0.3–0.8 ft) diameter. Polymer fibers used in commercial separators have very thin dense polymer layers as small as 35 nm that are supported on thicker porous walls. Air is compressed to about 160 psia (11 bars) and cooled, and condensed water is separated. The air then flows across one side of the membrane fibers. Oxygen permeates selectively through the fiber wall and is vented to the atmosphere. Some nitrogen permeates with the oxygen, depending upon the selectivity of the polymer used. Commercial polymers have permeation rate selectivities of about 6 for oxygen over nitrogen. Examples of polymers in use are polysulfone, polycarbonate, and polyimides.

Nitrogen is produced directly at pressures up to 150 psia (10.3 bars), with oxygen content down to 0.5% (**Fig. 3**). Higher nitrogen purity may be obtained and process efficiency improved by using additional membrane stages. The low oxygen content of the feed to the second stage produces a nitrogen-rich permeate that can be beneficially recycled to the feed to the first stage. Improved polymers will reduce the energy consumption required for gas separation and may allow the commercial production of oxygen as well as nitrogen. Small membrane units are used in oxygen enrichment for medical applications. *See* MEMBRANE SEPARATIONS; NITROGEN; OXYGEN.

Use for products. Air separation is a major industry. Nitrogen and oxygen rank second and third in the scale of production of commodity chemicals; and air is the primary source of argon, neon, krypton, and xenon. Oxygen is used for steel, chemicals manufacture, and waste processing. Important uses are in integratedgasification combined

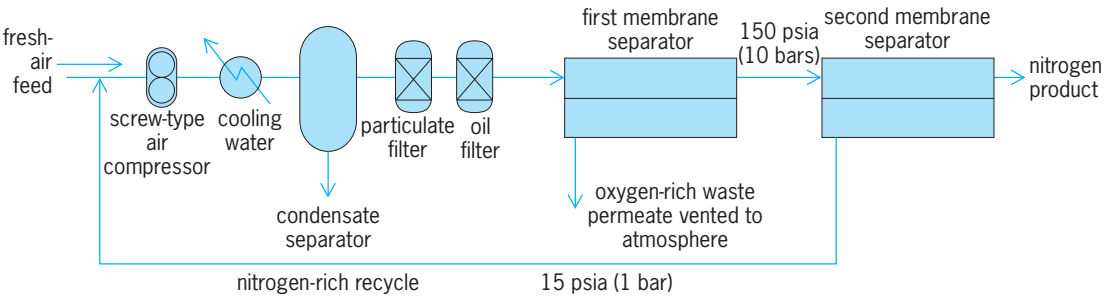


Fig. 3. Membrane process for nitrogen separation.

cycle production of electricity, waste water treatment, and oxygen-enriched combustion. Nitrogen provides inert atmospheres for fuel, steel, and chemical processing and for the production of semiconductors. Robert M. Thorogood

Bibliography. T. Flynn, *Cryogenic Engineering*, 2d ed., 2004; R. H. Perry and D. Green (eds.), *Perry's Chemical Engineers Handbook*, 7th ed., 1997; R. Rousseau (ed.), *Handbook of Separation Process Technology*, 1987; D. M. Ruthven, *Principles of Adsorption and Adsorption Processes*, 1984; *Separation and Purification Technology* (journal).

Air temperature

The temperature of the atmosphere represents the average kinetic energy of the molecular motion in a small region, defined in terms of a standard or calibrated thermometer in thermal equilibrium with the air. *See* TEMPERATURE.

Measurement. Many different types of thermometer are used for the measurement of air temperature, the more common depending on the expansion of mercury or alcohol with temperature, the variation of electrical resistance with temperature, or the thermoelectric effect (thermocouple). The electrical methods are especially useful for the automatic recording of temperature. The basic problems of ensuring that the temperature of the thermometer be as close as possible to that of the air are the same for all methods of measurement. For the atmosphere, probably the most serious difficulty is the heating or cooling of the thermometer by radiation to and from other bodies at different temperatures, the most obvious being the Sun. The representativeness of temperature measurements, meaning the degree to which they provide information about the temperature of the air over a region much larger than the thermometer, is also an important practical requirement. The well-known standard meteorological measurement of the air temperature in a louvered shelter, about 6.5 ft (2 m) above a natural ground surface, with a mercury-in-glass thermometer that averages the temperature over a period of about 1 min because of its thermal inertia, is designed to satisfy the above requirements. *See* METEOROLOGICAL INSTRUMENTATION; TEMPERATURE MEASUREMENT; THERMOMETER.

Causes of variation. The temperature of a given small mass of air varies with time because of heat added or subtracted from it, and also because of work done during changes of volume, according to alternate Eqs. (1) and (2). Here h represents heat added,

$$\frac{dT}{dt} = \frac{1}{C_v}\left(\frac{dh}{dt} - P\frac{d\alpha}{dt}\right) \quad (1)$$

$$\frac{dT}{dt} = \frac{1}{C_p}\left(\frac{dh}{dt} + \frac{1}{\rho}\frac{dP}{dt}\right) \quad (2)$$

P the pressure, ρ the density and α its reciprocal, and C_v and C_p the specific heat at constant volume and constant pressure, respectively. The heat added or subtracted may be due to many different physical processes, of which the most important are absorption and emission of radiation, heat conduction, and changes of phase of water involving latent heat of condensation and freezing. In the upper atmosphere, above about 12 mi (20 km), photochemical changes are also important; for example, those that occur when ultraviolet radiation dissociates oxygen molecules to atomic oxygen, which then recombines with molecular oxygen to form ozone. Because of the variation of air pressure with height, rising and sinking of air causes expansion and contraction and thus temperature changes due to the work of expansion, explicitly represented by the second term in parentheses in Eq. (1).

A spectacular example of temperature rise due to sinking of air from higher levels is the chinook, a warm wind that sometimes blows down the eastern slope of the Rocky Mountains in winter. The slower seasonal temperature changes are mainly due to a combination of radiational heat exchange and conduction to and from the ground surface, whose temperature itself changes in response to the varying radiational exchange with the Sun and the atmosphere. On a shorter time scale the diurnal variation of temperature throughout the day is caused by the same processes. *See* AIR PRESSURE; ATMOSPHERIC GENERAL CIRCULATION; CHINOOK.

The rate at which the temperature changes at a particular point, that is, as measured by a fixed thermometer, depends on the movement of air as well as the physical processes discussed above. Large changes of air temperature from day to day are mainly due to the horizontal movement of air, bringing relatively cold or warm air masses to a particular point, as large-scale pressure-wind systems move across the weather map. *See* AIR MASS.

Temperature near the surface. Temperatures are read at one or more fixed times daily, and the day's extremes are obtained from special maximum and minimum thermometers, or from the trace (thermogram) of a continuously recording instrument (thermograph). The average of these two extremes, technically the midrange, is considered in the United States to be the day's average temperature. The true daily mean, obtained from a thermogram, is closely approximated by the mean of 24 hourly readings, but may differ from the mid-range by 1 or 2°F (0.6 or 1°C), on the average. In many countries temperatures are read daily at three or four fixed times, chosen so that their weighted mean closely approximates the true daily mean. These observational differences and variations in exposures complicate comparison of temperatures from different countries and any study of possible climatic changes.

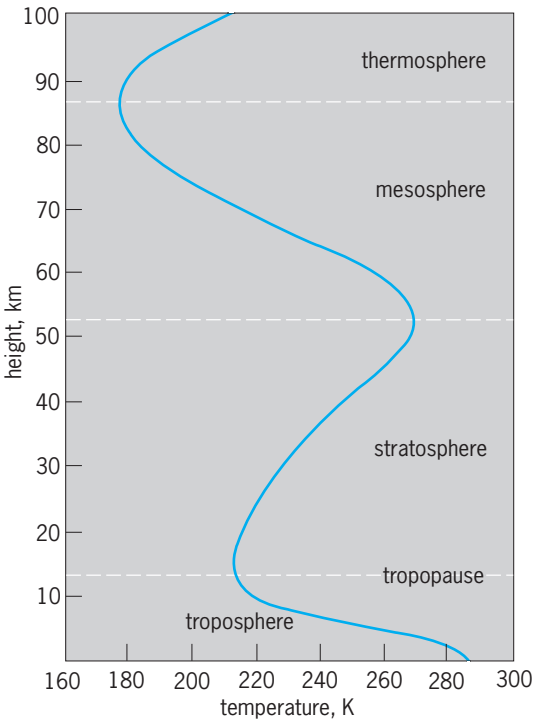
Averages of daily maximum and minimum temperature for a single month for many years give mean daily maximum and minimum temperatures for that month. The average of these values is the mean monthly temperature, while their difference is the mean daily range for that month. Monthly means, averaged through the year, give the mean annual temperature; the mean annual range is the difference between the hottest and coldest mean monthly values.

The hottest and coldest temperatures in a month are the monthly extremes; their averages over a period of years give the mean monthly maximum and minimum (used extensively in Canada), while the absolute extremes for the month (or year) are the hottest and coldest temperatures ever observed. The interdiurnal range or variability for a month is the average of the successive differences, regardless of sign, in daily temperatures.

Over the oceans the mean daily, interdiurnal, and annual ranges are slight, because water absorbs the insolation and distributes the heat through a thick layer. In tropical regions the interdiurnal and annual ranges over the land are small, because the annual variation in insolation is relatively small. The daily range also is small in humid tropical regions, but may be large (up to 40°F or 22°C) in deserts. Interdiurnal and annual ranges increase generally with latitude, and with distance from the ocean; the mean annual range defines continentality. The daily range depends on aridity, altitude, and noon sun elevation.

Extreme temperatures arouse much popular interest and often are cited uncritically, despite their possible instrumental, exposure, and observational errors of many kinds. The often given absolute maximum temperatures of 134°F (57°C) for the United States in Death Valley, California (July 10, 1913), and 136°F (58°C) for the world in Azizia, Tripoli (September 13, 1922) are both questionable; in the subsequent years, Death Valley's hottest reading has been only 127°F (53°C), and the Azizia reading was reported by an expedition, not a regular weather station. Lowest temperatures in the Northern Hemisphere are −90°F (−68°C) at Verkhoyansk (−89.7°F or −67.7°C on February 5 and 7, 1982) and Oimekon (−89.9°F or −67.7°C on February 6, 1933), Siberia; −87°F (−66°C) at Northice, Greenland (January 9, 1954); −81°F (−63°C) at Snag, Yukon Territory, Canada (February 3, 1947); −70°F (−57°C) at Rogers Pass, Montana (the current United States record, on January 20, 1954). The first winter at Vostok, 78°27′S, 106°52′E, encountered a minimum temperature of −125°F (−87.2°C) on August 25, 1958, and the third winter a minimum of −127°F (−88.3°C) on August 24, 1960, much lower than the lowest at the United States station at the South Pole. At Vostok on July 21, 1983, a new global minimum temperature record for Earth's surface was recorded, −129°F (−89.4°C). *See* ANTARCTICA.

Vertical variation. The average vertical variation of temperature in the atmosphere is shown in the **illustration.** The atmosphere is seen to consist of layers, each of which has a characteristic variation of temperature with height. The decrease of temperature with height in the lowestlayer, the troposphere, is basically due to the presence of a heat source resulting from the solar radiation absorbed at the Earth's surface, giving an excess of heat that is carried away from the surface mainly by convection currents and lost to space by reradiation. Heating a compressible fluid such as air from below results in a decrease in temperature with height because rising masses of air cool as they expand, according to Eq. (2). This is the main reason for the decrease of temperature with height in the troposphere. *See* ATMOSPHERE; HEAT BALANCE, TERRESTRIAL ATMOSPHERIC.



Average temperature distribution with height, from the International Reference Atmosphere. Note that the nomenclature for stratosphere and mesosphere varies, the upper part of the stratosphere being considered part of the mesosphere by some authorities. (*After R. G. Fleagle and J. A. Businger, An Introduction to Atmospheric Physics, Academic Press, 1963*)

In the mesosphere and higher layers the exchange of heat energy between layers of air by emission and absorption of infrared radiation is the most important factor determining the distribution of temperature with height. *See* INSOLATION; RADIATION.

Raymond J. Deland; Edwin Kessler

Bibliography. R. A. Anthes, *Meteorology*, 7th ed., 1996; R. G. Fleagle, *An Introduction to Atmospheric Physics*, 2d ed., 1980; D. D. Houghton (ed.), *Handbook of Applied Meteorology*, 1985; A. Miller and R. A. Anthes, *Meteorology*, 4th ed., 1980; H. Riehl, *Introduction to the Atmosphere*, 3d ed., 1978.

Air-traffic control

A service to promote the safe, orderly, and expeditious flow of air traffic. Safety is principally a matter of preventing collisions with other aircraft, obstructions, and the ground; assisting aircraft in avoiding hazardous weather; assuring that aircraft do not operate in airspace where operations are prohibited; and assisting aircraft in distress. Orderly and expeditious flow assures the efficiency of aircraft operations along the routes selected by the operator. It is provided through the equitable allocation of system resources to individual flights, generally on a first-come-first-served basis.

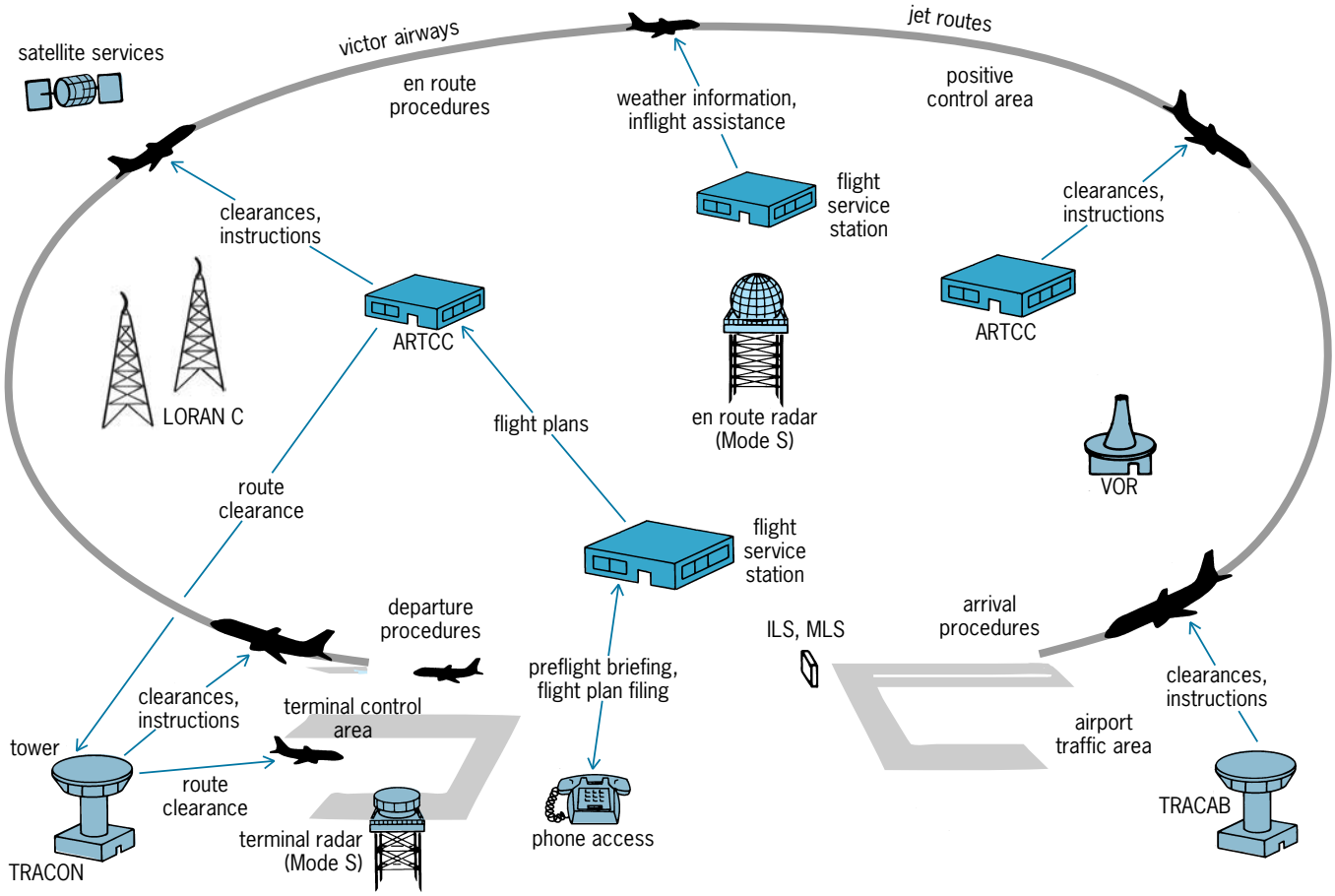


Fig. 1. Elements of air-traffic control.

In the United States, air-traffic control (ATC) is the product of the National Airspace System (NAS), comprising airspace; air navigation facilities and equipment; airports and landing areas; aeronautical charts, information, and publications; rules, regulations, and procedures; technical information; and personnel (**Fig. 1**).

Rules

The rules governing air-traffic control in the United States are published in Federal Aviation Regulations (FARs); two principal categories of rules are visual flight rules (VFR) and instrument flight rules (IFR).

Visual flight rules. These govern the procedures for conducting flight where the visibility, the ceiling, and the aircraft distance from clouds are equal to or greater than established minima. Ceiling is the height above the Earth's surface of the lowest layer of clouds or obscuring phenomenon that significantly restricts visibility. The minima for operation under visual flight rules vary by airspace. In controlled airspace, the ceiling must be at least 1000 ft (305 m) and the visibility must be at least 3 statute miles (4830 m). The aircraft must remain clear of clouds, at least 500 ft (150 m) below, 1000 ft (305 m) above, and 2000 ft (610 m) horizontally.

Instrument flight rules. When the meteorological conditions expressed in terms of visibility, distance from clouds, and ceiling are less than the minima specified for visual flight rules, that is, less than the minima for visual meteorological conditions (VMC), instrument meteorological conditions (IMC) prevail. Instrument flight rules govern flight under such conditions. To operate under these rules, the pilot must pass an instrument flight examination and have an adequately instrumented aircraft.

Flight plans. Instrument flight rules and visual flight rules also refer to types of flight plans. A flight plan is filed with the authority providing air-traffic control services [in the United States, the Federal Aviation Administration (FAA)] to convey information about the intended flight of the aircraft. Both types of flight plans contain essentially the same information, that is, aircraft identification number, make and model, and color; planned true airspeed and cruising altitude; origin and destination airports; planned departure time and estimated time en route; planned route of flight, fuel, and number of people on board; pilot's name and address; navigation equipment on board; and the aircraft's radio call sign, if different from the aircraft identification number.

Generally, a flight plan is not required for a flight under visual flight rules. However, if a flight plan is filed for such a flight and the aircraft is overdue at its destination, search and rescue procedures will be initiated. Hence the flight plan under visual flight rules provides a significant safety benefit. An IFR flight plan

is required for operation in controlled airspace when instrument meteorological conditions prevail.

Aircraft operating under visual flight rules (VFR aircraft) maintain separation from other aircraft visually; that is, the flight crew visually acquires other aircraft in the vicinity, and the pilot maneuvers the aircraft to maintain separation (a procedure called see-and-avoid). IFR aircraft in controlled airspace operate in accordance with clearances and instructions provided by air-traffic controllers for the purpose of maintaining separation and expediting the flow of traffic. (Formally, an air-traffic control clearance is an authorization by air-traffic control, for the purpose of preventing collisions between aircraft, for an aircraft to proceed under specified traffic conditions within controlled airspace. An air-traffic control instruction is a directive issued by air-traffic control for the purpose of requiring a pilot to take specific actions, for example, "turn left to heading two five zero.") Flight crews operating under instrument flight rules are responsible for seeing and avoiding other aircraft, but the air-traffic control clearances they receive provide substantial added assurance of safe separation. Consequently, flight crews often will operate under instrument flight rules even though the weather satisfies visual meteorological conditions.

Airspace

The two principal categories are controlled and uncontrolled airspace. In controlled airspace some or all aircraft are required to operate in accordance with air-traffic control clearances in order to assure safety, to meet user needs for air-traffic control, or to accommodate high volumes of traffic. Air-traffic control services including air-to-ground communications and navigation aids are provided in controlled airspace.

Two specific examples of controlled airspace are class A (the positive control area or PCA) and class B (the terminal control area or TCA). The positive control area is, with a few exceptions, the airspace within the conterminous 48 states and Alaska extending from 18,000 to 60,000 ft (5490 to 18,290 m) above mean sea level. All aircraft in a positive control area must operate in accordance with instrument flight rules and carry prescribed equipment.

Terminal control areas are centered on primary airports and extend from the surface to specified altitudes. An air-traffic control clearance and prescribed equipment are required prior to operating within a terminal control area regardless of weather conditions. Uncontrolled airspace simply is airspace that has not been designated as controlled; air-traffic control services may not be available in such airspace.

Equipment

An array of sophisticated equipment is required for providing air-traffic control services. Principal categories are air-to-ground communications, radio navigation aids, surveillance systems, and air-traffic control automation.

Air-to-ground communications. Two-way air-to-ground voice communications between civil pilots and air-traffic controllers are conducted in the very high frequency (VHF) band. In the United States there are approximately 2400 communications facilities providing access to air-traffic control services from terminal and en route facilities as well as from flight service stations (FSSs). In addition, certain radio navigation aids [nondirectional radio beacons (NDBs) and very high frequency omnidirectional range (VOR) stations] can provide one-way communications from controllers to aircraft. These channels generally are used to broadcast weather and aeronautical information to pilots. *See* RADIO SPECTRUM ALLOCATIONS.

Air-to-ground data communications (that is, data link) increasingly are used to transfer information to and from the cockpit. A principal advantage is that many of the communications errors associated with humans incorrectly reading, speaking, and hearing text are eliminated by communications protocols that detect errors in data transmissions, by electronically displaying the information received in the cockpit and the air-traffic control facility, and by storing the received information so that it can be recalled and reviewed by pilots and controllers. Data link also permits large quantities of data to be exchanged between ground-based and airborne computers. Civil aviation is exploiting three data-link media: some VHF voice channels have been converted to data-link channels; Mode S provides a data-link capability, both air to ground and air to air; and communications satellites provide data transmission capabilities.

Radio navigation aids. Radio navigation aids are used to determine the plan position of the aircraft (that is, the position in the horizontal plane) in coordinates referenced either to the navigation aid or to the Earth (that is, latitude and longitude). For most operations, the aircraft vertical position is determined by sensing atmospheric pressure on board and converting this pressure to altitude, based on a standard model of the atmosphere. For the landing phase of flight, precision landing aids provide horizontal and vertical position referenced to the runway. *See* ALTIMETER.

VOR. This is a principal system used for determining plan position, with approximately 1000 ground stations nationwide. The system provides the magnetic azimuth from the VOR station to the receiving aircraft accurate to $\pm 1^\circ$. Position determinations can be obtained from the intersection of radials from VORs with overlapping coverage volumes. A radial is a line of position extending from the VOR station at the indicated magnetic azimuth of the aircraft. With the addition of distance-measuring equipment at a VOR station, it is possible to obtain a position determination from a single station. *See* DISTANCE-MEASURING EQUIPMENT; RHO-THETA SYSTEM; VOR (VHF OMNIDIRECTIONAL RANGE).

Nondirectional radio beacon. This is an older technology, with few installations remaining. The system radiates a continuous signal from which direction-finding receivers can determine the azimuth to the ground station. *See* DIRECTION-FINDING EQUIPMENT.

Loran C. This is a pulsed system, with chains of ground stations each consisting of one master station and at least two secondary stations organized to transmit their signals at precisely defined time epochs. Loran C coverage in the United States includes the conterminous 48 states and southern Alaska. *See* LORAN.

Landing systems. In order to conduct approaches and landings in low-visibility conditions, it is necessary that an electronic glideslope (or glidepath) be provided as a reference for controlling the descent of the aircraft to the runway. In addition, a stable guidance signal is required to align the aircraft with the runway centerline.

The instrument landing system (ILS) has been the standard means for providing precision landing guidance to the runway, and is installed on approximately 1000 runways in the United States. The localizer antenna transmits the lateral (left and right) guidance signal over a 20° sector, 10° on both sides of the extended runway centerline. The glideslope antenna transmits the elevation guidance signal over a 1.4° sector, 0.7° on both sides of the glidepath, which is normally 3.0° above the horizontal. *See* INSTRUMENT LANDING SYSTEM (ILS).

A new standard system for providing precision approach guidance, the microwave landing system (MLS) has been designed to eliminate limitations of the instrument landing system. It utilizes scanning-beam technology to provide proportional landing guidance over 80° in azimuth (40° on both sides of the extended runway centerline) and 15° in elevation. In combination with a precision version of distance-measuring equipment, the system can provide three-dimensional landing guidance within the scanned volume, thereby permitting curved approaches and approaches at higher glideslope angles than those available from the instrument landing system. In the United States, the development of MLS for civilian applications has been terminated. A mobile version of MLS (MMLS) continues to be used for military operations. *See* MICROWAVE LANDING SYSTEM (MLS).

Global Positioning System. The constellation of Global Positioning System (GPS) satellites provides a highly accurate worldwide position determination and time transfer capability. In the horizontal plane, the position determined by a GPS receiver is within approximately 108 ft (33 m) of the true receiver position at least 95% of the time, averaged over the globe. The vertical position is accurate to within approximately 240 ft (73 m) on the same 95% probability basis. In addition, the receiver provides Coordinated Universal Time (UTC) with an accuracy of 50 nanoseconds (95% probability), averaged over the globe. During times of significant ionospheric activity, the timing accuracy degrades to approximately 120 nanoseconds at the worst location. Coordinated Universal Time is an internationally accepted time standard that never differs from Greenwich Mean Time by more than 1 s. The principal advantages of GPS are its accuracy and worldwide coverage. GPS is approved for oceanic, en route, area navigation (RNAV), and nonprecision approach operations. GPS receivers that incorporate correction signals from the Wide Area Augmentation System (WAAS) have accuracies of approximately 5–7 ft or 1.5–2 m (95% probability) in both horizontal and vertical dimensions. GPS/WAAS is approved for aircraft guidance down to 200 ft (61 m) above the runway surface for localizer performance with vertical guidance (LPV) approaches. LPV is an ILS-like capability without the need for ground equipment at the airport. *See* AIR NAVIGATION; ELECTRONIC NAVIGATION SYSTEMS; SATELLITE NAVIGATION SYSTEMS.

Surveillance systems. Air-traffic controllers use radar to monitor the positions of aircraft in their sectors of responsibility and to monitor areas of heavy precipitation. The radar information is used to develop clearances and instructions for separating aircraft operating under instrument flight rules, and to provide traffic advisories to IFR aircraft and to VFR aircraft receiving the traffic advisory service. Traffic advisories assist a pilot in seeing and avoiding other aircraft. They provide the ranges, bearings, and altitudes of aircraft in the pilot's immediate vicinity. The pilot is responsible for visually acquiring and avoiding any traffic that may be a collision threat. Two principal types of radar are used in civil air-traffic control: secondary, or beacon, radar and primary radar. *See* RADAR.

Secondary radar. This is an interrogate-respond system. The rotating directional antenna of the ground station transmits a pulse pair to the transponder in the aircraft. The pulse spacing encodes one of two messages, "transmit your altitude" (the Mode C interrogation) or "transmit your identity" (the Mode A interrogation). The aircraft transponder transmits an encoded pressure-altitude reply in response to the first interrogation and a four-digit identity code, assigned by air-traffic control and entered into the transponder by the pilot, in response to the second. The aircraft is shown on the controller's plan view display at the azimuth corresponding to the pointing direction of the antenna and the range corresponding to the round-trip time between transmission of the interrogation and receipt of the reply. Air-traffic control computers receive the encoded reply data from radar sites and place corresponding information in data blocks next to the symbols depicting the aircraft positions on the display. The identity code assigned by air-traffic control is correlated with the flight-plan database in the computer to display the radio call sign in the data block. The aircraft pressure altitude is displayed in hundreds of feet.

Primary radar. This operates by transmitting high-power, radio-frequency pulses from a rotating directional antenna. The energy is reflected from any aircraft in the directional beam and received by the antenna. The aircraft is displayed at the azimuth corresponding to the pointing direction of the antenna and the range corresponding to the round-trip time between pulse transmission and receipt of the reflected signal.

Comparison of radar systems. Primary radar has the advantage that aircraft without air-traffic control transponders can be detected, and energy reflected from heavy precipitation indicates to the controller areas of potentially hazardous weather. Secondary radar ground stations interrogate at 1030 MHz, and the transponder replies at 1090 MHz. The use of different frequencies eliminates extraneous returns (clutter) from surrounding buildings and terrain that can reduce the effectiveness of primary radar in detecting aircraft. Secondary radar also has the advantage that aircraft identity and altitude can be determined. At most air-traffic control radar sites, the secondary radar antenna is mounted on the primary radar antenna, and they are turned by a common drive system. In the United States, there are 220 radar systems operated in terminal areas and 116 long-range radars in the en route environment.

Mode S. The secondary radar system has been improved through the addition of Mode S, which employs more sophisticated signaling formats than Modes A and C. Each aircraft transponder is permanently assigned a unique address and interrogations therefore can be addressed to individual aircraft. Two benefits accrue. For a number of technical reasons, the accuracy and reliability of the surveillance data are improved. Second, it is possible to establish direct data communications with individual aircraft. The data link can be used to provide air-traffic control clearances to aircraft and to convey a wide variety of aeronautical and weather information in response to requests from the flight crew.

Oceanic surveillance. In the oceanic environment, the ground-based surveillance systems described above obviously cannot be used. Oceanic operations are now based on rigid procedures and high-frequency (HF) communications that sometimes are unreliable. With the advent of commercially available mobile satellite communication systems, the development of a technique called automatic dependent surveillance (ADS) has been undertaken to provide real-time position information from aircraft over the ocean. In the operation of this system, the position of the aircraft, as determined from on-board navigation sensors, is communicated to air-traffic control facilities when requested by satellite relay. This position information can be displayed to controllers as though it had been determined by a radar system. With automatic dependent surveillance and reliable pilot-controller communications via satellite, a number of improvements in oceanic procedures are possible, with corresponding benefits for safety and aircraft operating efficiency.

Automation. The principal elements of the controller's workstation are the plan view display, a track ball or mouse, the data-entry keyboard, printed flight strips showing the flight plans of aircraft for which the controller is responsible, and interfaces with communications facilities linking the controller with aircraft and with other controllers and air-traffic control facilities. The plan view display shows two principal types of data, map data and radar data (**Fig. 2**). Map data include the locations of airports and their runways, navigation aids, airways, obstructions, and the geographical limits of the facility's airspace. Radar data comprise the positions of aircraft, including their altitudes, ground speeds, and radio callsigns, as well as areas of precipitation. The

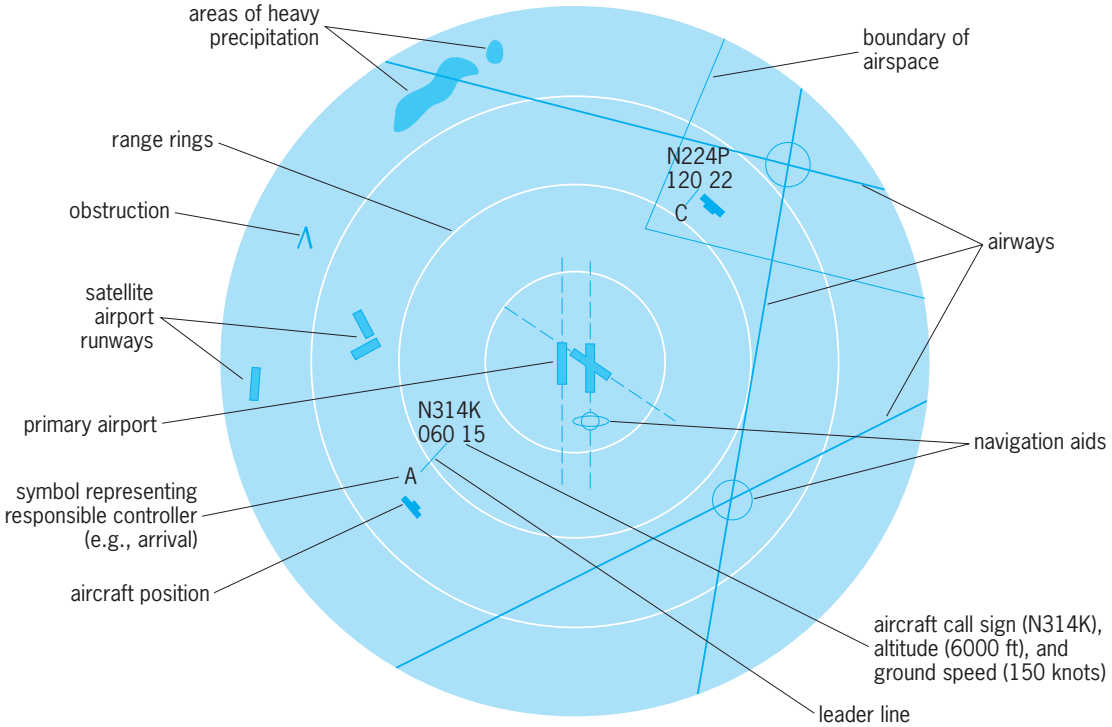


Fig. 2. Terminal radar controller's plan view display. Aircraft positions are shown by radar returns. Display is centered on primary airport with parallel and intersecting runways. Range rings are spaced 5 mi (9 km) apart. Airspace within boundary and above 10,000 ft (3050 m) belongs to the center.

data-entry keyboard allows the controller to modify data stored in the automation system, including flight plans. Extensive automation (computer) equipment is used in maintaining the flight-plan databases and processing radar data.

A number of automation aids have been developed to assist controllers in separating aircraft as well as in sequencing and metering aircraft into and out of busy terminal areas. The objective of these aids is to enhance the safety and efficiency of flight operations.

Flight management computer systems are installed in aircraft for the purpose of guiding the aircraft along its planned route of flight while minimizing operating costs by selecting optimum speeds and altitudes. Extensive databases are stored in the flight management computer system (FMCS), including the current flight plan, wind velocities and air temperatures along the planned route of flight, and the positions and operating frequencies of the radionavigation aids to be used. Interfaces with the FMCS for air-to-ground data communications permit changes to be made to the databases in flight and allow information to be extracted, such as automatic dependent surveillance position reports and estimated times of arrival at specific points along the planned route of flight. The integration of FMCS operations with ground-based air-traffic control operations is a principal theme in the further development of the art of air-traffic control. *See* AIRCRAFT INSTRUMENTATION.

Airborne collision avoidance systems. These systems are installed in aircraft to provide ground-independent protection from midair collisions, as a backup to the conventional air-traffic control system. Within the United States, the system is known as the Traffic Alert and Collision Avoidance System (TCAS). The TCAS equipment in the aircraft interrogates the secondary surveillance radar transponders in proximate aircraft and processes the replies to determine if any aircraft is on a collision course. Traffic advisories are displayed to the pilot to portray the range, bearing, and relative altitude of any aircraft that penetrates a protection volume around the TCAS-equipped aircraft. The pilot's response to a traffic advisory is to visually search for the intruder and to maintain visual separation once the intruder is acquired. If the intruder continues to close on the TCAS-equipped aircraft, a resolution advisory will be displayed to tell the pilot how to maneuver to avoid a collision. *See* AIRCRAFT COLLISION AVOIDANCE SYSTEM.

Airways and Procedures

Two fixed-route systems have been established for air navigation. From 1200 ft (360 m) above the surface up to but not including 18,000 ft (5490 m) above mean sea level, there are designated airways based on VORs and nondirectional beacons. The most prevalent are the so-called victor (V) airways defined by VORs. For example, in western New York, airway V115 extends from the Buffalo VOR to the Jamestown VOR. Jet (J) routes are defined from 18,000 to 45,000 ft (5490 to 13,710 m) above mean sea level, based solely on VORs.

Aircraft may navigate from one terminal area to another by following fixed routes from one VOR to the next. In this case, the route shown on the flight plan would list the airways to be flown by their V and J designations.

There are three principal categories of procedures: departure procedures for leaving terminal areas, arrival procedures for entering terminal areas, and en route procedures. Departure procedures prescribe the process for route clearance delivery to an aircraft, for providing takeoff runway and taxi instructions, and for defining or placing limitations on the climb-out route of the aircraft to the en route environment. Generally, pilots of IFR aircraft call the clearance delivery controller for their route clearance prior to taxiing. The route in the clearance may differ from the filed route because of system restrictions such as excess traffic, facility outages, and weather.

En route procedures deal principally with reporting aircraft flight progress to air-traffic control (position reporting) when the aircraft is outside radar coverage or is operating in holding patterns.

Arrival procedures prescribe the process for making the transition from the en route structure to the terminal area, for approaching the landing runway, and for executing a missed approach when a landing cannot be accomplished. An instrument approach procedure is a series of predetermined maneuvers by reference to flight instruments for the orderly transfer of an aircraft from an initial approach fix to a landing or to a point from which a landing can be made visually. Several procedures, using different navigation and approach aids, may be established for an airport.

Facilities

Air-traffic control facilities include flight service stations, air-route traffic control centers (ARTCCs), and terminal facilities.

Flight service stations. These provide preflight briefings for pilots, accept flight plans, broadcast aviation weather information, assist lost aircraft and aircraft in distress, and monitor the operation of radio navigation aids. The preflight briefing provides weather and aeronautical information pertinent to the proposed flight, including any facility outages that may be relevant and any known air-traffic control delays that may affect the flight. There are 75 flight service stations in the United States.

Air-route traffic control centers. These monitor all IFR aircraft not under the control of military or terminal facilities. In particular, they control all traffic operating in positive-control airspace. They assure separation of IFR aircraft by issuing clearances and instructions as necessary and issuing traffic advisories to aid see-and-avoid, provide weather advisories, accept amendments to flight plans from flight crews, and assist aircraft in distress.

Flight plans submitted to flight service stations usually are transmitted to the parent air-route

traffic control center, where they are processed and the route clearance is generated. This clearance is transmitted to the clearance delivery position at the departure airport about 30 min prior to the estimated departure time shown in the flight plan.

A typical center is responsible for more than 100,000 mi^2 (250,000 km^2) of airspace and hundreds of miles of airways. This geographical area is divided into 30 or more control sectors each with a controller team assigned. Each sector is defined by boundaries in the horizontal plane as well as by altitude bounds. There are 20 centers in the conterminous 48 states plus one each in Alaska, Hawaii, Puerto Rico, and Guam.

Terminal facilities. There are three principal positions staffed by tower controllers. The ground controller position is responsible for all ground traffic not on active runways (runways in use for takeoffs and landings). The local controller has jurisdiction over the active runways and the airspace close to the airport used by arrivals and departures. The clearance delivery position may be covered by the ground controller when traffic activity is low. The tower environment is essentially a visual environment. Controllers visually acquire and track aircraft and direct their movements by using radio or, when an aircraft has no operating radio, signal lights. In some locations, radar indicator equipment is installed in the tower to electronically display traffic that is being tracked by the local air-traffic control radar.

A radar terminal control facility serving a major airport and its satellite airports is frequently located in the tower building below the tower cab in the terminal radar approach control (TRACON) facility. At some low-activity airports, the radar displays for arrival and departure control are in the tower cab. This arrangement is known as the terminal radar approach cab (TRACAB). Terminal airspace where radar service is provided is divided into sectors which may include a number of departure sectors and arrival sectors each staffed by a team of one or more controllers. When activity is low, sectors may be combined to reduce personnel requirements.

There are approximately 5400 airports in the United States open to the public. Approximately 550 receive scheduled commercial passenger service, and 462 have towers staffed by the FAA. *See* AIR TRANSPORTATION.

Clyde A. Miller; Richard L. Greenspan

Bibliography. *Federal Aviation Regulations/Aeronautical Information Manual*, Federal Aviation Administration, published annually; *Federal Radionavigation Plan*, U.S. Department of Transportation, revised every 2 years; *Global Positioning System Standard Positioning Service Performance Standard*, Assistant Secretary of Defense for Command, Control, Communications, and Intelligence, October 2001; M. Kayton and W. R. Fried, *Avionics Navigation Systems*, 2d ed., Wiley-Interscience, New York, 1997; *National Air-space System Architecture*, Federal Aviation Administration, updated annually; M. S. Nolan, *Fundamentals of Air Traffic Control*, 4th ed., Brooks/Cole, Belmont, CA, 2003; C. D. Wickens et al., *Flight to the Future: Human Factors in Air Traffic Control*, vols. I and II, National Academy Press, 1997.

Air transportation

The movement of passengers and cargo by aircraft such as airplanes and helicopters. Air transportation has become the primary means of common-carrier traveling. Greatest efficiency and value are obtained when long distances are traveled, high-value payloads are moved, immediate needs must be met, or surface terrain prevents easy movement or significantly raises transport costs. Although the time and cost efficiencies obtained decrease as distance traveled is reduced, air transport is often worthwhile even for relatively short distances. Air transportation also provides a communication or medical link, which is sometimes vital, between the different groups of people being served.

Elements of the air system. The provision, continuance, and improvement of air transport services of all types require a complex integrated system whose main parts are the aircraft operator, the aircraft manufacturer, the airport terminal, and the air-traffic control system.

Aircraft operator. The aircraft operator provides the basic flight service to the customer, usually maintains the aircraft, and provides any needed customer services. The operator may be an airline providing for-hire service to customers on a commercial basis, or may operate the aircraft entirely for personal use. Scheduled and nonscheduled for-hire services are provided by a wide variety of airlines offering different classes of service. The type of route system served by the airline is often used to describe broad groups of carriers (that is, majors, nationals, large regionals, and medium regionals). More specialized airlines haul only cargo or operate only helicopters. Industry trade groups represent the combined interests of these carriers. Within the United States, the Air Transport Association represents the 21 largest American airlines from among approximately 93 airlines that provided scheduled service in 1987. Internationally, many of the major airlines are members of the International Air Transport Association, an organization primarily concerned with international tariff levels but also active in legal and safety matters. The International Civil Aviation Organization handles international air transportation matters, including technical standards and practices; provides an international language, including standard chart and identification codes; and collects and disseminates statistics and other data on the world's airlines.

Aircraft manufacturer. Aircraft manufacturers include the airframe manufacturer and the engine manufacturer, who together design and produce the airplane, fabricate the spare parts necessary to continue operation, and provide continuing technical help to the aircraft operator. The design, development, and fabrication of a major aircraft constitute a project of

tremendous scope which requires billions of dollars to complete. The airframe and engine manufacturer may be supported by thousands of subcontractors who produce different components of the airplane. The entire aircraft design, production, and assembly process may stretch across several nations. Governments may at times participate by providing direct financial support or by helping develop a steady stream of advanced technology which can be used to improve future aircraft. Such advances may come through spinoffs from advanced military aircraft or through research supported or accomplished by civil government agencies such as the National Aeronautics and Space Administration. In the United States, the transport airframe and engine manufacturers are represented by the Aerospace Industries Association of America, Inc., the general aviation industry by the General Aviation Manufacturers Association, and the helicopter industry by the Helicopter Association International.

Airport terminal. The airport terminal, characterized primarily by a runway and a passenger or cargo service facility, provides a central location for the airplane to take off and land and for passengers to begin and end the air portion of their journey. Major parts of the airport include the runways, taxiways, control tower, and buildings to service passengers, cargo, aircraft, and administrative needs. The largest airports have runways in excess of 8000 ft (2500 m) and are generally publicly owned and operated. Only one major airport has been built in the United States since 1970 and new construction has been confined to lengthening airport runways and improving the quality of smaller airports. Construction costs of a major new airport easily exceed several billion dollars, and land acquisition is both difficult and expensive. Local citizen opposition to airport construction is sometimes intense, with opponents citing objections to aircraft noise, ground congestion, land-use changes, and potential accidents. Airports may serve a wide variety of purposes or may be specialized (that is, passengers, cargo, general aviation, private, or military). Most airports of any size in the United States are paved, but in less developed parts of the world the vast majority of a nation's airports may have unpaved runways. *See* AIRPORT.

Air-traffic control system. Much of the air-traffic control (ATC) system is located at the airport terminal. Within the United States the ATC system is operated by a government agency, the Federal Aviation Administration (FAA; part of the Department of Transportation). The ATC system controls the aircraft's movements both on the ground and during flight, and provides a means of allocating available runways and air space between user aircraft. Equipment includes radios, radars, light signals, displays, and other types of electronic instrumentation. The airport control tower directs air traffic in the vicinity of the airport and on the runways and taxiways. Voice communication, navigational aid, weather information, and other services are also provided by the ATC operators to the crew flying the aircraft. The United States has over 17,000 public and private airports (including those with unpaved runways), of which over 400 have FAA towers. *See* AIR-TRAFFIC CONTROL.

Evolution and characteristics. The United States is the world's leading user and operator of aircraft, and has also produced most of the commercial airplanes used throughout the world. The foundation for this was laid with the Air Commerce Act of 1926, which began the development of commercial aviation in the United States. Regularly scheduled transcontinental service originated with a single route between New York and Los Angeles in 1930 (**Fig. 1**). To help

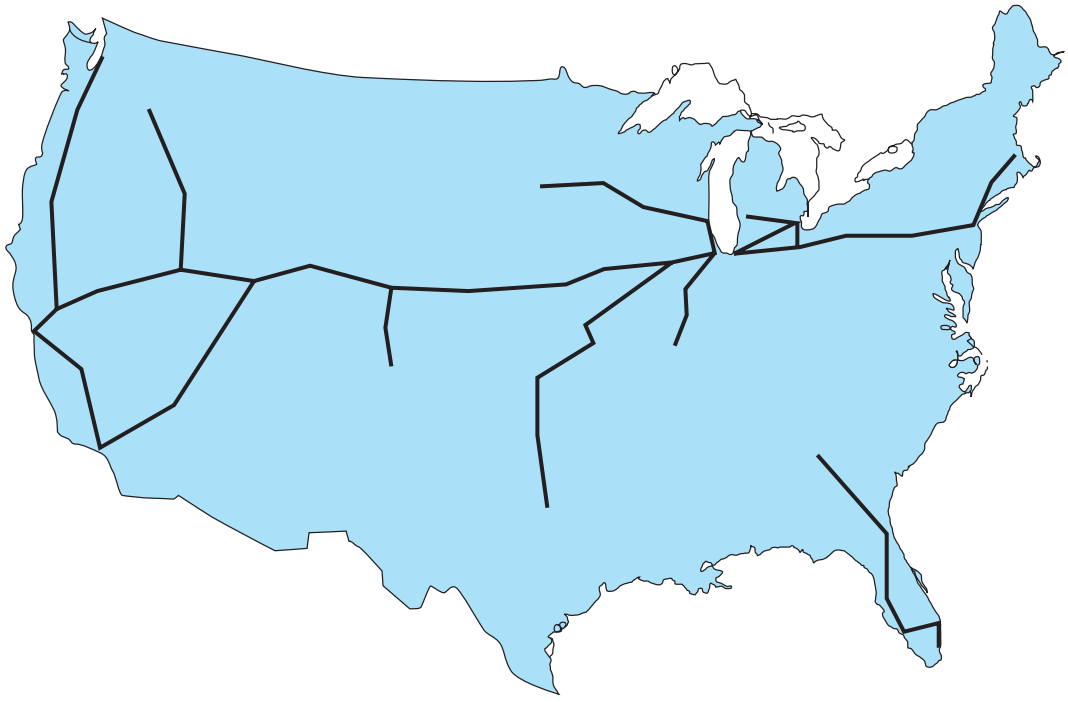


Fig. 1. Early airline route structure in the United States.

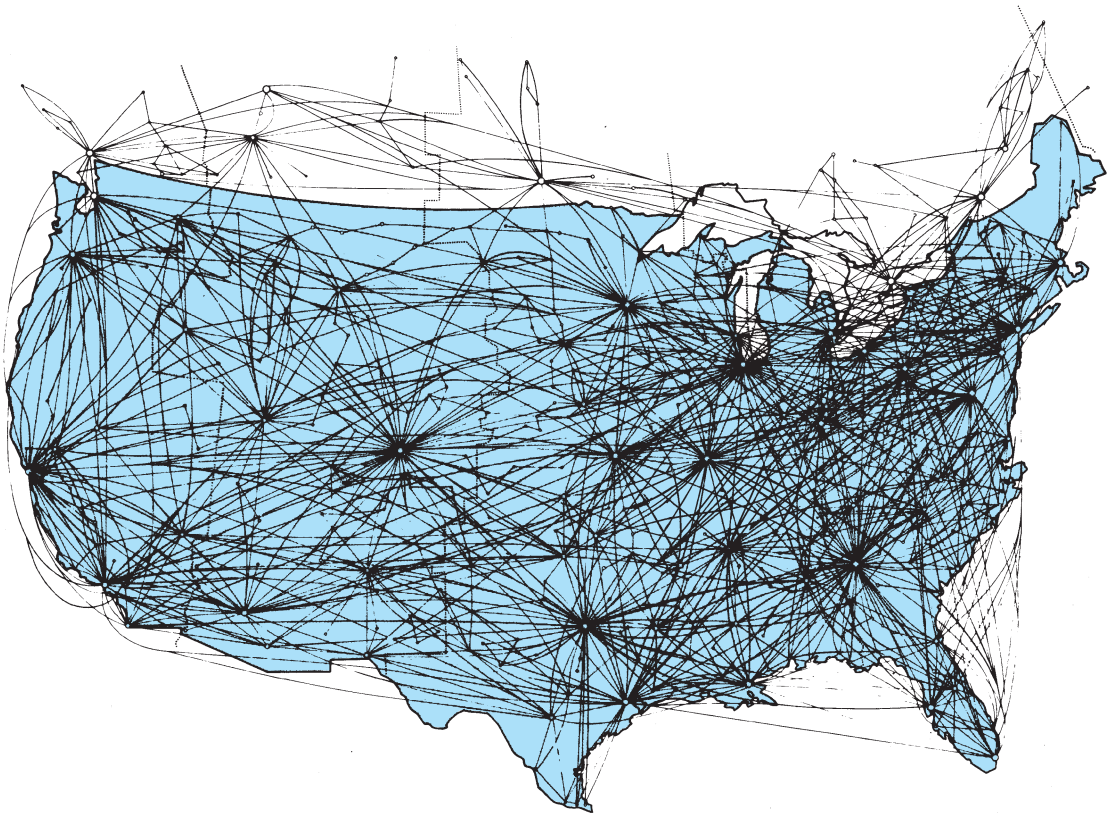


Fig. 2. Present airline route structure in the United States.

establish widespread air service, the government built and maintained navigational routes and provided subsidies to the airline operators through Post Office contracts to carry mail and through direct cash subsidies paid to cover losses generated between various city pairs. Subsidy payments to airlines were discontinued as the lines became able to support themselves.

Since the late 1940s, the number of passengers traveling by air between the United States and the rest of the world has exceeded the number of passengers traveling by sea. In domestic service, the revenue passenger-miles moved by air has exceeded those by rail since 1956. (A revenue passenger-mile is defined as one fare-paying passenger transported 1 statute mile or 1.6 km.) Air transportation is the dominant means of intercity common-carrier passenger service, moving over 92% of this traffic. The complex airline route system (**Fig. 2**) is a dramatic change from the original system (Fig. 1). Figure 2 also illustrates the so-called hub cities in the United States; air traffic is funneled from the smaller cities to the larger hub cities. Airlines are primarily carriers of people, obtaining over 87% of their revenue from passenger fares. Fares charged by airlines have historically increased at a considerably lower rate than consumer prices as a group.

Passenger composition. For a long time, airline service was oriented toward the business traveler, who made up the great bulk of demand. Over the years, however, the market has changed to the point where the business traveler now represents only about half of all passengers carried, with the pleasure or private traveler accounting for the other half.

Aerospace industry. The aerospace industry (aviation and missiles) is the leading United States exporter of manufactured products, consistently making an important contribution to the nation's balance of payments. Other countries have recognized the significance of this industry to their national economy, and some now produce aircraft similar in quality to those of the United States.

Safety. Safety has always been an overriding air transportation concern. Considerable media attention is given to an air crash of almost any sort, and such unfavorable publicity can be financially disastrous to a commercial carrier perceived as negligent. In 1958 Congress vested safety regulation in the Federal Aviation Administration. Before a new type of aircraft can be introduced to service, the FAA must certify the ability of the aircraft to meet airworthiness standards. Accident investigations are handled by the National Transportation Safety Board (NTSB). United States scheduled airlines have improved their safety performance with time: from the mid-1950s to the mid-1980s the fatality rate dropped from approximately 0.50 to 0.05 fatality per 100,000,000 passenger-miles. Much of this improvement is due to the introduction of the jet airplane and its almost complete replacement of propeller aircraft, as discussed below. Air transportation safety is comparable to that of other public modes of transportation and superior to that of the automobile.

Fig. 3. Advanced aircraft. (*a*) Boeing B-747-400 (*Boeing Company*). (*b*) McDonnell-Douglas MD-11 (*McDonnell-Douglas Corp.*).

Air fleet. The subsonic jet aircraft was introduced to United States commercial service in 1958 and revolutionized the air transport industry as well as the composition of the airline fleet. Within 10 years, jet aircraft had almost completely replaced propeller-driven airplanes. The advantages of the jet transport include superior economics, speed, range, capacity, comfort, and safety, which, taken together, forced the replacement of most propeller aircraft. Propeller aircraft are still used in certain types of airline service mainly for short-haul, low-passenger-capacity, and lower-altitude flights. *See* AIRCRAFT PROPULSION; JET PROPULSION.

Major transport airplanes include the B-767/757/747/737/727, the MD-80 series, and the DC-10/9/8 (**Fig. 3**). Such aircraft generally cruise at speeds of approximately 540 mi/h or 870 km/h (240 m/s) and at altitudes high enough that aircraft pressurization is required. Principal characteristics of three selected transport aircraft are given in the **table**.

Technological advances. Since its introduction, the transport airplane has steadily increased in speed, size, and range. Successive technology advances have continuously improved the quality and cost effectiveness of the air fleet. The most important of these advances have been the development of the jet and fan-jet engine, the high-bypass-ratio jet engine, major increases in aircraft size, and improvements in aerodynamics, materials, design techniques, and manufacturing. Many years (even decades) of research may be required before a new advance can be safely and economically incorporated into a production airplane. Despite the continuous improvements in technology, commercial aircraft are long-lived vehicles, and are often used for periods exceeding 20 years.

Supersonic aircraft. In the late 1970s, a new type of high-speed airplane, the Concorde, entered service on certain international air routes. Although it is capable of flight at speeds more than twice that of sound, its relatively high costs, limited range, small payload, and restricted flight capabilities resulted in a total of only 16 aircraft built. Widespread use of supersonic commercial aircraft therefore requires the development of a more technologically advanced airplane able to operate economically with a much greater range and payload capability, possibly over land as well as over water, and with acceptable environmental characteristics.

General aviation fleet. The general aviation fleet is far larger in numbers than the fleet operated by the commercial airlines. These aircraft are used for almost every conceivable transport purpose (personal, commercial, recreational, agricultural, and training). Many of these aircraft operate over short distances and do not require the elaborate communications, navigation, and safety equipment found in the typical commercial transport. The flexibility and utility of general aviation vehicles, however, are so great that the fleet has experienced major growth. Helicopters are also used for an increasing number of purposes, with industrial uses such as construction, mining, offshore oil drilling, and laying pipelines joining the traditional use of transporting passengers. *See* GENERAL AVIATION; HELICOPTER.

Development of new aircraft. New aircraft conceptual designs incorporating varying degrees of advanced technology are continually evaluated by the manufacturers of both transport and general aviation aircraft. Leading airlines work closely with transport manufacturers during the development stage to determine exactly what specifications and features the proposed aircraft should have. The transport manufacturers produce a new aircraft approximately every 12 years, and also supplement their product line with variations of the basic airplane called derivatives. Derivative aircraft are usually produced by "retching" the original airplane fuselage to permit the fabrication of airplanes of different length and passenger size. Sometimes the original airplane is "shortened" or "shrunk" for the same reason. Manufacturers thus increase the number of units eventually built and achieve economies of scale. General aviation manufacturers are not burdened by the

Characteristics of three transport aircraft

Characteristic	Boeing 747-400	McDonnell-Douglas MD-11	McDonnell-Douglas MD-81
Maximum weight: lb	870,000	602,500	140,000
kg	394,625	273,289	63,503
Passenger capacity (approx.)	600	320	155
Range: nmi	6600	6880	1650
km	12,222	12,741	3056

tremendous development costs of major transport aircraft, and are able to produce new aircraft more often than do the major manufacturers. Thus, despite the relatively limited resources of the general aviation manufacturers, they have been able to introduce new-technology aircraft before the major manufacturers.

CRAF. The airline fleet can also support the military logistic capability of the U.S. Air Force by its participation in the Civil Reserve Air Fleet (CRAF). These aircraft are able to carry certain types of military equipment and can be pressed into such service should the need arise. The CRAF fleet consists of several hundred aircraft, many of which are modern, long-range, high-payload vehicles.

Outlook. Air transportation is expected to experience steady growth in demand. In the developed countries, increases in economic growth and disposable income have brought air travel within the reach of large numbers of people and greatly lessened the industry's former dependence on the business traveler. In the developing countries, air transportation is being used to provide rapid development of transportation capabilities at low initial cost, and also to promote understanding and cooperation between segments of a country's population previously isolated from one another.

Air transportation, however, is also beset by many increasingly severe problems. These problems will require both institutional change and more use of advanced technology if the expected growth in passenger demand is to be satisfied while still providing adequate service. Though it is always difficult to establish which problems will be most important in the future, concerns now include energy, noise pollution, ground and air congestion, and the economic deregulation of the industry (required by the Airline Deregulation Act of 1978).

Energy. The importance of energy was first brought to public attention by the Arab oil embargo of 1973, which severely disrupted the air transportation system. Government regulatory authorities took strong actions aimed at curtailing fuel usage, including fuel allocations and approval of route capacity agreements aimed at reducing duplicate airline flights. Large numbers of commercial flights were canceled, airplanes sat idle, fuel prices increased by an order of magnitude, several major airlines went bankrupt, and some were forced to sell aircraft at bargain prices to foreign competitors. As a result of this experience, many actions were taken to improve the fuel performance of aircraft. Airlines increased the number of passengers carried on each flight, instituted drag cleanup control measures, flew more fuel-efficient flight profiles, and began many other fuel-use control efficiencies. Manufacturers developed a new series of fuel-conserving airframes and engines, mainly utilizing advanced aerodynamics, materials, and engine technologies, which permits more than a 35% improvement in the airplane miles obtained per gallon of fuel. Airlines purchased hundreds of these new-generation aircraft, which began appearing in 1982. Future airplanes will show even greater long-term fuel efficiency improvements as increased government and industry research efforts in most areas of aeronautics pay off. *See* AIRCRAFT ENGINE; AIRFRAME; COMPOSITE MATERIAL.

Aircraft noise. Concern over aircraft noise and its impact on people living near airports led to major improvements in the noise characteristics and the operation of jet aircraft. Advances include "quiet" acoustic nacelle liners, improved internal and external engine and nacelle designs, and operational procedures on takeoff and approach that minimize the impact of the aircraft on the community surrounding the airport. Some of these advances were retrofitted to the existing fleet. Other advances are an integral part of the design of aircraft introduced to service since 1982. Such aircraft are capable of meeting much more restrictive government noise standards, and should result in the confinement of excessive aircraft noise to the boundaries of the airport in most areas.

Congestion. The congestion problem (both ground and air) has become increasingly severe. It has long been recognized that ground congestion reduced the time and cost savings obtained by air transport, and airport officials have continuously tried to improve access and egress to airport facilities. The air congestion problem, however, has been largely overshadowed by other problems. In the early 1970s, declining traffic growth rates, introduction of high-capacity wide-body aircraft, and a diffusion of service caused by a partial shift from the predominant hub-spoke system (where regional air traffic is funneled through a major airport) to more direct point-to-point air service temporarily masked the disruption that can result from congested air service. In the later 1970s and 1980s, air traffic surged, causing long delays, "stacking" of airplanes over airports, crowding, and sometimes chaotic conditions at the busiest terminals. Airport curfews (used as a noise control measure) also contributed to the congestion problem. Maintaining adequate air service becomes progressively more difficult as traffic continues to increase and the construction of new airports lags over long periods of time. Long-term solutions have not as yet been identified, let alone implemented. Near-term approaches include the use of additional cities as hubs.

Economic deregulation. Economic deregulation has had a far-reaching impact on the air transportation system. With deregulation (established in 1978) fully effective since 1985, major airlines either left or reduced service to the smaller cities, added capacity between larger cities (where such carriers could better utilize their existing fleet), offered significant discount fares subject to certain conditions, and, in general, increased demand for air service. *See* AIRPLANE; AVIATION.

Dal V. Maddalon

Bibliography. Aerospace Industries Association of America, Inc., *Aerospace Facts and Figures,* annually; G. L. Donohue and A. G. Zellweger (eds.), *Air Transportation Systems Engineering*, 2001; R. M. Kane, *Air Transportation*, 14th ed., 2002.

Airborne radar

Radar equipment carried by commercial and military aircraft. These aircraft use airborne radar systems to assist in weather assessment and navigation. Military systems also provide other specialized capabilities such as targeting of hostile aircraft for air-to-air combat, detection and tracking of moving ground targets, targeting of ground targets for bombing missions, and very accurate terrain measurements for assisting in low-altitude flights. Airborne radars are also used to map and monitor the Earth's surface for environmental and topological study.

Airborne radars present unique design challenges, mainly in the severe nature of the ground echo received by the radar and in the installation constraints on the size of the radar. The peculiar clutter situation governs the nature of the signal processing, and the installation limitations influence the antenna design and the radio frequency to be used (the two being strongly related) as well as the packaging of the rest of the radar. Similar considerations influence the design of space-based radars as well.

Weather assessment. A particularly valuable use of airborne radar is weather assessment (**Fig. 1**), contributing to safer and smoother navigation of severe weather areas. Two of the most serious threats to aircraft operations are turbulence and hail, both products of thunderstorms. Radars generally operating in the C or X bands (around 6 GHz or around 10 GHz, respectively) permit both penetration of heavy precipitation, required for determining the storm's extent, and sufficient reflection from less intense precipitation (stronger for the shorter wavelengths of the higher frequencies). Regions of varying but heavy precipitation extending more than 15 mi (24 km) can be clearly displayed, even if they include areas of very heavy rain (greater than 2.4 in./h or 60 mm/h). The experienced pilot can recognize structures to avoid, such as hail shafts.

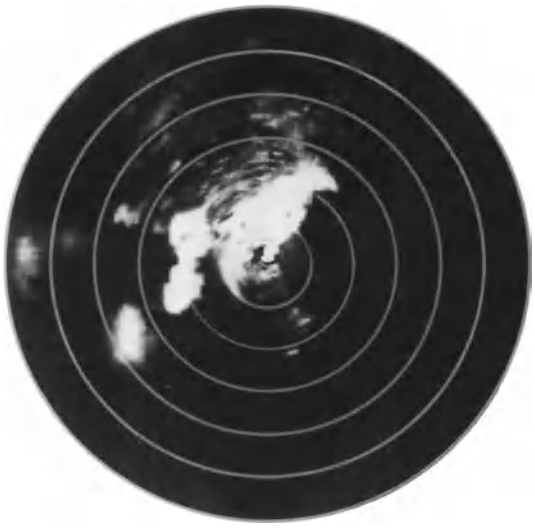


Fig. 1. Radar presentation showing a squall line extending from northwest to southeast. The range marks, indicating 5-mi (8-km) units, give estimation of extent of disturbance. (*RCA Corp.*)

It is important in such radars to detect the varying density of rain since the greatest turbulence is to be expected in the regions of the greatest rate of change or gradient of radar reflectivity, where there is a large change of reflectivity (and hence precipitation density) in a very short distance. Processing to enhance the reflectivity gradient and the resulting displays of this property, as well as the use of color coding to alert the operator, are valuable features of these radars. *See* METEOROLOGICAL RADAR; RADAR METEOROLOGY.

Radar navigation. Another basic and valuable airborne radar function is altimetry. The aircraft's altitude can be continuously measured, using (generally) C-band frequencies (around 6 GHz), low-power transmission, and a downward-oriented antenna beam. Sometimes, information from additional beams (looking somewhat forward, for example) is combined with measurements of the Doppler shift of the ground echo received to further aid in navigation. *See* ALTIMETER; DOPPLER EFFECT.

Another type of radar used in navigation is the radar beacon, in which a ground-based receiver detects an interrogation pulse from the aircraft and sends back a so-called reply on a different frequency, to which the receiver on the aircraft is tuned. The transmission of the reply at a different frequency than the interrogating signal avoids the reception of clutter echo from weather and the ground. Examples of this type of system are distance-measuring equipment (DME) and air-traffic control radar beacon systems (ATCRBS). The DMEs are used in conjunction with the network of very high frequency omnirange (VOR) stations, at each of which a transponder (transmitter-responder) has been located. By interrogating it with an inexpensive interrogator, an aircraft determines its distance from the station; a series of such measurements permits determination of the aircraft's position by trilateration. *See* DISTANCE-MEASURING EQUIPMENT; RHO-THETA SYSTEM; VOR (VHF OMNIDIRECTIONAL RANGE).

The ATCRBS involves interrogators at ground stations, generally at the sites of the major air-traffic control radars. The antennas of the two systems are aligned and rotate together to facilitate combining the radar and beacon data. When a target is detected by the radar, an interrogate pulse can be sent to the aircraft. The transponder in the aircraft sends a signal not only confirming its location but also containing other information (such as altitude and identifying number) of value to the air-traffic management system. The military "identification, friend or foe" (IFF) system operates in a similar way. These beacon systems operate in the L band (using 1.03 and 1.09 GHz frequencies for interrogation and response, respectively) and are called secondary surveillance radar (SSR) to distinguish them from ordinary two-way reflective radar, which is called primary. *See* AIR NAVIGATION; AIR-TRAFFIC CONTROL; SURVEILLANCE RADAR.

High-resolution mapping. Airborne radars are used effectively to provide high-resolution mapping of Earth's (or other planetary) surface, with a technique

Fig. 2. High-resolution radar map of a golf course near Stockbridge, New York. (*M.I.T.-Lincoln Laboratory*)

called synthetic aperture radar (SAR) [**Fig. 2**]. The processing uses the fact that surface objects produce a Doppler shift (due to the aircraft's flight) unique to their position as the aircraft passes by; this Doppler history is indicative of the scatterer's lateral, or cross-range, position at the particular range determined by the usual echo timing. With very stable radars and well-measured flight characteristics (and other focusing methods), picture cells (pixels) of 1 ft × 1 ft (0.3 m × 0.3 m) can be formed in the processed images from radars tens or hundreds of miles away. The resolution is somewhat like that possible had the flight path itself been used as a huge antenna, the synthetic aperture.

Modern SARs provide image quality that rivals aerial photography while being far less subject to conditions restricting optical visibility. For varying reasons, SARs use a wide range of frequencies, some in even the very high frequency (VHF) band, to be more sensitive to certain targets of interest; many use the familiar frequencies of 1 to 10 GHz, and some of the more resolute operate at Ka band (around 33 GHz). In some SARs, analysis of the polarization of the return (polarimetry) assists in target or surface texture (vegetation) classification. In others, interferometric analysis of the signals from two slightly different positions reveals topological (height) information for each pixel, useful in important environmental surveying and even in earthquake prediction. *See* REMOTE SENSING.

Target detection and tracking. Airborne radar provides military aircraft with an essential capability, the detection of targets, strategic and tactical, to be engaged. Strategic targets are those associated with an enemy's ability to wage war, such as munitions factories, shipyards, railroads, and other resources for delivering essential products. Tactical targets include the enemy's weapons and weapon systems brought to battle, such as fighter and bomber aircraft, warships, tanks, and other battlefield elements.

Strategic targets. Strategic targeting is accomplished by using high-resolution ground-mapping radar systems, such as those described above, to augment conventional strategic target location systems. Radar provides the ability to generate required imagery during either day or night and under all weather conditions. Typically, the strategic targets of interest for mapping radars are those that are movable (for example, deployed military equipment such as aircraft on the ground at air fields) or have been built after the most recent surveillance information was generated.

Tactical targets. Two major types of military airborne radar that deal with tactical warfare are the airborne early warning (AEW) radars and the airborne intercept (AI) radars. Of the former, the U.S. Air Force Airborne Warning and Control System (AWACS) is a good example; this S-band (around 3 GHz) system uses a large disk-shaped rotating antenna mounted atop the E-3A aircraft. AWACS provides from its typical flight altitudes a large-radius view of surface and air traffic, comprising targets of great tactical significance, and needs, therefore, excellent antenna performance (low sidelobes) and Doppler-sensitive signal processing (to detect targets moving only slightly relative to the heavy surface clutter background). The U.S. Navy's aircraft-carrier-based airborne early warning radar is borne by the E-2C Hawkeye aircraft, uses a similar but smaller top-mounted "rotodome" antenna, and operates at ultrahigh frequencies (UHF; around 400 MHz) to reduce the intensity of the surface clutter somewhat; it represents a design older than the AWACS, but has served very well in extended-area sea-surface tactical surveillance. *See* INSULATION RESISTANCE TESTING; MILITARY AIRCRAFT.

In higher-performance tactical aircraft, the radar antenna cannot possibly be mounted outside the aircraft body. Consequently, airborne intercept radars, used generally for targeting other aircraft in air-to-air combat, operate with very limited antenna dimensions (a flat disk antenna mounted behind a radome at the nose of the aircraft, for example). Therefore, to achieve useful narrow beamwidths for angle-measurement accuracy, they typically use the shorter wavelengths (about 3 cm or less) of the X band (around 10 GHz) for which the ground clutter is quite intense, placing great dependence upon sophisticated Doppler-sensitive signal processing.

Missiles are often guided by radar seekers mounted in the missile. Antenna size, limited to the diameter of the missile, suggests the use of shorter wavelengths. The X band and above (about 10 GHz or more; wavelengths less than 3 cm) are common for such radars to make best use of such small antennas. Some use continuous-wave (cw) waveforms rather than pulses, emphasizing angle sensing only, and some only receive the transmissions from a cooperating "illuminator" and are called semiactive seekers (bearing no transmitter in the missile). *See* GUIDED MISSILE.

Terrain following. For certain missions, military aircraft and attack missiles attempt to remain hidden from enemy radar defenses by flying at very low altitudes. In general, the lower the altitude, the better the shielding from the ground-based radars; but there is a limit below which the pilot's or the guidance system's capabilities may place the craft at

risk. Therefore, before the mission, speed and altitude are carefully chosen, considering the defenses at hand, foreknowledge of the terrain, and the dynamics of the aircraft. Then, during the flight, radar altimetry and a forward-looking radar are both used to establish angles to obstructions and the flight's relationship to an offset terrain (a line parallel to the terrain at the required clearance altitude); and with allowance for the flight dynamics involved, the system gives either directions to the pilot in an aircraft or inputs to the autopilot in a crewless missile. *See* GUIDED MISSILE.

Robert T. Hill

Bibliography. S. A. Hovanession, *Radar System Design and Analysis*, Artech House, Dedham, MA, 1984; W. C. Morchin, *Airborne Early Warning Radar*, Artech House, Norwood, MA, 1990; M. I. Skolnik (ed.), *Radar Handbook*, 2d ed., McGraw-Hill, New York, 1990; G. J. Sonnenberg, *Radar and Electronic Navigation*, 6th ed., 1988; G. Stimson, *Introduction to Airborne Radar*, 2d ed., SciTech Publishing, Mendham, NJ, 1998.

Aircraft

Any vehicle which carries one or more persons and which navigates through the air. The two main classifications of aircraft are lighter than air and heavier than air.

The term lighter-than-air is applied to all aircraft which sustain their weight by displacing an equal weight of air, for example, blimps and dirigibles. The weight of such aircraft is sustained by buoyant forces similar to the forces which sustain a ship in water. *See* AIRSHIP; ARCHIMEDES' PRINCIPLE; BLIMP.

Heavier-than-air craft are supported by giving the surrounding air a momentum in the downward direction equal to the weight of the aircraft. Aircraft with fixed wings impart a small downward momentum to a large quantity of air because they have a large forward velocity. Aircraft with rotating wings, such as helicopters, are also sustained by the downward momentum which they impart to the surrounding air. Because they operate at slower speeds, the momentum change is imparted to a smaller mass of air but the change in velocity is correspondingly greater. *See* AIRPLANE; BERNOULLI'S THEOREM; HELICOPTER; VERTICAL TAKEOFF AND LANDING (VTOL).

In general, aircraft have a means of support, a propulsion system to impart forward velocity, and a means of directional control so that navigation can be accomplished. *See* AIR NAVIGATION; AIRCRAFT COMPASS SYSTEM; AIRCRAFT PROPULSION.

Richard G. Bowman

Aircraft collision avoidance system

A device that reduces the risk of midair collision by providing advisories to the flight crew. It is known in the United States as a Traffic Alert and Collision Avoidance System (TCAS), and operates independently of the ground-based air-traffic control system to provide a safety net should normal separation standards be jeopardized. *See* AIR-TRAFFIC CONTROL.

TCAS advisories are of two types: traffic advisories (TAs) which aid the pilot in visually acquiring other aircraft, and resolution advisories (RAs) which recommend a vertical escape maneuver. TCAS II provides both traffic advisories and resolution advisories, and is mandated worldwide on aircraft with more than 30 passenger seats or more than 15,000 kg in weight. TCAS I provides only traffic advisories and is required by the U.S. Federal Aviation Administration (FAA) for aircraft with 10–30 passenger seats.

TCAS consists of three major subsystems: surveillance, collision avoidance system (CAS) logic, and pilot displays.

Surveillance. TCAS surveillance is based on the aircraft transponders carried by all commercial passenger and most private aircraft. The reception of a Mode A interrogation from an FAA ground radar will cause the transponder to transmit a Mode A reply that contains a 12-bit identity (squawk) code (**Fig. 1**). Mode C replies can also be solicited that contain automatically encoded 12-bit barometric altitude expressed in hundreds of feet. *See* SURVEILLANCE RADAR.

Newer Mode S transponders and ground interrogators also exist and are compatible with Mode A/C air and ground equipment, but provide a unique 24-bit address for every aircraft and an integral (that is, part of or built into) data link. TCAS interrogates both kinds of transponders to develop a surveillance picture once each second.

For Mode A/C aircraft, TCAS transmits only Mode C interrogations to obtain altitude. Range is estimated by measuring the roundtrip interrogation/reply time, and relative bearing is estimated using a small (approximately 100-cm-square or 40-in.-square) antenna. In this way, a three-dimensional picture of all nearby Mode A/C aircraft is developed each second.

Since TCAS antennas are small and have an omni or four-quadrant pattern, a "whisper-shout" interrogation sequence is used to avoid receiving multiple, overlapping replies from Mode A/C transponders at similar ranges. In this technique, TCAS first transmits a low-power P_1P_3 (Fig. 1). interrogation, resulting in replies from aircraft that have the most sensitive roundtrip link—a combination of transmit power, range, cable losses, antenna gains, and receiver sensitivities. TCAS then transmits a suppression pulse pair, P_1P_2, to suppress the transponders that previously replied, immediately followed by another P_1P_3 interrogation at a slightly higher power. The first set of aircraft will then be suppressed and not respond to the second interrogation. This sequence of suppressions and interrogations is repeated multiple times, ending in a full-power interrogation. The result is that the replies are separated, that is, do not overlap, even in dense airspace.

In contrast, Mode S surveillance interrogations are directed to individual aircraft. TCAS begins by passively listening for Mode S squitters, which are replies that are spontaneously transmitted each second by Mode S transponders and include the aircraft's unique 24-bit address. Once an aircraft's squitters are

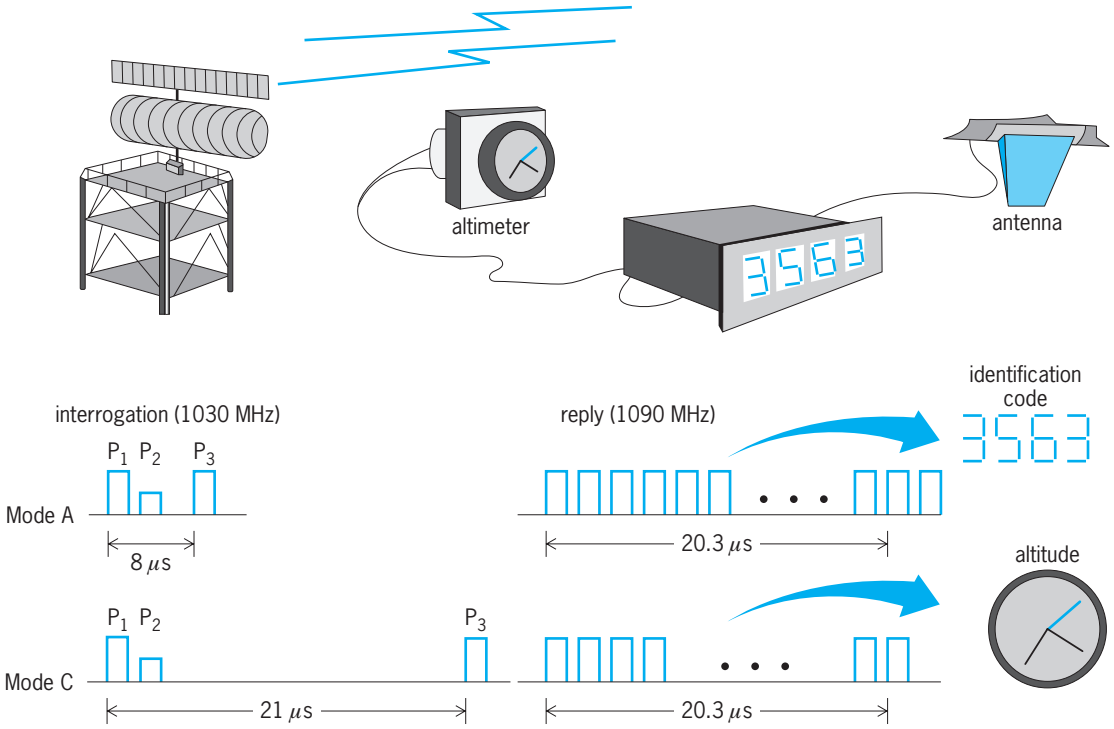
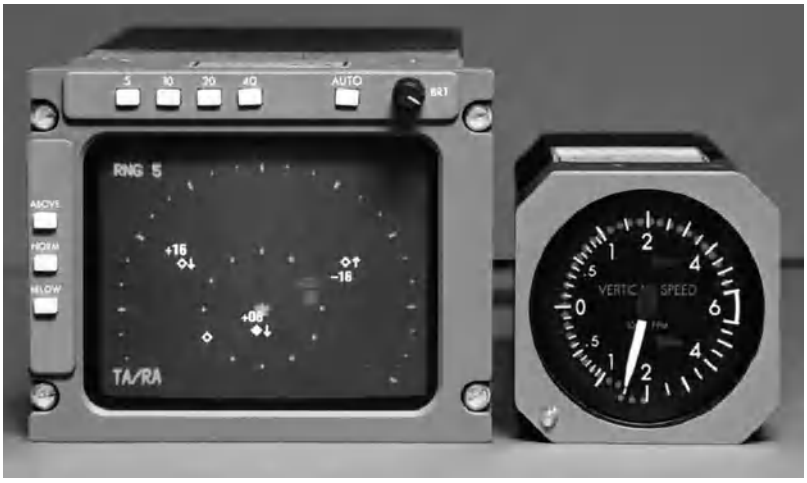


Fig. 1. Surveillance with Mode A/C transponders. (*Reprinted with permission of MIT Lincoln Laboratory, Lexington, MA*)

reliably received, TCAS begins regular interrogations of that aircraft using its address.

TCAS must provide reliable surveillance to at least 14 nautical miles (26 km) so that a flight crew can safely respond to TCAS resolution advisories for aircraft that are approaching as fast as 1200 knots (2200 km/h). In practice, surveillance is routinely achieved out to more than 50 nautical miles (93 km) in en route flight altitudes and 20 nautical miles (37 km) in the terminal area. TCAS avoids degrading the performance of air-traffic control ground interrogators by reducing the TCAS power and operating range in busy airspace and for distant intruder aircraft.



(a) (b)

Fig. 2. TCAS pilot displays. (*a*) Traffic advisory (TA) display. (*b*) Vertical resolution advisory (RA) display. (*Reprinted with permission of MIT Lincoln Laboratory*)

Collision avoidance system (CAS) logic. The CAS logic examines the position of each intruder aircraft provided by the surveillance system to determine the time to closest point of approach (CPA) for that intruder. This time is referred to as tau and is computed, in its most simple form, as range divided by range rate. Tau thresholds are based on the altitude of the encounter—the lower the altitude, the smaller the threshold. If tau is "small" (typically less than 25 seconds), TCAS computes the projected vertical separation at the closest point of approach. If the separation is also "small" (typically less than 750 ft or 230 m), TCAS issues a resolution advisory, or vertical maneuver. Traffic advisories, or warnings, are determined by the same process as above but use larger tau thresholds, typically 40 seconds.

Should a resolution advisory be deemed necessary, the least disruptive maneuver that will achieve adequate vertical separation is selected for display to the flight crew. If the encounter involves another TCAS-equipped aircraft, air-to-air coordination takes place using the Mode S data link to ensure compatible maneuvers.

The encounter is reevaluated each second to determine whether the advisory should be strengthened, sustained, weakened, or canceled. In addition, if the encounter geometry has changed and the advisory is no longer predicted to provide adequate separation, a sense reversal—climb instead of descend, or vice versa—may occur. This is most likely if a

non-TCAS-equipped intruder maneuvers unexpectedly after the resolution advisory has been issued, or if a TCAS-equipped intruder maneuvers contrary to its resolution advisory.

Since TCAS does not know the current flight plan or recent changes initiated by the flight crew or air-traffic controllers, "nuisance" advisories can occur. Early test flights and operational use led to a variety of CAS logic enhancements designed to minimize nuisance alarms while not compromising the collision avoidance performance. Pilots following a TCAS resolution advisory are instructed to notify air traffic control as soon as is practical and return to their clearance when the conflict is resolved.

Pilot displays. TCAS installations require two types of displays (**Fig. 2**). The traffic advisory display aides the flight crew in visually acquiring intruders and other nearby aircraft. Aircraft are tagged to show relative altitude and vertical rate trend, and color-coded to indicate degree of threat. The resolution advisory display indicates the vertical maneuver needed to achieve separation from the intruder. In Fig. 2*a*, the intruder at 2 nautical miles (3.7 km) and 2 o'clock is a threat, and the resolution advisory (Fig. 2*b*) tells the pilot to descend at 1500 ft (450 m) per minute. Resolution advisories may also be shown on vertical speed tapes; as pitch cues on the Primary Flight Display (PFD), using Flight Director guidance; or on a heads-up display (HUD). *See* AIRCRAFT INSTRUMENTATION.

Aural annunciations accompany both traffic advisories and resolution advisories. The annunciation for traffic advisories is "traffic, traffic." Annunciations for resolution advisories range from the basic ("climb, climb" and "descend, descend") to phrases that advise vertical rate increases or decreases, sense reversals, and altitude crossings.

Development areas. Automatic Dependent Surveillance-Broadcast (ADS-B), developed since the mid-1990s, provides for self-reporting of aircraft position, flight identification, and other aircraft-derived information at ranges up to 100 nautical miles (185 km). A Mode S extended squitter has been defined to broadcast ADS-B information and is compatible with TCAS receivers.

A method for allowing TCAS to make use of ADS-B data and still retain its independence as a safety system is referred to as TCAS Hybrid Surveillance. It allows use of extended squitter data for target surveillance only if those data have been validated by comparison with data from an active interrogation.

As a result of a midair collision near Namibia, Africa, and a heightened awareness of military safety concerns, the Department of Defense is equipping military transport aircraft with civil safety technologies, including TCAS. Special military versions of TCAS have also been developed to aid in tanker rendezvous and formation operations. The latest of these military versions make use of TCAS Hybrid Surveillance. TCAS Hybrid Surveillance is expected to make its way into the civil TCAS fleet in the future.

Raymond LaFrey

Bibliography. Federal Aviation Administration, *Aeronautical Information Manual*, FAA DoT, 4-4-10:4-4-15, 2004, published annually; International Civil Aviation Organization, *International Standards and Recommended Practices, Annex 10 to the Convention on International Civil Aviation, Aeronautical Telecommunications*, vol. IV: *Surveillance Radar and Collision Avoidance Systems*, 3d ed., 2002; Radio Technical Commission for Aeronautics, *Minimum Operational Performance Standards for Air Traffic Control Radar Beacon System/Mode Select (ATCRBS/Mode S) Airborne Equipment*, RTCA/DO-181C, 2001; Radio Technical Commission for Aeronautics, *Minimum Operational Performance Standards for Traffic Alert and Collision Avoidance System II (TCAS II) Airborne Equipment*, RTCA/DO-185A, 1997; T. Williamson and N. A. Spencer, Development and operation of the Traffic Alert and Collision Avoidance System, *Proc. IEEE*, 77(11):1735–1744, 1989.

Aircraft compass system

An instrument that indicates the bearing, or angle of the direction in which an aircraft is pointing in the horizontal plane. A compass may indicate magnetic heading or bearing, bearing referenced to a radio signal source, or bearing with respect to an inertially maintained line of position.

Inertial type. In modern commercial jet transport or military fighter aircraft, the primary sensor of the aircraft compass system is the inertial reference system (IRS). This system provides a gyroscopically derived reference to an inertial reference axis by sensing the linear and angular accelerations of the aircraft and continuously integrating these values to provide angular and linear velocities. These velocities are, in turn, integrated to develop estimates of the position and attitude of the aircraft. *See* GYROSCOPE; INERTIAL GUIDANCE SYSTEM.

The inertial reference system derives its position and attitude in an inertial frame referenced to true north. This approach represents an increase in precision over the original compass systems, which developed their lines of position referenced to magnetic north. Magnetic north is located at approximately longitude 104° west and latitude 74° north. The difference between the inertially derived north reference and the magnetically derived reference is called the magnetic declination, or magnetic variation. The magnetic variation is subject to local magnetic anomalies. Additionally, at extreme latitudes the magnetic variation can become very large, so as to make the magnetic reference unusable. *See* GEOMAGNETISM.

As navigation charting is usually referenced to magnetic values except near the poles, the aircraft converts from true to magnetic reference by accessing a magnetic variation model at the current aircraft position. As the magnetic variation value changes with time, this model must be revised approximately every 10 years.

Display. The magnetic and true bearings of the aircraft are presented on computer-generated cathode-ray tubes or flat-panel displays. These displays present aircraft heading in the form of a compass rose on both the primary flight display and the navigation display. The compass information can be presented to the flight crew in terms of either magnetic or true heading. The display can be oriented to present the data in a north-up or in a flight track-up mode (that is, north or the flight path of the aircraft points to the top of the display). References to radio navigation sources are also presented on the navigation display.

Magnetic type. The magnetic compass remains a simple, inexpensive, and reliable instrument for indicating the aircraft bearing. Limitations in the accuracy of this compass are due to the accelerations and vibrations of the aircraft, the local induced magnetic field of the aircraft, and lack of knowledge of local magnetic variation.

The magnetic compass is a secondary sensor of bearing on jet transports and most military aircraft, although it remains the primary instrument for many small, general aviation aircraft. This compass is typically a stand-alone instrument consisting of a card indicating the bearing installed in a liquid-filled case. The liquid serves to dampen rapid aircraft movements or oscillations. Magnets within the compass card align themselves with the magnetic field of the Earth. Compensators are installed to remove errors due to nearby instruments. A compass deviation card allows the navigator to correct for local magnetic variation. *See* MAGNETIC COMPASS.

In some aircraft, the magnetic compass has been coupled with a gyroscopic element to provide the gyroslaved, or gyrosynchronized, magnetic compass. The gyrosynchronized compass provides a means to overcome one limitation of the basic magnetic compass: the error in the presence of accelerations and maneuvers caused by the effect of gravity on the compass magnetic element. The gyrosynchronized compass uses a directional gyro to determine the local horizon and to sense accelerations and thus correct for the gravity effects on the compass output. *See* AIRCRAFT INSTRUMENTATION.

Robert W. Schwab

Bibliography. M. Kayton, and W. R. Fried (eds.), *Avionics Navigation Systems 2d ed.,* 1997; E. H. J. Pallett, *Aircraft Instruments, Principles and Applications,* 2d ed., 1981; G. M. Siouris, *Aerospace Avionics Systems: A Modern Synthesis,* 1993.

Aircraft design

The process of designing an aircraft, generally divided into three distinct phases: conceptual design, preliminary design, and detail design. Each phase has its own unique characteristics and influence on the final product.

Design Phases

Conceptual design activities are characterized by the definition and comparative evaluation of numerous alternative design concepts potentially satisfying an initial statement of design requirements. Conceptual design engineering is primarily analytical, making use of the body of knowledge from past aircraft designs and of computer programs for conceptual analysis based on correlations of prior design data. Experimental work, such as wind-tunnel testing, may be initiated during conceptual design if unique concepts are being considered. The conceptual design phase is iterative in nature. Design concepts are evaluated, compared to the requirements, revised, reevaluated, and so on until convergence to one or more satisfactory concepts is achieved. During this process, inconsistencies in the requirements are often exposed, so that the products of conceptual design frequently include a set of revised requirements. *See* AIRCRAFT TESTING; WIND TUNNEL.

During preliminary design, one or more promising concepts from the conceptual design phase are subjected to more rigorous analysis and evaluation in order to define and validate the design that best meets the requirements. Extensive experimental efforts, including wind-tunnel testing and evaluation of any unique materials or structural concepts, are conducted during preliminary design. The end product of preliminary design is a complete aircraft design description including all systems and subsystems. All major configuration features, size, shape, weight, and systems definitions are established and are not expected to change significantly during the final detail design. Performance and cost data are generated in sufficient depth to support the decision to proceed to the relatively lengthy and costly detail design phase.

During detail design the selected aircraft design is translated into the detailed engineering data required to support tooling and manufacturing activities.

The major design decisions determining the ultimate capabilities, production cost, and operating costs of the new aircraft are made during the conceptual design and early preliminary design phases.

Requirements

The requirements used to guide the design of a new aircraft are established either by an emerging need or by the possibilities offered by some new technical concept or invention.

Needs-based requirements, in the cases of military and commercial (airline) aircraft, are ordinarily established by the customer or user of the projected aircraft in a process which involves much interaction with aircraft companies in order to assure technical and economic feasibility of the requirements. These interactions also provide the basis for early conceptual design studies by aircraft companies. Needs-based requirements in the case of general aviation (private and business aircraft) are developed by classical market analysis techniques which involve economic projections and surveys of past and potential customers. *See* AIR TRANSPORTATION; GENERAL AVIATION; MILITARY AIRCRAFT.

Innovation-based requirements emerge from some new concept or invention which offers capabilities

not previously possible. This is illustrated by the development of the turbojet engine after World War II. This propulsion technology made it possible to design aircraft which flew higher and faster than previous generations, and at the same time provided engines which were inherently simpler and easier to maintain than the reciprocating engines which they replaced. *See* AIRCRAFT PROPULSION; JET PROPULSION; RECIPROCATING AIRCRAFT ENGINE; TURBOJET.

Requirements can be divided into two general classes: technical requirements (speed, range, payload, and so forth) and economic requirements (costs, maintenance characteristics, and so forth). A first-order list of aircraft design requirements is as shown below:

- **Technical**
  - Flight performance
    - Speed-altitude envelope
    - Range-payload
    - Takeoff and landing distances
    - Maneuvering
  - System requirements
    - Aircraft systems
    - Military mission systems
    - Environmental constraints
  - Airworthiness
    - Flying qualities
    - Structural integrity
    - Reliability
  - Military
    - Survivability
- **Economic**
  - Procurement cost
    - Aircraft
    - Spares
    - Support equipment
    - Training systems
  - Operating costs
    - Fuel usage
    - Maintainability
    - Training
  - Design life

Technical requirements. Flight performance requirements define the characteristics of speed, altitude, and payload carriage that will largely determine the size, aerodynamic configuration, and propulsion concept of the aircraft. Military combat aircraft have, in addition, stringent maneuvering requirements (for example, to permit a fighter to turn more rapidly than its opponent) that also strongly influence the size and shape of the aircraft.

The requirements for on-board aircraft systems, such as communication and navigation systems, are either defined in functional terms or may be specified in terms of selected existing systems. In the case of military aircraft, the system requirements are considerably more complex because of the demands of target detection, secure communications, special offensive and defensive avionics systems, and weapons capabilities. System requirements include the description of any constraints set by the aircraft's operating environment.

Airworthiness requirements comprise those issues having to do with safety. Flying-qualities requirements are concerned with the dynamic motions of the aircraft and its tendency to maintain a state of equilibrium, in response to pilot control inputs or to disturbances due to gusty or turbulent air. Structural-integrity requirements deal with the ability of the aircraft structure to withstand the maximum expected airloads, the life of the structure under repeated loads (fatigue life), and related issues such as corrosion resistance. Reliability requirements focus on the probabilities and consequences of failure of the many individual aircraft systems—mechanical, hydraulic, electrical, or electronic. Special military aircraft requirements include vulnerability and survivability criteria under specified hostile conditions. *See* AIRFRAME; FLIGHT CHARACTERISTICS; RELIABILITY, AVAILABILITY, AND MAINTAINABILITY.

The task of designing a set of technical requirements for a new aircraft is made easier by the existence of many requirements specification documents established over the years by government agencies. In the United States, civil aircraft are covered by various Federal Aviation Administration (FAA) specifications, and counterpart military specifications are issued by the Department of Defense. These specifications contain the basic criteria which must be met in order for a new aircraft to be permitted to operate, that is, to be certified by the FAA in the case of civil aircraft or approved for production by the military in the case of military aircraft.

Economic requirements. These are not as susceptible to quantitative statements as technical requirements but are more often stated as cost goals or as requirements related to cost, such as maintainance worker-hours per flight hour. Both procurement and operating cost requirements usually include the costs of any special or nonstandard support equipment for the new aircraft design, and also the costs of equipment and systems that may be required to train flight crew or maintenance personnel. Both commercial and military operators may also invoke some form of mission success criteria in the area of operating costs. Commercial operators, for example, might state requirements for dispatch reliability, while the military operator might require some overall measure of probability of mission success. Requirements for design life influence both procurement costs (since longer life costs more but is amortized over a longer period) and operating costs (since longer life results in more scheduled maintenance and overhaul actions).

Design Process

The design phases discussed above all involve aerodynamic, propulsion, and structural design, and the design of aircraft and mission systems. Each of these areas will be discussed, and the detail design phase will then be considered.

Aerodynamic design. Initial aerodynamic design centers on defining the external geometry and

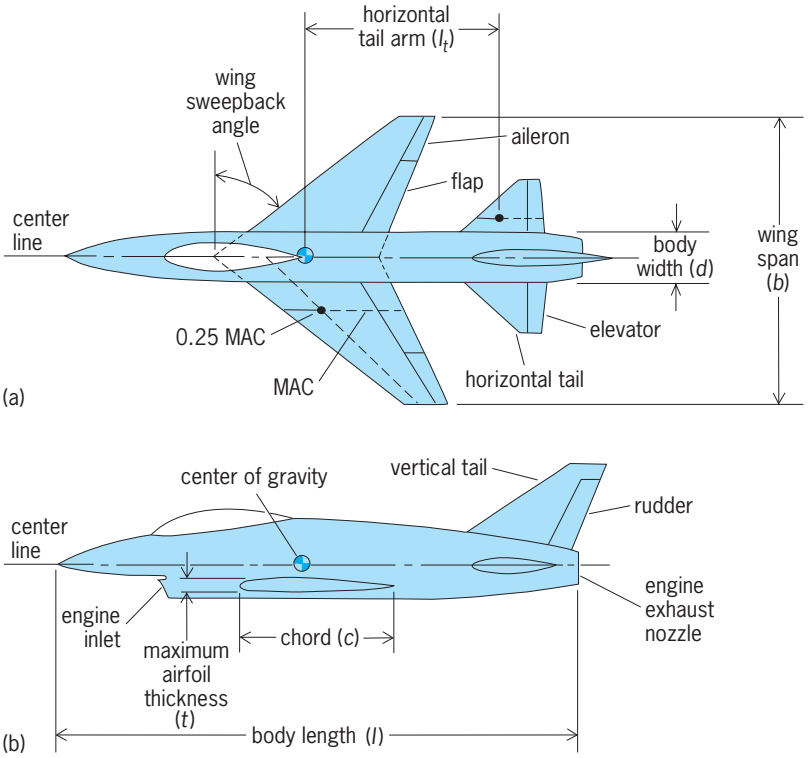


Fig. 1. Aircraft geometry and nomenclature. (*a*) Plan (top) view of aircraft. (*b*) Side view of aircraft. The terms MAC and 0.25 MAC point are explained in the text. Other important geometric parameters are the wing area S_w, the wetted area S_{wet}, and the wing aspect ratio, also defined in the text.

general aerodynamic configuration of the new aircraft. **Figure 1** defines the principal geometric parameters involved in aerodynamic design.

Lift and drag. The aerodynamic forces that determine aircraft performance capabilities are drag and lift. The basic, low-speed drag level of the aircraft is conventionally expressed as a term at zero lift composed of friction and pressure drag forces plus a term associated with the generation of lift, the drag due to lift or the induced drag. Since wings generally operate at a positive angle to the relative wind (angle of attack) in order to generate the necessary lift forces, the wing lift vector is tilted aft, resulting in a component of the lift vector in the drag direction (**Fig. 2**). *See* AERODYNAMIC FORCE; AIRFOIL; WING.

The magnitude of the friction drag force is directly related to the total surface area of the configuration exposed to the airflow, the wetted area S_{wet}. The level of pressure drag forces is determined by the relative slenderness of the body or fuselage ($1/d$ in Fig. 1) and by the relative thickness of wing and tail airfoils (t/c). The magnitude of the drag due to lift is influenced by wing area S_w, wing thickness (t/c), airfoil shape, and wing aspect ratio. [The wing area S_w is the total area of both wings, projected on the plan (top) view, and measured to the center line. The wing aspect ratio is the quantity b^2/S_w, where b is the wing span.] All of these influence the lifting efficiency at a given angle of attack and thus the size of the drag-due-to-lift vector. *See* ASPECT RATIO.

Aircraft that fly near or above the speed of sound must be designed to minimize aerodynamic compressibility effects, evidenced by the formation of shock waves and significant changes in all aerodynamic forces and moments. Because of the influence of compressibility effects, aerodynamic characteristics are usually referenced to the Mach number, the ratio of the airspeed to the speed of sound. Flight at the speed of sound is at Mach = 1.0 or sonic speed. Lower speeds or higher speeds are correspondingly termed subsonic or supersonic. Compressibility effects are mediated by the use of thin airfoils, wing and tail surface sweepback angles, and detailed attention to the lengthwise variation of the cross-sectional area of the configuration (**Fig. 3**). Smoothness of the area variation and the lowest possible peak value are crucial. *See* COMPRESSIBLE FLOW; MACH NUMBER; SHOCK WAVE; SUBSONIC FLIGHT; SUPERSONIC FLIGHT.

Those design parameters, S_w, t/c, wing aspect ratio, and airfoil shape, which were noted above as important to lifting efficiency (that is, in developing lift with minimum drag due to lift), are also the parameters which govern maximum lift capabilities. **Figure 4** illustrates two of these effects: that due to aspect ratio and that due to one of the airfoil shape parameters, camber, or the curvature of the airfoil center line (mean line).

Wing area S_w is strongly influenced by the selection of high-lift devices, wing components located in the leading-edge or trailing-edge regions or both. They augment lift by increasing the wing area or by changing the effective wing camber. **Figure 5** illustrates some typical effects of high-lift devices.

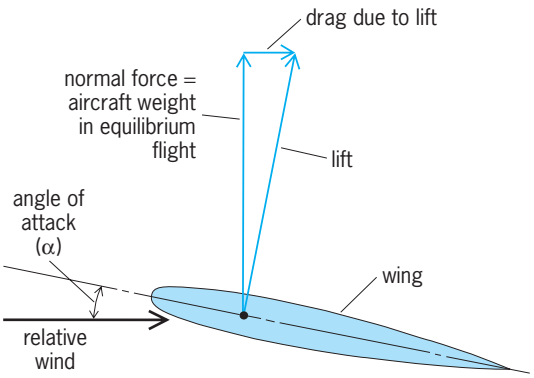


Fig. 2. Wing lift and drag due to lift.

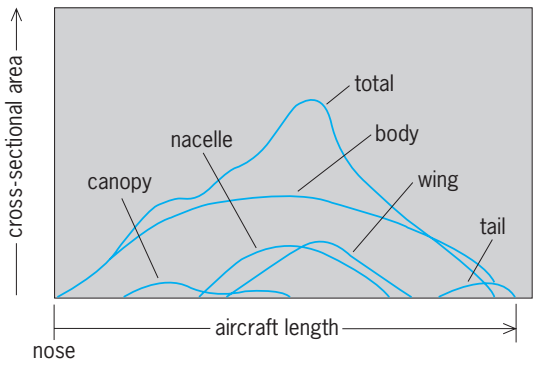


Fig. 3. Typical lengthwise variation of the cross-sectional area of an aircraft and its components.

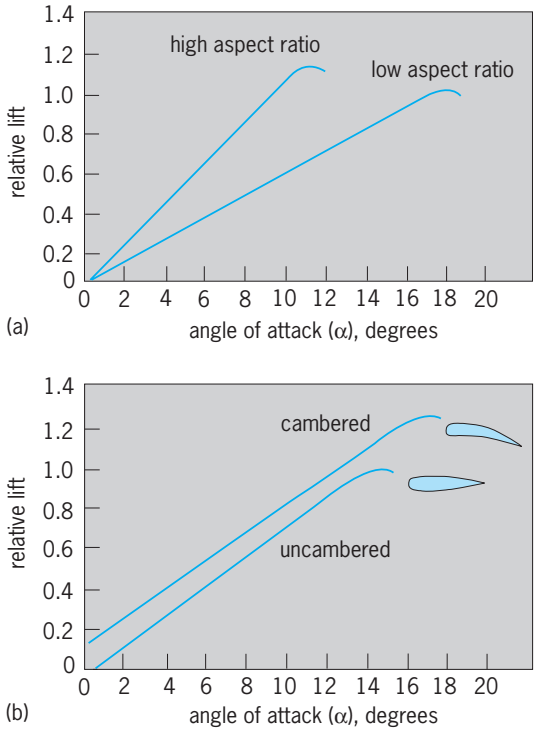


Fig. 4. Effects of wing geometry on lift. (*a*) Aspect ratio effects. (*b*) Camber effects.

Stability and control. The size and location of vertical and horizontal tail surfaces are the primary parameters that determine aircraft stability and control characteristics. In considering the conditions for longitudinal stability and control (relating to nose-up and nose-down motions), it is useful to introduce the concept of the 0.25 MAC point (Fig. 1). The mean aerodynamic chord (MAC) is defined to be the theoretical wing (or tail) chord line which divides the wing (or tail) into two areas having equal lift. The 0.25 MAC point is the intersection of the mean aerodynamic chord with the 25% chord line of the wing (or tail). Wing lift can be analyzed as a single force which typically acts at this point. If the wing is located so that this point is aft of the aircraft center-of-gravity location, then the aircraft is said to be statically stable, tail-off (without any tail contribution to stability). That is, if the aircraft is upset in the nose-up direction, then the increased wing lift (due to increased angle of attack) provides a moment about the center of gravity in the direction to restore equilibrium (that is, nose-down), and the converse is true for a nose-down upset. In this case the horizontal tail area and tail moment arm (l_t in Fig. 1) about the center of gravity are sized by the requirements for trim (to balance the moments about the center of gravity to zero in equilibrium flight) and for control to provide the required maneuvering capabilities. If the wing is located so that its lift center is forward of the center-of-gravity location, then the aircraft is statically unstable, tail-off, and sufficient horizontal tail area must be provided to restore static stability, that is, to make the tail moment about the center of gravity larger than the destabilizing wing moment. Analogous considerations apply for the case of directional (nose left–nose right) stability and control and the sizing and location of the vertical tail.

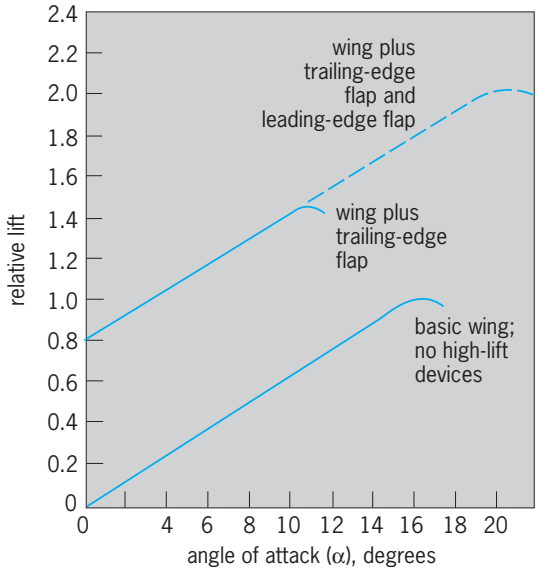


Fig. 5. Effects of high-lift devices.

Developments in digital computing and flight-control technologies have made the concept of artificial stability practical. Artificial stability systems continuously measure the aircraft's motions and activate its controls to provide inputs to change these motions to a desired state. *See* STABILITY AUGMENTATION.

To achieve the required control power, the designer must also decide how to provide for variation of the tail-surface lift forces. This is typically done by the use of the trailing-edge control surface (elevator or rudder) or by making the entire tail surfaces movable. Controls to effect rolling motions must also be provided, for example, wing trailing-edge surfaces (ailerons) or spoilers (wing upper-surface panels which are raised to deliberately spoil the lift on one wing). *See* AILERON; ELEVATOR (AIRCRAFT).

Design activities. The aerodynamic geometric parameters discussed above must be selected early in the design process and to reasonable accuracy. By the end of the conceptual design phase, these parameters should be known to within about 10%. This is necessary not only because the basics of size and performance are crucial to the technical and economic feasibility of the postulated designs, but also because the other design functions—structures, systems, and so forth—cannot proceed into much detail without accurate knowledge of aircraft size and shape constraints.

During preliminary design, the aerodynamic design activities focus on more rigorous methods of analysis together with wind-tunnel testing to refine all of the geometric parameters, to establish aerodynamic loads and pressure distributions for the structural designer, and to extend the analysis of flight performance capabilities. Through the use of a variety of analytical approximations and mathematical computational methods, aerodynamic estimates

of excellent accuracy are possible, and, with modern supercomputers, are practical in terms of time and cost. These methods have begun to replace the traditional use of wind-tunnel testing to provide aerodynamic data. However, until computational methods are developed to the point where they can provide solutions for arbitrary combinations of aircraft configuration and flight conditions, large-scale wind-tunnel testing will continue to be necessary to provide credence to the aerodynamic data.

Propulsion design. Propulsion design comprises the selection of an engine from among the available models and the design of the engine's installation on or in the aircraft.

Engine selection. Selection of the best propulsion concept involves choosing from among a wide variety of types ranging from reciprocating engine-propeller power plants through turboprops, turbojets, turbofans, and ducted and unducted fan engine developments. The selection process involves aircraft performance analyses comparing flight performance with the various candidate engines installed. In the cases where the new aircraft design is being based on a propulsion system which is still in development, the selection process is more complicated. Then the basic engine size is subject to variation, as are certain internal engine components and engine control systems, so that some degree of tailoring is possible to provide the best performance match between engine and airframe. Analysis of these expanded possibilities requires much interaction between aircraft and engine manufacturer design teams. *See* AIRCRAFT ENGINE PERFORMANCE; TURBOFAN; TURBOPROP.

Engine inlet design. Once an engine has been selected, the propulsion engineering tasks are to design the air inlet for the engine, and to assure the satisfactory physical and aerodynamic integration of the inlet, engine, and exhaust nozzle or the engine nacelles with the rest of the airframe.

Figure 6 illustrates the design parameters that govern the aerodynamic efficiency of a typical turbine engine inlet system. The major parameters to be chosen include the throat area, the diffuser length and shape, and the relative bluntness of the inlet lips. The throat area is the minimum cross-sectional area in the inlet. This area is chosen to match the engines' airflow demand with the flow of incoming air at a selected flight condition, for example, the flight condition at which the aircraft will spend the bulk of its flight time (for example, the cruise speed and altitude for a commercial airliner). At speeds or engine settings where airflow demand is either more or less than at the match point, drag penalties occur due to the nonuniformity of the incoming airflow at the inlet entrance.

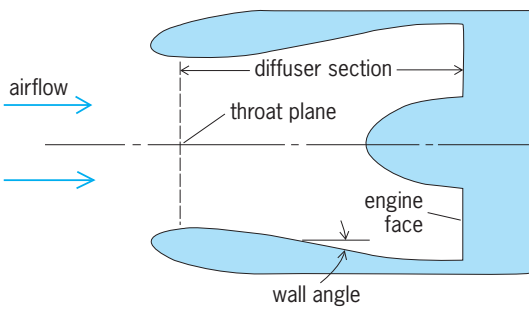


Fig. 6. Cross section of engine inlet, showing geometry and nomenclature.

Pressure distortions at the engine face adversely affect the inlets' pressure recovery (ratio of total pressure at the engine face to the total pressure available in the incoming stream). This, in turn, reduces engine thrust and in extreme cases can affect engine performance sufficiently to stall the engine. Inlet lips with a blunt leading-edge radius provide smoother airflow than sharper lip designs and operate well through a larger range of incoming flow angles (corresponding to a larger range of aircraft flight attitudes) than do sharp lip designs. The blunter lip designs, however, suffer a drag penalty at high speeds.

Diffuser geometry is governed by inlet airstream speed considerations. Turbine engines operate most efficiently with airflow velocities at the engine face of around 30–40% of the speed of sound (Mach = 0.3–0.4). In aircraft that fly at speeds much above this, the inlet must be designed to reduce the inlet stream velocity to match the engines' needs. The diffuser section accomplishes this; the airflow passing through the throat section (the minimum cross-sectional area) slows down in proportion to the increase in duct area through the diffuser to the engine face. Diffuser length becomes a function of the magnitude of aircraft design speed since the rate of increase of duct cross-sectional area is limited. Duct wall angles cannot exceed about 6° or the flow will separate (pull off) from the wall, causing unacceptable flow distortion at the engine face. *See* BOUNDARY-LAYER FLOW; DIFFUSER.

Supersonic inlet design is much more complex. The configuration shown in Fig. 6 can be varied to function as a supersonic inlet up to Mach numbers of 1.3–1.5. At higher Mach numbers, inlet pressure recovery is degraded to the point that more complicated inlet designs are necessary. In a supersonic version of the configuration of Fig. 6, the inlet lips are relatively sharper (than for subsonic design) to minimize high-speed drag, and the diffuser relatively longer because of the greater speed difference between throat and engine face conditions. At aircraft speeds of Mach = 1.0 and above, a shock wave forms just ahead of the inlet entrance. This shock wave serves to provide the initial slowdown of the incoming stream from supersonic to high subsonic speeds at the inlet entrance. Inlets designed for higher maximum Mach numbers require some form of variable geometry such as movable internal surfaces ahead of and at the throat to vary the throat area and set up multiple shock-wave systems at the inlet entrance, in order to provide the necessary subsonic flow at the diffuser entrance. *See* SUPERSONIC DIFFUSER.

Activities. As is the case with the aerodynamic design of the aircraft, the major propulsion design parameters must be established to within about 10% by the end of the conceptual design. During preliminary

design the propulsion design parameters are refined, supported by inlet model wind-tunnel tests or complete aircraft and inlet (or nacelle) model wind-tunnel tests. These tests frequently include electrically driven model engines to provide the most complete possible simulation of all airflow conditions.

Structural design. Structural design begins when the first complete, integrated aerodynamic and propulsion concept is formulated. The process starts with preliminary estimates of design airloads and inertial loads (loads due to the mass of the aircraft being accelerated during manuevers). These design loads, called static loads, are calculated at peak or maximum loading cases (combinations of speed, altitude, maneuvering condition, and aircraft weight) which are critical for a particular structural component.

Conceptual design activities. During conceptual design, the structural design effort centers on a first-order structural arrangement which defines major structural components and establishes the most direct load paths through the structure that are possible within the constraints of the aerodynamic configuration. An initial determination of structural and material concepts to be used is made at this time, for example, deciding whether the wing should be constructed from built-up sheet metal details, or by using machined skins with integral stiffeners, or from fiber-reinforced composite materials. These initial decisions are based on the new aircraft's design criteria and on weight and cost considerations from prior experiences. *See* WING STRUCTURE.

At this point there is sufficient information for mass properties analysis to begin. Initial estimates of weight and balance (center-of-gravity location) are made based on statistical correlations of data from past aircraft designs. These data support calculations to assess flight performance versus requirements. The initial weight estimates also provide a rough indicator of aircraft cost, and the initial balance estimates are critical to the accurate sizing of wing and tail areas.

Preliminary design activities. During preliminary design, the structural design effort expands into consideration of dynamic loads, airframe life, and structural integrity. Dynamic loading conditions arise from many sources: landing impact, flight through turbulence, taxiing over rough runways, and so forth. In the case of military aircraft, additional dynamic loads occur due to abrupt maneuvers and to gun firing and forced ejection of bombs and other weapons. The general effect of various dynamic load conditions is to add to material thickness requirements in localized regions. However, these loads may be severe enough to change the basic structural arrangement or choice of materials. For example, Navy aircraft designed to operate on aircraft carriers experience dynamic loadings during catapult launch and arrested landing which dominate the design of landing gears and fuselage structure. *See* LOADS, DYNAMIC.

Airframe life requirements are usually stated in terms of desired total flight hours or total flight cycles. (One takeoff, flight, and landing equals one cycle.) To the structural designer this translates into requirements for airframe fatigue life. Fatigue life measures the ability of a structure to withstand repeated loadings without failure. Design for high fatigue life involves selection of materials and the design of structural components that minimize concentrated stresses. Fatigue damage starts with a crack which propagates to the point that the affected structure cannot carry its design load. Particulars of the metal alloy selected influence resistance to crack propagation, as do the details of material processing (for example, heat treatment). These factors also influence the susceptibility to corrosion. Corrosion damage is one of the mechanisms that can initiate structural cracks (and can also influence airframe life in other ways, for example, loss of strength in a joint due to fastener corrosion). One of the benefits of fiber-reinforced composite materials for aircraft is the inherent resistance to both crack propagation and corrosion damage. Structural design to minimize stress concentrations includes the avoidance of structural discontinuities (for example, access doors cut in load-bearing structures such as wing skins) and the use of structures with multiple load paths. *See* COMPOSITE MATERIAL; CORROSION; METAL, MECHANICAL PROPERTIES OF; PLASTIC DEFORMATION OF METAL.

Structural integrity design activities impose requirements for damage tolerance, the ability of the structure to continue to support design loads after specified component failures. Fail-safe design approaches are similar to design for fatigue resistance: avoidance of stress concentrations and spreading loads out over multiple supporting structural members. Typical concepts are multispar wings, and the use of multiple bolts at splice joints rather than one or two large fasteners. Military aircraft have additional structural requirements in order to be able to survive specified damage due to hostile gunfire or missile firings.

As these structural design refinements proceed during preliminary design, the structural load and stress analysis techniques used become increasingly more detailed and accurate. The structural arrangement and selection of structural concepts are completed in sufficient detail to permit accurate sizing of structural members (**Fig. 7**). Weight and balance estimates for the structure with a statistical basis are replaced by estimates based on direct calculation of sized structural components.

As preliminary design of the aircraft structure progresses, it may become desirable to conduct laboratory testing of any unique structural components or unusual combinations of materials to verify estimated strengths and to provide confidence to proceed to detailed design. *See* STRUCTURAL DESIGN.

Aircraft systems design. Aircraft systems include all of those systems and subsystems required for the aircraft to operate. The major systems are power systems, flight-control systems, navigation and communication systems, crew systems, the landing-gear system, and fuel systems.

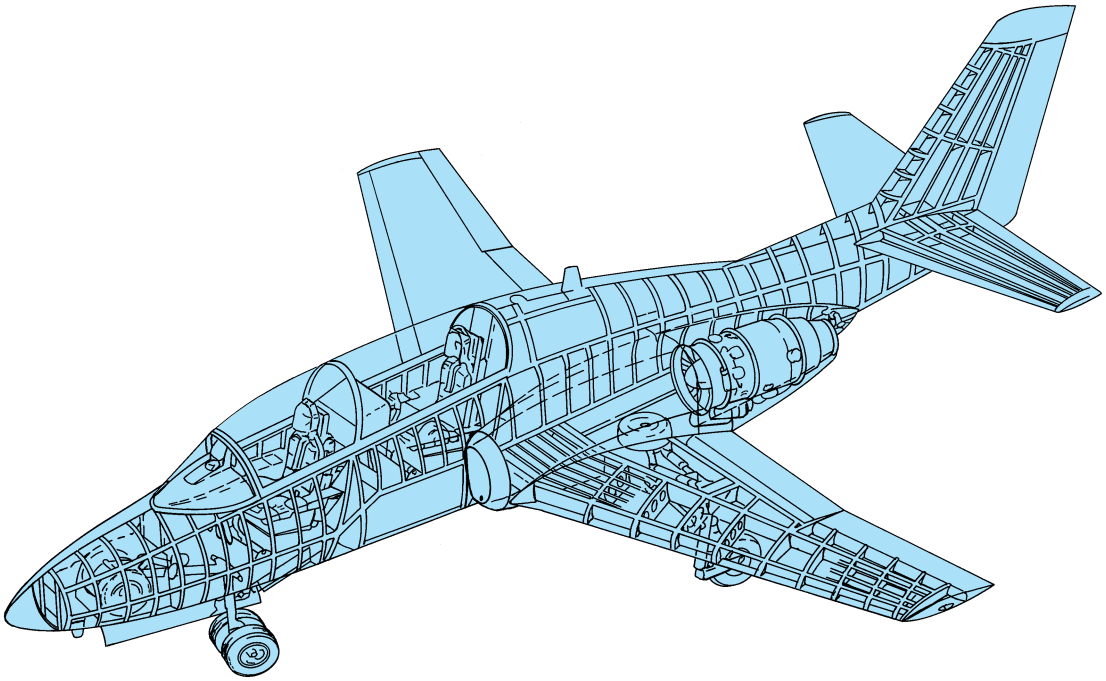


Fig. 7. Perspective drawing of a hypothetical two-place trainer aircraft, illustrating the structural detail developed during preliminary design.

Power systems comprise the electrical, hydraulic, or pneumatic power generating and power distribution subsystems to serve the other aircraft systems.

Flight-control systems comprise the subsystems which operate movable control surfaces, and provide direct (mechanical) or indirect (electrical or electronic) transmissions of pilot control signals to control surfaces. The design of autopilots and stability augmentation subsystems is also part of flight-control system design activities. *See* AUTOPILOT; FLIGHT CONTROLS.

Navigation and communication systems comprise the subsystems which provide the pilot or flight crew with information as to aircraft position and directive information to the planned destination, and the subsystems to communicate with ground stations and other aircraft as desired. *See* AIR NAVIGATION.

Crew systems comprise the cockpit or flight-deck design, including the human-factors engineering aspect of controls and displays, seat design, the subsystems which provide the crew with information as to aircraft flight and system status, and so forth. *See* AIRCRAFT INSTRUMENTATION.

The landing-gear system comprises the landing-gear structure, wheels and tires, and the means to activate the gear. *See* LANDING GEAR.

Fuel systems comprise the tanks and plumbing to distribute the fuel, pumps, and devices that measure the fuel condition.

Conceptual design. Design of these major subsystems must begin relatively early in the conceptual design phase, because they represent large dimensional and volume requirements which can influence overall aircraft size and shape or because they interact directly with the aerodynamic concept (as in the case of flight-control systems) or propulsion selection (as in the case of power systems). Early estimates must be made, to guide the aircraft design, of aircraft system features such as the size of nose-mounted radar dishes, the volume and location of fuel tanks, and the landing-gear design and location. Since many of the subsystems and much of the equipment in the various aircraft systems are purchased from suppliers, an important activity during conceptual design is preparation of technical and cost comparisons of available subsystems and equipment.

Preliminary design. During preliminary design, the aircraft system definition is completed to include the additional subsystems not yet defined. Examples include environmental-control subsystems which provide conditioned air to crew and passengers, fire detection and suppression subsystems, and various system-monitoring and alerting subsystems which detect system failures and alert the crew. The installation of the many aircraft system components and the routing of tubing and wiring through the aircraft are complex tasks which are often aided by the construction of partial or complete aircraft mock-ups. These are full-scale models of the aircraft, made of inexpensive materials, which aid in locating structural and system components.

A major systems design tool is the flight simulator. This system typically comprises a cockpit with instruments, displays, and controls; and a means for projecting an outside world view to the pilot in the cockpit. A digital computer mechanizes these hardware items and provides for simulation of arbitrary aircraft performance and system characteristics. The systems designer, by deliberately varying system characteristics, can develop optimal designs.

Selection of subsystems and equipment should be essentially complete at the end of preliminary design,

along with system mechanization concepts, installation drawings, and system reliability estimates.

Mission systems design. Mission systems are those additional systems and subsystems peculiar to the role of military combat aircraft. Offensive systems include target detection and location, weapon carriage, weapon arming and release, weapon guidance, and so forth. Defensive systems comprise various electronic countermeasures designed to confuse hostile tracking and fire-control systems, and subsystems which provide information about the location and activities of threatening aircraft or ground systems. The design process for military systems is similar to that for aircraft systems. *See* AIR ARMAMENT; ELECTRONIC WARFARE.

Detail design. Detail design engineering verifies and completes the technical definition of the aircraft configuration developed during the conceptual and preliminary design phases; translates the aircraft design data into the formats required to accomplish tooling, fabrication, and assembly processes; and, finally, provides support in resolving any difficulties that may arise in fabrication or assembly of the aircraft.

Aerodynamic and propulsion design activities center on completion of a final set of aerodynamic loads as a basis for completion of structural and flight-control system design. A final round of large-scale, powered-model wind-tunnel testing is usually conducted.

Structural design work peaks during detail design. Final analyses of static and dynamic loads, including all fatigue-life and fail-safe design considerations, are completed.

Verification of structural integrity is accomplished by extensive laboratory testing of structures, including static testing of a complete airframe during which it is loaded to simulate various critical load conditions. Additional structural integrity verifications include drop testing to demonstrate the integrity of landing gear and gear attachment structures under dynamic impact loads, and fatigue-life demonstration. These latter tests, which may involve millions of repeated load applications, usually extend from late in the detail design phase to well beyond the first deliveries of production aircraft.

Detail part design is completed, and all structural component and assembly drawings are completed and released for manufacture. The application of computer aided design and manufacturing (CAD/CAM) to this phase of the design process has grown rapidly. These technologies replace the use of drawings and various intermediate steps between drawing release and actual parts fabrication with the direct translation of engineering digital data describing a component or subassembly into software instructions to machining or assembly tools of various kinds. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

Aircraft and mission systems detail design efforts include preparation of technical specifications used for procurement of purchased subsystems and equipment, and preparation of final system installation drawings or computer databases, as appropriate. Piloted simulation studies generally continue into this detail design phase with real equipment items progressively tied into the simulation in place of software representations. This provides improved verification of end-to-end performance of the simulated systems.

Detail design normally is complete with the release of all information for production. Subsequent events may still cause further design activity. These range from design changes to facilitate manufacturing improvements and changes due to problems uncovered in flight tests, to problems that may emerge late in the aircraft's service life such as premature fatigue failures. *See* AIRPLANE. Peter L. Marshall

Bibliography. G. Corning, *Supersonic and Subsonic, CTOL and VTOL, Airplane Design*, 4th ed., revised, 1979; H. D. Curtis, *Fundamentals of Aircraft Structural Analysis*, 1997; J. P. Fielding, *Introduction to Aircraft Design*, 1999; L. M. Nicolai, *Fundamentals of Aircraft Design*, 1984; C. D. Perkins and R. E. Hage, *Airplane Performance, Stability and Control*, 1949; D. P. Raymer, *Aircraft Design: A Conceptual Approach*, 1999; F. K. Teichmann, *Fundamentals of Aircraft Structural Analysis*, 1968; K. D. Wood, *Aircraft Design*, 3d ed., 1968.

Aircraft engine

A component of an aircraft that develops either shaft horsepower or thrust and incorporates design features most advantageous for aircraft propulsion. An engine developing shaft horsepower requires an additional means to convert this power to useful thrust for aircraft, such as a propeller, a fan, or a helicopter rotor. It is common practice in this case to designate the unit developing shaft horsepower as the aircraft engine, and the combination of engine and propeller, for example, as an aircraft power plant. In case thrust is developed directly as in a turbojet engine, the terms engine and power plant are used interchangeably. *See* TURBOJET.

The characteristics primarily emphasized in an aircraft engine are high takeoff thrust and low specific weight, low specific fuel consumption, and low drag of the installed power plant at the aircraft speeds and altitudes desired. Reliability and durability are essential, as is emphasis on high output and light weight, so that a premium is placed on quality materials and fuels, as well as on design and manufacturing skills and practices. *See* AIRCRAFT ENGINE PERFORMANCE; AIRCRAFT FUEL.

Air-breathing types of aircraft engines use oxygen from the atmosphere to combine chemically with fuel carried in the vehicle, providing the energy for propulsion, in contrast to rocket types in which both the fuel and oxidizer are carried in the aircraft. Air-breathing engines suffer decreased power or thrust output with altitude increase, due to decreasing air density, unless supercharged. *See* AIRCRAFT PROPULSION; INTERNAL COMBUSTION ENGINE; JET PROPULSION; PROPELLER (AIRCRAFT); PROPULSION;

RECIPROCATING AIRCRAFT ENGINE; ROCKET PROPULSION; TURBINE PROPULSION. Ronald Hazen

Aircraft engine performance

The power or thrust, the specific fuel consumption (mass of fuel consumed per hour per unit of delivered thrust or power), and any other important operating parameters of an aircraft engine, reported in graphical, tabular, or computerized format as a function of a thrust (or power) selection parameter, of ambient air pressure and temperature, of flight speed, and of other environmental, operational, and installation variables. *See* POWER; SPECIFIC FUEL CONSUMPTION; THRUST.

Characterization. The ultimate function of an aircraft engine is to supply the thrust for aircraft propulsion and lift. An engine's performance is most conveniently reported as the net thrust force that the engine supplies. However, in certain types of engines, such as reciprocating engines, turboshaft engines, and turboprop engines, the propulsor, that is, the thrust-producing device (for example, the rotor of a helicopter or the propeller of an aircraft), is treated as a separate system. For these cases, the aircraft engine's performance is reported as the power provided by the engine to drive the propulsor, rather than as the thrust provided by the total propulsive system. Any residual thrust from the engine exhaust is reported separately, or converted to an equivalent amount of power. For turbojet engines, where the thrust is supplied by reaction forces in the basic power unit, or for turbofan engines, where there is a propulsor integrated into the engine, the net thrust of the engine is reported directly, without reference to the power generated within the engine. *See* AIRCRAFT PROPULSION; HELICOPTER; PROPELLER (AIRCRAFT); RECIPROCATING AIRCRAFT ENGINE; TURBINE PROPULSION; TURBOFAN; TURBOJET; TURBOPROP.

Operating parameters. Thrust selection is categorized by various thrust or power ratings. These may include special high ratings whose usage is limited to short duration or under specified extraordinary situations, such as maximum contingency or automatic power reserve, invoked when one engine becomes inoperative during takeoff in a multiengine aircraft. Use of these ratings may entail consequent engine inspection or overhaul. Another category of high ratings, which includes takeoff, maximum climb, and military or combat or intermediate in a military aircraft, may be limited to a specified duration or cumulative usage in order to conserve the life of the engine. Still other ratings such as maximum cruise, normal rated, maximum continuous, and cruise are permitted for extended use and are designed to afford long life and minimum fuel consumption. Additional ratings termed flight idle and ground idle are specified to accommodate operation where power or thrust and fuel consumption are to be minimized, while maintaining the engine in a self-sustaining condition where it is available for rapid application of power.

Other reported operating parameters may be starting time, acceleration time, engine rotative speed (or speeds for multirotor engines), and pressures and temperatures within the engine that may be monitored. Engine operation is affected by inlet and exhaust system losses, ambient humidity, power extracted from the engine to run aircraft accessories, and compressed air extracted from the engine for use in aircraft systems.

Power producer operation. At the heart of any reciprocating or gas turbine aircraft engine is a thermodynamic device that converts the energy from the combustion of fuel with air to useful mechanical energy. The fuel is usually a petroleum fraction such as aviation gasoline for reciprocating engines, and kerosine or jet fuel for gas turbine engines. A typical heating value of petroleum fuels is 18,650 Btu/lb (43,350 kilojoules/kg). If 100% conversion to mechanical energy were feasible, 1 lb/h of fuel would give 7.3 horsepower or a specific fuel consumption of 0.14 lb/(h)(hp), and 1 kg/h of fuel would give 12 kW or specific fuel consumption of 0.083 kg/(h)(kW). Basic thermodynamic principles imply that no more than 35 to 40% of this theoretical value of useful power is actually achievable, implying a lower limit on specific fuel consumption of about 0.4 lb/(h)(hp) [0.24 kg/(h)(kW)]. Inefficiencies in the components of any real power producer include leakage of high-pressure working fluid (air or products of combustion); leakage of heat from the high-temperature working fluid to the lower-temperature engine bay; pressure losses in the passages through which the working fluid moves from component to component; diversion of a portion of the air to cool high-temperature parts; and inefficiencies in the compression, combustion, and expansion components of the engine. *See* AIRCRAFT FUEL.

The basic principle on which all engines operate is first to compress the air drawn into the engine inlet, then to heat the compressed air by burning fuel in it, and then to extract work from the high-pressure, high-temperature gas, to drive the compression device and then to provide useful thrust for aircraft propulsion.

Reciprocating engine. In a reciprocating engine (the Otto cycle, **Fig. 1**), the movement of working fluid through the engine is a batch process, with each cylinder accepting a slug of air once every two revolutions of the shaft. The compression is accomplished by the piston reducing the volume of the slug of air trapped in the cylinder. The pressure rise afforded by compression in the cylinder may be supplemented by a supercharger, a compressor mounted in the engine's air inlet which is usually driven by a turbine mounted in the engine's exhaust stream. Combustion is accomplished at approximately constant volume with the air fully compressed. Work extraction is accomplished by allowing the hot compressed gas to expand against the retreating piston. The cycle is the same as that of an automobile engine. *See* INTERNAL COMBUSTION ENGINE; OTTO CYCLE; SUPERCHARGER.

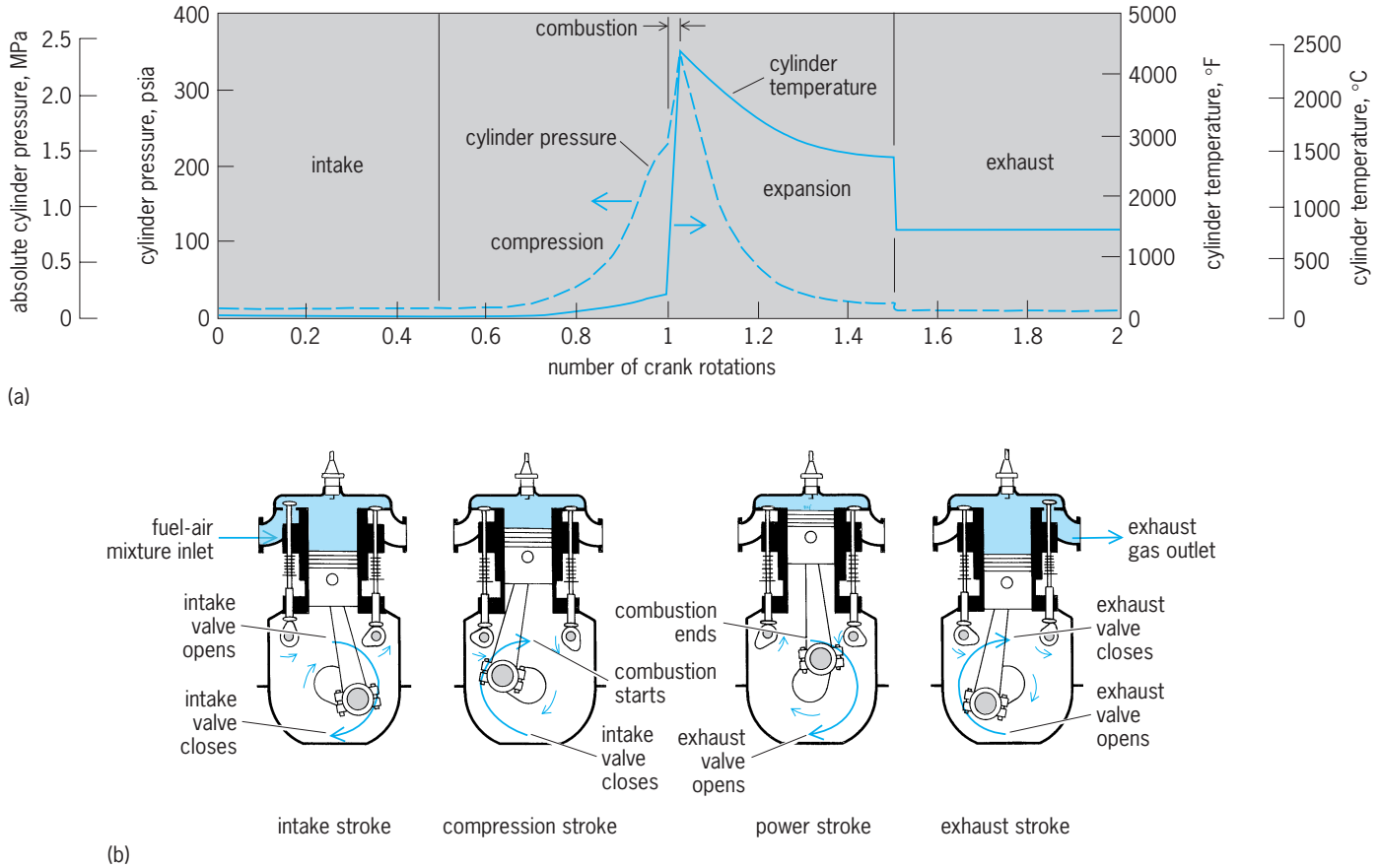


Fig. 1. Cycle of reciprocating engine. (*a*) Cylinder temperature and pressure during cycle of typical engine. (*b*) Engine configuration at five stages of the cycle. (*After R. D. Bent and J. L. McKinley, Aircraft Powerplants, 5th ed., McGraw-Hill, 1985*)

Gas turbine engine. In a gas turbine engine (the Brayton cycle, **Fig. 2**), the process is one of steady flow of the air through the components. A compressor increases the pressure of the continuous flow of air. The pressure rise may be supplemented by the propulsor (the propeller or fan) if the air passes through the propulsor on its way to the compressor. In order to achieve very high pressure ratios, some

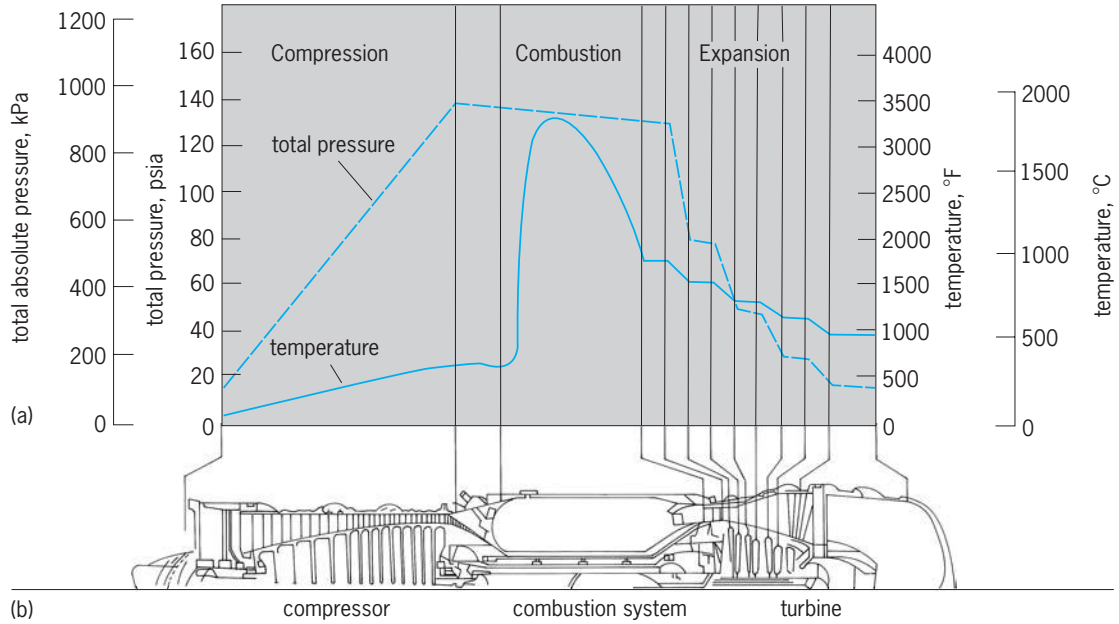


Fig. 2. Cycle of typical gas turbine (turbojet) engine. (*a*) Plot of temperature and pressure of air flowing through engine. Positions on horizontal scale are identified with locations on (*b*) diagram of engine. (*After I. E. Treager, Aircraft Gas Turbine Technology, 2d ed., McGraw-Hill, 1979*)

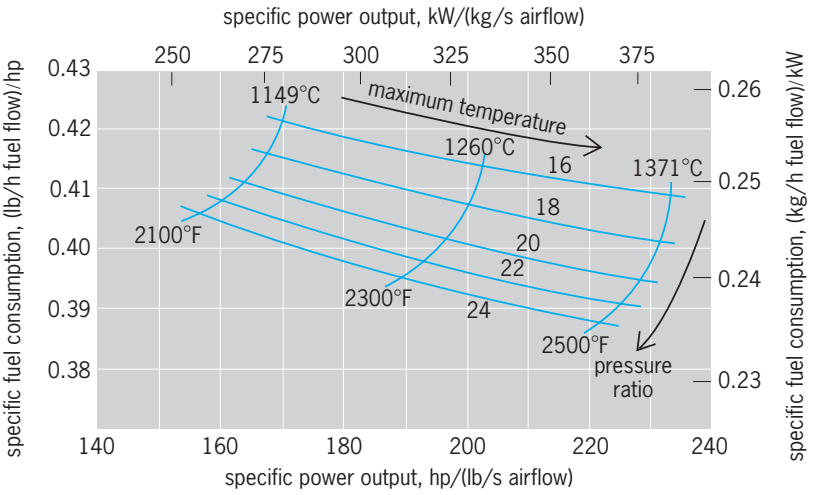


Fig. 3. Influence of cycle pressure ratio and maximum cycle temperature on performance (specific fuel consumption and specific power output) of a gas turbine (turboshaft) engine.

gas turbines use two compressors in tandem, each with its own driving turbine and each turning at its own rotative speed. Heat addition is accomplished at constant pressure in a steady-flow combustion chamber. Work extraction to drive the compressor is accomplished by steady-flow expansion through a turbine. *See* BRAYTON CYCLE; GAS TURBINE.

Ram effect. The pressure rise provided by the engine's compression device is supplemented by the ram effect of slowing down the high-velocity air stream (as seen on the aircraft) into the engine's inlet. This effect can almost double the pressure and the density of the air brought aboard at transonic flight speed, and results in considerably greater increases at supersonic flight speed.

Water injection. In certain applications, the power of the engine can be augmented for short durations by injection of water into the engine inlet. The evaporation of the water cools the compressed air to allow more fuel addition and increases the effective mass flow rate of the working fluid.

Power producer performance. The performance of a power producer is generally measured and assessed in terms of two parameters, the specific fuel consumption and the specific power. The specific fuel consumption is the mass rate of fuel flow per unit of thrust or power produced. A very low specific fuel consumption is desirable in order to minimize the cost of operation and maximize the range and performance of the aircraft by minimizing the weight of fuel it must carry. The specific power is the power produced per unit of airflow into the engine. A very high specific power is desirable since the bulk of the engine (its weight, length, and drag-inducing frontal area) is strongly dependent on the airflow of the engine. For a given required power, the bulk is minimized by increasing the specific power. These parameters are a primary function of two key variables that are selected by the engine's designers: the overall compression ratio of the cycle; and the maximum temperature of the products of combustion of the fuel and air.

Effects of cycle temperature. High maximum cycle temperature is of primary importance in achieving high specific power, and is of moderate importance in achieving low specific fuel consumption at low values of the cycle pressure ratio (**Fig. 3**). Modern gas turbines have maximum cycle temperatures between 2000 and 3000°F (1090 and 1650°C). Since the gas turbine is a continuous flow process, and the turbine is continuously bathed in air at these high temperatures, and since these temperatures often exceed the melting temperatures of the materials of which the engine is built, there is a considerable challenge to develop efficient means of cooling the structural parts, and to develop new materials that will withstand these high temperatures and still provide the durability and integrity demanded by the engine user. In a reciprocating engine, the peak cycle temperature is experienced only instantaneously one time in each cylinder for every two rotations of the crankshaft. Cycle temperatures of 4000–5000°F (2200–2760°C) are therefore feasible.

Effects of pressure ratio. High cycle pressure ratios are of considerable importance in achieving low specific fuel consumption, and have a moderately deleterious effect in reducing specific power (Fig. 3). Modern gas turbine power-producer sections, particularly those used in subsonic aircraft that do not benefit from the large ram pressure ratio experienced at supersonic flight speed, may have cycle pressure ratios of 40:1 or more, requiring considerable ingenuity and skill in the design of the compressor. Reciprocating engines may have pressure ratios of 10:1 or 11:1.

Effects of fuel flow. Thrust or power selection is usually accomplished by manipulating the fuel flow to the engine. The power producer responds to a reduction in fuel flow by reducing its rotative speed, which reduces the airflow into the engine and the power produced. In a gas turbine engine, this reduction in rotative speed is accompanied by a reduction in the pressure ratio produced by the compressor, and a considerable increase in specific fuel consumption at low power. On the other hand, the compression ratio built into the cylinder and piston of a reciprocating engine does not vary with the rotative speed of the engine, and it does not experience the steep increase in specific fuel consumption experienced by the gas turbine at low power.

Effects of ambient conditions. Ambient pressure and temperature and aircraft flight speed also have major effects on aircraft engine performance. Average ambient pressure and temperature vary considerably with altitude (**Fig. 4**) and also vary with latitude, with time of year, and from day to day. Decreasing ambient pressure at increasing altitude is accompanied by proportionate reductions in air density. Since engine geometry generally accommodates a fixed volumetric flow of air, the reduced density implies a reduction in mass flow of air and a consequent reduction in the power produced. An increase in ambient temperature also reduces air density, but more importantly, high temperature of the inlet air has a magnified effect in increasing the temperature of the compressed air, limiting the amount of fuel that may

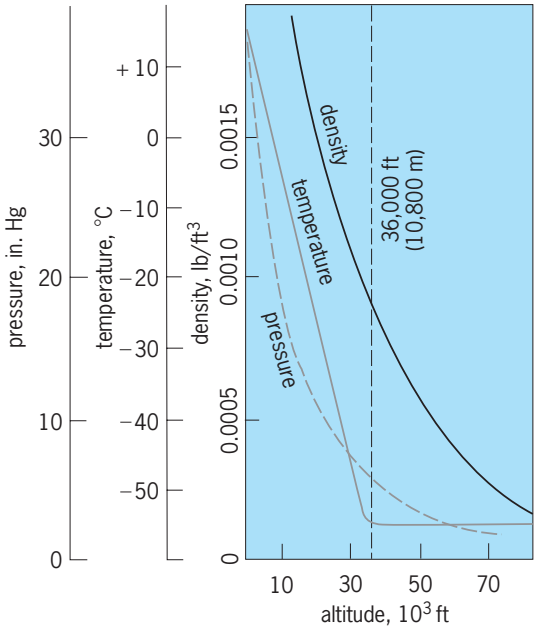


Fig. 4. Variation of ambient pressure, temperature, and density as functions of altitude. 1 in. Hg = 3.4 kPa; °F = (°C × 1.8) + 32; 1 lb/ft³ = 16.0 kg/m³; 10 M ft = 10,000 ft = 3000 m. (*After I. E. Treager, Aircraft Gas Turbine Technology, 2d ed., McGraw-Hill, 1979*)

be introduced without exceeding the limiting temperature of the turbine. Consequently, aircraft engine size must often be specified to accommodate the limiting condition of achieving the required takeoff power on the hottest day anticipated at the highest-altitude airport that might be used. Otherwise, an aircraft must have its cargo or passenger capacity restricted when these adverse conditions are encountered. *See* ATMOSPHERE.

Effects of flight speed. Aircraft flight speed also affects power producer performance, primarily by way of the ram pressure and ram temperature rise in the inlet air. In addition to the ram effects on the power producer, there are important flight speed effects on the conversion of the power to propulsive thrust, as discussed below.

Propulsor operation. After the expansion process, where work has been extracted to power the compression process—in the retreating piston of the reciprocating engine, or the turbine of the gas turbine engine—there is a surplus of energy residing in the high-pressure, high-temperature gases that is exploited to provide propulsion in either or both of two additional processes. (1) The high-pressure, high-temperature gases can be exhausted through a jet nozzle to produce thrust, as in the simple jet engine. The high-pressure gases, acting on the engine structure, provide a forward force (or thrust). The rearward force that would normally balance the forward force in a closed pressure vessel is absent in the jet nozzle by virtue of the nozzle orifice. (2) The high-pressure, high-temperature gases may be expanded in the gas turbine engine's second turbine or the reciprocating engine's cylinder to provide mechanical energy to an engine shaft which can then power a separate propulsor—a propeller or a helicopter rotor. *See* JET PROPULSION.

The turbofan exploits both of these processes. The high-pressure, high-temperature gases from the power producer are partially expanded through a turbine to power the fan (propulsor), and the residual energy is sent through a jet nozzle to produce thrust.

Afterburner. The thrust of a jet engine may be augmented by as much as 50% by burning additional fuel in a special auxiliary combustion chamber (that is, an afterburner) located just upstream of the jet nozzle. The increase in temperature of the exhaust gases to the range of 3200°F (1760°C) causes a considerable increase in exhaust jet velocity and, as discussed below, a consequent increase in thrust. Although afterburners are a relatively simple, lightweight, and inexpensive means of thrust augmentation, their specific fuel consumption is considerably higher than the basic jet engine's fuel consumption, and their use is generally limited to combat or short-duration portions of an aircraft's mission. *See* AFTERBURNER.

Newton's second law. The two propulsion processes discussed above are actually two aspects of one principle, Newton's second law of motion. That law, applied to aircraft propulsion (**Fig. 5**), provides a direct relationship between the net thrust developed by a propulsive device and the change in velocity imparted to the flow of air through the device. In Eq. (1), F is the resultant force or net thrust of the

$$F = M_e V_e - M_i V_i \tag{1}$$

propulsor; M_e is the mass flow rate and V_e is the the velocity of the gas stream exiting from the propulsor or the engine; M_i is the mass flow rate and V_i is the velocity (that is, the flight speed) of the gas stream entering the engine or propulsor; and the forces and velocity components are measured in the same direction. For engines like turbofans that have more than one exhaust stream, the jet momenta of the several streams must be added. For isolated propulsors like propellers or helicopter rotors, thrust is usually assessed by focusing on the left side of Eq. (1), by summing up the pressure forces on the blades of the propulsor. For jet engines and turbofans, it is more usual and convenient to infer the forces by

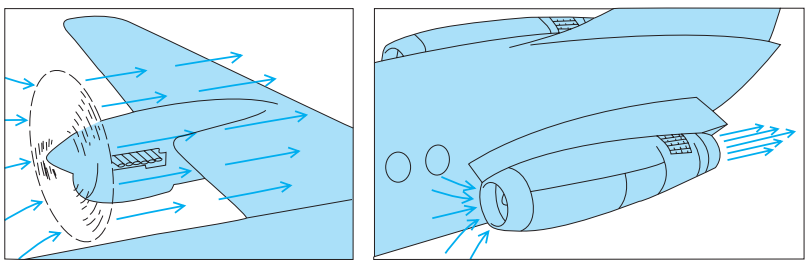


Fig. 5. Newton's second law of motion applied to two classes of aircraft propulsion. (*a*) Propeller gives a small acceleration to a large mass of air. (*b*) Turbojet engine gives a large acceleration to a small mass of air. (*After I. E. Treager, Aircraft Gas Turbine Technology, 2d ed., McGraw-Hill, 1979*)

evaluating the right side of Eq. (1). *See* FORCE; NEWTON'S LAWS OF MOTION.

Equation (1) also illustrates another important aspect of aircraft engine performance: the effect of flight speed on net thrust. The first term on the right side of Eq. (1), called the gross thrust, is the hypothetical thrust that would be developed at zero flight speed. The second term on the right side of Eq. (1), called the ram drag or induced drag, is the drag on the engine induced by the ram pressure developed on the front face of the engine. For supersonic and hypersonic aircraft, the effect is considerable, so that net thrust is a very small fraction of gross thrust, and the whole system's effectiveness is very sensitive to small changes in engine efficiency.

Propulsor performance. To complete an assessment of the performance of the total aircraft propulsion system, it is necessary to supplement the evaluation of the power producer with an evaluation of the propulsor.

Propulsive efficiency versus size. If the fuel flow rate is assumed negligible with respect to the airflow rate (it is indeed just a few percent), then, from Eq. (1), the specific thrust is given by Eq. (2), where a is the

$$\frac{F}{M_i} = V_i(a - 1) \tag{2}$$

ratio of exhaust velocity to flight speed ($a = V_e/V_i$).

Multiplication of the net thrust from Eq. (1) by the flight speed provides a measure of the useful energy developed by the engine. Comparison of this energy with the energy which the engine has imparted to the airstream provides a measure of the propulsive efficiency η_p of the device, which is given by Eq. (3).

$$\eta_p = \frac{2}{1 + a} \tag{3}$$

Equations (2) and (3) embody the basic dilemma of propulsor design. Equation (2) indicates that, for positive thrust, the velocity ratio a must be greater than 1.0. The greater the exhaust velocity, the less will be the airflow required for a given thrust, which implies increasingly lighter and smaller propulsors. On the other hand, Eq. (3) indicates that the greatest propulsion efficiency is obtained with the lowest exhaust velocity. An exhaust velocity just equal to the flight speed ($a = 1$) gives 100% propulsive efficiency but requires an infinitely large propulsor. Different propulsion systems have therefore evolved with a balanced compromise between propulsive efficiency and size.

Bypass ratio. An important approach to achieving this compromise is encompassed in the concept of the bypass ratio, β, given by Eq. (4) where M_b is that

$$\beta = \frac{M_b}{M_p} \tag{4}$$

portion of the airflow through the propulsion system which bypasses the power producer, and M_p is that portion of this airflow which goes through the power producer.

Typical bypass ratios and operational flight speed ranges of various aircraft engines

Engine type	Typical bypass ratio	Operational flight speed (Mach number)
Ramjet	0	Higher than 2
Turbojet	0	0.8–2
Low-bypass turbofan	0.1–2	0.8–1.8
High-bypass turbofan	3–8	0.7–0.9
Ultrahigh-bypass turbofan	9–12	0.5–0.85
Propfan or Unducted Fan (UDF*)	25–40	0.4–0.8
Conventional turboprop (including propeller)	50–150	0.2–0.6
Helicopter turboshaft (including helicopter rotor)	500–1700	0–0.2

*UDF is a registered trademark.

For a power producer of given size, generating a given amount of energy, the size of the propulsor may be varied. The bypass ratio can thereby be varied to optimize the propulsive efficiency for the flight speed regime of the aircraft which the engine is to power. A wide range of bypass ratios are in use (see **table**). At one extreme, for very high supersonic flight speeds (Mach numbers much greater than 1.0), where high exhaust velocities are required, turbojets are used which bypass none of the inlet airflow (that is, the bypass ratio is 0). At the other extreme, for very low-speed aircraft such as helicopters, which fly at zero speed (hover) to 0.2 Mach number, the propulsor (the main rotor used for lift and propulsion) may pass several hundred times the airflow of the turboshaft engine which powers it. As the bypass ratio increases through this range, the airflow and engine diameter also increase, while the specific fuel consumption and specific thrust both decrease. *See* MACH NUMBER.

Integrated performance. Because of its inherent mechanical, thermodynamic, and aerodynamic limitations, an aircraft engine is generally limited to operation over a flight envelope, most often specified

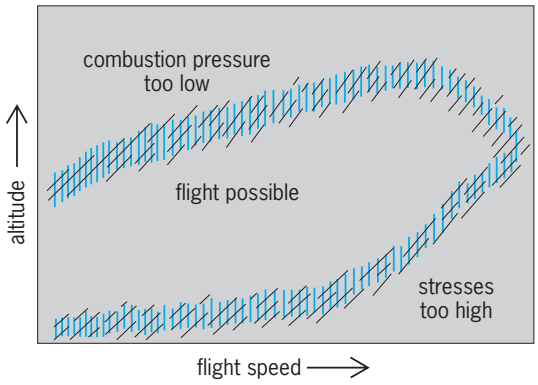


Fig. 6. Typical flight envelope for an aircraft engine. (Numerical values of quantities depend on particular engine.) (*After P. J. McMahon, Aircraft Propulsion, Barnes and Noble, 1971*)

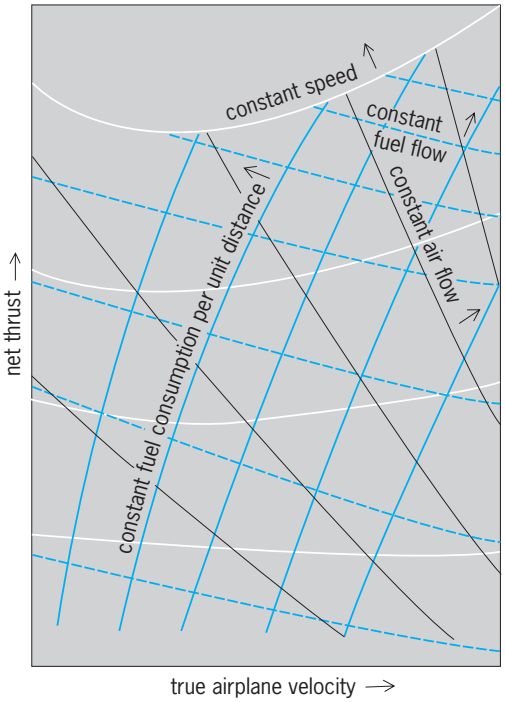


Fig. 7. Typical performance chart for a turbojet engine. (*After W. Kent, Mechanical Engineers' Handbook, vol. 4, 12th ed., Wiley, 1950*)

as the combinations of altitude and flight speed that are available for operation (**Fig. 6**). For a given engine type, of a given detailed design, it is then possible to assemble a comprehensive prediction of the engine's performance—its power or thrust and its fuel consumption—over the entire flight envelope by using a combination of data obtained from tests of the individual engine components; tests of the engine itself in a test cell which has the facility for simulating different flight speeds and ambient conditions; data from similar components or engines which may be scaled or otherwise modified to reflect the specific design being analyzed; and theoretical or analytical simulations of engine components. An excerpt from a typical display of the performance of a turbojet engine is shown in **Fig. 7**. Such performance summaries, more often in computerized format, are essentialelements in the design of aircraft, the prediction of the aircraft's performance, and the specification of the aircraft's flight plan to best accomplish its specified mission. *See* AIRCRAFT ENGINE; AIRCRAFT TESTING; GENERAL AVIATION.

Prospects. Future engines, particularly those intended for very high-speed supersonic or hypersonic flight, may be designed to use fuels with a higher ratio of hydrogen to carbon molecules, such as methane or pure hydrogen gas, to exploit their higher heating value and consequent lower required fuel weight. These gaseous fuels would be cryogenically liquefied for use as aircraft engine fuel. In engines for such high-speed aircraft, the ram pressure rise is so great that it is sufficient to provide an efficient thermodynamic cycle without being supplemented by a mechanical compressor. Such a compressorless jet engine is called a ramjet. Since they are ineffective at low flight speeds, ramjet-powered aircraft must be launched at high speed or require additional propulsion systems to take off and accelerate to high speed.

There has been considerable interest in applying the regenerator or recuperator to the gas turbine engine. This involves the use of a heat exchanger to transfer waste heat from the exhaust to the high-pressure air exiting the compressor, thereby reducing the fuel required to heat the air to the peak temperature required by the thermodynamic cycle.

Much experimental and development effort has been directed at variable-cycle, hybrid, and convertible engines, which contain mechanical devices for varying the bypass ratio and for converting turboshaft engines to turbofan engines, to power versatile airplanes which are required to perform complex missions that would otherwise require more than one type of engine.

In a completely separate class are human-powered aircraft, where the propeller or helicopter rotor is driven by bicycle pedals. *See* AIRCRAFT ENGINE; GENERAL AVIATION.

Fredric F. Ehrich

Bibliography. P. Hill and C. Peterson, *Mechanics and Thermodynamics of Propulsion*, 2d ed., 1992; J. L. Kerrebrock, *Aircraft Engines and Gas Turbines*, 2d ed., 1992; M. J. Kroes et al., *Aircraft Powerplants*, 7th ed., 1995; J. D. Mattingly, W. H. Heiser, and D. H. Daley, *Aircraft Engine Design*, 1987; G. C. Oates (ed.), *Aerothermodynamics of Gas Turbine and Rocket Propulsion*, 3d ed., 1997; Rolls Royce PLC, Technical Publications Department, *The Jet Engine*, 1986; I. E. Treager, *Aircraft Gas Turbine Engine Technology*, 3d ed., 1995.

Aircraft fuel

The source of energy required for the propulsion of airborne vehicles. This energy is released in the form of heat and expanding gases that are products of a combustion reaction that occurs when fuel combines with oxygen from ambient air. The exhaust gases are water vapor formed from hydrogen in the fuel, carbon dioxide formed from carbon in the fuel, traces of carbon monoxide and nitrogen oxides, and heated but uncombusted components of the intake air. Aircraft fuel is burned with ambient air and is thereby distinct from rocket propellants, which carry both fuel and oxidant. An important criterion for aircraft fuel is that its energy density, or heat of combustion per unit of weight, be high. This allows reasonable expenditures of fuel during takeoff, efficient performance in flight, and long range of flight duration. *See* PROPELLANT.

There are two general types of aircraft fuels in conventional use: gasolines for reciprocating (piston) engines, and kerosinelike fuels (called jet fuels) for turbine engines.

Piston engine fuels. Piston engine fuels, or aviation gasolines, are special blends of gasoline stocks and additives that produce a high-performance fuel that is graded by its antiknock quality. The gasoline blending stocks are virgin (uncracked) naphtha,

alkylate, and catalytically cracked gasoline. Naphthas are mixtures of hydrocarbons distilled directly from crude oil; alkylates are branched paraffin compounds synthesized by refining processes; and catalytically cracked gasolines contain ring compounds called aromatics (such as benzene). In general, the chemical composition of aviation gasoline can be approximated as $C_xH_{1.9x}$, where the number of carbon atoms x is between 4 and approximately 10. Tetraethyllead (Tel) is a common additive used in concentrations of up to 4 ml/gal (1.057 ml Tel/liter) of fuel to increase the antiknock quality of the fuel. *See* GASOLINE; NAPHTHA; PETROLEUM PROCESSING AND REFINING.

Antiknock quality. In reciprocating aircraft engines the fuel-to-air ratio can be varied from lean (for maximum economy) to rich (for maximum power). Fuel combustion is more likely to detonate—knock—under fuel-lean conditions. Knocking, if allowed to occur extensively, can harm an engine. Aviation gasolines, like automotive gasolines, are rated according to their antiknock quality as compared with a reference fuel (isooctane or isooctane plus specified amounts of Tel). *See* OCTANE NUMBER.

Aviation gasolines are rated by the (minimum observed) octane number for both lean and rich conditions. The American Society for Testing and Materials (ASTM) and the American National Standards Institute (ANSI) have specified standards and testing procedures for determining lean and rich knock values. For example, a fuel rated as 115/145 has a minimum octane rating of 115 when tested under fuel-lean conditions and 145 under fuel-rich conditions. Often, aviation gasolines are graded and designated by their lean octane number of 80 and a rich octane number of 87; grade 100 and grade 100LL have a lean octane number of 100 and a rich octane rating at least that of isooctane plus 1.28 ml Tel/gal (0.338 ml Tel/liter). Grade 80 may contain up to 0.5 ml Tel/gal (0.132 ml Tel/liter) and is dyed red. Grade 100 may contain up to 4 ml Tel/gal (1.057 ml Tel/liter) and is dyed yellow. Grade 100LL may contain up to 2 ml Tel/gal (0.528 ml Tel/liter) and is dyed blue.

Volatile. Aviation gasolines must be sufficiently volatile to evaporate quickly and blend with air in the engine manifolds and must be distributed evenly among all cylinders. They cannot be too volatile, however, or the fuel will boil in the tanks or lines. Gasolines boiling over a range of about 100°F (43°C) to 325°F (163°C) meet these requirements, and all grades of aviation gasolines have identical ANSI/ASTM distillation specifications. The tendency of a gasoline to boil is characterized by its Reid vapor pressure (RVP), which is approximately the absolute pressure that the gasoline will exert at 100°F (37.778°C). The RVP is between 5.5 and 7.0 lb/in.^{2} (38 and 48 kilopascals) for all grades of aviation gasoline. Aviation gasolines must also have low freezing points to be stable in storage; the ANSI/ASTM specification is −72°F (−58°C).

Heat of combustion. The heat of combustion of all grades of aviation gasoline is about 18,700 Btu/lb (43.5 megajoules/kg). This is the net or low heating value at which all combustion products are gaseous. A gallon of aviation gasoline weighs about 6.1 lb, and thus has an energy content of about 114,000 Btu (1 liter weighs about 0.73 kg and has an energy content of about 31.8 MJ).

Turbine engine fuels. Turbine engine fuels are distillate hydrocarbon fuels, like kerosines, used to operate turbojet, turbofan, and turboshaft engines. While all piston engine fuels have the same volatility but differ in combustion characteristics, jet fuels differ primarily in volatility; differences in their combustion qualities are minor. The volatility characteristics of several grades of jet fuel are shown in **Table 1**. For fuels in which the RVP is too low for accurate measurement, the flash point is given. This is the temperature to which a fuel must be heated to generate sufficient vapor to form a flammable mixture in air. The characteristics listed for Jet A and Jet B are the 1978 ANSI/ASTM standard specifications. *See* KEROSINE.

Fuel JP-1, which is no longer used, was the kerosine first used by the military and is substantially the same as Jet A or Jet A-1, which is now the most widely used commercial jet fuel. JP-4 and Jet B are military and commercial fuels, respectively, with nearly the same specifications. They are sufficiently volatile that explosive mixtures are present at most ground storage conditions and many flight conditions. JP-4 is used by the Air Force in subsonic aircraft, but Jet B has seen little commercial use in the United States. JP-5, the least volatile of the turbine fuels, is the Navy service fuel. JP-6 is the Air Force fuel for supersonic aircraft. *See* JET FUEL.

Composition. Production of distillate turbine fuel uses up to about 5% of crude oil input to a refinery. This percentage could be increased at added incremental costs and with a concurrent reduction in the

TABLE 1. Volatility characteristics of jet fuels

Jet fuel grade	Distillation range, °F (°C)	RVP, lb/in.^{2}, absolute (kPa, absolute)	Flash point, °F (°C)
JP-1	325–450 (163–230)	—	120 (49)
JP-3	100–500 (38–260)	6 (41)	—
JP-4	150–500 (65–260)	2.5 (17)	—
JP-5	350–500 (177–260)	—	150 (65)
JP-6	300–500 (149–260)	—	100 (38)
Jet A	—	—	100 (38)
Jet B	—	—	100

TABLE 2. Comparison of cryogenic fuels and Jet A fuel

Fuel	Chemical composition	Heat of combustion, Btu/lb (MJ/kg)	Boiling point, °F (°C)	Density, lb/ft^3 (kg/liter)	Specific heat capacity, Btu/lb-°F(kJ/kg-°C)
Jet A	$CH_{1.9}$	18,400 (42.8)	572 (300)*	51 (0.82)	0.47 (1.97)
Liquid hydrogen	H_2	51,600 (120.0)	−423 (−253)	4.4† (0.070)	2.3 (9.6)
Liquid methane	CH_4	21,500 (50.0)	−258 (−161)	26* (0.12)	0.84 (3.52)

*Final maximum boiling point.
†At normal boiling point.

output of motor gasoline and diesel fuel. Turbine fuel contains aromatic hydrocarbons; limits are placed on this content owing to concerns about smoke and coke formation. For military jet fuels the limit on aromatics is 25% by volume, and for commercial fuel the limit is 20% (except by mutual agreement between supplier and purchaser, in which case the content may not exceed 25% for Jet A or 22% for Jet A-1 or Jet B). Smoke can be an atmospheric pollutant, but its formation does not represent an appreciable loss in combustion efficiency. Coke is a carbonaceous deposit that adheres to the internal parts of the combustor and can reduce engine life.

High-temperature stability. An important requirement is to provide a fuel that is stable at relatively high temperatures. In subsonic jets the fuel is used to cool the engine lubricant, and the temperature of the fuel can be raised by about 200°F (110°C). In supersonic jets the fuel is used as a heat sink for the engine lubricant, for cabin air conditioning, and for cooling the hydraulic systems. For very high speed flight, the fuel may be used to cool additional engine components and critical air frame areas, such as the leading edges of wings. Therefore, depending on flight speeds and aircraft design, turbine fuels can be heated from 300°F (150°C) to 500°F (260°C) before they are burned. When they are heated to this degree, small amounts of solids may form, and foul the heat exchangers and clog the filters and fuel injectors. There are specifications to indicate the temperature at which solids are first formed and the amount of solids formed with time. In a specification test, fuel is preheated and passed through a heated filter for 5 h. No significant amount of solids may form in the preheater, and the pressure drop across the filter must stay within limits. For JP-5 and Jet A, A-1, and B, the test temperatures are 300°F (148.9°C) for the preheater and 400°F (204.4°C) for the filter. For JP-6, these temperatures are 425°F (218.3°C) and 525°F (273.9°C), respectively.

Freezing point. Turbine fuels must have low freezing points: −40°F (−40°C) for Jet A and −58°F (−50°C) for Jet A-1 and Jet B. There is also a limit on sulfur content: 0.3% by weight.

Heat of combustion. The heat of combustion of all jet fuels is about 18,400 Btu/lb (42.8 MJ/kg). This is the net low heating value. A gallon of turbine fuel weighs about 6.7 lb and thus has an energy content of about 123,000 Btu (1 liter weighs about 0.80 kg and has an energy content of about 34.4 MJ).

Alternative fuels. Alternative fuels, made from coal, oil shale, or solar or nuclear energy plus a suitable raw material, have been under consideration by the National Aeronautics and Space Administration (NASA) and several aircraft manufacturers.

Liquid hydrogen and liquid methane. Two alternative fuels are cryogenic liquid hydrogen and liquid methane. **Table 2** compares some pertinent properties of these fuels with those of the conventional Jet A fuel. As indicated in the table, liquid hydrogen has an energy density (heat of combustion) 2.8 times that of Jet A. Its volume is relatively large, and for the same energy content, the volume of liquid hydrogen would be four times that of Jet A. This, however, is of less importance than the high energy density, which is the predominant benefit of liquid hydrogen. Methane, with properties intermediate between those of Jet A and liquid hydrogen, could be more attractive than hydrogen on the basis of cost.

Hydrogen can be produced from water through electrolysis, with commercially available electrolyzers. For such production to make sense in terms of energy utilization, the electric power for operating the electrolyzers should be generated with a nonpetroleum energy source like falling water or perhaps nuclear heat. Both hydrogen and methane can be made from coal. *See* COAL GASIFICATION; HYDROGEN; METHANE.

Liquid hydrocarbons from oil shale or coal. Another option is to produce a synthetic crude or mixture of liquid hydrocarbons from oil shale or coal. When followed by refining steps, these liquefaction processes might be used to produce a synthetic turbine fuel similar to Jet A, as well as other fuels and chemicals. *See* AIRCRAFT PROPULSION; COAL LIQUEFACTION; JET PROPULSION; METAL-BASE FUEL; OIL SHALE; PROPULSION.

John B. Pangborn

Aircraft icing

Aircraft icing encompasses a range of conditions during which frozen precipitation forms on an aircraft. It is usually separated into two broad classifications, ground icing and in-flight icing. Icing can compromise flight safety by affecting the performance, stability, and control of the aircraft, and as a result the ability of the pilot to maintain the desired flight path. The primary effect of icing is its adverse impact on the aerodynamics of the airplane. Ice accretion results in

increased drag and reduced maximum lift which reduce the performance and safety of the flight. While ice does add weight to the aircraft, the amount is usually a very small percentage of the aircraft gross weight, and its effects are insignificant compared to the aerodynamic effects.

Ice accretion. Ground icing occurs when the aircraft is on the ground, and becomes significant when it affects the aircraft's ability to take off. This form of icing occurs when ice, snow, or freezing rain collects on the upper surfaces of the aircraft or when frost forms on the aircraft.

The formation of in-flight icing is much more complex. In-flight icing forms when an aircraft flies through a cloud of supercooled precipitation. Icing clouds usually contain water droplets with diameters of 2–50 micrometers and have concentrations, or liquid water content, of up to 2.5 grams of water per cubic meter of air. Water droplets approach the aircraft, approximately following the air streamlines. Near the surface of the aircraft, large changes in velocity exist. The droplets, because of their inertia, cannot change velocity rapidly enough to follow the air around the aircraft, and strike or impinge on the aircraft surface. Because of the role played by the aerodynamic velocity gradients and the droplet size in this process, more droplets impinge when they are large and the forward-facing aircraft surface has a small leading-edge radius. Ice forms on the leading or forward-facing edges of the wings, tail, antennas, windshield, radome, engine inlet, and so forth. *See* CLOUD PHYSICS.

Once the droplets have impinged, the freezing process determines the actual size and shape of the ice accretion. The two most common types of ice accretions are rime ice (**Fig. 1***a*) and glaze ice (Fig. 1*b*). Rime ice occurs at low temperature, low liquid water content, and low flight velocity so that the droplets freeze on impact. As a result, rime ice grows into the direction of the incoming droplets and forms an opaque, streamlined shape. Glaze ice occurs at temperatures near freezing with high liquid water content and high flight velocity. In this type of accretion, the impinging water droplets do not freeze on impact with the surface. Liquid water forms as a film or as hemispherical beads on the surface and eventually freezes, resulting in ice that is clear in color. The formation of beads, and sometimes larger ice roughness due to water flowing on the surface, augments the droplet impingement and accretion process, causing the large ice protrusions referred to as horns (Fig. 1*b*). Mixed ice accretions contain features of both rime and glaze ice, with glaze accretion occurring near the leading-edge stagnation point (where the air is stationary with respect to the wing), and rime ice accretion occurring farther back on the airfoil.

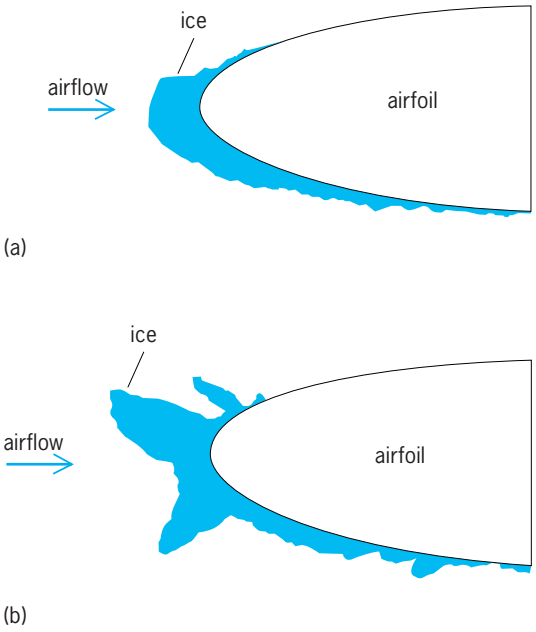


Fig. 1. Ice accretion on an airfoil leading edge. (*a*) Rime ice. (*b*) Glaze ice. *(NASA)*

Droplets much larger than 50 μm may exist in some meteorological situations, including freezing drizzle and freezing rain. These larger drops are referred to as supercooled large droplets. Ice formations from these droplets can extend very far back on the wing or other aircraft components and can pose serious safety hazards. *See* AERONAUTICAL METEOROLOGY.

Aerodynamics. Icing affects the propulsion system by reducing the thrust of the engine and thus the ability of the aircraft to climb and maintain speed or altitude. This thrust reduction is usually due to the restriction of airflow into the engine by ice formation in the carburetor of piston-engine-powered light aircraft or ice formation on the inlet of larger jet aircraft. On a propeller-driven aircraft, ice formation on the propeller can also result in reduced thrust.

Probably the most dangerous way that ice acts on an aircraft is through its effect on the aerodynamics, which results in degraded performance and control. Small amounts of ice or frost add roughness to the airplane surfaces. The roughness increases the friction of the air over the surface; this is called skin friction. Higher-than-usual levels of skin friction cause additional drag and also affect the lift (**Fig. 2**). The effect of small ice accretions or frost on lift is a reduction in the maximum lift coefficient and a reduction in the airplane angle of attack at which it is reached. Higher drag is also seen, which reduces the ability of the aircraft to climb and reduces its maximum speed. Reduced maximum lift coefficient increases the stall speed of the aircraft and thus increases the takeoff distance. Safety is compromised if ice is present and landing and takeoff speeds are not increased, since the margin between flight speed and stall speed is reduced. *See* AERODYNAMIC FORCE.

Large accretions can drastically alter the shape of the wing (Fig. 1). Then, in addition to skin friction, flow separation results in a further reduction in aerodynamic performance of the aircraft. For larger ice accretions, there is a larger reduction in maximum lift coefficient and a larger increase in drag coefficient (Fig. 2). Large ice accretions may also reduce the lift coefficient at angles of attack below stall.

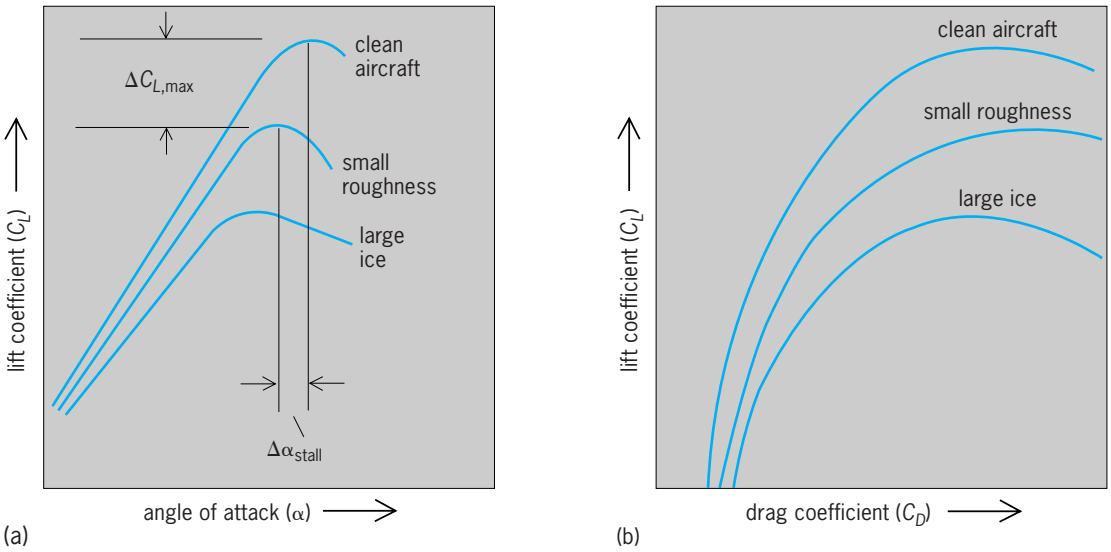


Fig. 2. Effect of ice on lift and drag coefficients. (*a*) Lift coefficient (C_L) versus angle of attack (α). $\Delta C_{L,max}$ is the reduction in maximum lift coefficient. $\Delta\alpha_{stall}$ is the reduction in the angle of attack at which stall occurs. Both are shown for small roughness case only. (*b*) Lift coefficient versus drag coefficient (C_D). (*After Effects of Adverse Weather on Aerodynamics, Proceedings of the AGARD Fluid Dynamics Specialists Meeting, AGARD-CP-496, 1991*)

Aircraft control can be seriously affected by ice accretion. Ice accretion on the tail can lead to reduced elevator effectiveness, reducing the longitudinal control (nose up and down) of the aircraft. In some situations, the tail can stall or lose lift prematurely, resulting in the aircraft pitching nose down. Similarly, ice on the wing ahead of the aileron can result in roll upset. Ice is often asymmetric and affects one wing more than another. This leads to unexpected roll, which can be difficult to control. For aircraft with unpowered controls, the pilot directly opposes the hinge moment produced by aerodynamic forces through the force applied to the control wheel. The ice-induced flow separation causes a redistribution of the air pressure on the wing or tail surface, which results in a large change in hinge moment and, therefore, in the control force required to maintain the desired flight condition. Both tail stall and roll upset are thought to be the cause of recent aircraft icing accidents. *See* AERODYNAMICS; FLIGHT CONTROLS.

Ice protection. All large aircraft and many light aircraft are equipped with in-flight ice protection systems to reduce the effect of ice. Ice protection systems are classified as de-ice or anti-ice systems. De-ice systems allow some ice to accrete, and then they periodically remove the ice. The most common system is the pneumatic boot, in which inflatable tubes are placed on the wing leading edge or other surfaces. The inflating boot expands, breaking the bond between the ice and the surface. The ice fractures into small pieces, which are carried downstream by aerodynamic forces and away from the aircraft. Anti-ice systems prevent ice from forming either by heating the surface above 0°C (32°F) or through the use of freezing-point depressants. The most common systems are electrothermal, used to protect small surfaces, and hot-air systems, used on most jet aircraft where engine bleed air is available. Wing and tail leading edges, engine inlet lips, propellers, air data system components, and windshields are typically ice protected.

Ground icing is usually dealt with by ground-based de-icing systems. Freezing-point depressant fluids are applied to the upper surface of the aircraft to remove ice and frost, and a coating remains to prevent the formation of additional ice. The fluid is effective for some period of time after application, referred to as holdover time, which depends on the type of fluid, atmospheric temperature, and rate of precipitation. The fluid is designed to flow off the aircraft wing due to aerodynamic shear during takeoff so that the fluid itself does not act as surface roughness and reduce the performance and safety of the flight. *See* ANTIFREEZE MIXTURE.

Supercooled large droplets. Until very recently, aircraft ice protection systems were tested and certified only for icing cloud droplet sizes between 2 and 50 μm. The area of the aircraft to be protected was set by the impingement location of these drops. However, during the investigation of a commuter aircraft accident which occurred in 1994, the probable cause was determined to be the resulting aerodynamic effect of ice accretion due to supercooled large droplets, which have diameters from 50 to 500 μm. Supercooled large droplets impinge much farther back from the leading edge and can cause ice to accrete aft of the ice protection systems. Research is under way to characterize the atmosphere, improve computer models, understand aerodynamic effects, and improve ice protection systems for the supercooled-large-droplet icing environment.

Michael B. Bragg

Bibliography. M. B. Bragg et al., Effect of underwing frost on a transport aircraft airfoil at flight Reynolds number, *J. Aircraft*, 31:1372–1379, 1994; A. Heinrich et al., *Aircraft Icing Handbook*, DOT/FAA/CT-88/8-1, 1991; A. Khodadoust and M. B.

Bragg, Aerodynamics of a finite wing with simulated ice, *J. Aircraft*, 32:137–144, 1995; *Proceedings of the FAA International Conference on Aircraft Inflight Icing*, DOT/FAA/AR-96/81, vol. 2, 1996.

Aircraft instrumentation

A coordinated group of instruments that provides the flight crew with information about the aircraft and its subsystems. Together with the controls, aircraft instruments form the human–machine interfaces that enable the flight crew to operate the aircraft in accordance with the flight plan. These instruments provide flight data, navigation, power plant performance, and aircraft auxiliary equipment operating information to the flight crew, air-traffic controllers, and maintenance personnel. While not considered as instrumentation, communication equipment is, however, directly concerned with the instrumentation and overall indirect control of the aircraft.

Situation information on the operating environment, such as weather reports and traffic advisories, has become a necessity for effective flight planning and decision making. The prolific growth and multiplicity of instruments in the modern cockpit and the growing need for knowledge about the aircraft's situation are leading to the introduction of computers and advanced electronic displays as a means for the pilot to better organize and assimilate this body of information.

Types of Instruments

Instrumentation complexity and accuracy are dictated by the aircraft's performance capabilities and the conditions under which it is intended to operate. Light aircraft may carry only a minimum set of instruments: an airspeed indicator, an altimeter, an engine tachometer and oil pressure gage, a fuel quantity indicator, and a magnetic compass. These instruments allow operation by a pilotage technique, that is, operation where weather conditions and visibility permit visual reference to a horizon for attitude control and navigation by topographic observation in relation to a map or the pilot's knowledge of local terrain. Operation at low altitude and under visual flight rules (VFR) allows use of very basic instruments that sense a parameter and convert that sensory information through direct mechanical means to a desired instrument reading (for example, static air pressure is converted to an altitude reading). *See* AIRSPEED INDICATOR; ALTIMETER; PILOTAGE; RATE-OF-CLIMB INDICATOR.

Operation under low visibility and under instrument flight rules (IFR) requires this same information for a more precise form and also requires attitude and navigation data. An attitude-director indicator (ADI) presents an artificial horizon, bank angle, and turn-coordination data for attitude control without external visual reference. The attitude-director indicator may contain a vertical gyro within the indicator, or a gyro may be remotely located as a part of a flight director or navigational system.

Flying through a large speed range at a variety of altitudes is simplified if the indicated airspeed is corrected to true airspeed for navigation purposes and the Mach number (M) is also shown on the ADI for flight control and performance purposes. Mach number is the ratio of the aircraft speed to the speed of sound under the existing temperature and pressure conditions. Rate-of-climb is provided by an instantaneous vertical-speed indicator (IVSI). Heading data are provided by a directional gyro or data derived from an inertial reference system. The basic flight instruments are usually located directly in front of the pilots in the form of a T. The type of information and the arrangement of the indicators on a "full panel" have evolved over the years as the airplane has increased in performance capability and complexity. *See* GYROSCOPE; INERTIAL GUIDANCE SYSTEM.

Navigation instruments. Navigation instruments primarily relate to the position of the aircraft with respect to specific locations on the Earth. Navigational aids include very high-frequency omnidirectional radio ranges (VOR) that transmit azimuth information for navigation at specified Earth locations; distance-measuring equipment (DME) that indicates the distance to radio aids on or near airports or to VORs; automatic direction finders (ADF) that give the bearing of other radio stations (generally low-frequency); low-range radio altimeters (LRRA) which by radar determine the height of the aircraft above the terrain at low altitudes; and instrument landing systems (ILS) that show vertical and lateral deviation from a radio-generated glide-path signal for landing at appropriately equipped runways. Some inertial navigation systems include special-purpose computers that provide precise Earth latitude and longitude, ground speed, course, and heading. Integration of this equipment with a coupled automatic pilot allows automatic controlled flight and landings under low visibility conditions. The pilot may eventually control the airplane manually with augmented control and advanced displays, even through landing in fog conditions. Today, however, the pilot must see outside to land manually or to monitor the automatic landing. *See* AIR NAVIGATION; AIR-TRAFFIC CONTROL; AUTOPILOT; DIRECTION-FINDING EQUIPMENT; DISTANCE-MEASURING EQUIPMENT; ELECTRONIC NAVIGATION SYSTEMS; INSTRUMENT LANDING SYSTEM (ILS); MICROWAVE LANDING SYSTEM (MLS); VOR (VHF OMNIDIRECTIONAL RANGE).

Engine instruments. Engines require specific instruments to indicate limits and efficiency of operation. For reciprocating engines, instruments may display intake and exhaust manifold pressures, cylinder head and oil temperatures, oil pressure, and engine speed. For jet engines, instruments display engine pressure ratio (EPR), exhaust gas temperature (EGT), engine rotor speed, oil temperature and pressure, and fuel flow. Vibration monitors on both types of engines indicate unbalance and potential trouble. Engine instruments tend to be clustered together. In large aircraft they are usually located in the center front instrument panel between, and visible to, both pilots. Special-purpose computers may also be

Fig. 1. Cockpit displays and flight engineer's instruments in Boeing 747. (*Boeing Co.*)

used to compute and indicate engine performance limits for the existing environmental conditions and aircraft flight mode. *See* JET PROPULSION; RECIPROCATING AIRCRAFT ENGINE; TACHOMETER.

Auxiliary instruments. Depending on the complexity of the aircraft and the facilities that are provided, there is also an assortment of instruments and controls for the auxiliary systems. Pressurized cabins require cabin altitude, cabin differential pressure, cabin rate-of-climb or descent, and cabin temperature. Secondary power system instruments and controls are provided for electrical, hydraulic, and pneumatic power and distribution systems. Fuel flow rates, fuel remaining in specific tanks, and total fuel remaining are also shown on other indicators. Many of these system instruments and control switches are strategically located on a schematic diagram on the instrument panel to provide an overview of the system configuration and operational status. Engine instruments are also included when auxiliary ground power or in-flight emergency power is provided by an auxiliary power unit (APU). Indicators show the positions of functional equipment such as landing gear, flaps, ailerons, spoilers, elevator, rudder, and trim devices. Other indicators show aircraft condition; these include fire detection and warning, acceleration monitors, overspeed deviation devices, stall warning indicators, approach-speed deviation indicators, and configuration and altitude deviation devices. These devices may provide indication through round dial or vertical tape indicators, warning lights, or audible signals. Modern computerized ground proximity warning systems (GPWS) provide warning of altitudes too close to the terrain under different flight regimes by computer-synthesized voice signals.

Large commercial transport aircraft also have flight recorders, cockpit voice recorders, and maintenance recorders. The more modern aircraft also include special built-in test equipment (BITE) that when activated on the ground conducts a prescheduled test and evaluates the response of a system or component in order to isolate problems to a specific item that is replaceable as a line replaceable unit (LRU). **Figure 1** shows the flight deck of a Boeing 747 jet transport. This flight deck is typical of those commercial jet transports built in the 1960s which used two pilots and a flight engineer.

Applications of Electronic Technology

Electronic technology developments during the 1960s and 1970s included ring laser gyros, strapdown inertial reference systems, microprocessor digital computers, color cathode-ray tubes (CRT), liquid crystal displays (LCD), light-emitting diodes (LED), and digital data buses. Application of this technology allowed a new era of system integration and situation information on the aircraft flight deck and instrument panels. Commercial jet transports developed during the 1980s use digital electronics to improve safety, performance, economics, and passenger service. The concept of an integrated flight management system (FMS) includes automatic flight control, electronic flight instrument displays, communications, navigation, guidance, performance management, and crew alerting to satisfy the

Fig. 2. Flight deck mock-up of Boeing 767 configuration for crew of three. (*Boeing Co.*)

requirements of the current and future air-traffic and energy-intensive environment. **Figure 2** shows the flight deck mock-up for a crew of three in a Boeing 767 jet transport. The following discussion is representative of the concepts used on this generation of aircraft instrumentation and controls. *See* ELECTRONICS; INTEGRATED CIRCUITS.

Flight management system. To be effective in the role of flight manager, the pilot must have ready access to relevant flight information and suitable means to accomplish aircraft control within reasonable work-load bounds. The extensive data-processing capabilities and integrated design of a flight management system provide the pilot with access to pertinent information and a range of control options for all flight phases. The elements of an integrated flight management system are shown in **Fig. 3**.

The avionics may be subdivided into three groups: sensors, computer subsystems, and cockpit controls/displays. The cockpit controls operate the sensors and computer subsystems, and the displays are supplied with raw and processed data from them. The cockpit-mounted elements of the system include caution-advisory display, attitude-director indicator (ADI), horizontal-situation indicator [HSI, including map and weather radar (WXR)], and controls for the sensors and computer subsystems.

Especially notable among the avionics systems for innovation and special benefits are the inertial reference system (IRS), the flight management computers (FMC), the electronic flight instrument system

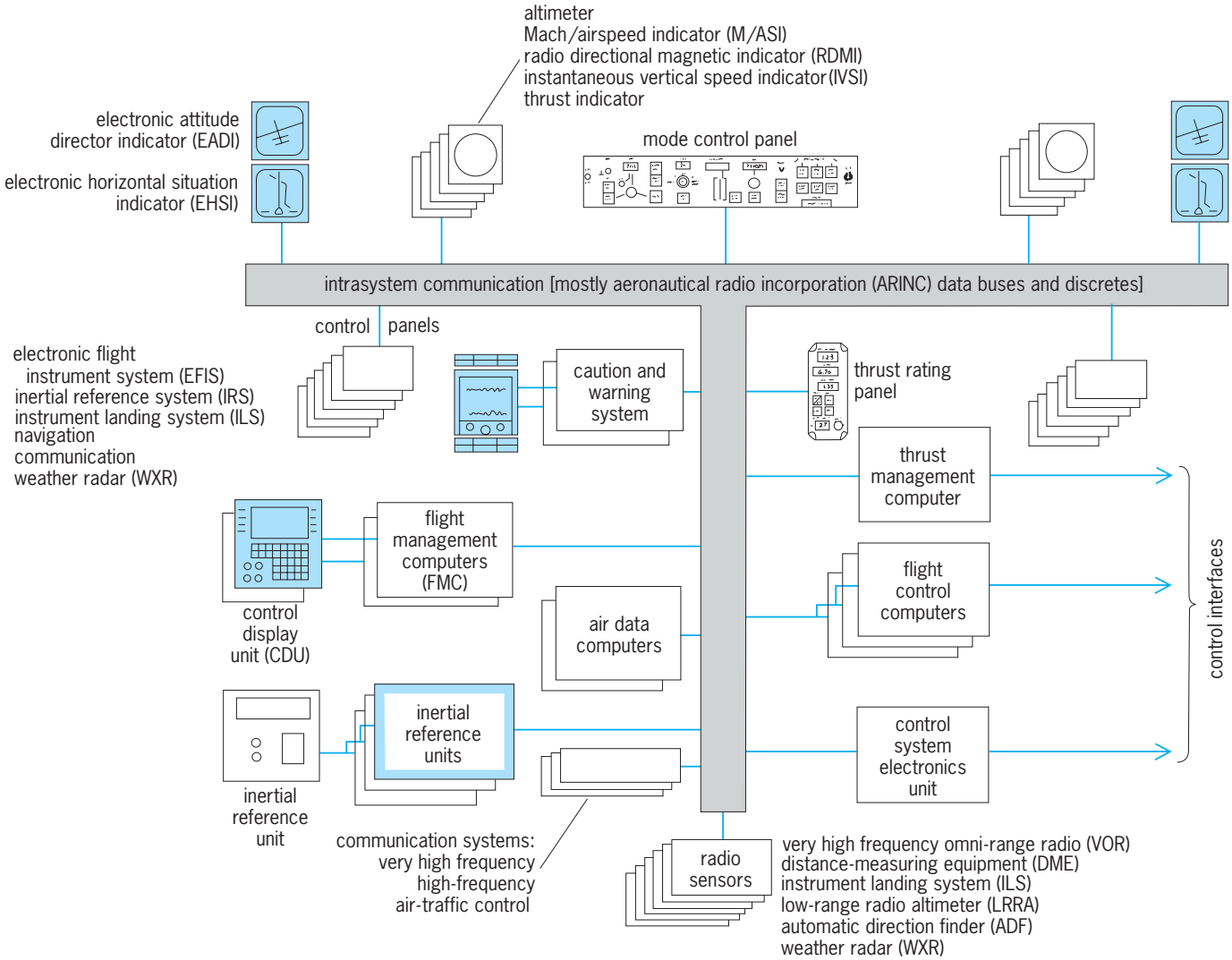


Fig. 3. Block diagram of Boeing 757/767 flight management system. (*Boeing Co.*)

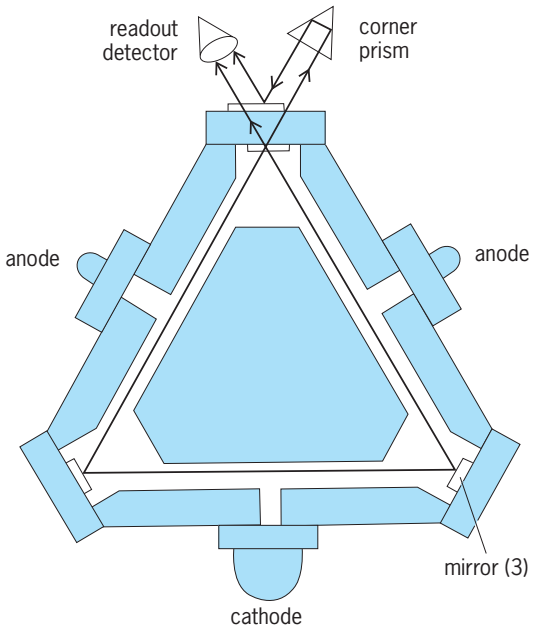


Fig. 4. Ring laser gyro operation. (*Honeywell, Inc.*)

(EFIS), and the caution and warning system.

Inertial reference system. Inertial navigation systems have benefited from two technologies developed in the 1970s: strap-down inertial techniques and the ring laser gyro. Strap-down inertial techniques eliminate the costly and bulky gimbaled stable platform previously used in high-accuracy inertial navigation systems. The laser gyro is unconventional since it does not have a spinning wheel. It detects and measures angular rates by measuring the frequency difference between two contrarotating laser beams. **Figure 4** shows how the two laser beams circulate in a triangular cavity simultaneously.

Flight management computers. Flight management computers integrate the functions of navigation, guidance, and performance management. Accurate path guidance is dependent upon display of the aircraft's current position and velocity to the crew. Position and velocity are determined by combining data from the inertial reference system with range and bearing from VOR/DME stations. The appropriate VOR/DME stations are normally automatically tuned as the aircraft progresses through its flight, but can be manually tuned at the pilot's option.

The control-display unit (CDU) is the interface between the pilot and the flight management computer. It provides the means for manually inserting system control parameters and selecting modes of operation. The control-display unit incorporates an alphanumeric keyboard for data entry and dedicated mode keys that select navigation and performance functions. In addition, it provides computer readout capability and enables pilot verification of data entered into storage. Flight plan, performance, and advisory data are continuously available for display on the control-display unit. The data are displayed on a cathode-ray tube, and control of the unit is provided by a microprocessor.

Electronic flight instrument system. Effective flight management is closely tied to providing accurate and timely information to the pilot. The nature of the pilot's various tasks determines the general types of data which must be available. The key is to provide these data in a form best suited for use. If the pilot is not required to accomplish extensive mental processing before information can be used, then more information can be presented and less effort, fewer errors, and lower training requirements can be expected. Computer-generated displays offer significant advances in this direction.

Color cathode-ray-tube displays have replaced the conventional attitude-director indicator and horizontal-situation indicator. The high-resolution, sunlight-readable, electronic displays selectively display more information with less clutter than is possible with electromechanical instruments. *See* CATHODE-RAY TUBE; ELECTRONIC DISPLAY.

An electronic attitude-director indicator (EADI; **Fig. 5**) provides a multicolor cathode-ray-tube display of information such as that found on previous attitude-director indicators. This gives attitude information showing the airplane's position in relation to the instrument landing system or a very high-frequency omnirange station. In addition, the EADI indicates the mode in which the automatic flight control system is operating and presents the readout from the radio altimeter. Ground speed is displayed digitally at all times near the airspeed indicator.

The electronic horizontal-situation indicator (EHSI; **Fig. 6**) provides an integrated multicolor map display of the airplane's position, plus a color weather radar display. The scale for the radar and map can be selected by the pilots but is always the same for both. Wind direction and velocity for the airplane's present position and altitude, provided by the inertial reference system, are shown at all times. Both the horizontal situation of the airplane and its deviation from the planned vertical path are also provided, thus making it a multidimensional situation indicator.

The EHSI operates in three primary modes—as a map display, a full compass display, and a VOR mode

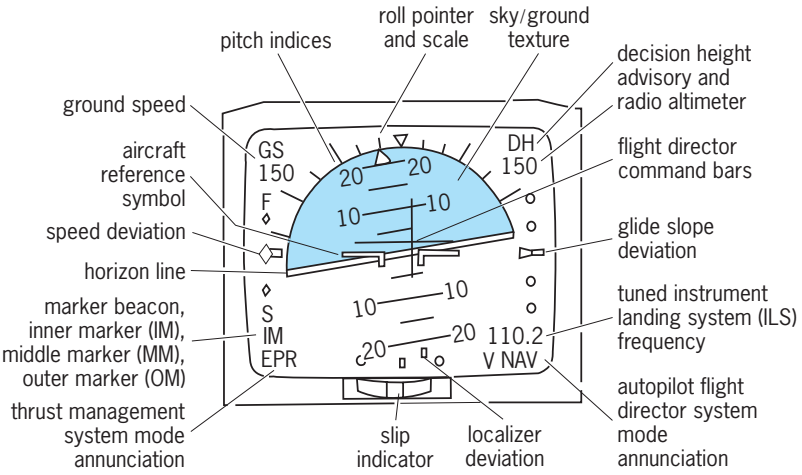


Fig. 5. Electronic attitude-director indicator (EADI). (*Boeing/Collins*)

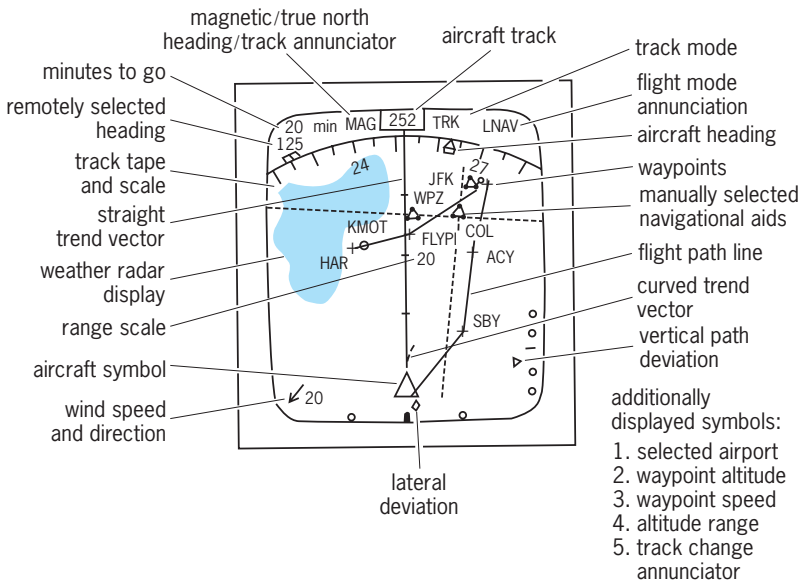


Fig. 6. Electronic horizontal-situation indicator (EHSI) map mode. (*Boeing/Collins*)

that displays a full or partial compass rose. The map displays are configured to present basic flight-plan data, including such parameters as the route of flight, planned way points, departure or arrival runways, and tuned navigational aids. Additional data such as navigation stations, airports, way-point altitude and speed targets, intersections, and weather radar returns can be called up by the pilot as the need arises, using any of several display-data push buttons.

A powerful display tool is the presentation of predictive information in the context of the basic display. The map display features two such predictions. One combines current ground speed and lateral acceleration into a prediction of the path over the ground to be followed by the airplane over the next 30, 60, and 90 s. The second prediction, an altitude range arc used for climb or descent, shows where the airplane will be when the target altitude is reached. This feature allows the pilot to quickly assess whether or not a target altitude will be reached before a particular location over the ground.

The capabilities of cathode-ray-tube displays for application to future aircraft generations include computer-generated runway imagery on the EADI, combined with other information, which may permit manually controlled landings without seeing outside at all. Also, other pertinent traffic may be portrayed on the EHSI which might allow the pilot to assure safe separation from other aircraft, control the spacing in trailing and merging situations, and avoid conflicts by early corrective action. *See* COMPUTER GRAPHICS.

Caution and warning system. The essential display elements of a typical alerting system for airplanes with a crew of three are portrayed in **Fig. 7**. The alert message display is a cathode-ray tube with multicolor capability located on the pilot's forward main engine instrument panel. Two colors are used: warnings (emergency operational or aircraft system conditions that require immediate corrective or compensatory action by the crew) are presented in red alphanumerics; cautions (conditions that require immediate crew awareness and eventual corrective or compensatory action) and advisories are presented with amber alphanumerics. The advisories (conditions that require crew awareness and may require action) are indented to distinguish them from cautions. Cancel and recall switches enable the pilot to control the density of information on the alert message display and redisplay all messages that are still valid.

The discrete warning display is a matrix of dedicated red lights that serves as a backup to the cathode-ray-tube display and provides a display of warnings when only standby electric power is available. The discrete caution display consists of

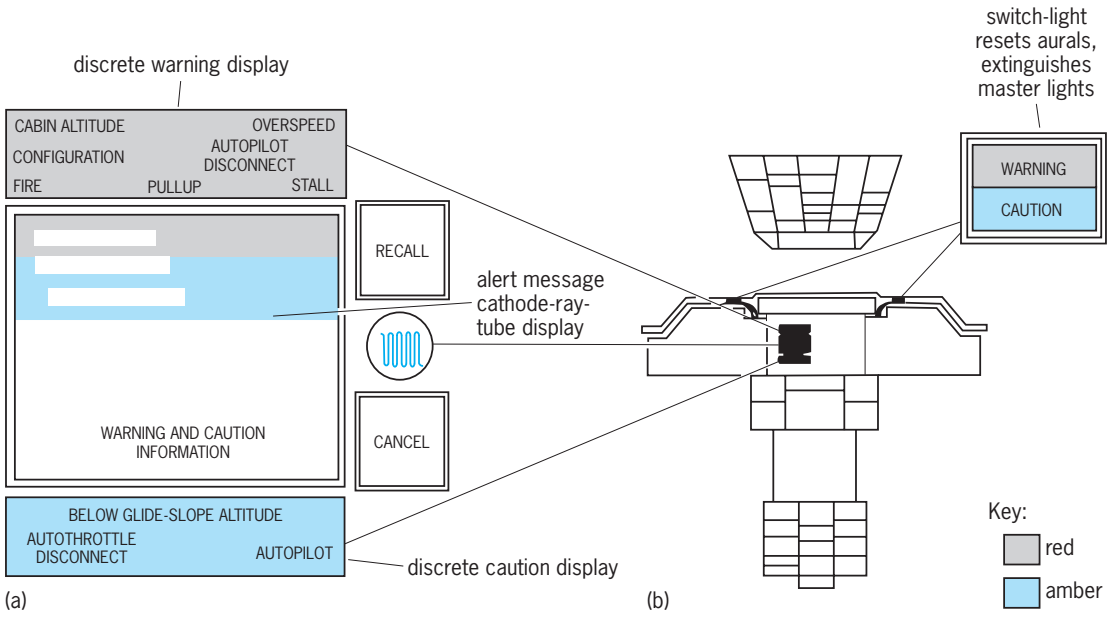


Fig. 7. Caution and warning display. (*a*) Configuration of elements. (*b*) Location of elements on aircraft instrument panel. (*Boeing Co.*)

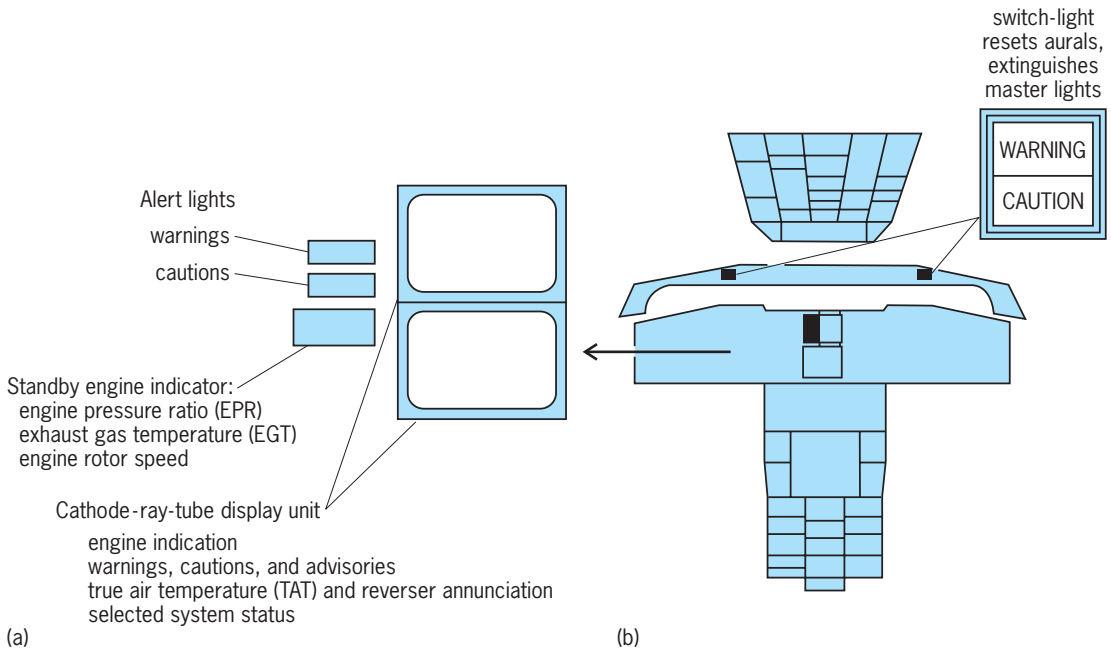


Fig. 8. Two-crew engine-indication and crew-alerting system (EICAS). (*a*) Configuration of the elements. (*b*) Location of the elements as they occur on an aircraft instrument panel. (*Boeing Co.*)

miscellaneous, not duplicated, amber lights normally located on the pilot's front panel. Master warning/master caution push-button lights illuminate whenever any warning or caution situation occurs. Both lights can be extinguished by pushing the light. Pushing the master warning light also cancels certain warning-related aural alerts.

Engine indication. An engine-indication and crew-alerting system (EICAS) is a careful consolidation of many formerly dedicated displays of propulsion parameters and subsystem indications, plus the full caution and subsystem indications, plus the full caution and warning system described above (**Fig. 8**). It provides additional automation for propulsion control and subsystem monitoring on airplanes with a crew of two.

The EICAS consolidates information on centrally located cathode-ray-tube and other displays and automatically monitors engine and subsystem parameters. During normal flight, the crew can perform any required operation with only two or three primary thrust parameters on display. Should a problem develop, the out-of-tolerance parameter is automatically called up on the display, together with other related parameters, to assist the crew in assessing the problem. A high exhaust-gas temperature, for example, will call up the closely related fuel flow indication. Any problem with engine oil pressure, quantity, or temperature will call all three of these parameters to the display. The particular item which is out of tolerance is identified because its normally white pointers and numerals are shown in amber or red, depending on the extent of the problem. These automatically displayed parameters cannot be turned off by the crew until the problem is alleviated. At any time, however, the crew can call up all engine indications via the EICAS display select panel.

A portion of one cathode-ray tube is reserved for caution and warning indications. This area is not "time-shared" by any other function; thus the appearance of any information in this area is an immediate alert to the crew.

The EICAS also features status and maintenance formats. The status format provides information concerning the airplane's status for dispatch, such as hydraulic quantity and control surface positions, and a listing of any equipment failures which could affect dispatch or routing of the flight. The maintenance format contains those parameters of use only to the maintenance crew and is not available in flight. It lists all equipment failures regardless of whether or not they affect dispatch. Bernard C. Hainline

Military Aircraft

Two types of military instrumentation requirements exist, one essentially for the instrumentation described above and one for special mission needs. Many display and control requirements for military aircraft are the same as those for commercial aircraft, including flight control, altimetry, navigation, engines, and auxiliary/systems management instruments. Additionally, a large percentage of military aircraft has much the same noncombat flight regime as the same aircraft types found in the civil sector. Accordingly, such aircraft feature much the same primary instrumentation.

Distinctive military aircraft instrumentation relates to the military mission and weapons operations. As management of these operations becomes more complex, instrumentation requirements increase in order to present the added information and control parameters. However, this results in increased workloads and attention has then focused on automation, on the simplification of display and control requirements, on the integration of information,

and on simplified display formats for easier, more accurate use.

Use of electronic displays and controls in aircraft weapon systems has led to a variety of changes in the presentation formats, which are now tailored to specific aircraft mission requirements. Situation comprehension requirements change with the mission and with emerging avionics systems, navigation aids, weapons operation, data-processing capability, and battle-management requirements. Key uses of electronic displays in this context are (1) the change of formats during the mission according to need (time sharing); (2) the tailoring of complex information into more interpretable and usable formats; and (3) the prioritizing of information (for example, caution-warning) according to critical, present need.

Types of display. Two major forms of electronic display were used initially in military systems. In one, the head-up display, a collimating lens is used in the forward field of view. (The effects of collimation are similar to those of far-focusing eyeglasses.) The resulting flight symbology is thus focused at infinity and is mixed with the infinity focus of out-the-window vision, so that near-to-far vision accommodation problems are avoided. The other type, a radar map display, presents radar reflections of ground imagery and targeting information. Two types of displays that were subsequently introduced are infrared displays, based on object temperatures, and multifunction displays, with changing, time-shared formats that present data on, for example, aircraft systems, weapons, battle situations, and integrations of tactical and strategic information. *See* AIRBORNE RADAR; INFRARED IMAGING DEVICES.

Examples. Current military instrumentation (**Fig. 9**) ranges from somewhat antiquated round-dial instruments through the sophisticated use of data-processing and electronic display and control

Fig. 9. Military aircraft instrumentation. (*a*) Soviet Mikoyan MiG-29 cockpit. (*b*) United States Navy/Marine Corps McDonnell Douglas F/A-18 cockpit (*parts a and b from Aviat. Week Space Technol., 129(12):32, September 19, 1988*). (*c*) Operators' stations for on-board mission management system of advanced strategic aircraft (*Boeing Co.*). (*d*) Advanced multirole, multifunction cockpit system (*Boeing Co.*).

systems to integrate and combine information.

For example, the cockpit of the Soviet Mikoyan MiG-29 (Fig. 9*a*) features round-dial instruments for engine and flight data and numerous stick- and throttle-mounted actuators. Primary flight-control data are on the left and weapons management information on the right. (Traditionally, primary flight information has been on the centerline.) The aircraft has a radar, an infrared display, and a head-up display (to the top of the figure, not shown). The complexity of the conventional approaches is evident, and Soviet pilots have indicated that integrating weapons operation is, in fact, difficult.

Alternatively, the United States Navy/Marine Corps McDonnell Douglas F/A-18, developed in the same period as the MiG-29, incorporates one of the most advanced cockpits of United States fighter attack aircraft. It has a head-up display, three format-interchangeable cathode-ray tube displays, and small round-dial instruments for backup. Primary flight control and aiming data are presented on the head-up display. Communication and navigation information is directly below, on the control panel. The lowest central display is a moving map, an electronic map of the area moving below the airplane. The other displays present weapons management and radar data. Format integration and color coding of the electronic displays improve pilot performance time, accuracy, and workload.

Figure 9*c* shows the work stations of an on-board mission management system for the offensive and defensive operations of an advanced strategic aircraft. This system employs artificial intelligence and expert systems software to recommend tactics and courses of action, and full-color, high-resolution graphic displays with programmable touch panels. Mapping, targeting, weapons management, flight reference data, and control modes are shown. *See* ARTIFICIAL INTELLIGENCE; EXPERT SYSTEMS.

Development has been undertaken of an advanced multirole, multifunction cockpit system that features a head-up display with advanced graphic pictorials (Fig. 9*d*). Lower displays show threat zones ahead, and an overall situation display of area, threat, path, and target information. The weapons management displayed on the left represents stores, location, and ready status. To the right is an integrated engine and energy management display.

Donald L. Parks

Aircraft propulsion

Aircraft generally derive their propulsion from fuel-fed heat engines whose power is fed to a propulsor. The propulsor accelerates a stream of air through the engine (as in the case of turbojets, turbofans, and ramjets) or around the aircraft (as in the case of helicopter rotors and propeller rotors) in a direction rearward of the flight direction. The integrated pressure forces impacting on the surfaces of the propulsive machinery required to accelerate the propulsive stream ultimately react on the aircraft to propel it in the flight direction. This is an application of Newton's second and third laws: The force required to accelerate a mass flow is proportional to the quantity of the mass flow multiplied by the rate of its acceleration, and for every action (of accelerating the mass flow) there is an equal and opposite reaction (on the engine and ultimately on the aircraft). *See* NEWTON'S LAWS OF MOTION.

Types of propulsion systems. The two most common types of heat engines used as power producers in aircraft propulsion are (1) a reciprocating- or piston-engine (Otto cycle) power producer, often mechanically or turbo-supercharged, which, produces mechanical energy to drive a propulsor; and (2) a gas turbine (Brayton cycle) power producer (compressor/combustor/turbine), where the energy produced is in the form of high-pressure, high-temperature products of combustion which are fed into a power turbine to drive a propulsor or into a jet nozzle which serves as a propulsor. A wide variety of propulsors are driven by these power producers, tailored to the aircraft range of speed and its unique mission requirements. In the case of rocket engines, the heat engine is simply a combustion chamber into which fuel and oxidizer (or a single monopropellant) are fed and burned, and the resultant stream of high-temperature, high-pressure air is fed to a propulsion nozzle. The most common types of aircraft propulsion systems are listed in **Table 1**, and diagrams of basic types are shown in **Fig. 1**. *See* AIRCRAFT ENGINE PERFORMANCE; HELICOPTER; PROPELLER (AIRCRAFT); RECIPROCATING AIRCRAFT ENGINE; ROCKET PROPULSION; SCRAMJET; TURBOFAN; TURBOJET; TURBOPROP.

Other propulsion systems have been proposed and studied and have undergone partial development, but have played a lesser role in the development of the airplane. For instance:

1. The pulsejet engine involves intermittent, cyclic combustion in a cylindrical chamber fitted with a propulsive jet nozzle and with a grid of flapper valves (acting as check valves) at its forward and aft ends to permit entry of air and exit of heated, pressurized gas. The pulsejet engine was initially developed by the Germans for use on the V-2 missile during World War II. It has no long-term utility for aircraft because of its extreme noise, internally destructive vibrations, limited flight speed, low efficiency, and short life.

2. The use of nuclear reactors instead of chemical fuels as the source of energy for aircraft engines was thought to offer the unique and attractive capability of providing spectacularly long unrefueled flight endurance. However, efforts to develop nuclear aircraft engines were terminated because of the unacceptable hazard to life and crops posed by the release of radioactive material in the event of an aircraft crash.

3. The turboramjet engine is a hybrid engine that includes a ramjet for high-speed flight and a turbojet for takeoff and acceleration to high speed. The complexity of the engine and the failure to identify important missions that might utilize its capabilities relegated it to a marginal, historical role in aircraft propulsion development.

4. The propfan is composed of a gas turbine power producer whose discharge is fed to a power turbine

TABLE 1. Types of propulsion systems

Engine type	Power producer, propulsor, and usage	Typical cruise speed*
Reciprocating- or piston-engine–driven helicopter rotor	Reciprocating- or piston-engine (Otto cycle) power producer that drives a helicopter rotor; used in small systems only	Low subsonic
Turboshaft-engine–driven helicopter rotor	Gas turbine power producer whose discharge is fed to a power turbine that drives a helicopter rotor; in general usage	Low subsonic
Reciprocating- or piston-engine–driven propeller	Reciprocating- or piston-engine (Otto cycle) power producer that drives a propeller; used in small systems only	Medium subsonic
Turboprop engine	Gas turbine power producer whose discharge is fed to a power turbine that drives a propeller; in general usage	Medium subsonic
Ultrahigh- or high-bypass turbofan engine or afterburning turbofan engine	Gas turbine power producer whose discharge is fed to a power turbine that drives a fan, a much larger turbocompressor that passes 2 to 10 times the airflow of the power producer (that is, a bypass ratio of 2 to 10) and is generally mounted in front of the power producer; The high-bypass turbofan is in general usage. The ultrahigh-bypass turbofan has been the subject of experimental demonstration and development.	High subsonic to transonic, $M < 1.2$
Medium- or low-bypass turbofan engine or afterburning turbofan engine	Gas turbine power producer whose discharge is fed to a power turbine that drives a fan, a somewhat larger turbocompressor that passes up to two times the airflow of the power producer and is generally mounted in front of the power producer. The engine may include an afterburner to augment the engine thrust for takeoff, combat, or transonic acceleration. In general usage.	Transonic to low supersonic, $M < 2.5$
Turbojet engine or afterburning turbojet engine	Gas turbine power producer whose discharge is fed to a jet nozzle to produce thrust. The engine may include an afterburner to augment the engine thrust for takeoff, combat, or transonic acceleration. In general usage.	Transonic to low supersonic, $M < 3$
Ramjet or scramjet engine	Engine with the same (Brayton) cycle as a jet engine, but with the compression produced by diffusion of the supersonic inlet airstream without the use of turbomachinery. The scramjet includes combustion of the airstream without prediffusion to subsonic speed. Crewless missile and experimental usage only.	Medium supersonic, $3 < M < 8$
Rocket	Combustion chamber and exhaust jet nozzle with not only the fuel but also the oxidizer carried onboard (rather than using captured air from the surrounding environment as oxidizer). Missile and extraterrestrial usage. Historical usage for assisted takeoff in aircraft.	High supersonic

*Where "sonic" is the velocity of sound or a Mach number (M) of 1.0 and the speed is 762 mi/h (1226 km/h) at an ambient temperature of 60°F (15.6°C) and 669 mi/h (1077 km/h) at an ambient temperature of −60°F (−51.1°C).

which drives a set of coaxial counterrotating high-speed propellers, mounted aft of the power producer. The specific fuel consumption at high subsonic flight speed is significantly lower than that of the high-bypass turbofan, and approaches that of the turboprop. In 1988, a flight demonstration was conducted of a unique version of this type of engine (which was called the unducted fan or UDF), involving a counterrotating power turbine that obviated the need for a gearbox to drive the propellers. In the same time period, a geared prop fan was demonstrated. Neither engine attracted enough interest to justify further development, primarily because of the difficulty of mounting the large-diameter propellers under the wing, and the unfavorable impact on aircraft weight balance in mounting the engines at the aft end of the fuselage. *See* TURBOPROP.

Propulsion system installation arrangements. In transport aircraft, the most usual approach to mounting the engines on the aircraft is enclosing each of the engines in a aerodynamically refined pod, called a nacelle, which is suspended from the aircraft on a pylon. **Figure 2** shows some typical arrangement for two-, three-, and four-engine transport aircraft.

In modern military combat aircraft, it is most usual to blend the engine enclosure into the underside of the aircraft's fuselage, with the engine's inlet just forward of the aircraft wing's leading edge. **Figure 3** illustrates such an arrangement for one- and two-engine military combat aircraft.

The major issues involved in the mounting of engines on aircraft are: (1) locating the engine inlet so that boundary layer or separated flow from the aircraft's fuselage, wings, or empennage (particularly in extreme flight maneuvers) is not ingested into the engine; (2) locating the engine inlet so that, during takeoff roll, sand, dust, and debris will not be ingested into the engine; (3) in any case where the engine is mounted in a nacelle hung by a pylon from the wing or fuselage of the aircraft, keeping appropriate separation between the nacelle and the wing or fuselage so that the nacelle does not engender deleterious interference drag; (4) locating the center of gravity of the engine or engines so that the center of gravity of the aircraft (to which the engines' weight makes a major contribution) is as close as possible to the center of lift of the aircraft's wings; and (5) in a multiengine application, keeping the engines' center lines as close as possible to the aircraft's center line so that, in an emergency one-engine-out situation, unbalanced yawing moments of the operative engines are minimized.

Propulsion system technical requirements. The requirements that may be placed upon the propulsion system for a particular aircraft may include any or all of the following:

Performance. Performance requirements include a definition of the flight regime or flight envelope as a function of altitude and flight speed in terms of Mach number (**Fig. 4**). Also to be specified are the thrust or power delivered by the engine and the rate of fuel consumed over the entire flight regime as a function of throttle setting (maximum, takeoff, climb, cruise, idle, and so forth), flight speed,

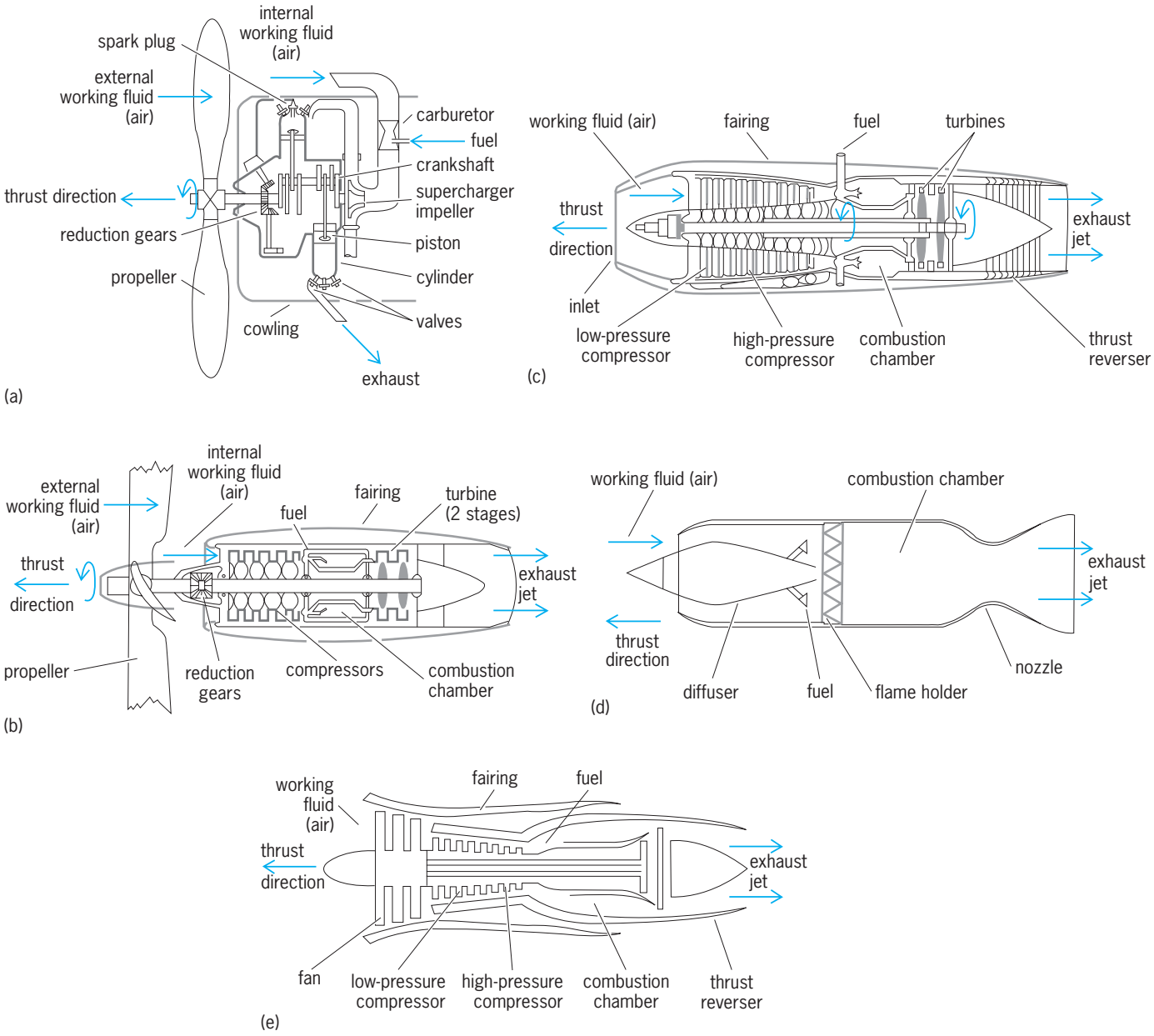
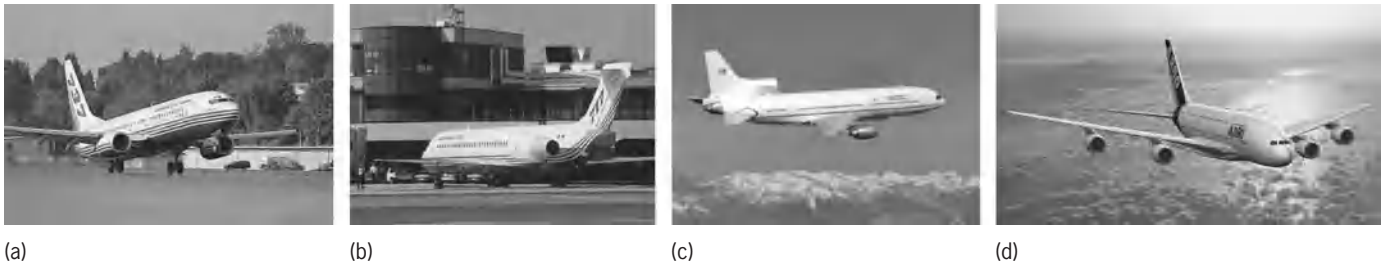


Fig. 1. Operational diagrams of the basic types of airplane engines. (*a*) Reciprocating. (*b*) Turboprop. (*c*) Turbojet. (*d*) Supersonic ramjet. (*e*) Fanjet (turbofan).

altitude, ambient pressure, temperature, humidity, and so forth; a description of additional modes of operation such as afterburning-augmented operation and reverse thrust; time limits on the usage of the high-power settings imposed to prolong the life of the engine; the maximum elapsed time allowed for transients in engine power to accommodate emergencies such as the sudden need for a burst of power



(a) (b) (c) (d)

Fig. 2. Typical installation arrangements for multiengine transport aircraft with pylon/nacelle-mounted engines. (*a*) Boeing 737 with two-engine, under-the-wing mounted installation (*Boeing*). (*b*) Boeing 717 with two-engine rear-fuselage mounted installation (*Boeing*). (*c*) Lockheed Tristar with three-engine installation with tail-mounted middle engine (*NASA*). (*d*) Airbus 380 with four-engine under-the-wing installation (*Airbus*).

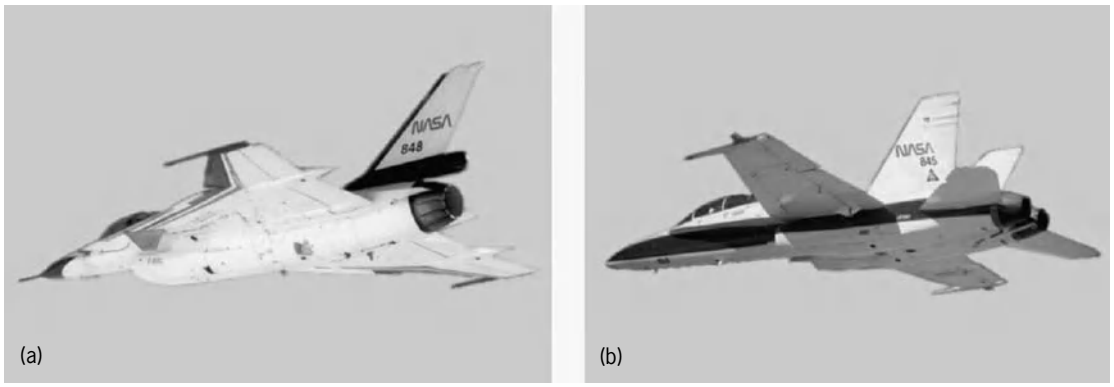


Fig. 3. Typical installation arrangements for one- and two-engine military combat aircraft. (*a*) Lockheed Martin F-16 with one-engine fuselage-mounted installation. (*b*) Boeing F-18 with two-engine fuselage-mounted installation. (*NASA*)

in the event of an aborted landing; the time for an engine to reach idle in the course of a starting sequence; the maximum deterioration of thrust or power that one might encounter in the course of the engine's life; the amounts of engine mechanical and electrical auxiliary power available to drive aircraft accessories, and the compressed air available from the engine (as well as the effects of their extraction on engine performance); and fuel types and limiting pressures and temperatures of the fuel delivered to the engine. *See* AIRCRAFT FUEL.

Weight. An aircraft's drag, and hence its thrust and fuel requirements, is strongly dependent on its weight since the two quantities are closely related by the characteristic lift/drag ratio of the wings. In virtually all aircraft systems, the engines' weight constitutes a major fraction of the total aircraft system weight, so the engines' weight is a critical parameter to be specified. The location of the engine's center of gravity may also be specified. *See* AERODYNAMIC FORCE.

Mechanical installation. The description of the mechanical installation includes the physical envelope of the engine, the mounting points of the engine where it is structurally mated to the aircraft's structure, and the details of the mount (whether

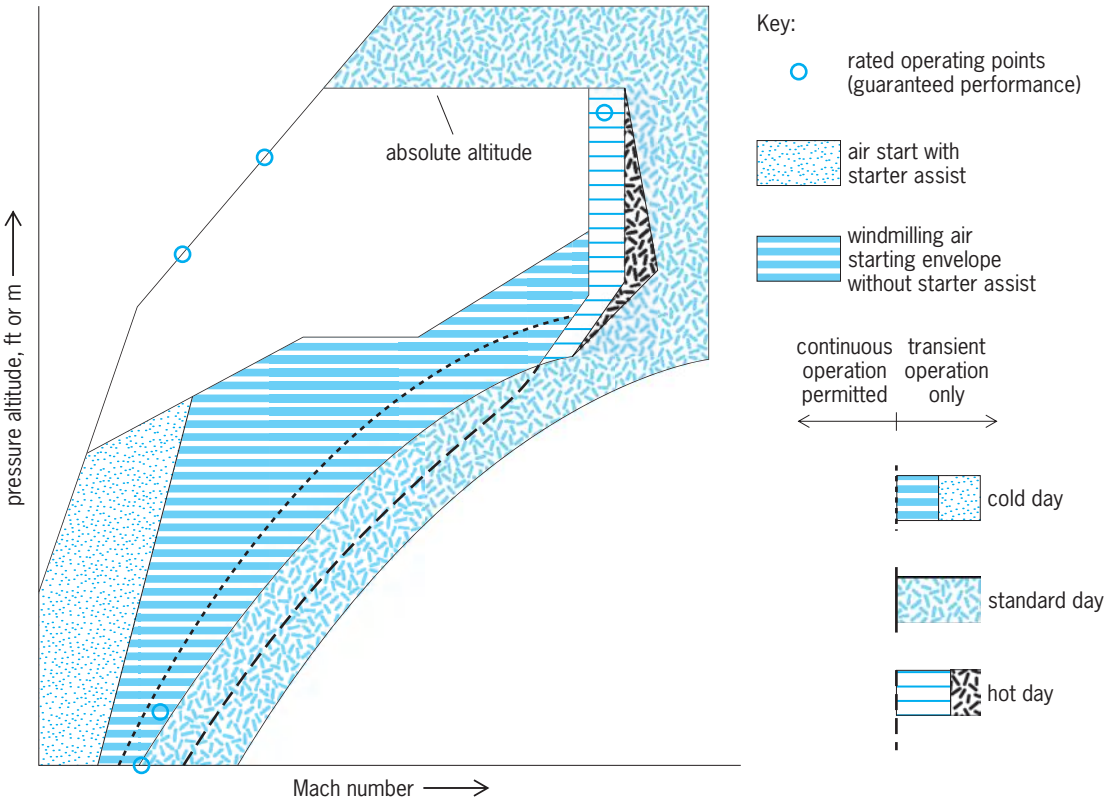


Fig. 4. Typical flight envelope of an aircraft in terms of the aircraft propulsion system limitations. Cold day, standard day, and hot day refer to U.S. Standard Atmosphere (1976).

hard-mounted or mounted in vibration and shock isolators, and any restrictions on the direction of the forces to be transmitted). Any limitations on the accelerations and gyroscopic moments associated with aircraft maneuvers or hard landings also must be identified. The interconnections between the aircraft and the engine—such as instrumentation and engine control information channels; fuel connections; pneumatic, electrical, and hydraulic system connections; and so forth—may be characterized. Surface temperatures of the engine must be specified to enable the design of the surrounding aircraft structure.

Environmental impact limitations. In particular, these include limitations on propulsion system noise and propulsion system exhaust emissions. In addition to the environmental impact of these emissions, the effect of noise and vibration on the aircraft and its passengers must be limited to acceptable values.

Engine durability, reliability, and safety. The engine durability in such terms as time between overhauls must be specified and the criteria for parts replacement identified. A schedule of periodic inspections may be specified. The engine's safety in the event of ingestion of hail and ice, birds, or other foreign objects must be ensured by specifying amounts of remaining thrust or power in the event of a minor incident or assurance of damage containment in the event of a major incident. The amount of pressure and temperature distortion in the inlet airflow to the engine that can be tolerated by the engine without its stalling or surging must be specified. Limitation must be placed on the leakage of flammable fluids, and any fire-suppression or inhibition devices (such as fire-walls) may be specified.

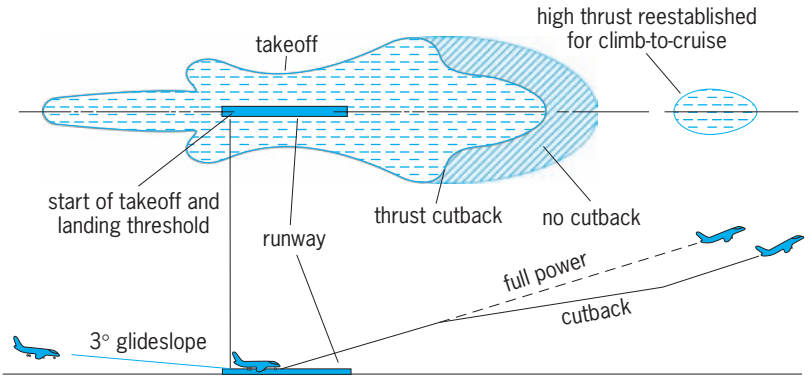


Fig. 5. Typical aircraft composite noise footprint for landing, full-power climb, and climb with power cutback. (*After M. J. T. Smith, Aircraft Noise, 2d ed., Cambridge University Press, 2004*)

Engine accessibility and maintainability. The ease of engine removal, the simplicity of making engine inspections and adjustments while on-the-wing, and the capability of making on-the-wing change-out of accessories are important features of any engine design.

TABLE 2. Aircraft/propulsion system noise emissions

Noise category	Cause	Type of noise generated	Mitigation or remediation
Jet noise	Random turbulence generated by the mixing layer at the periphery of high-velocity exhaust jets from jet engines and from low- and medium-bypass turbofan engines	Broadband or "white" noise (roaring or hissing sound)	Exhaust mixer nozzles
Combustor and afterburner noise	Random turbulence generated by mixing and flame instability in the combustion zones of the combustor or afterburner	Broadband or "white" noise (roaring or hissing sound)	In afterburner engines, limiting usage in populated areas
Combustor screech	Afterburner acoustic instability	Pure-tone noise	Perforated screech liner in the afterburner
Duct and turbomachinery blade surface noise	Random turbulence generated by flow over the surfaces of engine ducts, struts, and turbomachinery blading	Broadband or "white" noise (roaring or hissing sound)	Minimization of surface velocities where possible and necessary
Rotating blade and blade interaction noise	Interaction of rotating pressure fields with static parts or of static pressure fields with rotating turbomachinery blade rows at engine inlet and at engine exhaust	Discrete frequency or pure-tone noise at rotor-blade-passing frequency or its integer multiples	(1) Minimization of blade row-to-row interaction of blading pressure fields upstream of blade rows and trailing-edge wakes downstream of blade rows by maximizing spacing between succeeding blade rows; (2) selection of blade number to achieve "cutoff" for maximum acoustic attenuation; (3) employment of duct liners in inlet and exhaust ducts
Fan rotor "buzz-saw" noise	Upstream propagation of oblique shock from the tip of the fan rotor blade whose rotating velocity with respect to the incoming flow is supersonic	Discrete frequency or pure-tone noise at rotor-blade-passing frequency or its integer multiples	Employment of duct liners in inlet and exhaust ducts
Propulsor (that is, helicopter rotor or propeller) noise	Interaction of rotating propulsor pressure fields with static parts or static pressure fields with rotating propulsor blades	Discrete frequency or pure-tone noise at rotor-blade-passing frequency or its integer multiples	—
Sonic boom	Generated by the oblique shock cone emanating from the forward tip and wing leading-edge of the aircraft traveling at transonic and supersonic speed	Large pressure pulse as heard by the stationary observer on the ground	(1) Constraining aircraft to operation at subsonic speed over populated areas; (2) constraining aircraft to high enough altitude that cone shock is sufficiently attenuated when it reaches the ground

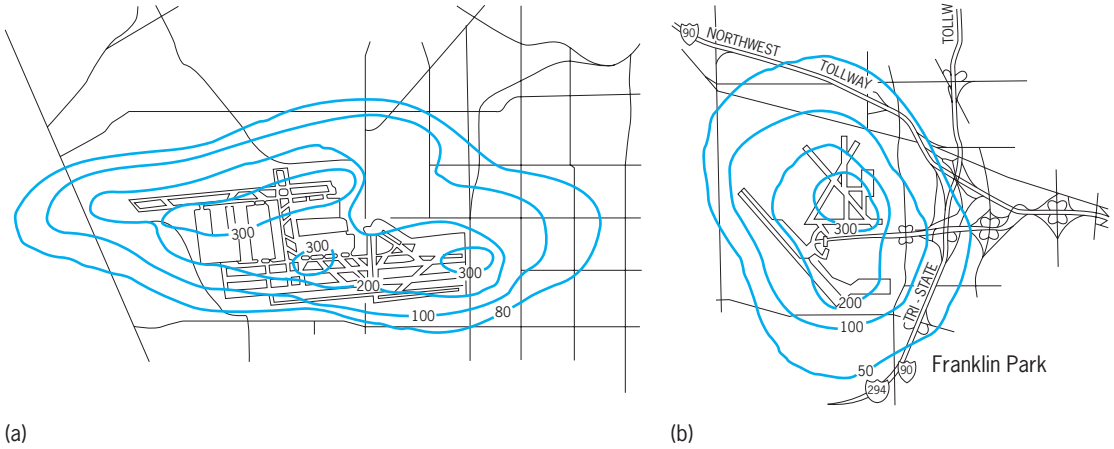


Fig. 6. Typical contour maps of NO_X concentration (measures in micrograms per cubic meter) in major airports—annual averages for 1980. (*a*) Los Angeles International Airport. (*b*) O'Hare Chicago International Airport. (*After U.S. Environmental Protection Agency, Aircraft Emissions: Impact on Air Quality and Feasibility of Control, Report to Congress, undated*)

Unique military requirements. The engine requirements may also include the capability of ingestion from ordnance gun gas or rocket exhaust, minimization of radar cross section and infrared emissions, invulnerability to and survivability from weapons damage, hardening of electronics to damage from nuclear radiation, and so forth.

Propulsion system noise emissions. The noise generated by the propulsion system is a major consideration in the design and installation of the engine and propulsor. **Table 2** summarizes some of the most important constituents in the generated noise, the specific causes, the type of noise generated, and the measures that are taken to mitigate or remediate their impact on the public. *See* AERODYNAMIC SOUND; SONIC BOOM; TURBINE ENGINE SUBSYSTEMS.

Operation of the aircraft can have a profound effect on the noise "footprint" of the aircraft. In the case in **Fig. 5**, the power of the engine is cut back from full power in the midst of its climb after takeoff to reduce the noise heard by the public under the flight path, substantially reducing the footprint. But an additional area of intrusive noise is included at a further point where high thrust climb is reestablished for continuing the full-power climb to cruise altitude. *See* ACOUSTIC NOISE; AIRCRAFT NOISE.

Propulsion system exhaust emissions. A second major impact on the environment of heat engines fueled by fossil fuels for aircraft propulsion is in the engines' emission of products of combustion into the atmosphere. Although the mechanism of production of pollutants is similar to that in all vehicular and stationary power plants, the specific consequences and issues are unique to aircraft propulsion because of the aircrafts' use in the troposphere and stratosphere and the intense concentration of their use on and near the ground at airports and their environs (**Fig. 6**). The major constituents of aircraft engine exhaust pollution emissions are listed in **Table 3** along with the engine operation at which they are encountered, their impact, and approaches to their mitigation or remediation. *See* AIR POLLUTION.

TABLE 3. Propulsion system exhaust emissions

Emission category	Primarily associated with engine operation at	Impact	Mitigation or remediation
Smoke (soot)	Takeoff and climb-out	Visibility nuisance around airports	Lean primary combustion zone fuel–air mixtures; more effective fuel-air mixing within the primary zone
Unburned hydrocarbons	Low power, especially ground idle	Contribution to urban smog burdens	Improvement of fuel atomization at low power; enhancing primary zone fuel-air mixing quality; minimizing fuel impingement on the combustor liner; reduction of idle power usage
Carbon monoxide (CO)	Low power, especially ground idle	Contribution to urban CO burdens	Improvement of fuel atomization at low power; enhancing primary zone fuel-air mixing quality; minimizing fuel impingement on the combustor liner; reduction of idle power usage
Nitrogen oxides (NO_X)	All high power, including cruise	Contribution to urban smog burdens in environs of airports	Reduction of peak flame temperatures; leaner primary zone mixtures at high power
Subsonic aircraft engines		Possible contribution to global warming	
Future supersonic aircraft engines		Stratospheric ozone depletion	
Carbon dioxide (CO_2)	All power settings	Contribution to global warming	Improvement of engine fuel efficiency
Sulfur oxides (SO_X)	All power settings	Contribution to urban smog burdens	Use of low-sulfur fuel

The disparate requirements of improving combustor operation at both very low power conditions and at very high power conditions have dictated that more complex combustion systems with variable geometry or variability in the fuel injection system have had to be employed. Typically, a modern aircraft gas turbine engine combustion system might include two sets of fuel injection nozzles and employ a single set at low power settings and both sets at high power settings.

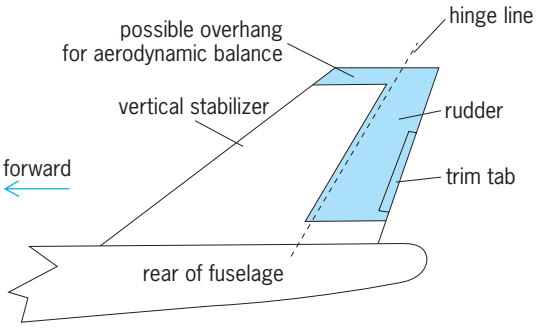
Limitations on emissions. Limitations on aircraft noise emissions and on aircraft exhaust emissions have been promulgated by the International Civil Aviation Authority (ICAO). These emissions are regulated in the United States by the Federal Aviation Administration (FAA), in the United Kingdom by the Civil Aviation Authority (CAA), and in Europe by the Joint Aviation Authority (JAA). Fredric F. Ehrich

Bibliography. M. Barrett, *Aircraft Pollution—Environmental Impacts and Future Solutions*, WWF Research Paper, August 1991; R. D. Bent and J. L. McKinley, *Aircraft Powerplants*, 5th ed., 1985; J. L. Kerrebrock, *Aircraft Engines and Gas Turbines*, 2d ed., 1992; M. J. Kroes and T. W. Wild, *Aircraft Powerplants*, 7th ed., 1995; J. D. Mattingly, W. H. Heiser, and D. T. Pratt, *Aircraft Engine Design*, 2d ed., AIAA, Reston, VA, 2002; G. C. Oates, *Aircraft Propulsion Systems: Technology and Design*, 1989; Rolls-Royce PLC, *The Jet Engine*, Derby, United Kingdom, 5th ed., 2005; M. J. T. Smith, *Aircraft Noise*, 2d ed., Cambridge University Press, 2004.

Aircraft rudder

The hinged rear portion of an airplane's vertical tail. The vertical tail is composed of the vertical stabilizer and the rudder. The vertical stabilizer is mounted to the fuselage and is fixed relative to it. The rudder is hinged to the rear of the vertical stabilizer (see **illus.**) and moves to the left or right in response to control inputs from the rudder pedals or from an automatic stability and control system.

The rudder provides an aerodynamic moment about the aircraft's center of gravity for the purpose of yaw control. When the rudder turns clockwise, for example, as viewed from above, its trailing edge moves to the left, effectively adding camber to the vertical tail. The result is that an aerodynamic side force is produced on the vertical tail to the right. This force, in turn, produces a counterclockwise yawing moment about the airplane's center of gravity, resulting in a turn to the left. Conversely, to turn to the right, a pilot pushes on the right rudder pedal, causing the rear of the rudder to swing to the right and the airplane to turn in that direction. *See* AIRFOIL.



Vertical tail of an airplane, showing location of the rudder.

An additional, small, movable surface, known as a trim tab, may be hinged to the rudder. When deflected to a fixed position, the tab causes the rudder to deflect to, and hold, a desired angle. Thus a steady yawing moment is produced which can relieve the pilot from having to make a sustained push to offset any disturbing moments such as from a propeller.

For some supersonic applications, the vertical tail is a one-piece configuration, with the entire tail rotating to provide yaw control. For vee-tails, the rudder and elevator control surfaces are one and the same. For pitch control (which is provided by the elevators in a conventional configuration), the hinged surfaces on either side move together. For yaw control, they move differentially; that is, one of the surfaces moves up while the other one moves down. The result is a net side force to one side or the other while the net vertical force is zero. *See* ELEVATOR (AIRCRAFT); TAIL ASSEMBLY.

Some airplanes employ twin, or even multiple, vertical tails. For this configuration, rudders are usually placed on each vertical tail and move in unison to the pilot's control input to yaw to the left or right. The total effectiveness of multiple rudders is equal closely to the effectiveness of one rudder multiplied by the number of vertical tails. If the tails are close together, there is some amount of interaction which can be determined from aerodynamic analyses. *See* FLIGHT CONTROLS. Barnes W. McCormick, Jr.

Aircraft testing

Subjecting a complete aircraft or its components (such as wings, engines, or electronics systems) to simulated or actual flight conditions in order to measure and record physical phenomena that indicate operating characteristics. Testing is essential to the design, development, and acceptance of any new aircraft.

Aircraft and their components are tested to verify design theories, obtain empirical data where adequate theories do not exist, develop maximum flight performance, demonstrate flight safety, and prove compliance with performance requirements. Testing programs originate in laboratories with the evaluation of new design theory; progress through extensive tests of components, subsystems, and subsystem assemblies in controlled environments; and culminate with aircraft tests in actual operational conditions. *See* AIRCRAFT DESIGN.

Laboratory Testing

Instrument testing, in controlled conditions of environment and performance, is used extensively during the design performance assessment of new

Fig. 1. Scale model of fighter aircraft being mounted inside test section of large wind tunnel. (*NASA*)

aircraft to avoid the costly and sometimes dangerous risks of actual flight. Each method of testing contributes to the development of an efficient aircraft.

Wind tunnel tests. A wind tunnel is basically an enclosed passage through which air is forced to flow around a model of a structure to be tested, such as an aircraft. Wind tunnels vary greatly in size and complexity, but all of them contain five major elements: an air-drive system, a controlled stream of air, a model, a test section, and measurement instruments. The drive system is usually a motor and one or more large fans that push air through the tunnel at carefully controlled speeds to simulate various flight conditions. A scale model of an actual or designed aircraft is supported inside the test section (**Fig. 1**), where instruments, balances, and sensors directly measure the aerodynamic characteristics of the model and its stream of airflow.

Wind tunnel tests measure and evaluate airfoil (wing) and aircraft lift and drag characteristics with various configurations, stability and control parameters, air load distribution, shock wave interactions, stall characteristics, airflow separation patterns, control surface characteristics, and aeroelastic effects (**Fig. 2**). Tunnels are also used to calibrate airflow-sensing and pressure-sensing devices, engine inlet performance, engine fan airflow–nacelle (engine housing) interference, weapon drop trajectories, and engine performance tests. *See* AIRFOIL.

Test speeds are generally divided into subsonic (Mach numbers 0–0.95), transonic (Mach 0.95–1.3), supersonic (Mach 1.3–5), and hypersonic (Mach 5 and above). The hypersonic range includes the transatmospheric vehicle concept that merges aeronautics and space technologies across the speed ranges involved in the conventional ground takeoff of the vehicle to Earth-orbital velocities and the vehicle's return to ground landing. *See* HYPERSONIC FLIGHT; MACH NUMBER; SUBSONIC FLIGHT; SUPERSONIC FLIGHT; TRANSONIC FLIGHT.

Different kinds of wind tunnels simulate the airflow of these speed ranges, including two- and three-dimensional, return and nonreturn flow, intermittent and continuous flow, atmospheric and variable density, and vertical free flight. Hypervelocity facilities have been developed to better simulate conditions at hypersonic speeds, including shock tubes and tunnels, hotshot and impulse tunnels, ballistic ranges, and light-gas guns.

A relatively new design of transonic wind tunnel uses cryogenic (supercold) nitrogen gas as a test medium (instead of high-pressure air) in a continuous-flow, pressurized, closed circuit. Liquid nitrogen is vaporized inside the tunnel at temperatures as low as $-149^{\circ}C$ ($-300^{\circ}F$) to reduce the viscosity and increase the density of the gas moving through the tunnel. The cryogenic concept provides almost exact simulations of the flight conditions of advanced design aircraft that will fly in or through the transonic speed range.

Several aircraft configurations generally evolve during preliminary design, and the best candidates are chosen for testing. Scale models of each configuration are installed in a wind tunnel for short tests to determine if a model's aerodynamic shape produces the desired performance. Many tests may be made of various configurations before the most satisfactory aerodynamic design is selected for further development.

After a general configuration has been established, a more detailed design is developed, and a complex series of wind tunnel tests is begun. Aircraft lift and drag curves are determined throughout the design speed and Mach number range. Aircraft static stability forces and moments are measured for expected variations in flight conditions. Wing air-load distributions are found by measuring the static air pressure

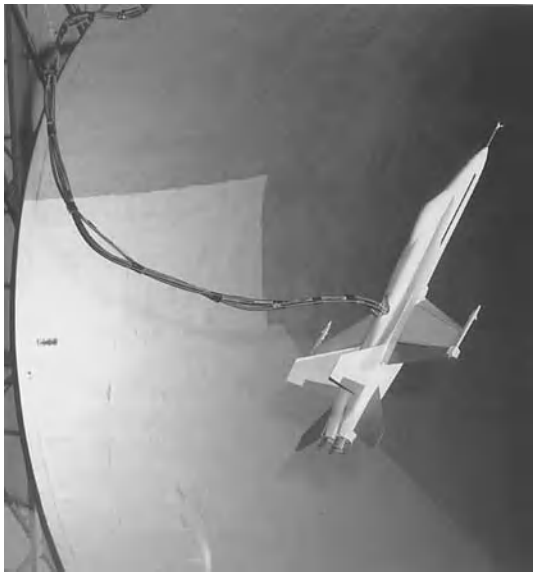


Fig. 2. Free-flying scale model aircraft in wind tunnel; attached cable contains sensors and other measurement instruments. (*NASA*)

on the upper and lower wing surfaces, fuselage, and tail surfaces, at various points along the aircraft's length and width.

Patterns of airflow are made visible through several techniques, including neutrally buoyant luminescent bubbles of helium, smoke filaments, and tufts of thread. Schlieren photography or shadowgraphs can visualize large changes in air density, such as shock patterns. Two techniques make use of lasers. (1) A laser hologram uses relatively simple optics to split the light from a powerful laser into two beams, one of which passes through a wind tunnel test section. When recombined on a photographic plate, the beams form a hologram (an interference pattern) that captures the pattern of density gradients within the test section. (2) A laser velocimeter operates like a radar to acquire detailed measurements of flow velocity. A laser beam is split into two beams that cross at adjustable points in a tunnel, allowing two velocity components in two different planes to be determined at the same time. Free-flight vertical wind tunnels are used to evaluate an aircraft's dynamic spin characteristics. Flexible models simulate the structural flexibility of an aircraft to evaluate aeroelasticity characteristics. *See* AEROELASTICITY; FLOW MEASUREMENT; HOLOGRAPHY; SCHLIEREN PHOTOGRAPHY; SHADOWGRAPH; WIND TUNNEL.

Integration testing. Aircraft components are integrated into subsystems, system elements, and complete operational systems to help resolve interface problems. Tests establish functional and operational capability and evaluate complete system compatibility, operation, maintenance, safety, reliability, and best possible performance (**Fig. 3**).

Typical examples of integration testing are tests of control systems and avionics (aviation electronics) systems. Testing the integration of a control system involves a completely rigged primary control system that is installed on a structural frame similar to that of an actual aircraft. The arrangement can test friction, force feel, fatigue wear, and response, while subjecting the system to flightlike air loads. An automatic flight control system checks compatibility, and pilots are included in tests to obtain significant results. Avionics systems tests can range from a simple antenna test with a receiver to such complex tests as integrating a system of electrical power, communications and guidance equipment, and mission avionics, which includes airborne warning radar and command and control equipment. *See* AIRBORNE RADAR; CONTROL SYSTEMS; FLIGHT CONTROLS.

Dynamic ground tests. Rocket-propelled sled tests evaluate crew ejection escape systems for high-performance aircraft. A fuselage section, mounted on a sled, is propelled by rockets along fixed tracks. When a desired speed is reached, the ejection mechanism is automatically triggered, firing rockets that propel crew seats (containing instrumented mannequins) clear of the fuselage, and activating parachutes to limit the free-flight trajectory of the mannequins and allow safe descent to the ground. Water-propelled sled tests study landing gear systems and runway surface materials. High-pressure water

Fig. 3. Scale model of E-7 advanced fighter aircraft used for tests of experimental component called augmenter (rectangular opening in wing with open flap). (*NASA*)

can propel sled-mounted test equipment at speeds comparable to those of high-speed aircraft (**Fig. 4**). The sled technique provides accurately controlled test conditions that are documented with high-speed photography. Results are studied to verify that all parts of a test system are within human or mechanical tolerances.

Other dynamic ground tests include acceleration and arresting tests of aircraft fuel system venting,

Fig. 4. Test sled propelled down runwaylike track by high-pressure water to test aircraft components during simulated landing. (*NASA*)

transfer, and delivery, which are evaluated while the system is subjected to flightlike forces and attitudes. Engines are operated in highly instrumented test stands that use aircraft mountings, controls, and inlet and exhaust cowlings containing thrust reversers. Tests substantiate satisfactory cooling, vibration, and installed engine performance. *See* AIRCRAFT ENGINE PERFORMANCE.

Ground vibration tests are done before the first flight of a new model aircraft, subjecting the entire aircraft to vibrations induced from electromagnetic exciters. Structural response is analyzed with magnetic pickups whose outputs are read on oscilloscopes. The tests verify that fuselage structures and empennage (aft control structures) are not responsive to vibrations that will be encountered by engines and equipment during gust air loads anticipated in actual flight.

Static load tests. Proof load tests of actual aircraft are usually done on one or more of the first airframes built. An airframe, mounted in a laboratory, is fitted with thousands of strain gages, the outputs of which are recorded on an automatic data-recording system. Simulated air and inertia loads are applied to airframe components, which are loaded simultaneously, in specified increments, to simulate loads encountered during takeoff, maneuvering flight, and landing. Structural stress and deflection data are recorded at each load level, starting with loadings equal to 80% of design limit. Loadings are increased to design limit and then to ultimate failure to locate possible points of excessive yield. Flight air loads, including those encountered during gust upsets, are repeatedly applied to determine the structure's fatigue life. Critical parts may be redesigned to withstand required loads and ensure satisfactory service life. *See* AIRFRAME; STRAIN GAGE.

Environmental tests. Component parts and system subassemblies are tested with various loadings while operating under expected extremes of temperature, humidity, and vibration to determine service life. Scale models containing communication antennas are tested for radiation properties to verify transmission performance. Mockups that simulate electrical and avionic components are operated under load to verify that their installation is free from electromagnetic interference. Environmental chambers verify the operation of rain-removal, de-ice, and anti-ice equipment. Maurice Parker

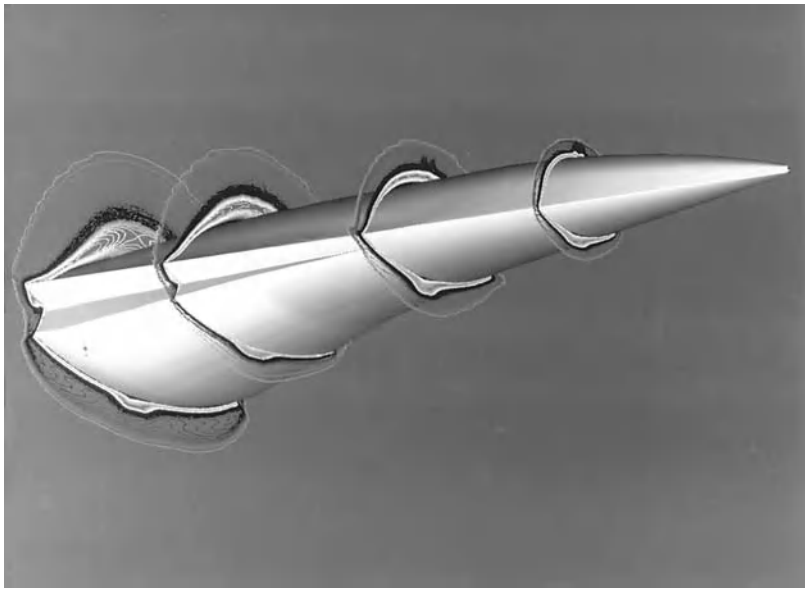


Fig. 5. Supercomputer-aided study of hypersonic lifting body, showing computer-generated lines of constant pressure during Mach 19 flight. (*NASA*)

Flight Simulation

Aircraft flight and systems characteristics are represented with varying degrees of realism for research, design, or training simulation purposes. The representation is usually in the form of analytic expressions programmed on a digital computer. Flight simulation may be performed with or without a human pilot in the loop. The pilot imposes additional constraints on the simulator such as requiring a means of control in a manner consistent with the means provided in the aircraft being simulated. Flight simulation requires representation of the environment to an extent consistent with the purposes of the simulation, and it further requires that all events in the simulator occur in real time. Real time is a term which is used to indicate that all time relationships in the simulator are preserved with respect to what they would be in the airplane in flight. *See* REAL-TIME SYSTEMS. Frank Cardullo

Simulators range in size and complexity from actual aircraft, outfitted with special flight decks that can be reconfigured to test different systems, to desktop simulators that can test individual or integrated components.

Use of computers. Highly sophisticated computers are the primary element in most aircraft simulation tests. Analog computers provide accurate and repeatable discrete simulations, such as systems analysis and human–machine flight simulations. Hybrid computers combine analog and digital computers to work on a problem simultaneously, with each computer doing the task best suited to its special capability. Realistic simulation testing provides flight crews with exact flight deck representations (on either fixed- or moving-base platforms), visual displays, and functioning instruments. *See* ANALOG COMPUTER; DIGITAL COMPUTER.

Supercomputers, large, powerful scientific computers that contain memories of more than 250 million words and can perform 1 billion computational operations each second, are accelerating the rate of progress in computationally intensive disciplines such as aerodynamics, structures, materials, and chemistry, particularly for hypersonic and transatmospheric flight vehicles (**Fig. 5**). A relatively new testing technique, made possible by the increasing availability of supercomputers, is computational fluid dynamics. It provides powerful analytical, simulated, and predictive tools to study the basic physics of aerodynamic flow fields, especially to better understand the complex flow environment of advanced aircraft configurations, and allow the aerodynamic

Fig. 6. Controls of transport-systems research-vehicle simulator, used to develop automated pilot aids, being operated by engineer. (*NASA*)

perfection of new designs. Computations are made of three-dimensional viscous compressible flows that contain separated flow regions over wing and fuselage combinations. An example is computation of hypersonic wing–body surface pressure contours for a flight environment of Mach 25, a 5° angle of attack, and a 3000°R (1667 K) wall temperature. *See* SUPERCOMPUTER.

Applications. Simulators may be classified by their use in research, design, or training. Research simulators are usually employed to determine patterns of human behavior under various workloads or in response to different flight instrument display configurations or different aircraft dynamic characteristics. An example of a simulator-tested system is the takeoff performance monitoring system (**Fig. 6**), used to develop automatic pilot aids. Inside the flight deck of a transport-system research-vehicle simulator, information displayed on electronic screens as symbols and numbers is superimposed onto a scaled out-the-window runway scene. Predictions and advisories are updated in real time, based on sensed conditions and performance during a takeoff roll. The system predicts where on a runway important takeoff events (such as rotation or stopping) will occur and provides advisory information on whether the takeoff should be continued or aborted. The system should enhance takeoff safety by providing pilots with previously unavailable information. Maurice Parker

Design simulators are used to conduct tradeoff studies to evaluate different design approaches in the aircraft. Virtually all aircraft manufacturers in the United States employ simulators to a large extent in the aircraft design process. While the investment is substantial, many hours of flight tests are saved, and designs can be evaluated much earlier in the aircraft production schedule, resulting in considerable savings in retrofitting hardware. Simulators are also used by aircraft manufacturers to demonstrate aircraft characteristics to potential customers well before the first flight of the airplane.

The most pervasive use of flight simulators is for training operators of the aircraft and its systems and maintenance personnel. The flight simulator was patented in 1929 by E. A. Link to train pilots in instrument flying; it was responsible for training a half million pilots during World War II. The current flight simulator is quite different from Link's Blue Box and is used to train pilots of virtually every military and commercial aircraft in the United States and Europe.

Training simulators can be broken down into five categories. Operational flight trainers are used to train pilots, copilots, flight engineers, and navigators, as appropriate, in all aspects of flying a modern aircraft, either with a full crew or alone. A second type of device is a weapons systems trainer, employed on tactical and strategic aircraft. These devices add sophisticated avionics and weapons systems simulation to an operational flight trainer. A part task trainer is a device which provides simulation of part of the operational flight trainer or weapons systems trainer, such as the situations confronted by a navigator in a strategic aircraft or an antisubmarine warfare systems operator in a naval patrol aircraft. Cockpit procedures trainers are employed to train pilots in the procedures of operating the systems of an aircraft. Finally, maintenance trainers are employed to train mechanics and technicians in the diagnosis of problems and the procedures of routine maintenance on various aircraft systems such as engine, flight controls, and radar.

Training simulators have been used extensively by NASA for training astronauts in programs from Project Mercury to the space shuttle. They were also used during the *Apollo 13* mission to verify procedures and flight program modifications that helped to avert possible disaster after an inflight explosion partially disabled the spacecraft. *See* SPACE FLIGHT.

That training simulators save substantial amounts of money in fuel costs, weapons expenditure, and so forth is obvious. However, the simulator has much greater benefit in that it is in many cases a better training device than the aircraft. This is true because of the safety, versatility, and speed with which critical maneuvers may be performed. For example, in practicing landing in the aircraft, a student pilot may make an approach to either a touch-and-go landing or a full-stop landing or perhaps a missed approach. The student then must contend with air traffic and weather to reenter the landing pattern for another attempt. This takes considerable time, whereas in the simulator, after landing, the simulated aircraft may reset to the approach configuration at the top of the glide slope, ready to go again in seconds. The simulator also allows the student pilot to learn to deal with malfunctions in a safe environment. Practicing the

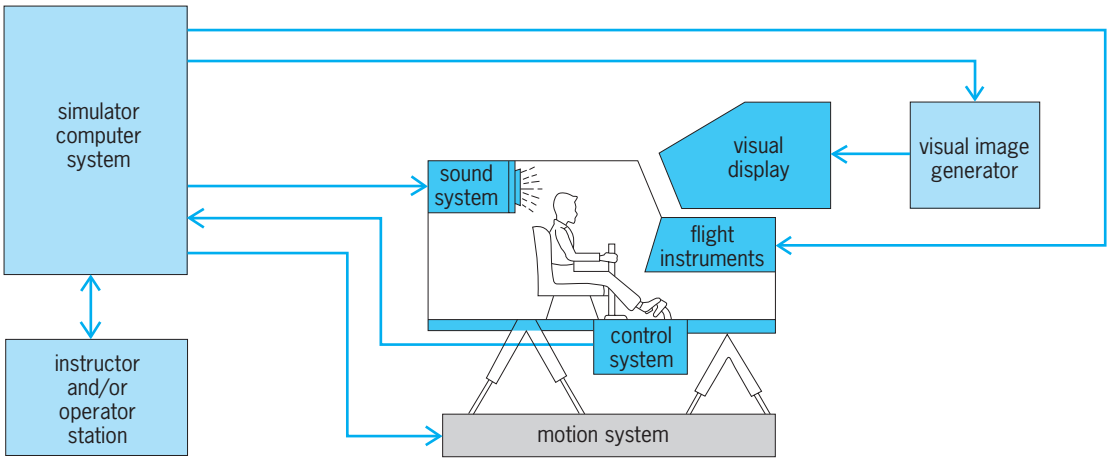


Fig. 7. Diagram of flight simulator illustrating major components and their interconnections.

engine-out approach and landing is very useful but also very dangerous to perform in the airplane. The simulator provides an excellent training medium for this and many other malfunctions.

Components. **Figure 7** illustrates the major components of a piloted flight simulator. The primary component is the cockpit with flight controls and instruments. The other elements are the visual display system, which replicates the out-the-window scene; the motion cuing system, which provides the bodily sensations of the motion of the aircraft; the sound system, which provides the normal and abnormal sounds of flight; the instructor or operator station, which provides the control of the device; and the computer system, which drives the entire complex.

The pilot of a flight simulator manipulates the controls in the simulator cockpit in the same manner as in the aircraft. Sensors in the control system measure this control manipulation and transmit the analog information to the simulator computer system, whereupon the appropriate dynamics of the aircraft are computed by the algorithms in the computer system. These dynamics then are used by the cockpit instrument drive algorithms to provide the appropriate commands to each of the instruments indicating the state of the simulated vehicle, including its engines and other systems. The vehicle state information is also used by the simulator visual system to display the resulting out-the-window imagery and by the simulator motion systems to provide the appropriate motion cues. The sound system is driven by the vehicle and system state information. Algorithms in the computer also determine the proper force which should be felt by the pilot due to the aerodynamic effects on the control surfaces. The calculated force is then transmitted to a hydraulic force-feel system at the cockpit which alters the forces experienced by the pilot.

Software. A key element in this simulation process is the software which controls and performs the simulation. **Figure 8** illustrates the software found in a typical simulator and also shows the hardware interfaces. The flight simulator software comprises three parts: the system software, the instructional or operator software, and the mathematical models which simulate the various aircraft functions. The system software includes the real-time operating system, input-output processing, task scheduling, and so forth. The instructional software is utilized only by training simulators and provides the features necessary for training such as performance monitoring, recording and playback of various scenarios, and maneuver demonstration. The operator software allows the various modes of the simulator to be exercised, such as freeze, which stops the action in place; reset, which allows the operator or instructor to position the aircraft at predetermined locations in space; and fast time–slow time, which allows for faster than or slower than real-time operation.

Mathematical models are required of the aircraft dynamics, the engines, the control system, the avionics systems, the weapon systems, the atmosphere, and aircraft systems such as electrical, hydraulic, and fuel management. A mathematical model is a set of mathematical equations that describe the behavior of a physical system. Figure 8 illustrates the type of mathematical models typically found in a flight simulator and the information flowing among the various modules. The broken lines enclose the portion of the simulator which is implemented in software. The area to the right represents hardware at the flight station, and the area to the left is the instructor or operator station.

The vehicle dynamics simulation comprises the mathematical models contained within the innermost area of Fig. 8. The flight-controls module senses the pilot's control activity and interprets it as control surface deflections or engine commands. These parameters are then passed to the appropriate modules, such as the engine module or the aerodynamics module. The engine module contains the simulation of the engine that is installed in the particular aircraft being simulated. The output of this module is the thrust of each engine, which ultimately contributes to the forces and moments acting on the airplane, as well as to the revolutions per minute, fuel consumption, and other engine parameters which are displayed in the cockpit.

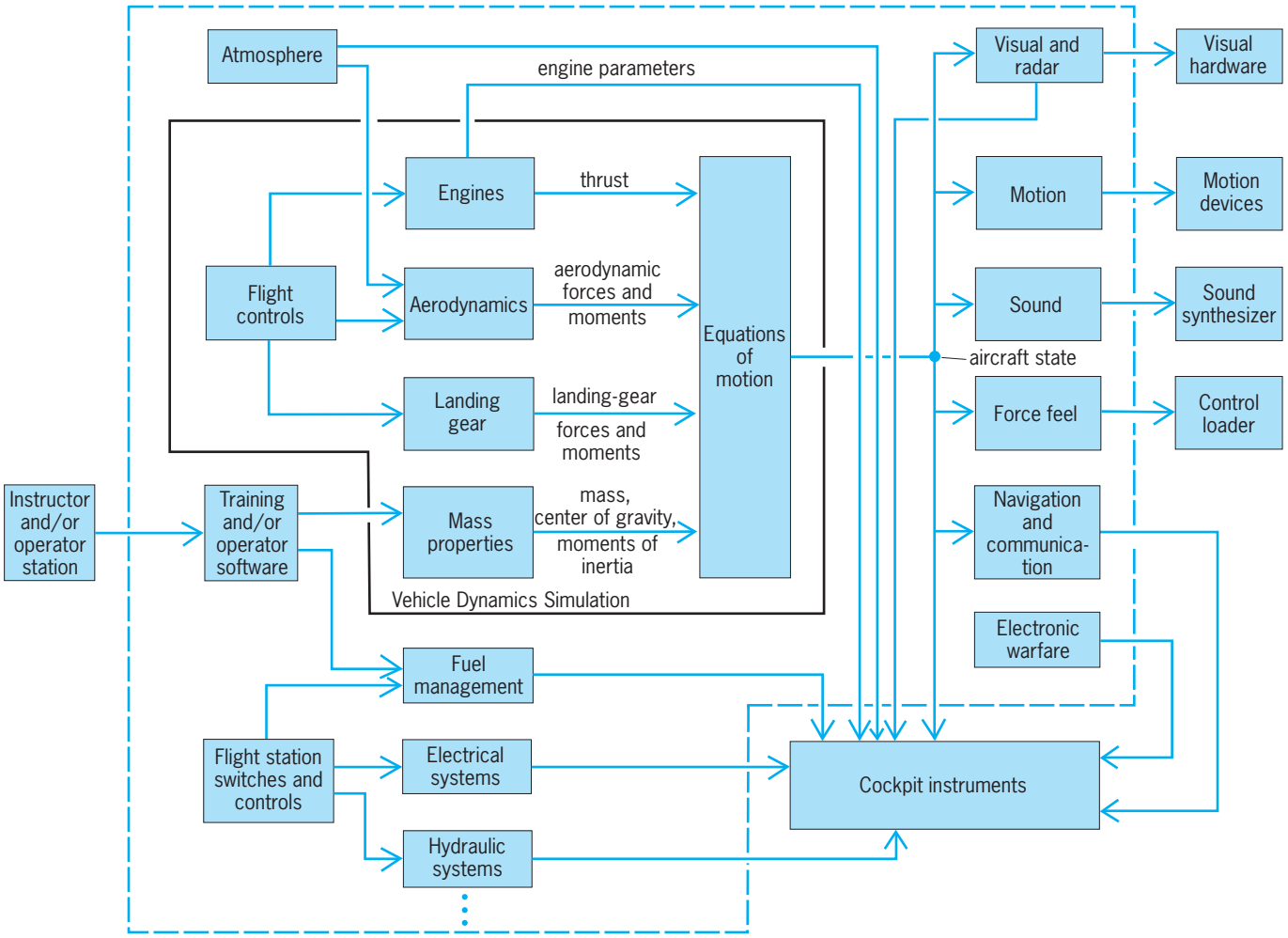


Fig. 8. Simulator block diagram indicating information flow among the software modules, within the broken line, and the interfaces to major hardware elements.

The aerodynamics module computes all the aerodynamic forces and moments acting on the aircraft as a result of control actions by the pilot and the aircraft's interaction with its environment. The forces are drag, side force, and lift. The moments are rolling moment, pitching moment, and yawing moment.

The landing gear module computes the forces and moments acting on the aircraft as a consequence of the interaction of the airplane and the ground in taxiing, landing, and taking off. The mass-properties module computes the vehicle mass and location of the center of gravity as well as the moments of inertia.

The outputs of these four modules are used by the equations-of-motion module to compute the simulated aircraft state vector, which is the aircraft's location in space (latitude, longitude, and altitude) and its orientation relative to the ground (pitch, roll, and heading). The aircraft state vector includes velocity and acceleration along the flight path as well as velocity and acceleration components in the above-mentioned six degrees of freedom. This computation is accomplished by summing the forces and moments due to engines, aerodynamics, landing gear, and weight; computing the acceleration components; integrating these components to obtain velocity components; and integrating again to obtain the spatial coordinates and orientation of the vehicle relative to the ground.

As is illustrated in Fig. 8, the simulated aircraft state vector is passed to various other software modules which utilize it to simulate other aircraft systems or to drive cuing systems such as visual, motion, and sound. The navigation and communication module simulates the aircraft's navigation aids, landing aids, and various communication devices. The electronic warfare simulation may actually be composed of several modules distributed between offensive systems and defensive systems, which may contain mathematical models of threats such as other aircraft or missiles. *See* ELECTRONIC WARFARE.

As shown in Fig. 8, these software modules interface with hardware systems that provide cues to enable the pilot to control the simulated aircraft as if it were a real aircraft. In most cases, the cuing devices synthesize the cues; therefore, the corresponding software modules contain algorithms to effect that synthesis.

Visual simulation. The visual and radar module, in modern flight simulators, formats the aircraft state

vector and environmental information, which is then passed to the image generator. The image generator employs sophisticated computer graphics techniques either to generate imagery for a radar display or to create the view the pilot would have out the windows of the cockpit. The technology employed for generating these scenes has in the past included film, camera–model, scanned transparency, and video disk. However, the most advanced technique is now computer-generated imagery. In addition to image generation, projectors and screens of some sort are required to present the scene properly to the pilot of the simulator. The projectors include cathode-ray-tube and light-valve projectors. The screens include flat or curved screens, dome projection surfaces, and virtual-image displays composed of large, complex arrays of optics. **Figure 9** shows a simulator used to train pilots of the E-3A (AWACS) aircraft. The exterior of the projection screen is at the front of the simulator cab. *See* COMPUTER GRAPHICS.

Motion simulation. Also shown in Fig. 9 below the cab is a platform supported by six hydraulic jacks which are controlled by the motion system software to provide cues to the pilot of the motion environment. While these motion systems are capable of providing motion in all six degrees of freedom in which the aircraft moves, that motion is constrained by the physical limitations of the actuation size and geometry. Therefore, highly sophisticated drive algorithms are required in the software to provide usable cues to the pilot and eliminate false cues. Flight simulators frequently use ancillary motion cuing devices such as seat vibrators to simulate stall buffet and other vibratory motion. In addition, g-seats are sometimes employed to provide sustained cues by stimulating pressure sensations across the back and buttocks of the pilot in the simulator. The anti-g suits commonly found in high-performance aircraft are frequently utilized in simulators to assist the pilot in withstanding the stress of high acceleration, and these have been found to produce very effective cues.

Fig. 9. Training simulator for the Boeing E-3A (AWACS) aircraft. Cabin containing flight deck, exterior of visual display system, and supporting six-degree-of-freedom motion system are shown. (*Rediffusion Simulation*)

Sound simulation. Modern flight simulators often include sound simulation, which is achieved by software control of digital sound synthesizers. These systems are capable of reproducing all aircraft sounds such as engine sounds, aerodynamic noise, and sounds from landing gear and flap deployment. In addition, messages can be given to pilots from simulated air-traffic controllers or on-board warning systems, or maneuver critiques from a surrogate controller.

Cockpit instrument simulation. All or some of the myriad of cockpit instruments are simulated, depending on the purpose of the simulator. In the past, simulator cockpit instrument panels were cluttered with numerous dials, indicators, switches, lights, and other devices by which the status of the simulated aircraft's systems could be monitored and controlled, reflecting the aircraft being simulated. Each of these devices was under computer control. However, modern aircraft have replaced many of these instruments with video display units, and the buttons, knobs, and switches by computer key pads. Hence, the simulator has followed suit, thereby simplifying the cockpit hardware substantially. *See* AIRCRAFT INSTRUMENTATION.

Validation. For simulators to be effective research, design, or training devices, their performance must be validated against the aircraft being simulated to the extent required for the mission of the simulator. This is frequently accomplished by ascribing tolerances to the simulator performance with respect to aircraft flight tests for the vehicle dynamics simulation and with respect to other test data for other systems. *See* SIMULATION.

Frank Cardullo

Flight Testing

Flight testing can be considered the final step in the proving of a flight vehicle or system as capable of meeting its design objectives. This definition applies whether the concept is a complete vehicle, a vehicle subsystem, or a research concept. Flight testing can be categorized as research, development, and operational evaluation. These categories apply both to aircraft and to spacecraft and missiles.

Research testing. The purpose of this form of flight testing is to validate or investigate a new concept or

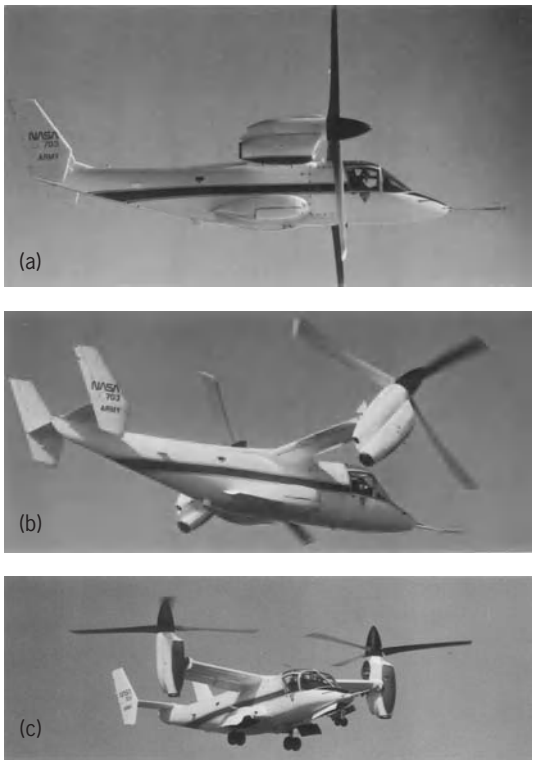


Fig. 10. XV-15 tilt-rotor research aircraft in (*a*) airplane, (*b*) conversion, and (*c*) helicopter mode. (*NASA*)

method with the goal of increasing the researchers' knowledge. Many times, the vehicle used is a one- or two-of-a-kind article designed specifically for the concept being investigated. One such concept aircraft is the XV-15 tilt-rotor (**Fig. 10**), of which two were made. This aircraft takes off with the rotors in a helicopter configuration and then undergoes a transition into a propeller configuration for high-speed forward flight. This concept and its testing proved so successful that it was adopted by the military for a new aircraft, the V-22 Osprey. *See* VERTICAL TAKEOFF AND LANDING (VTOL).

Research testing may also be done with an existing production aircraft to which modifications are made incorporating advanced technology. Such an aircraft is the F-111/AFTI, in which a production F-111 was modified to fly a mission-adaptable wing. Knowledge from research testing is used by designers in producing future generations of systems and is, in general, not applied to current in-service aircraft.

Development testing. A new vehicle or subsystem enters this phase of testing after it has been designed and the basic concepts proven in research flight testing. During this phase of testing, problems with the design are uncovered and solutions are developed for incorporation in the production aircraft.

Development testing involves the use of several preproduction aircraft, each assigned to specific testing tasks. These tasks can include stability and control, engine tests, performance evaluations, flutter and vibration investigations, air loads, avionics, weapons integration, and operational evaluation. These aircraft are highly instrumented, with specialized sensors designed to measure events associated with the task assigned. Testing is performed on all the aircraft simultaneously to expedite testing and field the system more rapidly.

Initial development testing of an aircraft is approached with caution: only after the aircraft has been shown to be controllable are maneuvering tests performed to prove structural or functional integrity to the limits of the design environment. These tests are performed carefully in a step-by-step process in which data from one step indicate if it is safe to proceed to the next step.

Aeroelastic stability testing includes both quasi-steady (divergence and control surface reversal) and dynamic (flutter and vibration) considerations. Quasi-steady aeroelastic effects are usually associated with loss of control effectiveness (the control surface provides less force for the same incidence angle) and a reduction in maneuverability (pitch, roll, or yaw) with higher speeds.

Dynamic flutter is a self-excited, often destructive response of the vehicle structure involving structural rigidity, inertias, and aerodynamic forces. This can be detected by an increase in strain or accelerometer readings as the flight condition for flutter is approached. In this way, with careful airspeed increases and close attention to the sensors, the condition of flutter can be detected before loss of the vehicle occurs. *See* FLUTTER (AERONAUTICS).

Vibration detection and reduction is an important consideration in testing since vibration affects not only the fatigue life of the structure but the reliability and accuracy of the on-board subsystems as well. Included in the fatigue considerations are the gross structural vibrations caused by such sources as power plants and aerodynamic forces and also the low-amplitude effects of acoustic fields. Frequency spectra from on-board accelerometers, strain gages, and microphones are used to determine how to reduce the forcing functions or where to modify the structure to withstand the vibratory forces and enhance fatigue life. *See* VIBRATION.

Air-load data are obtained from strain gages and are used in determining the adequacy of the structure and the accuracy of the designers' predicted loads and load paths. Such data are usually obtained during the initial development flight testing and determine the safe operational flight envelope for the airframe.

Performance testing. Once the final configuration is relatively fixed, performance testing is done to develop aircraft characteristic data. Such data include takeoff and landing distances, rate of climb, best-range airspeed, maximum level flight and limit speeds, and time to climb. Installed engine performance is also evaluated. These data are used to prepare the flight manuals for the vehicle and to verify that the design performance requirements were met.

Avionics testing. The term avionics is a contraction of aviation electronics and refers to the electronic devices and computers used in a flight system. With advances in electronics and computer science, the so-called black box systems have become more sophisticated and in some cases can include millions

of lines of code. Thorough testing of the avionics becomes increasingly important with the complex and information-intensive environment in which the pilot and vehicle have to perform. In general, avionics systems handle communication, identification, navigation, guidance and control, countermeasures, data processing, and information display. In most cases, each of these systems is linked through digital data buses with other systems, and they must interact with each other as well as with other aircraft subsystems such as controls and weapons.

The ultimate aim of avionics flight testing is to develop each system to effectively fulfill definite operational requirements and reliably interface with other systems to create a reliable and viable vehicle avionics system. The electronic subsystem is first tested in a laboratory environment to determine its standalone capabilities and correct any deficiencies found. It is then integrated into the vehicle and flight tested as a complete system to detect any deficiencies in interface design.

System demonstration is performed after the integration testing to determine if the applicable specifications and standards have been met. At this stage, not only is electronic behavior tested to determine correctness of programming, but the behavior of the electronics in the presence of environmental effects is tested as well. The durability and ability of the electronics to withstand the vibration, temperature, humidity, and electromagnetic environment of the operational envelope are determined and any limits are assessed.

Operational testing. This test phase involves customer participation to evaluate the capability of the fully equipped vehicle to meet its intended mission objectives. Testing is performed to determine system reliability, define maintenance requirements, and evaluate special support equipment. Military vehicles are also tested to determine weapon delivery techniques and effectiveness, including target acquisition capabilities, ability to perform in all weather conditions, operational behavior in battlefield conditions, and, in the case of naval aircraft, carrier suitability. Commercial aircraft are tested for blind landing-approach systems, passenger services, baggage and cargo loading, noise levels, and safety provisions. Crew training simulators, handbooks, and procedures are also tested in this phase to demonstrate the ability to maintain and operate the aircraft effectively.

Concurrent testing. An increase in the time necessary to fully test a system is required as the flight systems being fielded become more sophisticated and more electronic. This means that the traditional sequence of testing can delay fielding a system to the point that it is technologically outmoded by the time it is in service. One method of testing that is designed to save time in fielding a system is called concurrent testing. Concurrent testing involves starting production of a system before the development or operational testing is completed, with the idea that any modifications generated by the testing will be made as field upgrades to in-service systems.

Instrumentation. The quantities measured in flight testing have remained basically unchanged in character since the Wright brothers first flew at Kitty Hawk, North Carolina. Aircraft attitudes (such as pitch, roll, and yaw) and velocities (airspeed, pitch rate, roll rate, and yaw rate) are still the basic measurements. However, the accuracy, sophistication, and quantity

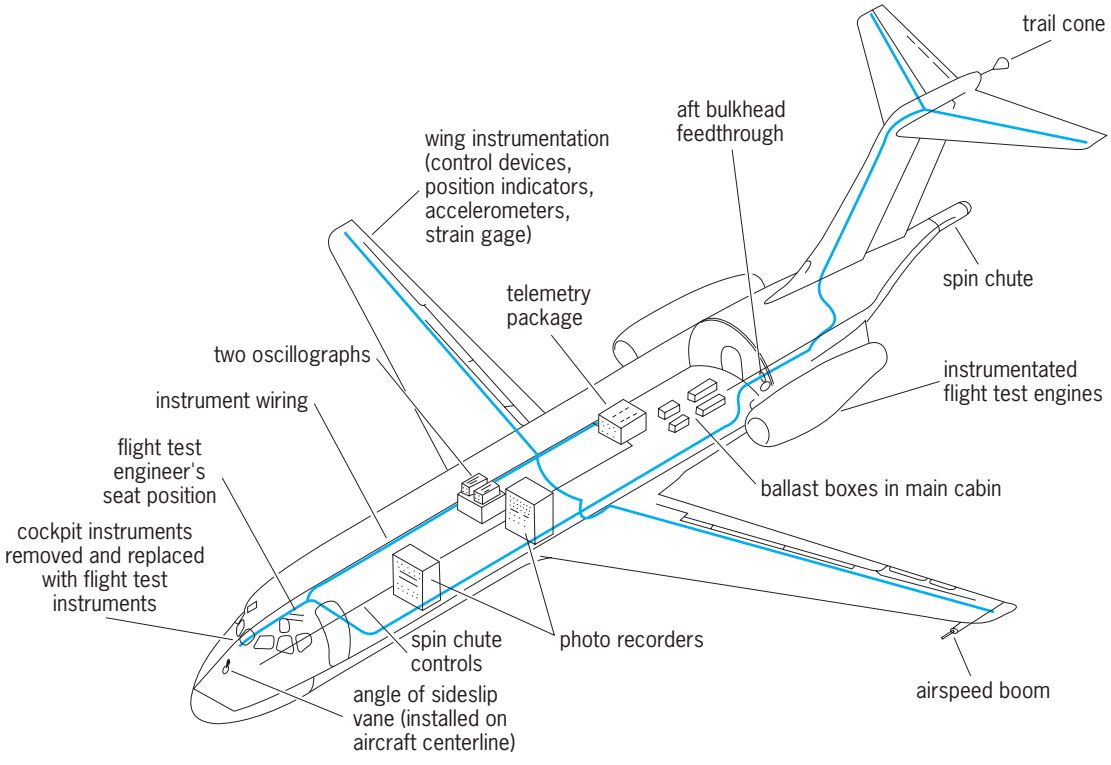


Fig. 11. Test instrumentation installed in a DC-9 aircraft for flight testing.

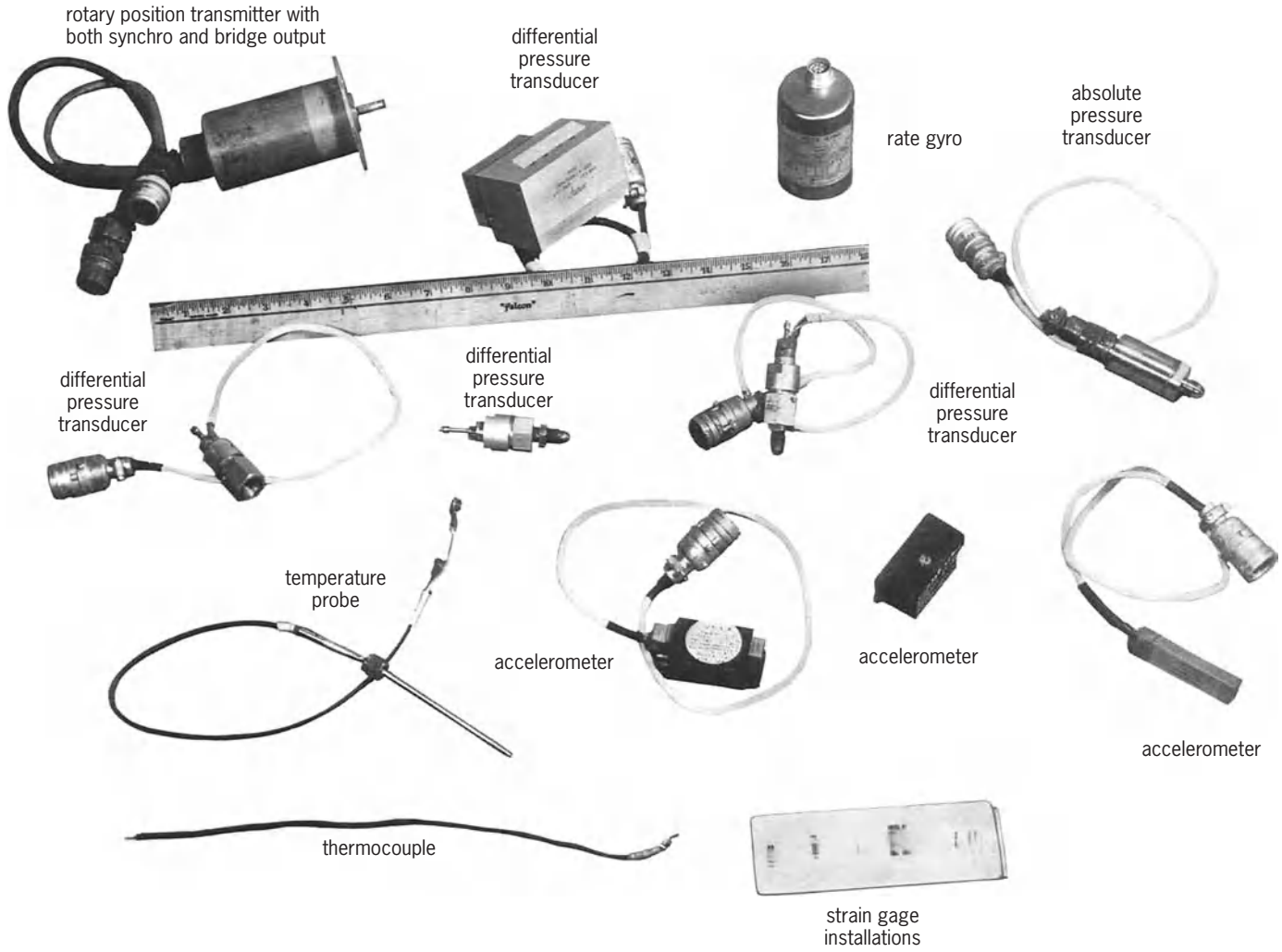


Fig. 12. Examples of miniaturized transducers used in flight testing.

of test data have made tremendous advances with the coming of the digital era. It is not uncommon for millions of bits of information to be present on data buses which transfer digital data from aircraft systems at high speed.

Special-purpose test equipment. To control conditions and to acquire specific data during test flights require the design, development, and installation of test equipment tailored to the constraints of the particular aircraft (**Fig. 11**). Hydraulically actuated aerodynamic vanes are installed at the wing tips to provide sinusoidal motions at controlled frequencies for flutter testing; a mechanically or ballistically deployed parachute is attached to the tail cone to ensure spin recovery during aerodynamic maneuvers; trailing cone airspeed static sources and wing- or nose-mounted airspeed booms are used to obtain distortion-free air samples for airspeed calibration; and dumpable water ballast systems are installed in large aircraft to control weight and center of gravity during flight tests. Specially designed pressure rakes measure inlet duct recovery efficiencies, and motion picture and television cameras document structural motions, information displays, and other test phenomena. Rain and icing environments are provided by flying the test aircraft in the path of water discharged from a tanker aircraft at altitudes where icing temperatures exist.

Test ranges. Test instrumentation can include equipment performing ground measurements as well as on-board systems. Ground instrumentation includes tracking systems (such as radar, laser, and optical) which give the location of the vehicle with reference to the ground. Such ground systems become extremely important in missile testing, in which the trajectory is as important as the vehicle state. *See* OPTICAL TRACKING SYSTEMS.

Transducers. The basic device used to measure phenomena is the transducer. Transducers provide a voltage output, change in impedance, or electric current proportional to the physical event that is being measured. In many cases this output is converted from analog (continuous in time) form to digital (comprising discrete time samples) form at the transducer. Common transducers are accelerometers, pressure sensors, rate gyros, attitude gyros, potentiometers, strain gages, and thermocouples. In general, these are electromechanical devices designed to measure such physical quantities as accelerations, airspeed, surface air pressure, altitude, angular rates about aircraft axes, aircraft attitudes, control and control surface positions, structural loads, and temperatures.

Because of severe weight and space constraints, most instrumentation used in flight testing has been miniaturized (**Fig. 12**).

Most electrical transducers employ the bridge principle, where the sensing element is in one or more legs of the bridge. Strain gages, for example, are strained through mechanical means, resulting in an unbalancing of the bridge with the current or voltage output from the bridge being proportional to the mechanical input. *See* BRIDGE CIRCUIT.

Transducers supply either analog or digital electrical outputs. An analog output consists of one signal which is continuous in time and amplitude. In a digital system the output is a sequence of discrete signals with finite differences in amplitude between them.

Data recording. Signals from the transducers either are transmitted to the ground for real-time data processing and recording or are recorded on on-board tape recorders or other instruments, or both. The real-time data processing and display include the traditional strip chart recorder, in which an ink pin or thermal print head is moved in response to an electrical signal on a moving strip of paper, giving a permanent record of the signals with time. Oscilloscopes can be used to capture events with a photographic record being taken. By far the method of choice is computer analysis and display of digital data. This method yields higher frequency responses and increased resolution. Additionally, digital analysis allows real-time reduction of the incoming data and display of derived parameters to aid the test director in evaluating how well the aircraft is achieving the desired conditions. The real-time ability saves repeating costly test flights to obtain hard-to-get data points, because the test director can ascertain if the data are good while the aircraft is in the air and does not have to wait for processing of the data until after the landing. *See* ANALOG-TO-DIGITAL CONVERTER; GRAPHIC RECORDING INSTRUMENTS; OSCILLOSCOPE; TRANSDUCER.

Oscillographs. The recording oscillograph is an automatic device using miniature galvanometers containing a small mirror that projects a light beam on photosensitive paper or film. As the galvanometer rotates in response to current output from the sensing device, the light beam is deflected across the photosensitive paper. Simultaneously, the paper is transported past an exposure slit at a fixed rate. The processed record thus contains a trace (line) that is a function of time with its deflection proportional to the physical stimulus. Timing lines, correlation counter readings, and a trace numbering system are also provided (**Fig. 13**).

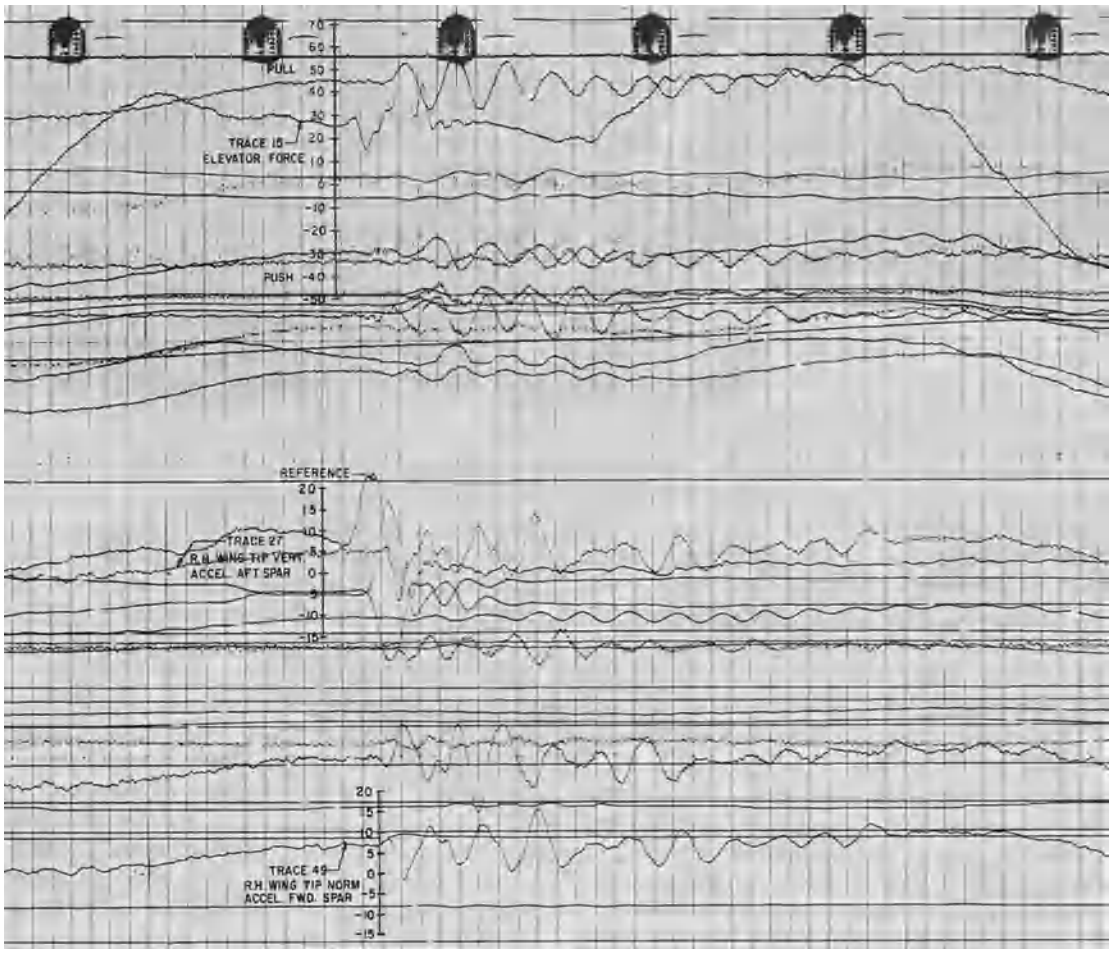


Fig. 13. Typical multichannel flight testing oscillograph.

Two types of recording oscillographs are available. One uses a tungsten lamp, requiring a closed magazine and wet chemical processing for the photosensitive paper. The other uses a mercury vapor lamp (direct write) producing paper lines that are brought into view by exposure to direct sunlight. This process may be accelerated by using a fluorescent latensifier. This oscillograph, because of its dry process and immediate delivery of data, is especially useful for airborne applications.

Oscillographs can record from 1 to 50 different signals. As many as nine recording oscillographs may be installed in one large aircraft.

Magnetic tape recorders. Magnetic tape recorders use the input signals from transducers to excite magnetic tape as a function of time. During playback, electronic equipment decodes the signals and produces a visual display. Alternatively, data from the tape can be processed automatically at high speed by using digital computers and special analyzers. *See* MAGNETIC RECORDING.

Analog magnetic tape recording systems are widely used both in the laboratory and in test aircraft. When a ground-based system is used, data are telemetered via one or more radio-frequency channels to a ground station recorder for future playback. Airborne systems record data signals directly, and the tape record is reproduced on the ground after the aircraft has completed its tests. Often airborne and ground tape recorders are used simultaneously to allow for limited analysis of the data while the aircraft is in flight and to provide a means of saving the data should the test vehicle be lost. Data are recorded in three different formats: direct analog, frequency modulation (FM), and pulse code modulation (PCM). Each format has its particular advantages, and for many tests all three techniques are used simultaneously. A single tape can simultaneously record output signals from several data systems through the use of multiple recording tracks, using a separate tape track for each input. Most instrumentation tape recorders (25-mm or 1-in. tape) contain 14 recording channels, while some have 28 recording tracks. *See* MODULATION.

The direct recording system preserves the varying amplitude input signal in the form available from the transducer. A recording amplifier provides the gain required to produce proper recording head currents. Frequency response of this technique is 100–100,000 Hz. A limitation is that each input signal requires a separate tape track.

FM recording involves converting the input amplitude signals to changes in a carrier frequency. Conversion is accomplished by an oscillator capable of frequency shifts induced by input amplitudes. Zero input signal causes the oscillator to assume a center frequency; plus and minus amplitude inputs produce frequencies above and below the center frequency, respectively. This technique thus results in recording a frequency-varying signal proportional to the amplitude of the input signal. Direct-current response (steady-state input signal) is achieved, and upper-frequency limits of 20,000 Hz can be obtained. As many as 12 separate signals can be recorded on one track of tape by using oscillators with different center frequencies; signals are then separated by filters during playback. Because the signals are in the form of a frequency analog, this system is free from errors caused by amplitude variations, but the technique requires exceptionally close speed regulation of the record and playback tape transport units. An adaptation to the system (called pulse amplitude modulation) is obtained by commutating many input channels into one frequency-modulated oscillator. This multiplexing provides more input channels, but at a drastic reduction in frequency response of each channel. *See* FREQUENCY MODULATION.

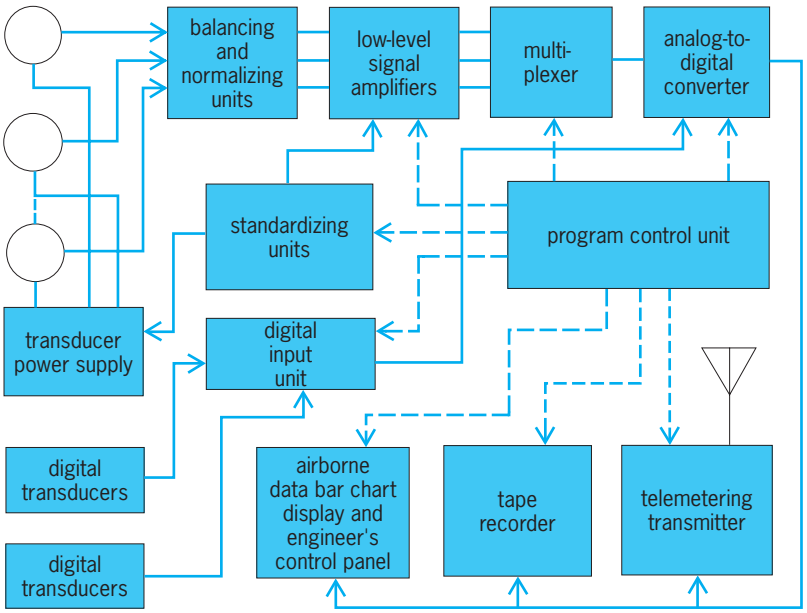


Fig. 14. Simplified functional diagram showing system for digital recording of data in aircraft and for PCM transmission to ground station during flight tests.

The airborne digital data system utilizing the PCM technique (**Fig. 14**) contains all equipment necessary to input, condition, format, record, and transmit data to the ground station. Low-level signal amplifiers multiply the millivolt output of each transducer for input to the analog-to-digital converter. Balancing and normalizing units provide for adjusting the electrical output to zero when zero physical stimulus is applied, and for attenuating the signal to the maximum acceptable input of the amplifier. Standardizing units enable the recording system and the transducer power supply to be calibrated for zero drift and sensitivity during flight. The multiplexor consists of electronic switches that gate transducer outputs in time sequence to the analog-to-digital converter. The analog-to-digital converter functions as an accurate voltage comparator and converts the transducer analog signal to a binary digital signal for recording on magnetic tape. The program control unit is the clock that time-sequences and coordinates all functions.

Sampling and digitizing rates of 1,000,000 data bits per second and tape recording densities of 6250 data bits per inch per track have been attained by using

parallel recording techniques, in which each binary bit level is recorded on a separate track. Developments in tape transport technology permit serial recording of the multiplexed data stream on one recorder track with packing densities up to 14,000 bits per inch. The remaining tracks of a standard 14-track tape transport can be used to record outputs of several data systems or to record direct analog, or FM, data. Digital data may be telemetered by pulse code modulation and recorded on the ground without loss of accuracy. *See* PULSE MODULATION.

Digital systems are characterized by a high resolution and accuracy. Because data are digitized as they are sensed, data contamination is held to a minimum, and once recorded in digital form, no errors are produced in recovery of the information. The system is highly flexible with capacities of 1 to 10,000 channels and frequency response up to 20,000 Hz. Digital systems have the advantage over analog systems that data are recorded in a form that is easily transcribed for entry into digital computers. These systems are normally designed with data-editing and -transcribing facilities as part of an integrated overall data acquisition and processing system.

Aircraft test data processing. Most final test data are processed and analyzed automatically in general-purpose digital computers. A major part of any test effort is the preparation of data for entry into the computer. When a photographic recorder is used, this preparation consists of reading and tabulating the instrument indications and keying them into a computer.

Semiautomatic means may be used to transcribe data from an oscillogram. These devices normally consist of an analog reading device and an analog-to-digital converter. The operator establishes a zero reference on the record and then moves cross hairs connected to potentiometers to intersect the recorded trace at the desired point. The electrical output of these potentiometers, proportional to the displacement of the cross hairs, is fed to the analog-to-digital converter, and the corresponding digital value is automatically entered into the computer. An average operator can prepare 400–600 data values per hour for computer entry by using this equipment.

The major advantage of the all-automatic magnetic tape recording system is the rate at which the desired data can be selected and written on computer magnetic tape for direct entry into the computer. In these systems it usually requires no more time to transcribe thousands of data values from a given test than it does a few. Typical transcription rates are in the order of 5000 data values per second. Thus these systems considerably reduce the cost of data reduction and processing.

Advanced cathode-ray tubes (CRT) for automatic data processing reduce the overall time of final data analysis by permitting a person-machine interface during either real time (telemetry) or post-flight reduction of airborne recorded magnetic tapes. Rather than requiring individual editing, transcribing, and preprogrammed computer data processing, this technique provides a single, high-speed process. The CRT displays, in conjunction with computer interactive hardware and software, provide direct communication with the tape signals by using the digital computer and peripheral processing equipment to receive and decode the incoming data stream. The data engineer controls display format through the use of control keyboards and photoelectric pens. These devices control the path of the data processing by switching to various routines within the computer program, and entering numeric or alphabetic information, such as channel number, scale ranges, and titles. Multiple CRT displays and interactive hardware allow simultaneous processing of more than one flight tape and hasten the data processing on any one flight tape or telemetry stream. As each graph or tabular listing is completed, a reproducible hard-copy print-out can be obtained through a slave CRT copier for a permanent record. These records are available for inclusion in formal test reports. *See* AIRPLANE; CATHODE-RAY TUBE.

Michael Watts; Donald W. Douglas, Jr.

Bibliography. *NASA Aeronautics*, annually; J. M. Rolfe and K. J. Staples (eds.), *Flight Simulation*, 1986; D. T. Ward, *Introduction to Flight Test Engineering*, 2d ed., 1998.

Airfoil

The cross section of a body that is placed in an airstream in order to produce a useful aerodynamic force in the most efficient manner possible. The cross sections of wings, propeller blades, windmill blades, compressor and turbine blades in a jet engine, and hydrofoils on a high-speed ship are examples of airfoils. *See* COMPRESSOR; PROPELLER (AIRCRAFT); TURBINE PROPULSION; WIND POWER; WING.

Geometry. The mean camber line of an airfoil (**Fig. 1**) is the locus of points halfway between the upper and lower surfaces as measured perpendicular to the mean camber line itself. The most forward and rearward points of the mean camber line are the leading and trailing edges, respectively. The straight line connecting the leading and trailing edges is the chord line of the airfoil, and the distance from the leading to the trailing edge measured along the chord line is simply designated the chord of the airfoil, represented by c. The thickness of the airfoil is the distance from the upper to the lower surface, measured perpendicular to the chord line, and varies with distance along the chord. The maximum thickness, and where it occurs along the chord, is an important design feature of the airfoil. The camber is the maximum distance between the mean camber line and the chord line, measured perpendicular to the chord line. Both the maximum thickness and the camber are usually expressed in terms of a percentage of the chord length; for example, a 12% thick airfoil has a maximum thickness equal to $0.12c$.

Aerodynamic forces. The airfoil may be imagined as part of a wing which projects into and out of the page, stretching to plus and minus infinity. Such a wing, with an infinite span perpendicular to the page, is called an infinite wing. The aerodynamic force on the airfoil, by definition, is the force exerted

on a unit span of the infinite wing. For this reason, airfoil data are frequently identified as infinite wing data.

The flow of air (or any fluid) over the airfoil results in an aerodynamic force (per unit span) on the airfoil, denoted by R (**Fig. 2**). The relative wind is the magnitude and direction of the free-stream velocity far ahead of the airfoil. The angle between the chord line and relative wind is defined as the angle of attack of the airfoil, denoted by α. By definition, the component of R perpendicular to the relative wind is the lift, L; similarly, the component of R parallel to the relative wind is the drag, D.

The airfoil may be visualized as being supported by an axis perpendicular to the airfoil, and taken through any point on the airfoil. The point that is one-quarter of the chord distance from the leading edge, the so-called quarter-chord point, is frequently chosen (Fig. 2). The airfoil has a tendency to twist about this axis; that is, there is an aerodynamic moment exerted on the airfoil. By definition, the moment is positive or negative if it tends to increase or decrease respectively the angle of attack (that is, if it tends to pitch the airfoil up or down, respectively). The moment about the quarter-chord point is designated $M_{c/4}$.

Airplane flight. Airplane wings are made up of airfoil sections. Airplanes are sustained in the air by the production of lift on the wings. The physical manner by with an airfoil generates lift is based on the fact that an increase in velocity of air is always accompanied by a decrease in pressure. This is frequently called the Bernoulli effect and is quantified by the differential equation $dp = -\rho V dV$, where ρ is the fluid density, V is the fluid velocity, dV is an infinitesimally small change of velocity, and dp is the corresponding infinitesimally small change in pressure. This equation is simply a statement of Newton's second law (force = mass × acceleration) applied to a moving element of the air. If the velocity increases (dV is positive), the pressure decreases (dp is negative). When an airfoil is producing lift (Fig. 2), the average flow velocity is higher over the top surface and lower over the bottom surface. In turn, the average pressure is lower over the top surface and higher over the bottom surface. The net difference between these pressures results in a net upward force, or lift. For most conventional airplanes in cruise, relative to the freestream pressure the average pressure over the top surface of the wing changes more than the average pressure over the bottom surface; hence typically about 70% of the lift of an airfoil is due to the lower pressure on the top surface. *See* COMPRESSOR.

The reasons that the flow over the top of the airfoil moves faster than that over the bottom are based on another fundamental principle, mass conservation. The lifting airfoil presents a type of obstruction to the flow, and the individual streamtubes of air that curl around the leading edge and flow downstream over the top are squashed, causing the flow in these streamtubes to speed up in order to preserve a constant-mass flow. This occurs even when the airfoil is a flat plate. Indeed, for a flat plate at an angle of attack, the streamtubes that curl around the leading edge are especially squashed, leading to a very large increase in velocity in that region. This is why paper airplanes can fly. However, for such a flat plate at a moderate angle of attack, the flow separates over the top surface and causes a high drag. The smooth, streamlined shape of an airfoil is designed to avoid flow separation; hence airfoils produce lift with only a small amount of drag.

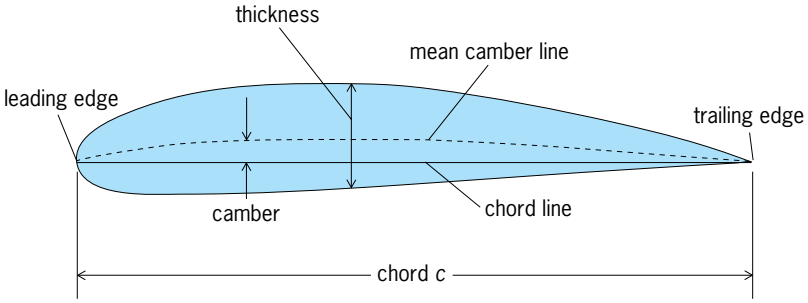


Fig. 1. Airfoil nomenclature. The shape shown is an NACA 4415 airfoil.

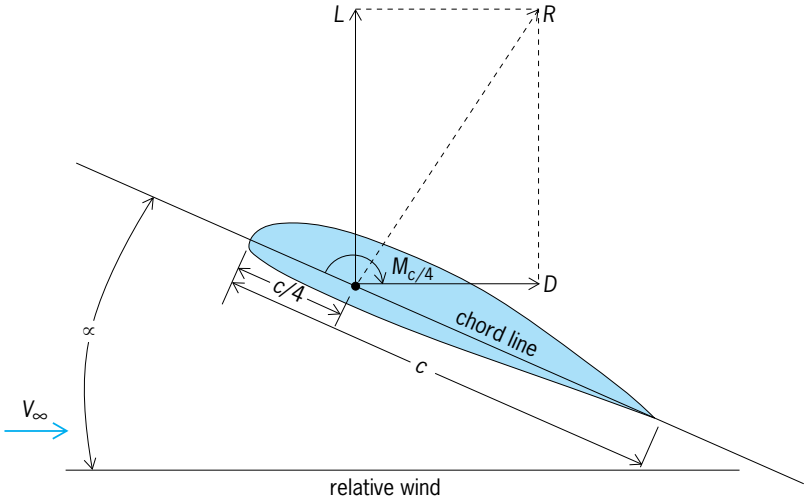


Fig. 2. Definitions of lift, drag, moments, angle of attack, and relative wind for an airfoil.

Dimensionless coefficients. The dynamic pressure q_∞ is defined by Eq. (1), where ρ_∞ and V_∞ denote

$$q_\infty = {}^1/_2\, \rho_\infty V_\infty^2 \tag{1}$$

the free-stream density and velocity. With L and D as the lift and drag per unit span, the airfoil lift and drag coefficients c_l and c_d are defined by Eqs. (2) and (3). Similarly, with M as the moment per unit span, the moment coefficient c_m is defined by Eq. (4).

$$c_l = \frac{L}{q_\infty c} \tag{2}$$

$$c_d = \frac{D}{q_\infty c} \tag{3}$$

$$c_m = \frac{M}{q_\infty c^2} \tag{4}$$

See AERODYNAMIC FORCE; AERODYNAMICS.

The aerodynamic characteristics of airfoils, as couched in terms of the lift, drag, and moment coefficients, are quite different depending on whether the flow is low speed (somewhat less than Mach 1) or high speed (near or above Mach 1). The reason is that the physical characteristics of subsonic flow are quite different from those of transonic or supersonic flow.

In low-speed, subsonic flow over a typical airfoil the lift coefficient c_l varies linearly with the angle of attack α at low values of α. At higher angles of attack, the lift coefficient reaches a maximum value denoted by $c_{l,\max}$, and then decreases as the angle of attack is further increased. In this region, where c_l the lift coefficient rather dramatically decreases, the airfoil is said to be stalled. Positive lift exists at zero angle of attack, and the airfoil must be pitched to some negative angle of attack for zero lift (called the zero-lift angle of attack). This behavior is characteristic of all positively cambered airfoils; positive camber occurs when the camber line lies above the chord line (Fig. 1).

The lift coefficient also depends on the Reynolds number Re, defined by Eq. (5), where μ_∞ is the coefficient of viscosity in the free stream. In the linear region, the Reynolds number has virtually no effect. However, at higher angles of attack, the value of $c_{l,\max}$ is dependent on the Reynolds number; higher values of $c_{l,\max}$ are achieved at higher Reynolds numbers. The drag coefficient c_d is also sensitive to both the angle of attack and the Reynolds number. *See* REYNOLDS NUMBER.

$$\mathrm{Re} = \frac{\rho_\infty V_\infty c}{\mu_\infty} \tag{5}$$

Compressibility effects. High-speed flow over an airfoil, but with a subsonic free stream, depends on the free-stream Mach number, which is the ratio of the velocity to the speed of sound in the free stream. When the free-stream Mach number is higher than 0.3 but still less than 1.0, compressibility effects occur which influence both the lift and drag coefficients, and the variation with Mach number becomes a primary consideration. For subsonic flow, the lift coefficient increases with the free-stream Mach number. A simple, and historically the first, compressibility correction to the lift coefficient is the Prandtl-Glauert rule: If $c_{l,0}$ denotes the low-speed value of the lift coefficient, then the value of the lift coefficient for the same airfoil at the same angle of attack, but with a free-stream Mach number Ma_∞, is given by Eq. (6). The rule holds reasonably accurately up

$$c_l = \frac{c_{l,0}}{\sqrt{1 - \mathrm{Ma}_\infty^2}} \tag{6}$$

to Mach numbers of about 0.7. *See* GAS DYNAMICS; SUBSONIC FLIGHT.

Transonic flow. Airfoil characteristics at high subsonic Mach number are associated with what is known as the transonic flight regime. When the free-stream Mach number approaches 1, the flow field over the airfoil begins to exhibit some dramatic changes. An index for the threshold of these changes is the critical Mach number. When the flow expands over the top surface of the airfoil, the local velocity increases above that for the free stream. Hence, even though the free-stream Mach number is subsonic, if it is close enough to unity there can be regions of locally supersonic flow over the airfoil. By definition,

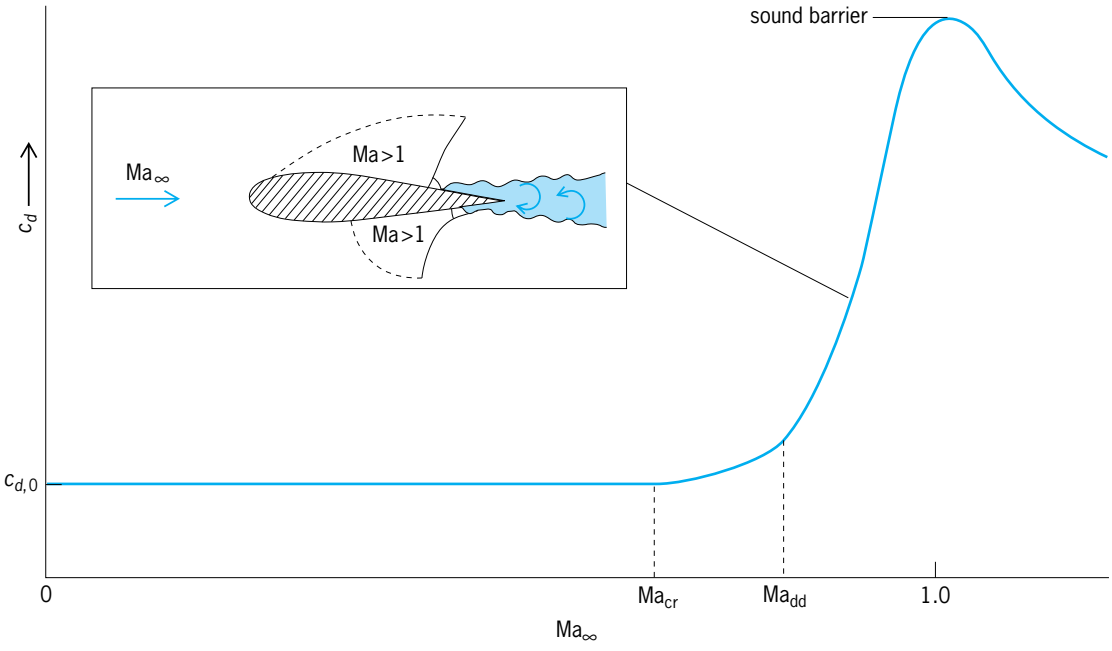


Fig. 3. Variation of the profile drag coefficient with the free-stream Mach number Ma_∞, illustrating the critical Mach number Ma_{cr} and drag-divergence Mach numbers Ma_{dd} and showing the large drag rise near Mach 1. Inset shows the flow field at velocities above that of the drag-divergence Mach number, with large regions of locally supersonic flow (Ma > 1).

for a given airfoil at a given angle of attack, that free-stream Mach number at which sonic flow is first encountered at a point on the airfoil is the critical Mach number. When the free-stream Mach number is increased slightly above the critical Mach number, whole regions of locally supersonic flow occur over the airfoil (**Fig. 3**). These supersonic regions are terminated by a shock wave. If the shock wave is strong enough, its interaction with the boundary layer separates the flow at the shock impingement point. Both the shock wave itself and the flow separation caused by the shock contribute to a massive increase in drag and a decrease in lift.

The onset of these phenomena is best described by considering the measurement of the drag coefficient of a given airfoil at a fixed angle of attack in a wind tunnel, as a function of Mach number (Fig. 3). The drag coefficient remains relatively constant as the Mach number is increased from an initial low value all the way up to the critical Mach number. As the Mach number is increased slightly above the critical value, a finite region of supersonic flow appears on the airfoil. With a further small increase in Mach number, a point is encountered where the drag coefficient suddenly starts to increase. The value of the Mach number at which this sudden increase in drag starts is defined as the drag-divergence Mach number. Beyond the drag-divergence Mach number, the drag coefficient can become very large, typically increasing by a factor of 10 or more. The peak value of the drag coefficient was so daunting to aeronautical engineers in the 1930s and early 1940s that it gave rise to the concept of the sound barrier, that is, that airplanes could not fly faster than sound (faster than Mach 1). Of course, there is no such barrier to supersonic flight. The supersonic transport Concorde cruises at Mach 2.2, and some high-speed military airplanes can fly at Mach 3 and above.

The advent of jet propulsion in the 1940s ushered in the era of high-speed flight and focused attention on designing airfoils with high critical Mach numbers, thus trying to delay the transonic drag divergence to higher subsonic Mach numbers. One well-known approach was to use thin airfoils on high-speed aircraft; everything else being equal, a thin airfoil has a higher critical Mach number than a thicker airfoil. In the 1960s, a new airfoil design philosophy was developed, the supercritical airfoil. For the same free-stream Mach number, the supercritical airfoil is shaped so as to produce a smaller region of supersonic flow with a lower local supersonic Mach number than the more conventional airfoil shape. Because of this behavior, the supercritical airfoil has a larger gap between the critical and the drag-divergence Mach numbers; that is, it can penetrate farther into that part of the free-stream Mach number regime above the critical Mach number before encountering the drag-divergence phenomena. *See* SUPERCRITICAL WING; TRANSONIC FLIGHT.

Supersonic flow. When the free-stream Mach number is greater than 1, the physical aspects of the flow over an airfoil are quite different from those for a free-stream Mach number less than 1. In supersonic flow, shock waves and expansion waves dominate the flow pattern. Consequently, the pressure distribution over an airfoil in supersonic flow is quite different than that for subsonic flow, giving rise to a major component of drag called wave drag. Supersonic airfoil shapes are characterized by sharp leading edges and thin profiles. Both features reduce the strength of the shock waves and hence reduce the wave drag. *See* SUPERSONIC FLIGHT.

John D. Anderson, Jr.

Bibliography. I. H. Abbott and A. E. von Doenhoff, *Theory of Wing Sections*, 1949; J. D. Anderson, Jr., *Aircraft Performance and Design*, 1999; J. D. Anderson Jr., *Fundamentals of Aerodynamics*, 4th ed., 2005; J. D. Anderson, Jr., *A History of Aerodynamics*, 1997; J. D. Anderson, Jr., *Introduction to Flight*, 5th ed., 2005.

Airframe

The structure consisting of the wings, fuselage, and fin and tail plane (vertical and horizontal stabilizer) of an airplane. This structure has a clearly defined role: to resist all the extreme loads that can be experienced with no danger to the passengers, whether the aircraft be civil or military. The temptation therefore could be to overdesign the airframe and "play safe." But an overweight civil aircraft would be commercially disastrous, and an overweight military aircraft could be outperformed. To put this matter into perspective, a civil airframe can be up to 40% of the total dry weight, whereas the payload may be as low as 20%. Thus a structure which is overdesigned by only 10% could reduce the payload by 20%. This represents the profit margin on many routes.

Design loads. An airframe has to withstand two different types of loading: a maneuver and a sudden gust. A maneuver is intentional, and when a pilot increases the aircraft incidence by applying a download on the tail plane the main wing experiences a sudden increase in lift and accelerates upward. The resulting inertia forces experienced by all the masses can be visualized as balancing the extra lift (**Fig. 1**). This extra inertia is conventionally referred to as the number of g, where 1 g is the normal weight due to gravity. For an agile combat military aircraft the specification may call for a factor of 9 g or more, depending on the assistance given to the pilot by pressurized suits and whether the maneuver lasts for

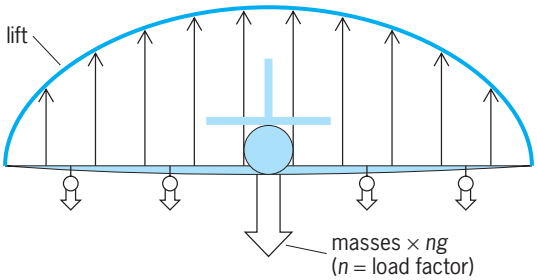


Fig. 1. Forces acting on the main wing during a maneuver.

many seconds. For a civil aircraft the current allowable acceleration is +2.5 *g* or −1 *g* "downwards." *See* AERODYNAMIC FORCE.

Civil aircraft also have to be designed to resist sudden gusts. A vertical upward gust of velocity u will effectively increase the incidence of the aircraft by u/V radians, where V is the aircraft velocity. Currently the maximum specified gust is 20 m/s (or 45 mi/h). This figure represents the maximum ever recorded in flight. However, all maneuver and gust loadings are multiplied by 1.5 as a safety factor when designing civil aircraft. This somewhat arbitrary figure has withstood the test of time. Structural failures are almost unheard of, provided the specified maintenance is followed. There have been a very small number of cases where pilots have overloaded an aircraft, but nowadays a computer between the pilot and the controls will impose software limitations on control forces. Certain parts of the airframe may be designed to withstand rather special loads, such as the undercarriage in heavy landings.

Materials behavior. To discuss airframe design, it is first necessary to describe the behavior of airframe materials. The simplest way to test a material (say, a simple rod) is to pull it and measure the extension due to the applied force. However, a better measure of the actual forces on the material molecules or crystals is the stress (σ), that is, force divided by cross-sectional area. Similarly, a better measure of the crystal deformation is the strain (ε) or fractional change in length of the rod. A final important property is the material's stiffness (E), which is the ratio of stress to strain for small strains. Some typical stress/strain curves for aircraft materials are shown in **Fig. 2**. *See* STRESS AND STRAIN.

The most common structural material in use worldwide is mild steel for which the strain is proportional to stress (a linear curve) until "yield," when the crystals start to move along their boundaries or dislocations. This material then deforms at virtually constant stress, and then when unloaded will do so linearly as shown in Fig. 2, leaving a permanent deformation of, say, 7%. This particular property is extremely valuable since it allows thin steel sheets to be pressed into everything from car bodies to domestic cutlery.

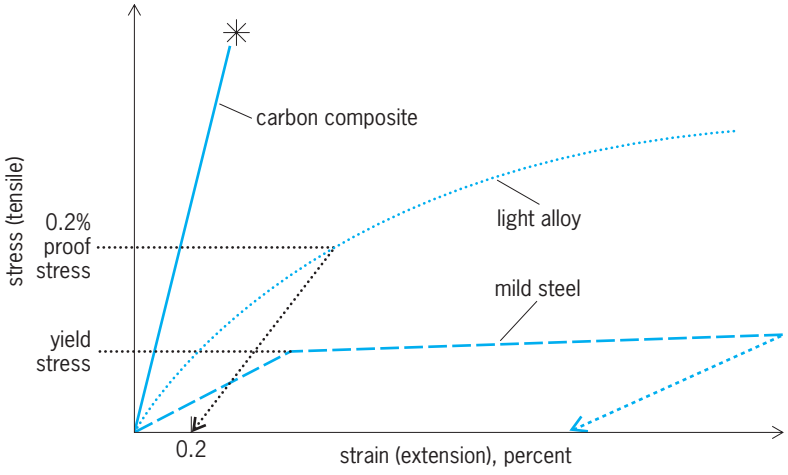


Fig. 2. Stress/strain curves for various types of material.

Fig. 3. Effect of wing box strains of 0.7%.

A similar behavior accurs in aluminum, but by adding small amounts of copper, zinc, magnesium, and so forth, the dislocations can be inhibited at key points. The maximum stress can therefore be increased as shown in Fig. 2. Light aluminum alloys have been the most common airframe material from the early days. To define the maximum allowable stress (in the absence of a true yield), the aircraft sector has concentrated on the allowable strain when the material is unloaded. (Clearly an aircraft should return from flight with the same shape as it took off.) The atypical allowable is a strain of 0.2%, and the corresponding stress is called the 0.2% proof stress. Although these strains may seem very small, an aircraft wing, for example, is not a very efficient structure in bending, and if the strains reached 0.7% the deformations would look something like **Fig. 3**. *See* METAL, MECHANICAL PROPERTIES OF; PLASTIC DEFORMATION OF METAL.

To obtain higher failure stresses, by avoiding dislocations and extending the linear range, it is necessary to turn to brittle materials like glass, carbon, boron, and various ceramics, all of which have a much purer crystalline structure. However, unlike ductile metals, these materials are very susceptible to flaws and minor cracks. In fact, the standard way that a glazier will cut glass is by inscribing a shallow surface scratch and then propagating it. One way of overcoming this 'notch sensitivity' is to manufacture extremely fine fibers and then embed millions of them in a resin matrix. Fibers still have surface flaws, but typically the distance between flaws is about 15–25 diameters for carbon fibers. Thus, when a single fiber breaks, the load is diffused onto the surrounding fibers since the chances of all fibers having cracks at the same location are zero. A common way of using carbon fiber composites is to make a thin lamina of about 0.13 mm (0.005 in.) thickness in which all the thousands of fibers lie in one direction. A simple laminar (or ply) is extremely stiff and strong in the direction of these fibers but very weak in a direction at right angles to them (like balsa wood). A composite wing skin, for example, is then built by stacking a large number of plies in various directions to suit the nature of the internal stress field. Composite structures are consequently "tailored," unlike structures of homogeneous metals.

Carbon composite structures have dominated combat military aircraft since the 1980s, since they are exceptionally stiff and strong. The **table** shows

Properties of some materials used in aircraft structures

Material	E, GPa	σ, MPa	ρ, Mg/m^3	E/ρ	σ/ρ
Aluminum alloy	71	480	2.8	25	172
Titanium	110	1000	4.4	25	226
High-tensile steel	207	1720	7.8	26	221
Carbon composite	200	1500	1.55	130	970
Glass composite	40	840	1.80	22	460
Kevlar composite	82	1500	1.39	60	1080

the values of strength (ultimate tensile strength, σ) of stiffness (E), and of density (ρ) of some materials used in aircraft structures. The specific stiffness and strength (ratios of stiffness and strength, respectively, to density) are the crucial measures since there is no point in having a superior material that is too heavy.

The values in the last two columns of the table indicate that the specific stiffnesses of the three metals are comparable, but clearly carbon composites are superior. Carbon also has better specific strength, matched only by Kevlar. This artificial aramid fiber is cheaper than the other structural materials but is competitive only in tension; it has a very poor compressive strength. *See* COMPOSITE MATERIAL; MANUFACTURED FIBER.

Wing design. An aircraft wing is not a very efficient structure in resisting bending due to gust or maneuver loadings. Clearly, aerodynamic efficiency is the driver, and hence a wing has a small thickness/chord ratio, as low as 1/20 for high-speed military aircraft. When a wing bends due to aerodynamic lift, the top surface goes into compression, the lower into tension. For maximum efficiency, most of the structural material should go to these surfaces, as far as possible away from the center of the section (the neutral axis), where the bending strains are zero. The ubiquitous I section, used in buildings and so forth, is a good example. However, a wing also has to resist torsion, and should twist as little as possible to avoid changing the effective angle of incidence. An I section is hopeless at resisting torsion. For maximum torsional stiffness, a thin-walled structure should be a closed "tube" with as large an enclosed area as possible. In a wing, a fuel tank also fulfills this specification.

A typical wing section, with a front and rear spar, is shown in **Fig. 4**. The trailing edge also contains the flaps and ailerons, and the leading edge may contain deployable slats for increased lift. The upper skin, when in compression, has to be designed to resist buckling. The stress at which a thin plate buckles is proportional to $(t/b)^2$, where t is the plate thickness and b is the width. This property explains the large number of stiffeners shown diagrammatically in Fig. 4 and intended to reduce the effective width b. The buckling stress is also proportional to $(1/l)^2$, where this time the plate length l is involved and is reduced by inserting ribs into the wing box. These ribs are quite light since their only function is to stop the outer skin's buckling and to maintain their shape for aerodynamic reasons; a large aircraft may have 40 or 50 of them along the wing. Some ribs are heavily loaded, however; for example, the engine loads have to be diffused into the wing box. **Figure 5** shows an engine mounting rib (for the Airbus 380), which is clearly heavy and has an array of stiffeners (machined from the solid) that inhibit buckling under the high shear forces coming from the engine. Where the top edge of the rib is flanged (to be bolted to the wing skin), the cut-outs are seen to allow the wing skin stiffeners to pass through without interruption. *See* AILERON.

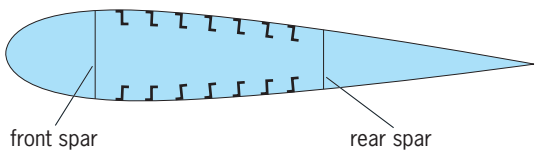


Fig. 4. Typical wing box section.

Fig. 5. Wing rib for taking engine mounting on Airbus 380 aircraft. (*Courtesy of Airbus UK*)

The fin (vertical stabilizer) and the tail plane (horizontal stabilizer) have very similar structures to the main wing, although they are much lighter. Actually, their structures should be exceptionally light since they are at the extreme end of the aircraft and if too heavy would adversely move the aircraft's center of gravity. For this reason all Airbus (and now Boeing) fins and tail planes have been made of carbon composites.

A final hazard for the wing designer is the interaction between the wing deflections and the aerodynamic forces: an aeroelastic effect. If the wing lacks sufficient torsional stiffness, it may twist and thereby increase the aerodynamic lift due to the increase in incidence. At a critical divergence speed, the twist may grow beyond limit and lead to structural failure. Another phenomenon, called flutter, occurs when the balance between aerodynamics, structural stiffness, inertia forces, and damping leads to oscillations that may grow without limit. In the absence of aerodynamic forces, the vibration of any structure will gradually subside due to natural material damping. However, the aerodynamic lift and pitching moment are proportional to the surface velocity and result in negative damping. One solution is to introduce mass balancing for controls like the aileron and rudder, so that the extra inertia forces dampen the

oscillations. On a main wing, the forward-placed engines fill this role and ensure a sufficiently high flutter speed. *See* AEROELASTICITY; AIRCRAFT RUDDER; FLIGHT CONTROLS; FLUTTER (AERONAUTICS); WING; WING STRUCTURE.

Fuselage design. The fuselage has a much larger cross section than a wing and is consequently much better able to resist bending and torsion. A fuselage skin may be only 1–2 mm (0.04–0.08 in.) thick, compared to wing skin thicknesses in excess of 20 mm (0.8 in.). Parts of the fuselage also experience bending compression during landing and takeoff, for example. Therefore, as with wings, there is a need for stiffeners and ribs, although the "ribs" are really frames of depth 10–20 mm (0.4–0.8 in.). Heavier frames are needed to take the wing loads into the fuselage, for example, or at undercarriage pickup points. At the rear of the fuselage, the fin or tail plane is often connected to the curved bulkhead, designed to take the cabin pressure loading.

Although the working stresses in a fuselage can be low, the design case is really one of fatigue. On every flight, the external pressure at altitude falls to a very low value, while the cabin pressure is maintained at about half an atmosphere for comfort. Thus, in an operating lifetime the fuselage stress field will be cycled thousands of times. All metallic materials can fail at a relatively low stress after a sufficiently large number of cycles. The response of carbon composites is much better since the loading and unloading seems to be reversible with no material damage even at the crystal scale.

The presence of windows and doors can exacerbate the fatigue problem by introducing local stress concentrations, but designers now know how to design window and door frames to minimize these concentrations. Unfortunately, this was not the case in the 1950s when the Comet jetliner, the first civil aircraft to cruise at altitudes in excess of 10,000 m (30,000 ft), had fuselage fatigue failures in three aircraft before it was grounded. *See* FUSELAGE.

Composite structures. The superiority of carbon composite materials has already been mentioned, that is, their strength, stiffness, and fatigue performance. In spite of their basic advantages, there are two disadvantages that have to be overcome, and indeed mostly have been.

1. Carbon composite materials are susceptible to stress concentrations. Laminated thin plates are strong in their own plane but have little resistance to peeling apart of the individual plies. They can easily delaminate at the matrix resin strength. Ideally, all laminates should be in a two-dimensional stress field. But stresses will be three-dimensional near any discontinuity of geometry or load. Careful design and local reinforcements can eliminate this problem. The Airbus 380, for example, has a totally composite center wing box where the two wings intersect the fuselage. This is a region full of joints and discontinuities, and the design is a tribute to the skill of manufacturers and designers who are able to model this problem and analyze the stress fields.

2. Carbon composites can be expensive. Carbon and boron fibers are costly to produce in small numbers. For a long time, the total output of carbon fiber for sporting goods exceeded that needed by the aerospace industry. Only military aircraft could disregard this expense since in that case a poor performer demanded rejection.

Not only was the basic material costly, but so were the manufacturing methods. Hand lay-ups of many thin plies (possibly up to 80 in military aircraft) are costly. This cost has gradually been reduced by replacing hand lay-ups with computer-controlled automatic machines, which lay a large number of overlapping tapes with great precision. Another breakthrough has been resin transfer molding, in which a dry fabric can be sandwiched between two faces of a mold and then the resin injected, before cure, under pressure or sucked through under vacuum.

The Boeing 787 is an impressive example of a civil aircraft with composite wings and fuselage, having overcome all these problems. Possibly the only threat to the economics of mass-produced airframes is the future legal requirement that manufacturers recycle carbon structures when their operating life is over.

Virtual testing. To satisfy either military or civil airworthiness authorities, aircraft have to be tested to ultimate design loads. Prior to this, the manufacturer will have made many component tests on panels, ribs, undercarriages, and so forth. These very expensive and time-consuming tests have been traditional since no theories were available for highly three-dimensional and detailed components. However, since the 1980s remarkable progress has been made in the ability to simulate in a computer model the behavior of structures loaded to failure. Basically there are only three sorts of equations to satisfy in theoretical structures: (1) the stresses everywhere have to be in equilibrium with themselves and the applied forces; (2) the strains have to be related to the displacement field (purely a geometrical argument); and (3) the two are linked by a stress/strain law based on material tests. The main problem has been the enormous geometrical complexity of aircraft structures at the detailed level, but the advent of the digital computer (and later

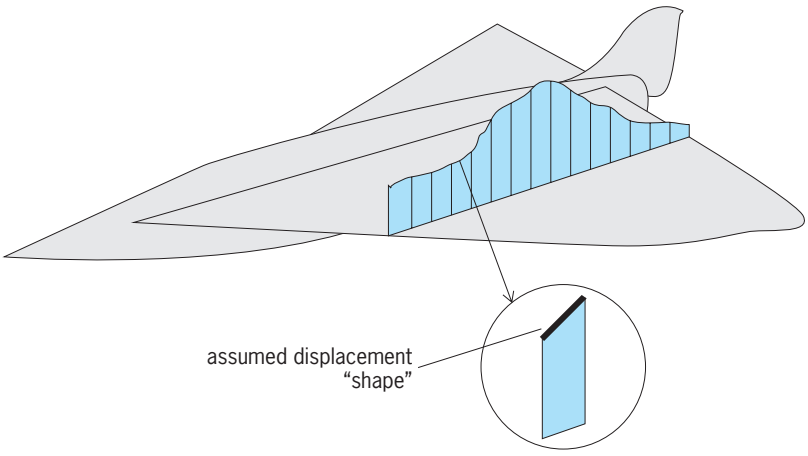


Fig. 6. Finite element approximation.

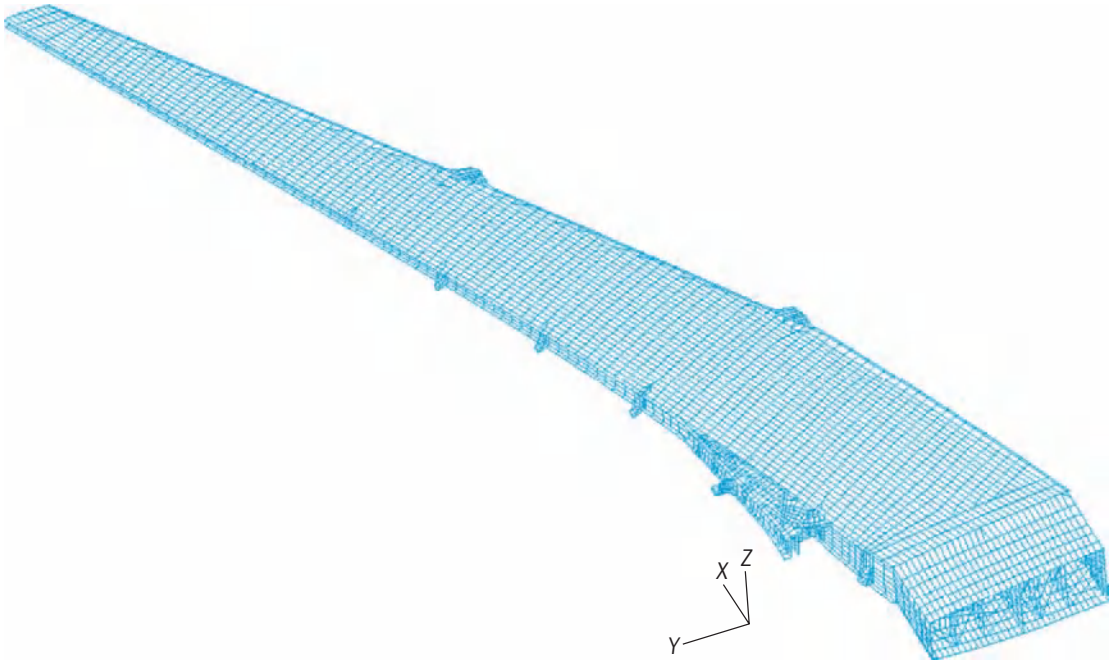


Fig. 7. Finite element model of a wing box. (*Courtesy of Airbus UK*)

computer-aided-design software) has changed this forever. The finite element method is conceptually simple. A typical displacement field across a wing from leading to trailing edge is illustrated in **Fig. 6**. The structure is then divided up into finite elements which are so small that a reasonable approximation can be made for the shape of the variation across any element. In Fig. 6 a linear variation is implied. Then, all that remains is the magnitude of the displacements at, say, the edges of the elements. The strains can then be expressed in terms of these magnitudes, the strain/stress law implemented, and finally the equilibrium equations. The result is a large number of equations giving the many displacements in terms of the applied forces. The number of displacement unknowns and equations can run into hundreds of thousands, but modern computers are powerful and inexpensive enough to cope with problems of this magnitude. A typical finite element display is shown in **Fig. 7**. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; FINITE ELEMENT METHOD.

Commercial computer codes have been available since the 1970s, but only recently has it been feasible to simulate the failure process, which is nonlinear as damage propagates and which modern codes can simulate. Virtual testing is now a reality and more informative than field testing since the internal stress fields are modeled in detail—for example, when a bird strikes an aircraft or is sucked into an engine fan. *See* AIRCRAFT TESTING. Glyn A. O. Davies

Bibliography. H. D. Curtis, *Fundamentals of Aircraft Structural Analysis*, Times-Mirror HE Group, Los Angeles, 1997; M. Davies (ed.), *The Standard Handbook for Aeronautical and Astronautical Engineers*, McGraw-Hill, New York, 2003; T. H. Megson, *Aircraft Structures for Engineering Students*, 3d ed., Edward Arnold, London, 1999.

Airglow

Visible, infrared, and ultraviolet emissions from the atoms and molecules in the atmosphere above 30 km (20 mi), generally in layers, and mostly between 70 and 300 km (45 and 200 mi).

The airglow, together with the ionosphere, is found in the uppermost parts of the atmosphere that absorb the incoming energetic radiations from the Sun. While the airglow consists of spectral features similar to those of the aurora, it is mostly uniform over the sky; and it is caused by the absorption of solar ultraviolet and x-radiations, rather than energetic particles. *See* AURORA.

Dayglow. The daytime airglow (dayglow) is caused mainly by fluorescence processes as molecules and atoms are photodissociated and photoionized. The photoelectrons that are produced in the ionization processes are a further source of airglow in their collisions with other atoms and molecules. Emissions produced by recombination of previously produced neutral species and ions are a minor source of dayglow, and resonant and fluorescent scattering of sunlight also contributes. *See* FLUORESCENCE.

Only during twilight is the blue sky glow that is caused by sunlight scattered on the lower atmosphere sufficiently weak that some dayglow emissions can be observed from the ground. Thus, twilight offers an opportunity to observe resonant scattering of sunlight on layers such as those of the alkali atoms sodium, lithium, and potassium. As the Earth's shadow scans through the layers, the changes of intensity allow their heights (near 90 km or 55 mi) to be measured. The atoms can be observed by lidar (laser radar) with similar results but more detail. *See* ALKALI EMISSIONS; LIDAR; SCATTERING OF ELECTROMAGNETIC RADIATION.

Nightglow. The nighttime airglow (nightglow) is predominantly due to recombination emissions. The ionospheric plasma recombines near the bottom of the F region (150–200 km or 90–120 mi) where the densities and thus collision frequencies are higher, producing bright atomic oxygen (O) spectral lines in the red (at 630 and 636 nanometers) and a weaker green (558-nm) line. The recombination of the neutral radicals that are generated in daytime photodissociation proceeds fastest in the mesosphere, where the densities are higher at the bottom of the region of production.

Atomic oxygen recombination forming excited dioxygen (O_2) molecules leads to several characteristic spectral-band systems. The recombination energy can alternatively be transferred to an oxygen atom in collision and can become the source of a strong green (558-nm) emission. Excited oxygen atoms that could produce the red (630- and 636-nm) lines in the mesosphere are nearly always rapidly quenched in another collision. *See* ATOMIC STRUCTURE AND SPECTRA.

Strong infrared hydroxyl (OH) emission arises in the mesosphere from the reaction of atomic hydrogen (H) with ozone (O_3), and the H is regenerated by the OH reacting with O to produce O_2, with the net result being the recombination of O and O_3 and the recycling of the H. Similar cyclic reactions excite sodium atoms. A weak continuum in the green is due to the recombination of nitric oxide (NO). The total of all the visible nightglow emissions, together with zodiacal light and scattered starlight, can be seen with the naked eye as the faint light between stars. From space, the edge-on view of airglow is of a band of light above the horizon. *See* ATMOSPHERIC OZONE; ZODIACAL LIGHT.

Observations. Observations made by spacecraft are free of the constraint of the blue sky glow, and can also detect the ultraviolet emissions that cannot be seen from the ground. Such observations of resonant ultraviolet scattering of oxygen ions (O^+) and helium ions (He^+) can be used to map the distribution of ions in the plasmasphere to distances to four times the Earth radius. Scattering in the Lyman series maps the so-called geocorona of atomic hydrogen out to 15 times the Earth radius.

A variety of remote-sensing techniques from the ground and from space utilize airglow to determine the composition and structure not only of the Earth's upper atmosphere but also of the atmospheres of other planets. Spectroscopic observations of relative intensities and Doppler profiles identify the composition, temperature, winds, and energy inputs, as well as ionospheric layer heights and ion concentrations. Images of extended areas of terrestrial airglow show wave structures caused by atmospheric gravity waves. These waves originate deep in planetary atmospheres, and their amplitude grows exponentially with height; their dissipation is an important source of heating of upper atmospheres.

Images made of the ionospheric recombination emissions, looking toward the magnetic equator from low-latitudes sites on Earth, show large-scale structures due to plasma instabilities.

The airglows from other planets are remarkably different from that of Earth. The airglows from Mars and Venus show strong emissions of carbon-oxygen species (CO, CO^+, CO_2^+), O, and H; from Titan emissions of molecular nitrogen (N_2) and atomic nitrogen (N); and from Jupiter and Saturn, emissions of molecular hydrogen (H_2) and H. These emissions confirm that CO_2, N_2, and H_2 are the dominant constituents in each case.

Brian A. Tinsley

Bibliography. J. W. Chamberlain, *Physics of the Aurora and Airglow*, 1961; J. W. Chamberlain and D. M. Hunten, *Theory of Planetary Atmospheres*, 2d ed., 1987; G. Paschmann et al., *Auroral Plasma Physics*, 2003; M. H. Rees et al., *Physics and Chemistry of the Upper Atmosphere*, 1989.

Airplane

A heavier-than-air vehicle designed to use the pressures created by its motion through the air to lift and transport useful loads. Although airplanes exist in many forms adapted for diverse purposes, they all employ power to overcome the aerodynamic resistance, termed drag, thereby achieving forward motion through the air. The air flowing over specially designed wing surfaces produces pressure patterns which are dependent upon the shape of the surface, angle at which the air approaches the wing, physical properties of the air, and velocity. These pressure patterns acting over the wing surface produce the lift force necessary for flight. *See* AIRFOIL.

To achieve practical, controllable flight, an airplane must consist of a source of thrust for propulsion, a geometric arrangement to produce lift, and a control system capable of maneuvering the vehicle within prescribed limits. Further, to be satisfactory, the vehicle should display stable characteristics, so that if it is disturbed from an equilibrium condition, forces and moments are created which return it to its original condition without necessitating corrective action on the part of the pilot. Efficient design will minimize the aerodynamic drag, thereby reducing the propulsive thrust required for a given flight condition, and will maximize the lifting capability per pound of airframe and engine weight, thereby increasing the useful, or transportable, load.

Air propulsion systems. Because the airplane moves through a fluid medium, its propulsion system must produce a change in the momentum of its working fluid to generate the thrust force required to overcome aerodynamic drag. Reciprocating and turboprop engines produce power by converting chemical energy of fuel into rotation of a shaft; a propeller fitted to this shaft accelerates the air passing through its rotational area, thereby changing engine power into useful thrust. Turbojet and ramjet engines produce thrust directly by adding heat to the air passing through the engines and ejecting the resulting hot exhaust gases at high velocity. A fanjet (turbofan) engine is a hybrid between the turboprop and the pure turbojet that combines a number of features common to both. The turbojet, ramjet, and fanjet en-

gines use the surrounding air both as primary working fluid and as oxidizer to support combustion. *See* JET PROPULSION; PROPELLER (AIRCRAFT); RAMJET; RECIPROCATING AIRCRAFT ENGINE; TURBOFAN; TURBOJET; TURBOPROP.

Thrust. The engine-propeller combinations, whether reciprocating or turboprop, are effective at low forward speeds, but their thrust falls off rapidly as flight speeds approach the speed of sound (Mach 1), largely because of the tremendous increase in power required to turn the blades when shock waves form at their tips. The turbojet thrust increases with speed because of aerodynamic compression, which becomes more marked at higher speeds. The fanjet, by combining many of the characteristics of the propeller and the turbojet, provides efficient propulsion in the range of speeds between those where the thrust of the propeller falls off rapidly owing to the compressibility drag rise at the tips, and the thrust of the turbojet picks up. The limit of turbojet operation is set largely by the temperature limitations of the materials used to fabricate the turbine and by shock patterns which may produce unsteady flow and separations at high speeds, interfering with efficient aerodynamic compression, creating difficulties with the intake system, and causing the engine to stall and flame out. *See* SUPERSONIC DIFFUSER.

The ramjet cannot produce thrust at low forward speeds because it depends upon an appreciable aerodynamic (ram) compression. It therefore requires a booster system to raise the airplane speed sufficiently that useful thrust can be produced. Because of improved aerodynamic compression, as with the turbojet, the thrust produced increases as the velocity of flight increases. Since a ramjet is designed for maximum performance at a given flight condition, its intake design may be such that its range of operating speeds is small.

Weight and fuel consumption. Ramjet engines, although light, have large fuel consumptions, whereas the heavy reciprocating engine has the lowest consumption, and turboprop, fanjet, and turbojet engines are intermediate in weight and fuel consumption (see **table**). Requirements of speed, range, size, and cost must be balanced to select the engine best suited for the specific airplane design. *See* AIRCRAFT ENGINE PERFORMANCE; SPECIFIC FUEL CONSUMPTION.

Aerodynamic resistance. Aerodynamic resistance is composed of four parts: (1) skin friction drag, which arises from the viscosity of the air; (2) pressure drag, which arises from the pressure field about the body; (3) wave drag, which arises from the compressible nature of air, and which occurs as flight speeds approach and exceed the speed of sound; and (4) induced drag, which is a resistance produced by the generation of lift.

Skin friction. The magnitude of skin friction drag depends upon the surface area of the airplane and the nature of the flow layer that is in immediate contact with it. If the boundary-layer flow is smooth and steady (laminar), the skin friction generated is much less than when this layer becomes turbulent. As a laminar boundary layer proceeds along a surface, depending upon the smoothness and shape of the surface, as well as the density, viscosity, and velocity of the air, the fluid momentum of the flow decreases. Simultaneously, disturbances in the flow or the surface produce motions perpendicular to the surface; these motions tend to grow as the fluid momentum decreases. Eventually a transition condition is reached at which these motions cause the laminar boundary layer to change to turbulent flow. Hence, unless special devices are employed, such as suction slots to remove the low momentum boundary-layer air before it becomes turbulent, the laminar flow will break down after it has passed over a sufficient length of surface. For this reason, on those portions of the airplane where flow initiates, such as the nose and wing leading edges, extreme smoothness to minimize viscous losses is desirable, while further aft, where the flow is turbulent, greater surface roughness can be tolerated. *See* BOUNDARY-LAYER FLOW.

Pressure drag. If sufficient momentum is lost in the boundary layer, the main flow breaks away from the surface, forming a region of separation. Separation results in considerably lower surface pressures than would result in the ideal streamline case. As a consequence, suction pressure on the downstream portions of the body produces pressure drag. This drag is minimized by shaping the body geometrically to reduce the separation region to a minimum by giving it a streamlined shape (**Fig. 1**). In modern high-speed aircraft, pressure drag has been reduced to a small percentage of the total drag. *See* STREAMLINING.

Wave drag. As flight speed of the airplane increases, there comes a point where the flow passing over

Fuel consumption and weight of principal types of airplane engine

Engine	Fuel consumption, lbm fuel/(h)(lbf thrust)*	Weight, lbm engine/lbf thrust†
Ramjet	1.7–2.6	0.4
Turbojet	0.75–1.1‡	0.8§
Fanjet (turbofan)	0.3–0.5	0.7§
Turboprop	0.4	1.2
Reciprocating	0.2	2.4

*Minimum values. Value for ramjet is for Mach 2; values for turbojet and fanjet engines are based on static ground tests; values for reciprocating and turboprop engines are for speeds below Mach 0.3. 1 lbm/(h)(lbf) = 0.1 kg/(h)(N).
†Based on total installed nacelle and power-plant weight, and on thurst at 375 mi/h (168 m/s) at sea level. 1 lbm/lbf = 0.1 kg/N.
‡Value for Olympus 593 engine aboard Concorde is 1.19 at Mach 2.
§Approximate cruise value.

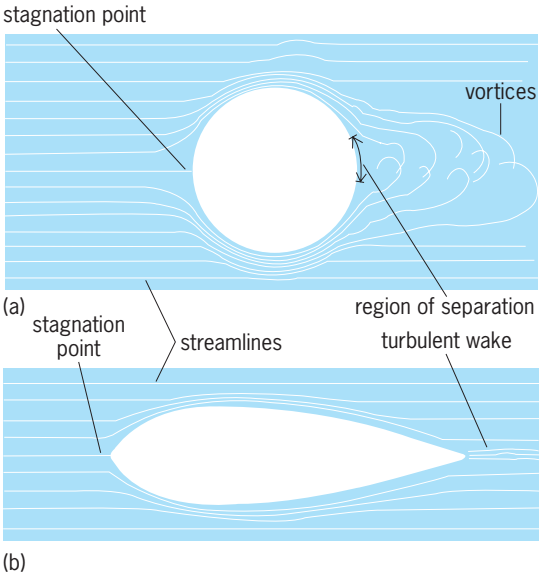


Fig. 1. Effect of streamlining. (*a*) Body without streamlining. (*b*) Body with streamlining, minimizing the region of separation.

a surface reaches the local speed of sound. At this speed (Mach 1) pressure waves propagate through the air. This wave propagation forms shock patterns that interact with the boundary layer to cause premature separation and a sudden drag rise. Because, to create low pressures necessary to generate lift, local velocities next to the upper surface of the wing can be appreciably higher than the flight speed, this separation usually takes place over the wing. Separation over a wing causes an abrupt increase in drag and loss of lift. As the speed continues to increase, shock patterns become steady, the bow and stern shocks changing location and becoming much stronger until they become analogous to the surface waves that are created by a ship. As with such waves, at supersonic speeds the generation of these shock waves represents the major contribution to the drag of the vehicle. *See* AERODYNAMIC WAVE DRAG.

Early experimenters discovered that it was possible to delay and alleviate somewhat the drag rise associated with the shock-induced separations that occur when local sonic speed is first reached at some point on the wing surface. They swept back the leading edge of the wing with respect to the flow direction. The velocity past the wing can be considered as comprising a component in the spanwise direction and a component normal to the span (**Fig. 2**). Only this latter component affects lift and is subjected to local accelerations produced by the airfoil shape. As a consequence, the free-stream velocity can be considerably increased over the value that would produce force divergence for a straight wing. When the flight Mach number becomes sufficiently large, the effect of sweep tends to vanish.

Other attempts to delay the onset and lessen the magnitude of shock-induced separations have concentrated on airfoil design. If the region in which the air passing over the upper surface exceeds the speed of sound can be distributed over more of the surface at a given flight speed, the magnitude of both the shock and the resulting separation can be reduced. So-called supercritical airfoils offer the possibility either of increasing the Mach number (speed of flight/speed of sound) by the order of 15% or, holding the Mach number constant, of increasing the thickness of the wing by as much as 40%. *See* SUPERCRITICAL WING.

Area rule. Studies, at first theoretical in nature but later confirmed by experiment, demonstrated that the drag rise associated with the initial formation and subsequent growth of the shock wave pattern about an aircraft traveling at transonic and supersonic speeds was a function of the rate of change of aircraft cross-sectional area. Furthermore, these studies showed that, if there were an abrupt change in the cross-sectional area perpendicular to the line of flight, the drag rise at a given speed would be greater than if the change were tailored to be more gradual. For this reason aircraft are now designed with the fuselage shape, as well as engine nacelle location and shape, so arranged to produce optimum variation of cross-sectional area for the given design speed; the design principle is termed the area rule. (**Fig. 3**). *See* TRANSONIC FLIGHT.

Induced drag. The final major type of aerodynamic resistance that must be considered arises from the fact that in creating lift the wing must impart a downward momentum to the air flowing past. This downward component of local velocity, termed downwash, rotates the resultant force of the wing in such a manner that a force component is directed in the downstream direction. This component, which depends directly upon the magnitude of the lift force, is termed the induced drag.

Lift of an airplane wing depends upon its shape, speed of flight, angle the wing makes with the free

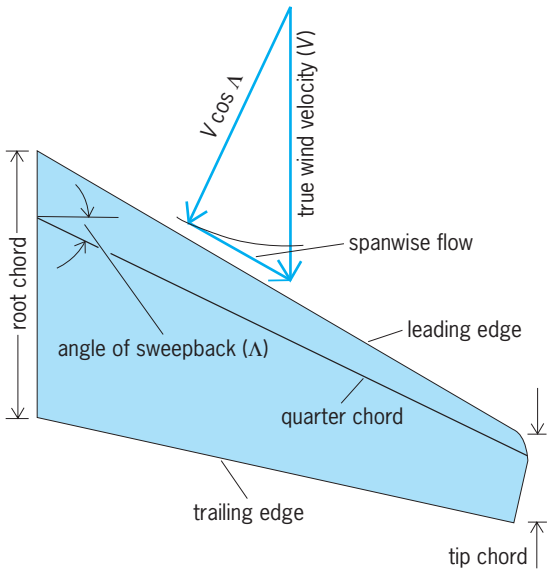


Fig. 2. Geometry of an airplane wing. Wind velocity component perpendicular to the wing span ($V \cos \Lambda$, where V is the true wind velocity and Λ is the angle of sweepback) produces lift.

stream, and properties of the air. These factors are all grouped in the nondimensional coefficient of lift C_L, which is related to the lift L by Eq. (1), where ρ and

$$L = {}^1\!/_2 \rho V^2 S C_L \qquad (1)$$

V are the density and speed of the air, and S is the wing planform area. The coefficient of lift is generally

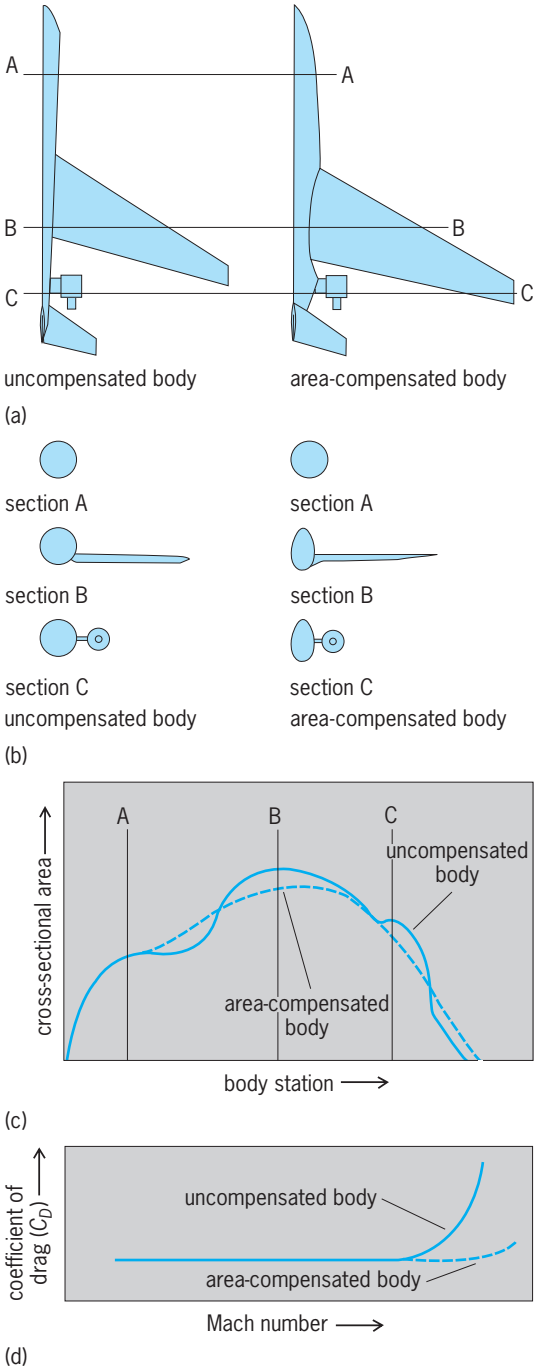


Fig. 3. Illustration of the area rule. (*a*) Half-plan views of airplanes with uncompensated body and area-compensated body. (*b*) Cross sections of these airplanes. (*c*) Variation of cross-sectional areas of these airplanes with body station (longitudinal distance along aircraft). (*d*) Resulting variation of drag coefficients C_D of these airplanes with Mach number.

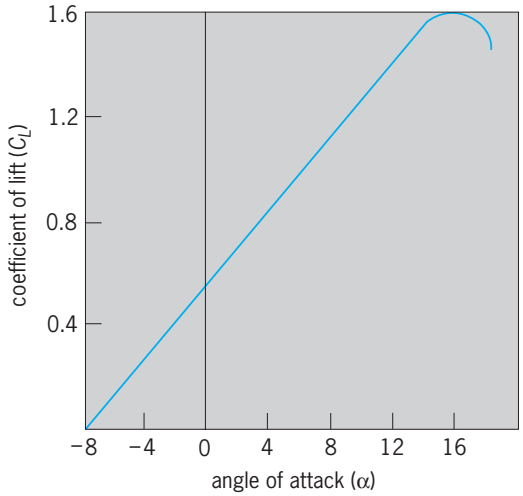


Fig. 4. Variation of coefficient of lift with the angle of attack, showing increase up to the stall angle.

presented as a function of the angle of attack (**Fig. 4**). To maintain level flight, the wing must produce a lift force equal to the weight of the airplane. For a given altitude this implies a low C_L at high speeds but, as the speed is decreased, C_L must increase if level flight is to be maintained. The pilot actually controls the speed of the airplane by changing the angle of attack.

There is a maximum angle of attack beyond which no lift increase is experienced (Fig. 5). As this angle is approached, the air in the boundary layer next to the wing surface experiences such energy losses that it can no longer flow over the wing, but separates, resulting in a loss of suction pressures and a decrease in lift. When there is no further increase in lift as the angle of attack increases, the wing is said to be stalled.

Because induced drag depends upon the magnitude of the lift coefficient, it is of concern primarily in those flight regimes where a large lift coefficient is encountered: relatively low speed or high altitudes. The magnitude of the angle through which the resultant force of the wing is rotated is a function of the aspect ratio, the ratio of the span to the chord of the wing. The higher this ratio, the smaller the angle. Hence airplanes designed for flight at high altitude or for cruising at relatively low speeds will have long, slender wings. Planes designed for higher speeds will be less affected by the induced drag and will generally have much smaller spans to minimize the other forms of resistance.

A measure of efficiency of the airframe is the ratio of lift to drag that it can produce. To achieve very low speeds, special devices such as flaps or slots are used. Such devices, as well as high-lift boundary-layer control, where separation of the boundary layer is delayed either by sucking the low-momentum air next to the surface away through slots or holes in the surface or by adding momentum to the layer by blowing into it through slots, increase the lift that can be achieved before the wing stalls. These devices are thus useful for landing and takeoff maneuvers. The increase of lift achieved by their deployment is

usually accompanied by an appreciable increase of drag and hence does not increase airframe efficiency. *See* AERODYNAMIC FORCE; AERODYNAMICS.

Range. The distance that an airplane can travel with a given amount of fuel is its range, expressed by Eq. (2), where C_t is specific fuel consumption in

$$\text{Range} = \int \frac{V}{C_t}\frac{L}{D}\frac{dw}{w} \qquad (2)$$

pounds of fuel consumed per pound of thrust per hour, V is flight velocity, L/D is lift to drag ratio, and dw/w is a measure of the rate at which total airplane weight changes during the flight.

If the total difference in weight during the flight is produced by fuel consumption, the term $(V/C_t)(L/D)$ should be maximized. In general, it is desirable to have low specific fuel consumption, high speed, and high L/D. Unfortunately, these factors do not maximize at the same point. With reciprocating engine-propeller combination aircraft, maximum range could be achieved at the speed for maximum L/D. With the advent of other forms of propulsion, however, the speed for maximum range has increased. With some aircraft it is possible to achieve high range even at supersonic speeds, for although L/D is reduced, the increase in flight velocity tends to compensate.

Airplanes traveling at supersonic speeds create shock waves which extend outward like the wake of a ship. As the speed and size of aircraft increase, the impact of the shock waves becomes increasingly noticeable and possibly destructive in the form of a sonic boom. Commercial airplanes designed to fly at supersonic speeds must, therefore, operate to reduce shock waves when flying over populated areas. The airplanes must fly either at extremely high altitudes or at speeds below the speed of sound. A significant amount of flight time of supersonic aircraft is spent flying at speeds below Mach 1. The L/D of typical optimum supersonic configurations with highly swept wings and low aspect ratio is low at subsonic speeds. The result of the combination of low L/D and low speed is a substantial reduction in range. To overcome this disadvantage, airplane configurations capable of changing the sweep of the wings in flight have been developed. The variable-geometry compromise provides both high-aspect-ratio, nearly straight wings at low speed and low-aspect-ratio, highly swept wings at supersonic speeds. *See* SONIC BOOM; SUPERSONIC FLIGHT.

Stability. A necessary element of a successful airplane is stability, which is the tendency of forces and moments to be set up that restore the vehicle to its equilibrium position if it is disturbed from this position. For an airplane to be in equilibrium, the sum of the forces acting at its center of gravity must equal zero, as must the sum of the moments. *See* FLIGHT CHARACTERISTICS.

To examine the condition of stability, suppose that the equilibrium of the airplane is suddenly disturbed in a manner that causes it to accelerate. This means that its lift coefficient decreases (**Fig. 5**). A stable airplane would immediately be subjected to a nose-up moment which would tend to increase the lift coefficient and reduce the speed. Similarly, a reduction in the speed of a stable airplane should produce a nose-down moment.

Stability is achieved by balancing the force and moment contributions of various portions of the airplane so that the summation is stable. In conventional configurations the major stabilizing contribution arises from the tail, but by careful design tailless or tail-first (Canard) configurations are possible. Because the balance is achieved for moments about the center of gravity, the location of this point is important. Its position depends upon the load condition and can change in flight as fuel is consumed or as disposable load is dropped. The airplane must be designed so that it reacts in a stable manner throughout the entire range of center-of-gravity locations encountered in service.

Controls. An airplane must be controllable. Major aerodynamic control surfaces include the elevators, ailerons, and rudder (Fig. 5*a*). The angle of attack, and consequently the speed in level flight, is controlled by the elevators. Deflecting these surfaces downward increases the lift of the horizontal tail (just as flaps increase the lift of the wing) and hence causes the airplane to nose downward. An upward deflection produces the opposite effect. *See* ELEVATOR (AIRCRAFT).

Lateral control is achieved by differential action of ailerons. One aileron goes down, increasing wing lift on that side, and the other goes up, reducing lift on that side. The effects couple to produce a rolling moment that tilts the wing. Ailerons also serve to turn the airplane by tilting the wing lift and producing a component normal to the flight path tending to curve it (Fig. 5*b*). *See* AILERON.

Occasionally ailerons are replaced by a spoiler, a device which projects upward on the downgoing

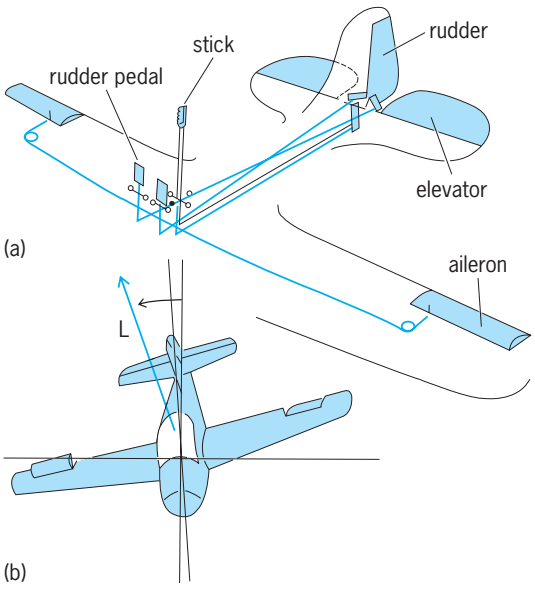


Fig. 5. Aerodynamic control of an airplane. (*a*) Basic controls and control surfaces. (*b*) Effect of tilting the airplane in the direction of the wing lift *L*.

wing, thus producing separation that reduces lift. The spoiler may be used for several reasons. It reduces the twisting moment, which for an aileron can be so great that the torsional resistance of the wing is exceeded, causing it to rotate and thereby changing its local angle of attack. The upgoing aileron, instead of reducing lift, can produce such a change in angle of attack that the control action is reversed. A spoiler is also used to overcome adverse yaw, the tendency of the higher lift on the upgoing wing to produce an increased induced drag on that side. This drag tends to turn the airplane in the wrong direction.

The vertical tail provides directional stability. The rudder prevents the airplane from slewing, or picking up too large an angle of yaw during maneuvers. Although the airplane could be turned with the rudder, such a maneuver would be uncomfortable, and the ailerons are always used. If the vertical tail were large enough to prevent slewing by itself, the rudder could be eliminated and a two-control machine built. *See* AIRCRAFT RUDDER; FLIGHT CONTROLS.

Augmentation systems. While it is both possible and desirable to build aircraft that are inherently stable at relatively low flight speeds, as these speeds increase into the transonic and supersonic ranges the magnitudes and nature of the pressure fields about the aircraft change, resulting in two effects. The first arises from the airloads becoming so great that not only do the forces required to move the controls exceed the pilot's strength but the aircraft, normally thought of as a rigid body, distorts and twists, changing its aerodynamic characteristics (aeroelastic effects). The second effect results from the change in the nature of the distribution of the loads over the airframe and control surfaces (compressibility effects). In combination the two effects alter both the stability and controllability characteristics of the airplane. *See* AEROELASTICITY.

To handle such cases it is often necessary to provide systems that not only supply the necessary power to move the control surfaces but also augment the stability characteristics of the basic aircraft by sensing motions and automatically applying compensating control inputs independent of pilot action. A common application of such a system is to augment yaw damping at high subsonic speeds. In many modern designs, particularly military ones, the pilot is no longer physically connected to the control surfaces. Instead, pilot control motions are interpreted by a computer that superimposes pilot commands on those control motions required to maintain stability. In some cases the airplane has been designed to be unstable if unaugmented in order to achieve high maneuverability. *See* STABILITY AUGMENTATION.

Structures. Airplane structures must provide the required strength and rigidity, within the shape envelope dictated by the aerodynamic requirements of the airplane, for the lowest possible weight. Other factors such as manufacturing ease, corrosion resistance, fatigue life, and cost are important in selecting both the materials and the nature of the structural system to be employed, but the strength-to-weight ratio is of critical importance.

Most materials used in aircraft construction display the characteristic that stress is essentially proportional to strain up to a point called the proportional limit. Removing the load before this point is reached returns the material to its original unstrained shape, while exceeding the proportional limit results in permanent distortion. Airplane structures are designed so that no stress exceeds the proportional limit if the airplane is maneuvered over its full range of speed and design accelerations: the ultimate stress is not exceeded even if the structure is subject to $1^1/_2$ times its design load. *See* AIRFRAME; FUSELAGE.

The major structural problem is design of the wing. Aerodynamic considerations demand that wing thickness be kept small with respect to wing chord. The air load is distributed along the wing, and the magnitude and configuration of the load are a function of the maneuver under consideration. The wing may thus be considered as a long, slender beam subjected to loads that during accelerations may total several times the weight of the aircraft. Because the depth of the beam (airfoil thickness) is small, the outer fiber stresses are large. *See* WING; WING STRUCTURE.

One solution to the problem of high bending stresses is the biplane, where the upper and lower wings, which are connected by struts and wires, act as the flanges of a beam having a depth equal to the separation between the wings. Another solution is the use of a pin connection at the wing root and an external strut.

High-performance airplanes cannot afford the drag penalties of such solutions and must employ cantilever wings. Because of the high bending and torsional stresses encountered, a cantilever structure implies a stressed skin, or outer covering of the wing. The skin may be thin, supported by many stringers, or relatively thick with fewer longitudinal members. Ribs along the span aid in distributing loads and in reducing column action in the stringers.

High flight speeds demand smooth skin surfaces. These are sometimes produced by using skin milled from metal blocks rather than sheet. Two thin sheets bonded to a low-density core form a sandwich, a type of construction frequently used.

Aircraft flying at high speeds encounter high temperatures due to the high aerodynamic compression that occurs at the nose and the leading edges of the wings. At high Mach numbers these temperatures become sufficiently high to reduce the strength of aluminum and its alloys, so other materials, such as stainless steel or titanium, are sometimes used in these areas. When flight speeds become so high that the temperature rise affects the characteristics of the construction material, it may be necessary to employ cooling in sensitive areas. *See* AEROTHERMODYNAMICS.

David C. Hazen

Bibliography. J. D. Anderson, Jr., *Introduction to Flight*, 5th ed., 2005; R. Jackson (ed.), *Jane's All the World's Aircraft*, revised periodically; M. J. H. Taylor and D. Mondey, *Guinness Book of Aircraft*, 6th ed., 1992.

Airport engineering

The planning, design, construction, and operation and maintenance of facilities providing for the landing and takeoff, loading and unloading, servicing, maintenance, and storage of aircraft. Facilities at airports are generally described as either airside, which commences at the secured boundary between terminal and apron and extends to the runway and to facilities beyond, such as navigational or remote air-traffic-control emplacements; or landside, which includes the terminal, cargo-processing, and land-vehicle approach facilities.

The design of airports involves many diverse areas of engineering. Their significant cost has resulted in the development of sophisticated techniques for planning, construction, and operation of airports.

Design considerations. Airport design provides for convenient passenger access, efficient aircraft operations, and conveyance of cargo and support materials. Airports provide facilities for changing transportation modes, such as people transferring from cars and buses to aircraft, cargo transferring from shipping containers to trucks, or regional aircraft supplying passengers and cargo for intercontinental aircraft. In the United States, engineers utilize standards from the Federal Aviation Administration (FAA), Transportation Security Administration (TSA), aircraft performance characteristics, cost benefit analysis, and established building codes to prepare detailed layouts of the essential airport elements. These elements include airport site boundaries, runway layout, terminal-building configuration, support-building locations, roadway and rail access, and supporting utility layouts. Airport engineers constantly evaluate new mechanical and computer technologies that might increase throughput of baggage, cargo, and passengers.

Design choices must enhance user orientation and comfort. Because airports and airlines are key transportation and economic facilities for major cities worldwide, airport engineers frequently integrate a wide variety of secondary design considerations, including air- and water-quality enhancement, noise concerns, citizen participation, wildlife impacts, and esthetic features.

Site selection for new or expanded airports. Site selection factors vary somewhat according to whether (1) an entirely new airport is being constructed or (2) an existing facility is being expanded. Few metropolitan areas have relatively undeveloped acreage within reasonable proximity to the population center to permit development of new airports. For those airports requiring major additional airfield capacity, and hence an entirely new site, the following factors must be evaluated for each alternative site: proximity to existing highways and major utilities; demolition requirements; contamination of air, land, and water; air-traffic constraints such as nearby smaller airports facilities; nearby mountains; numbers of households affected by relocation and noise; political jurisdiction; potential lost mineral or agricultural production; and cost associated with all these factors. Some governments have elected to create sites for new airports using ocean fills. The exact configuration of the artificial island sites is critical due to the high foundation costs, both for the airport proper and for the required connecting roadway and rail bridges.

New airports offer significantly increased runway capacity over airport reconstruction projects due to the larger space available for more runways, longer runways, and more widely spaced runways. New airports are rare: Dallas/Fort Worth International Airport opened in 1972, and Denver International opened in 1995. In 1998, two new airports constructed at Air Force bases opened: the Austin, Texas, airport; and MidAmerica, at Scott Air Force Base in Illinois. Killeen Airport in Texas opened as a joint-use facility with Robert Gray Army airfield in 2004. Denver capitalized on its geographic location in the midsection of continental America, its historic hub airline operations, and undeveloped land adjacent to the metropolitan area to build an unusually large airport (**Fig. 1**). This location is farther from the Rocky Mountains than the old airport, a significant aeronautical advantage. The number of citizens affected by the FAA's regulated noise level was significantly reduced by the new airport's location.

Denver's master plan and environmental approvals were completed between 1986 and 1989. Design, construction, training, testing, and airport certification were completed between 1989 and 1995.

For commercial airport sites being expanded or military bases being revamped for commercial use, many site factors must be considered. The possibility of acquiring adjacent commercial or residential properties requires extensive analysis of social and economic impacts and development costs. Mitigation measures of noise impact constitute a key element of any proposed airport expansion. Such measures can include restrictions on types of aircraft using particular runways, avoidance of noise-sensitive areas by air- traffic routing, acquisition of additional land adjacent to the airport as a noise buffer, improvements to existing occupied structures to dampen noise transmission, or construction of noise berms for ground operations.

Airfield configuration. Since the runways and taxiways constitute the largest portion of the airport's land mass, their layout is generally one of the first steps in the airport planning process. The optimum runway configuration, or layout, is based on long-range forecasts of the estimated numbers of aircraft landings and departures, airport elevation, local prevailing wind direction, and regional and national air-traffic patterns. A paved runway surface 12,000 ft (3660 m) long and 150 ft (45 m) wide is suitable for most applications. However, runway length requirements vary according to the type of aircraft, temperature, and elevation. A parallel taxiway is generally constructed 600 ft (180 m) from the runway (measured centerline to centerline). It is connected by shorter high-speed taxiways to allow arriving aircraft to leave the runway surface quickly so that another aircraft can land. This combination of paved

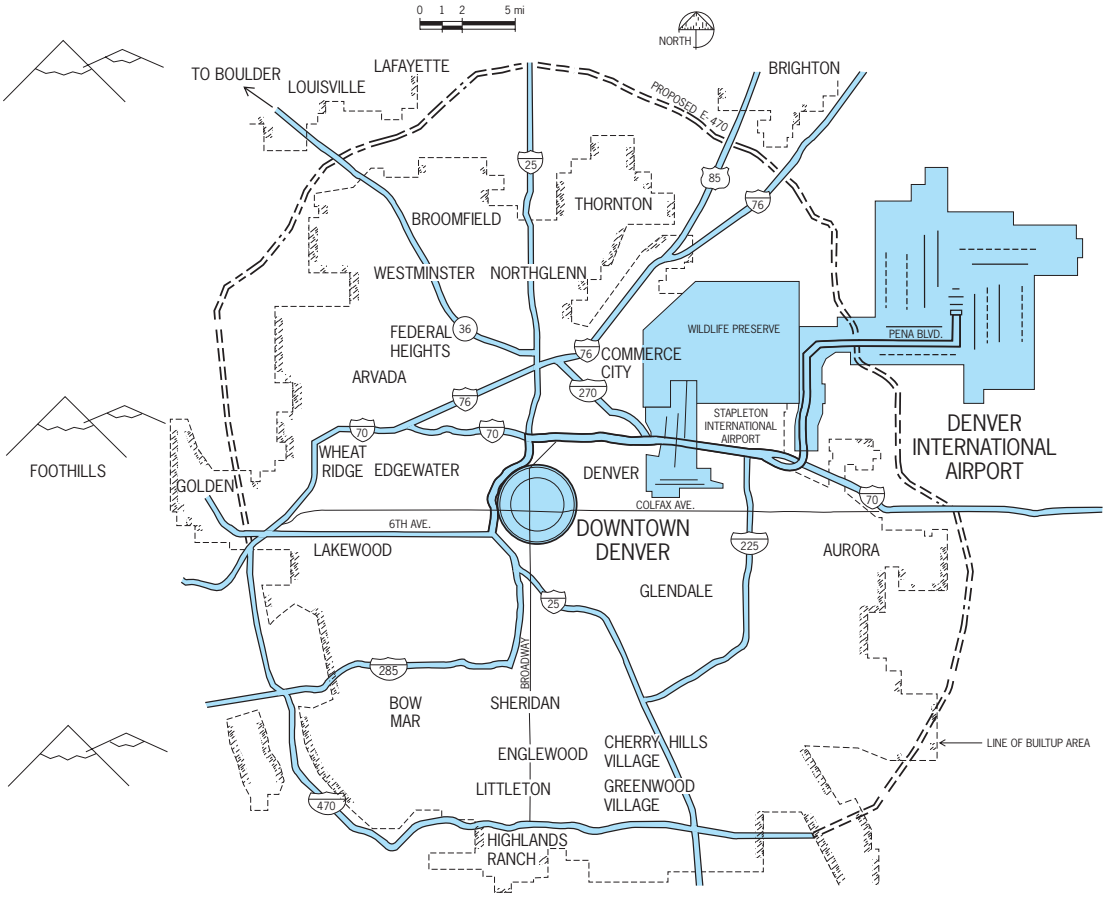


Fig. 1. Regional map of the Denver International Airport.

surfaces is generally referred to as a runway-taxiway complex.

Airports provide runway layouts so that incoming and departing aircraft can maintain required minimum separation for safe, simultaneous aircraft operations. For parallel runways, thresholds would be slightly staggered to avoid wake turbulence interference between incoming aircraft. Staggered thresholds might also be used to minimize crossing of active runways by taxiing aircraft. Each runway crossing is a potential aircraft delay and a safety hazard.

When airports have sufficiently high-velocity crosswinds or tailwinds from more than one direction, crosswind runways are located at some angle to the primary runway as dictated by a wind rose analysis. *See* WIND ROSE.

Alternative airfield configurations are generally analyzed under various operating conditions by using computer simulations. Typically, these simulations assume a certain number of aircraft arrivals and departures, based on actual airline schedules for an entire year, weather conditions based on historical data, and various combinations of runway usage by the air-traffic controllers; and then the required time for each aircraft operation is calculated (**Fig. 2**). The configuration with the least amount of arrival time and taxi time on an annual basis is preferred. Justification for additional runways would accrue as the construction and financing costs of the runway are compared to the operational savings of reduced aircraft delays.

Some runways, widely spaced so that the aircraft operations are independent, are used primarily for instrument flight rules (IFR) conditions, when visibility is low. In visible flights rules (VFR) conditions, FAA rules allow independent operations with more closely spaced runways. The result is shorter taxi distances after the aircraft lands. The IFR runways, which are farther from the terminal, might also be used during peak periods to increase the airport's capacity. If excess runway capacity is available, FAA air-traffic controllers have more discretion to assign aircraft landings and departures to minimize noise exposure and taxi distances.

In the United States, the FAA defines standard techniques for calculating the hourly capacity, annual service volumes, delays, average taxiing distances, and annual runway crossings.

After the runway configuration is determined, detailed construction drawings are prepared that dictate the runway profile and cross section, pavement depth and materials, subsurface remediation, drainage controls, lighting, and paint markings. A cost-effective design balances grading materials from the airport site to minimize trucking costs and the associated impact on air quality and traffic congestion, and specifies locally available materials for the paving mix. *See* AIR POLLUTION; PAVEMENT.

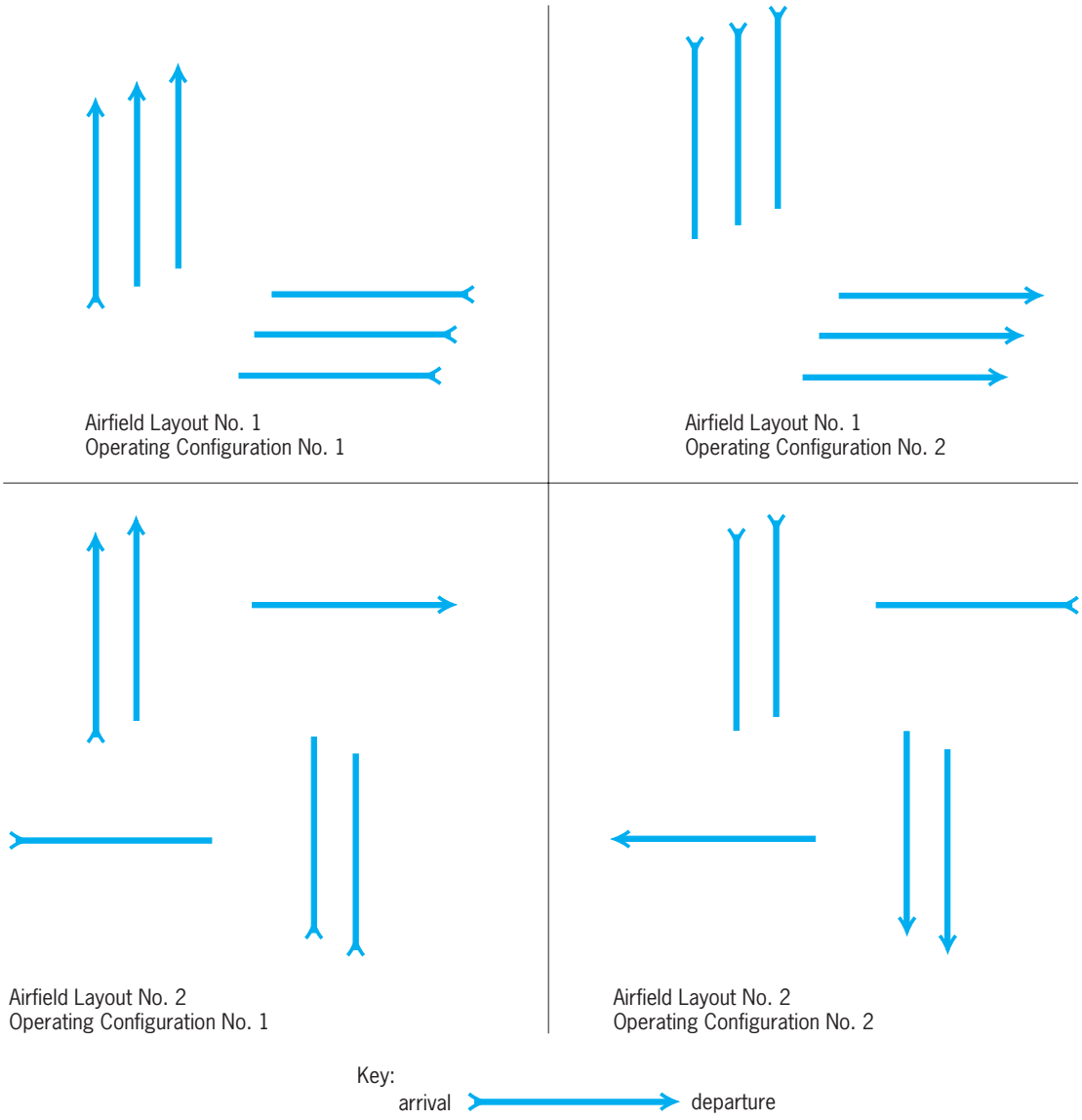


Fig. 2. Combinations of arrivals and departures on the same runway configuration.

Runways are paved with concrete, asphalt, concrete-treated base, or some combination of layers of these materials (**Fig. 3**). Runways for larger aircraft require thicker, more expensive pavement sections. Engineers design these pavements for long-term durability. The expected life of a concrete runway can be increased from 20 to 40 years, based on enhanced mix designs and sections. *See* CONCRETE.

The longitudinal profile of a runway requires an iterative design process. The objective is to minimize grade changes for the smoothest aircraft operations possible, while minimizing the total quantities of earthwork (cut-and-fill) to construct the runway.

The availability of computerized controls and improved lighting fixtures is increasing the use of technology known as surface movement guidance systems for pilots. These advanced airfield lighting systems provide positive control for taxiing aircraft to prevent aircraft collisions. Centerline lights provide positive guidance along predetermined routes for the pilot to follow into gates during low-visibility conditions. Stop bars consisting of rows of red lights prevent planes and vehicles from traveling onto an active runway without positive control from the air-traffic control tower. The controller has positive identification of the aircraft location through the use of ground pressure sensors and airport surface detection radar equipment. Design of these systems must mirror the airport's operating procedures. As an example, the lighting circuit design reflects certain groupings of taxiway segments and runways used for various operating configurations (**Fig. 4**).

A system of vehicle service roads must be provided around the perimeter of the airfield both for access to the runways and for security patrols of the perimeter fencing. Airfield security fencing with a series of access gates is monitored with patrols and, increasingly, a remote camera surveillance system.

Terminal configuration. The terminal building houses the ticketing, baggage claim, and transfer

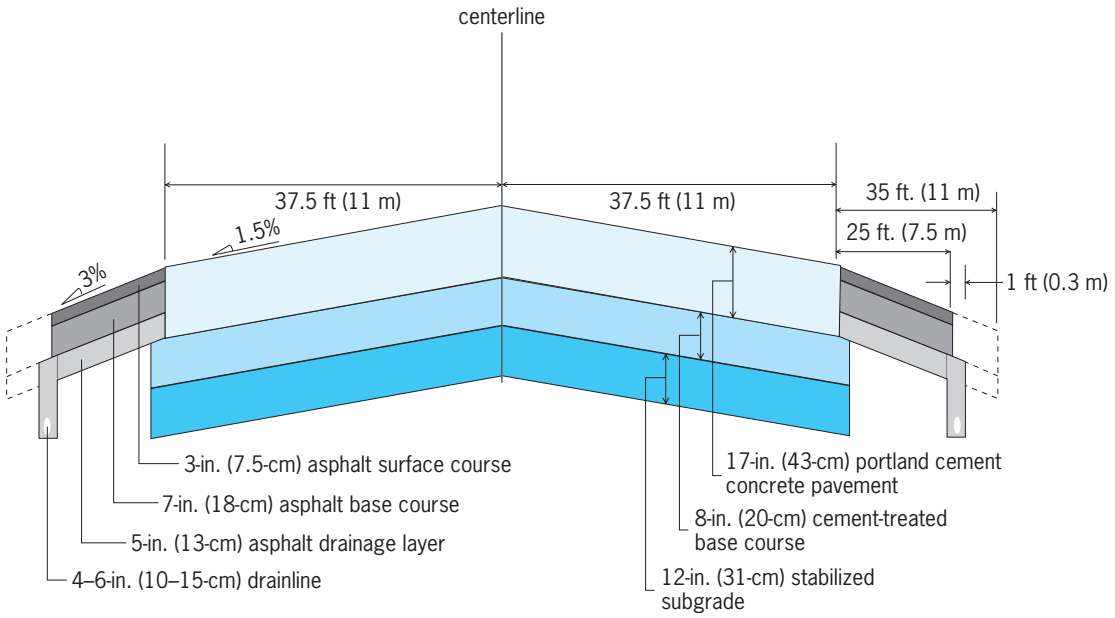


Fig. 3. Cross section of a runway-taxiway pavement.

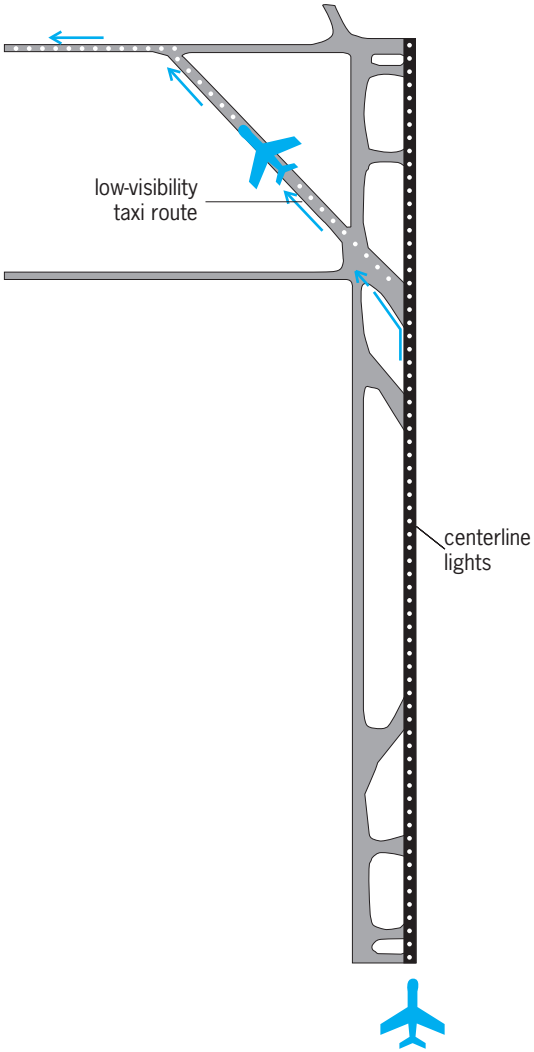


Fig. 4. Surface movement and guidance and control system plan for a runway and the associated low-visibility taxi route.

to ground transportation facilities. The concourse is generally the combination of facilities for passenger-boarding of aircraft, sorting baggage according to flight, and uploading cargo carried in commercial aircraft. The most popular terminals with passengers are those that offer amenities such as a hotel integrated into the terminal main area, extensive retail shops, and vibrant integrated artwork. Airport terminal and concourse configurations generally fall into three categories (**Fig. 5**): (1) a terminal contiguous with concourse satellite extensions (known as piers or fingers) used for boarding aircraft, such as at Miami International Airport; (2) unit terminals, which serve as transfer points both from the ground transportation modes into the building and from the building into the aircraft, such as at New York's John F. Kennedy Airport; (3) and a detached terminal and concourses, sometimes referred to as a landside and airside terminals, connected by a people-mover train system such as at Atlanta's Hartsfield International Airport, or an underground walkway such as at Chicago's O'Hare Airport, or a surface transport vehicle such as at Washington's Dulles Airport.

Contiguous. The contiguous terminal-concourse has a relatively low cost since no trains or special airport vehicles are used. Passengers walk through the complex, aided by moving sidewalks. Expansion of additional aircraft gates is convenient by lengthening one of the piers or adding a pier. However, expansion is fairly limited since continued lengthening of the concourse adds walking distance and can infringe on required airfield operational space. Another limitation is the maneuverability of aircraft between the piers. Aircraft can be delayed while waiting for an aircraft to push back from an adjacent or opposite gate.

Unit. Unit terminals are the most convenient for local passengers, allowing short distances from the

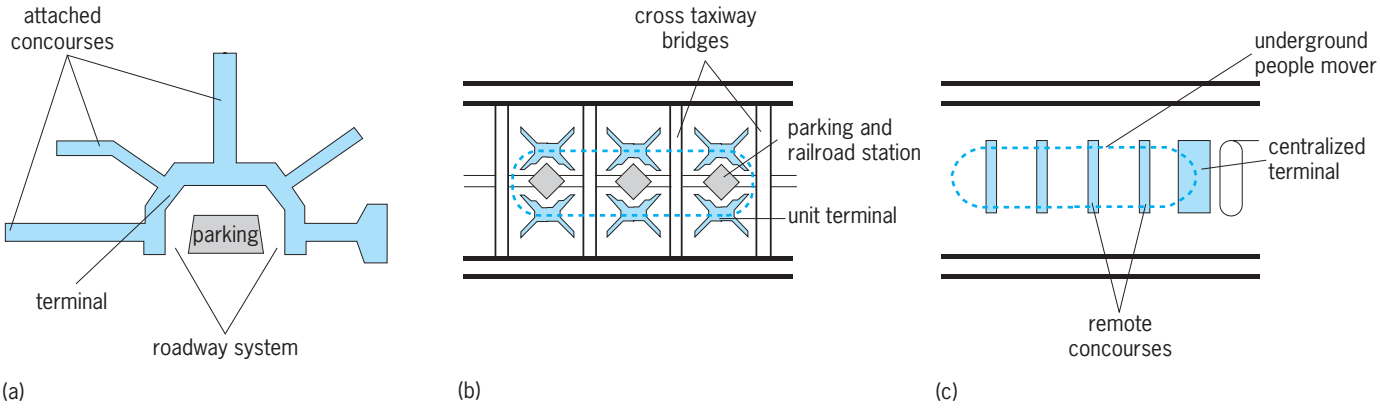


Fig. 5. Alternative terminal configurations. (*a*) Contiguous terminal and concourses. (*b*) Unit terminal. (*c*) Detached terminal and concourses.

curbside roadway system to the aircraft boarding area. This arrangement requires a more extensive investment in the curbside roadway infrastructure. Hubbing passengers, those transferring from one plane to another at the same airport, are disadvantaged in this configuration because they have to move from one building to another.

Detached. Detached terminal and concourse configurations allow airfield facilities to be widely spaced for more efficient aircraft movement, while keeping passenger walk distances short by providing automatic transport systems. Landside and airside building can be expanded independently as different functions require. This configuration allows for a single security screen pavilion at the entrance to the aircraft gate area, which is an advantage for both security and cost.

Shapes. A number of shapes can be used for the airside concourse, depending on the available space. The most common is a long, linear concourse with an underground train station in the middle. This concept was originally developed for Atlanta's Hartsfield International Airport, and it proved to be the most efficient shape for aircraft movement. This linear shape mirrors the parallel pattern of aircraft movement required for highly efficient operations. The hydraulic analogy is laminar flow, as opposed to turbulent flow, which is inefficient. The master plan for Denver International Airport recognized the efficiencies of the Atlanta-style concept and adapted to it. The Denver airport's linear concourses are more widely spaced than the Atlanta concourses, allowing simultaneous taxiing between concourses. Denver's linear concourses were widened to allow wide moving sidewalks. An expanded central core was also added to allow for additional food and retail services for passengers. This master plan concept has now been adopted for the new terminal at Salt Lake International Airport and the new Terminal 5 at London's Heathrow International Airport.

At the Pittsburgh airport, a remote concourse has been constructed in an X shape. The large middle houses the station for the underground people mover. A great advantage is that the average passenger walk distance is halved in comparison to the linear concourse. If each concourse contains 40 gates, a passenger on the linear must walk the length of 20 gates to reach the farthest gate. The farthest a passenger in an X concourse would walk is the length of 10 gates. Gates vary in size according to the type of aircraft scheduled for use. Wider wingspans require wider (longer) gates. Aircraft with higher seat counts must have gates with a commensurate amount of passenger seating.

Smaller facilities. Smaller terminal facilities are required for general aviation, charter companies, and commuter or regional aircraft. These are generally less system-intensive. Selection of locations must take into account both the number of passengers that connect from these terminals to the air carrier concourses and the preferred mode of transport between the terminals.

Dimensional requirements. Airports are among the most specifically designed and configured public facilities. An important part of airport engineering is optimizing the trade-off between providing sufficient dimensional space requirements for such functions as passenger queues, aircraft movement, seating areas, and circulation without requiring overly large spaces that drive up the cost of construction and operation. A detailed dimensional requirements study is prepared that determines the width of ticket lobby, size of train stations, mix of various-sized aircraft gates, width of concourse, number of roadway lanes, number of short-term parking spaces, number and size of food service outlets, linear feet of ticket counter, and space programs for support functions; and that otherwise defines as many elements the airport as possible. Reasonable delay criteria are defined for almost every step of the trip through the airport. These criteria are then translated into speed and size requirements.

Terminal subsystems. Because of the high volume of passengers and baggage in an airport terminal, mechanized subsystems are used extensively. Moving sidewalks shorten walk distances and passenger connect times. Double or extrawide moving sidewalks are desirable, provided the building is sufficiently wide, to allow some passengers to rest while others move briskly. Increasing passenger

preference to carry small items of luggage and shopping parcels has also required engineers to widen moving sidewalks.

Two devices are used to move passengers from one level in an airport building to another. Banks of elevators are required for every level change in order to accommodate all possible passenger needs. Escalators are required in order to move the peak numbers of passengers down from ticketing to a people mover station or from a concourse down to baggage claim level. These two types of devices are complementary and are always used in tandem. Their numbers and dimensions are determined by accessing the passenger demand and required hourly capacity. One level change, for example, might require one elevator, two down escalators, and one up escalator.

Use of electric trains to transport passengers within an airport building or between buildings is the most expensive alternative. Extensive structural requirements and fire code provisions add to the base system cost. Extensive signage and voice messages are required to ensure that passengers feel comfortable using such a system to move from gate to gate or concourse to concourse. This system is, however, the most comfortable for passengers.

Mechanized sortation systems are used extensively at airports for mail, cargo, parts, and baggage. They significantly decrease the time to sort and transport the aircraft load. They generally use tunnels and basements to move these materials into the aircraft. Each system has special requirements (such as maximum and minimum dimensions, heavy structural loading, and maintenance access requirements) that must be considered in the building design.

Security requirements. After the events of September 11, 2001, new security regulations enforced by the Transportation Security Administration have resulted in many changes to airport facilities and operational procedures. They include 100% screening of all checked baggage, matching all checked baggage to passengers, closer inspection of carry-on luggage, and prohibition of certain objects in the aircraft cabin, as well as secure cockpit doors, more rigorous background checks of employees, and more explosives detection. New airport security initiatives focus on gathering and analyzing information about passengers, cargo, vehicles, employees, and even the planes, both on the ground and in the air. An entirely new security-focused information management infrastructure is being created.

Information management/information technology. In recent years, computers and computerized control systems have been developed for virtually every component of airport design and operations. Therefore, it is critical to the success of every airport project that these systems be defined and planned at the earliest stage. Airports have an unusually high number of tenants and users. Each requires a different set of systems for up-to-date information and business controls.

Generally, a fiber-optic cabling system is provided for transmission of data and voice signals by all tenants. This has become an essential utility. For radio transmission, some airports are managing the high usage of radio with another utility system, a leaky coaxial cable, which picks up low-intensity transmissions by individual radios and transmits them for amplification for the end-receiver of the message. This avoids high-intensity transmissions, which cause interference between users.

Airports generally have a set of life-safety systems that include the fire alarm system, public address system, smoke evacuation, security system, and some aspects of transport systems such as underground train systems. Each of these systems must perform its role in communicating to the passengers and airport employees and controlling devices such as exhaust fans, doors, and trains.

Airport operations require a paging system, card access security, controlled-circuit TV security, ground-to-ground radio, ground radar, parking and ground transportation revenue and control system, broadcast TV, and Doppler radar weather system.

Some level of integration of these systems is essential to minimize the time required to communicate critical information. For instance, if a weather detection system senses a tornado or wind shear condition, the integrated control system can preprogram audible alarms.

Airports also have a set of maintenance systems that are best managed by an integrated control system. These systems manage devices such as airfield in-pavement lights, pavement temperature sensors, moving sidewalk controls and alarms, irrigation controls, fuel system monitoring data, environmental monitoring data, work orders, and spare parts inventory.

For engineering purposes, airports may use a computer-assisted drafting (CAD) database of all airports facilities. Information is generally stored in layers, with all structural and demising walls in one layer, utilities and electrical wiring in another, door hardware in another, and lease information in yet another.

Airlines are increasingly providing their own radar to gain up-to-date information on incoming aircraft locations for better management of gates. They use both air-to-air and air-to-ground radio systems. Each airline has a master computer system that controls reservations, ticketing, and flight management. This master computer system may control other computerized systems, such as flight information displays or baggage systems.

Some changes are occurring in the areas of responsibility for these various systems. Some airports are installing Common User Terminal Equipment, which provides complete passenger data-handling infrastructure to different airlines. Airports are also increasingly installing their own Flight Information Display Systems, so that multiple carriers' flights can be displayed on the same devices. Another change is that systems historically installed and operated by the FAA may transfer to airports. One example is approach lightning systems, which indicate the distance to the runway threshold to pilots on approach. As these systems are turned over to airports,

another integration opportunity is presented. Integration controls for both approach lights and inpavement lights would decrease the opportunity for delay or error and allow the controller to focus on communication with pilots.

A gradual change from radio technology to use of digital transmissions will affect many airport systems. Receipt of a radio transmission to a pilot cannot be verified by the air-traffic controller. However, receipt of a digitized message can be verified and can be printed in the cockpit. As older aircraft are replaced, new aircraft are equipped to receive such messages.

Industry groups encourage cooperation between airlines and airport authorities. They are also encouraging development of system standards and interfaces between airline and airport authority computer systems. Recent technology developments include machine-readable boarding passes and tags, which allow better security and faster passenger processing; self-service passenger check-in kiosks; and software that allows printing of boarding passes at home. These innovations decrease passenger waiting times on lines and reduce airline personnel costs.

Movement of cargo requires communication between different agencies such as airlines, customs agents, cargo shippers, cargo receivers, and trucking companies. Guidelines for standardizing these types of system are also being recommended by the industry groups.

Ground transportation facilities. Airline passengers use terminal and parking facilities to transfer to private cars, commercial vans and buses, taxis, rental cars, or rail transport. Large numbers of these vehicles must be able to park, process business transactions, and load passengers. Demand for short walk distances requires that many of these functions be incorporated into the front area of the terminal building.

Remote staging facilities are often constructed so that large commercial vehicles do not dwell in front of the terminal. Commercial vehicles are mobilized to the curbside by radio dispatch, loaded with passengers, and moved away from the terminal as quickly as possible, thereby maximizing the number of vehicles that can load in a peak hour, and minimizing the exposure of waiting passengers to vehicle emissions.

Rail connections are best provided by having a station incorporated directly into the terminal or conveniently accessible from the terminal. Stations providing airline passengers access to regional transit are becoming more common. Airports with such facilities include New York JFK, Atlanta Hartsfield, Washington Reagan National, St. Louis Lambert, Portland (Oregon), and Baltimore-Washington. Several airports are now planning such links, including Washington Dulles, Los Angeles, and Phoenix Sky Harbor.

Support buildings. The primary types of support buildings required by the airlines for their airport operations are flight kitchens to prepare meals for passenger, hangars to service aircraft, and ground support equipment buildings to service vehicles such as tugs, baggage carts, and service trucks. The high number of trips for support vehicles to travel from these buildings to load or service aircraft requires that the buildings be located in reasonable proximity to the aircraft gates. However, the buildings should be sufficiently distant to allow the concourses to be expanded without requiring demolition of these support facilities. Careful forecasting of growth, analysis of operational costs, and dimensional analysis are required to obtain the optimum location.

Airline companies may also require parts storage buildings, employee facilities, and administrative offices. Airlines frequently have an exclusive cargo building to support delivery and sortation of cargo carried in excess space on commercial flights.

An airport requires fire equipment to provide extremely fast primary and secondary responses to every runway. Locating the aircraft rescue and firefighting stations to meet these responses times stipulated by the FAA requires careful positioning with respect to the taxiway systems. The exact building layout must recognize the need for the fire equipment doors to exit onto the most direct response route. *See* FIRE TECHNOLOGY.

Other types of support buildings include vehicle maintenance buildings for snow removal and airport vehicles, sand storage buildings, roadway revenue plaza offices, and training facilities.

Fuel and deicing facilities. Economics of scale and safety considerations generally encourage the implementation of large, centralized common systems for aircraft fuel. The large storage tanks required to ensure adequate reserves of fuel are located in remote areas of the airport, generally in aboveground facilities. Underground distribution piping transports the fuel to hydrant pits or truck fueling stations close to aircraft operations. This system, like most utilities, is designed with backup capacity by looping piping around each service area. If a problem occurs in one section of pipe, valves are automatically closed and the supply direction is reversed. Fuel tanks require extensive structural, mechanical, and electrical design. These tanks are widely spaced to avoid the transmission of fire and to allow room for a surface detention area to store burning fuel.

Airlines use either propylene glycol or ethylene glycol, mixed with water, to deice aircraft and to provide coatings that retard development of ice accumulation on aircraft. Facilities are required to store the glycol, mix it, and sometimes heat it before distribution. It can then be loaded onto deicing trucks for spraying onto the aircraft or distributed to deicing gantries or booms that are fixed to a base but have an extendable arm with an operator cab on top from which to spray the fluid.

Glycols have serious water-quality impacts on receiving streams and treatment plants. Increasingly, airports are making provisions to grade the ramp areas where deicing operations occur so that the spent glycol can be captured. Once it is segregated from the normal storm collection system, it can be

stored for metering in treatments facilities, reprocessed for reuse off airport, or otherwise disposed of. Ginger Sunday Evans

Bibliography. A. Graham, *Managing Airports: An International Perspective*, 2d ed., 2003; R. M. Horonjeff, F. X. McKelvey, and B. Sproule, *Planning and Design of Airports*, 5th ed., 2007; R. de Neufville and A. Odoni, *Airport Systems: Planning, Design, and Management*, 2002; A. T. Wells, *Airport Planning and Management*, 4th ed., 2000.

Airport noise

The unwanted sound from airport operations, primarily from aircraft of all types. It affects neighbors both adjacent to and farther from the airport. It can be controlled using several approaches, and assessed using a number of tools. Airport traffic and construction, which also create noise, will not be discussed here.

Types of operations. Aircraft using an airport cause noise during several types of operations. The noisiest is when an aircraft takes off. The plane is fully loaded with passengers, cargo, and fuel and uses maximum thrust to take off from the runway. Neighbors behind and to the side of the runway experience the first noise from a take-off. At later times, people along the take-off path experience the noise from the plane as it passes overhead.

When the plane approaches its destination airport, it again causes noise for the people along its path. It is now much lighter, having used up fuel enroute. However, at United States airports, the planes are also much lower as they line up to land on the designated runway, coming in on a 3° glide slope. When the aircraft lands, the pilot applies the brakes and/or reverses the engines. This operation also causes noise for the neighbors adjacent to the airport.

Aircraft also produce noise when they taxi to or from the runways, gates, parking areas, or maintenance areas. In this case, the amount of thrust needed is much smaller than the amount needed to be airborne, and the associated noise is much less.

Finally, aircraft can cause noise when they perform maintenance run-ups. After aircraft maintenance, the Federal Aviation Administration (FAA) requires that the system involved be tested before flight. This may require running the engines to partial or full thrust. The run-up is done on airport property but without the plane actually taking off. A full-power run-up is nearly as noisy as a take-off.

Mitigation. Noise control can be applied at the source, along the sound path, or at the receiver. In the case of airport noise, the source is the aircraft and the receiver is the neighbor. Noise control at the source is preferred, since it reduces the total amount of sound released into the environment.

Source noise control of jets has progressed with time. The first jet aircraft, called stage 1, had no noise control on their engines. Federal law now prohibits use of stage 1 in the United States. Simple muffling was added to the jet engines in the 1970s to give stage 2 aircraft. At the same time, improvements were made to the engine design, resulting in about 10 decibels of noise reduction for a given aircraft type. This type of engine powers the stage 3 aircraft. As of January 1, 2000, all large turbojet aircraft operating in the United States are stage 3, by federal mandate. In anticipation of the stage 3 mandate, owners of stage 2 aircraft could sometimes use an engine retrofit, called a hushkit, to meet the stage 3 noise criteria, prolonging the useful life of the aircraft. Additional improvements, starting in 1992, have continued to reduce aircraft noise emissions. In June 2001, the International Civil Organization (ICAO) adopted the next quieter level, called Chapter 4 of their regulations (Annex 16). It effectively eliminates hushkitted older aircraft from operating in countries which adopt these rules. *See* AIRCRAFT ENGINE PERFORMANCE; JET PROPULSION; MUFFLER; TURBOJET.

Airports can reduce the noise from backing up, repositioning, or taxiing by having the aircraft towed to or from the gate. The noise from towing is much less than that from the aircraft engines. In addition, aircraft may be asked to use only the engine away from the neighbors, to cut the noise source level by 3–6 dB depending on the number of engines on the aircraft. Finally, aircraft can be connected to airport electrical service at the gates instead of running their auxiliary power units or they can use ground power units, especially at night.

The source noise during flight can be reduced by using lower thrust settings when planes fly over noise-sensitive areas. So, if neighbors are far from the airport, the aircraft can use maximum thrust to get as high as possible and lower thrust farther away. This method, encouraged by the FAA, is used at a number of United States airports.

The most common path noise control is by use of barriers or berms. For effective noise reduction (at least 5 dB), the top of the barrier must intercept the line-of-sight between the aircraft engine and the observer. Since the elevation of aircraft engines can be as high as 30 ft (9.1 m), noise reduction occurs only when the neighbors are close to the ground, when the aircraft are on the ground, and when the barriers are high. Noise from taxiing and run-ups can be reduced in this way. At some airports, buildings are constructed near the active taxiways to provide effective barriers.

A second type of path noise control is to increase the distance between the source and receiver. The sound level is decreased 6 dB for every doubling of distance. Airports, working with the FAA, can develop noise abatement departure or arrival paths to maximize the distance between the source and receiver or to take advantage of open space. The airport can also move or "buy out" the affected residents. The area left vacant can either be redeveloped as compatible land use, which may be commercial or industrial, or left undeveloped.

Path noise control of aircraft run-ups is achieved by use of a hush-house, a hangarlike structure in which the aircraft can perform its engine test. This partially

or fully enclosed structure reduces the amount of noise emitted to the environment. These facilities are used at military facilities, overseas, and at a small number of United States commercial airports.

Receiver noise control is usually by sound-insulating the structure. The windows are replaced with double or triple glazing, and doors may be replaced with solid core versions. Typical improvement in noise reduction is 5–25 dB depending on the original condition of the structure. Air conditioning may also be added. This approach, which is effective only when the neighbor is inside the structure, has been used in the United States since 1980.

Time-of-day considerations. Most airports receive complaints from their neighbors about aircraft noise. The complaints include not being able to hear the television, constant noise, sleep deprivation, inability to work, and vibrations. United States airports must (by federal law) be open 24 hours per day, 7 days per week, unless there is another facility which can take the traffic. Complainants may also be affected by through flights or other airports' operations.

To mitigate this situation, an airport can direct nighttime operations to occur on a runway which affects the minimum number of people, whenever operationally possible. Limitations can also be placed on runways which affect many neighbors or are very close to them, if other runways are available. Altitude restrictions may be added at night to maximize the distance between the aircraft and the neighbors.

At major airports, where the maximum hourly number of arrivals and departures may be larger than 100, there will be several runways which can be used for aircraft operations. In this case, an airport can request that the FAA alternate among them, consistent with operational requirements, to relieve communities from constant noise.

Assessment. Airport noise is measured with noise monitors, consisting of microphones and meters. Most large airports have a distributed network of permanent noise monitors which record observed noise levels and accumulate the data for future transfer to the airport's computer. Small airports are most likely to have only portable measuring equipment or to use consultants. *See* NOISE MEASUREMENT.

Observations of aircraft operations are either direct or automated. With an automated system, an airport may have access to the FAA's data or to passive radar data. These data give the exact location, identity, and time of every operation in the area. Direct observations of aircraft operations are very reliable but are labor-intensive.

The radar data can be used to determine which operation affected a given resident at a given date and time. It can also be accumulated to determine the number of operations in a year by aircraft type on each of the airport's runways by time of day. The information can be used by the airport for assessing conformance with established noise abatement paths and procedures and for verifying operational reports from the airlines.

Combined with information on airline engine types and typical performance data on the aircraft type, the radar data can be used in airport noise prediction models, such as the FAA's Integrated Noise Model, to determine contours of noise exposure. These contours are averages of day-night equivalent sound levels for the year and can be compared with measured averages of the same type for locations for which the airport has noise monitors. These contours, in turn, can be used by the airport for planning purposes or for obtaining federal funds for noise mitigation programs.

If an airport has regulations, the above information may be used for enforcement. If the regulations are noise-level-based, microphone data will be used for this purpose. At airports where the regulations are operations-based, observations by airport personnel are often used to ensure compliance with local regulations. Complaints from the community may assist an airport in identifying potential violations. The FAA also assists the airport in ensuring compliance with noise regulations.

In deciding to adopt new noise mitigation at an airport, the FAA's regulations in Part 150 of the Code of Federal Regulations govern. A study of existing noise impacts is prepared, and mitigation strategies such as those discussed above are proposed. The FAA reviews the document and may approve it for federal funding of the proposed projects. Any new regulations would also need to be approved by the FAA. In addition, many states have environmental reporting requirements for their airports. Measurements of the noise impacts and new initiatives are properly filed with the appropriate authority. *See* ACOUSTIC NOISE.

Nancy S. Timmerman

Bibliography. M. J. Crocker (ed.), *Encyclopedia of Acoustics*, Wiley, New York, 1997; C. M. Harris, *Handbook of Acoustical Measurements*, 3d ed., Acoustical Society of America, Sewickley, PA, 1998; C. M. Harris, *Handbook of Noise Control*, 2d ed., McGraw-Hill, New York, 1979; H. H. Hubbard, *Aeroacoustics of Flight Vehicles: Theory and Practice*, Acoustical Society of America, Woodbury, NY, 1995.

Airport surface detection equipment

A ground mapping system that uses analog radar equipment to provide surveillance of aircraft and other surface vehicles on an airport surface. It is used by air-traffic controllers in airport control towers to monitor and control the movement of aircraft and vehicles. A situation display of the targets includes a map identifying the runways and taxiways and a visual map of the airport features, created through the contrast on the radar display resulting from the absence of returns from smooth concrete surfaces and the ground-clutter returns from grassy areas. An important safety function of the airport surface detection equipment (ASDE) is to determine whether or not a runway is clear for the next departure or arrival operation. The prevention of runway incursions by aircraft or by service vehicles has become

more urgent because of several incidents since 2003. This runway clearance determination is aided by the ASDE's capability to display an image of the aircraft in which the target's extremities (nose, tails, and wing tips), especially for large aircraft, are evident to the eye. *See* AIRPORT.

Operation. The ASDE antenna revolves at 60 revolutions per minute, providing a rapid update of target movements. It is located on a high vantage point, typically on top of the tower cab or on a special remote tower, that provides line-of-sight coverage of the desired runway and taxiway areas. Older ASDE systems have several inherent weaknesses, especially equipment failures, and excessive clutter during precipitation conditions, and are typically operated at a fixed frequency (for example, the ASDE-2 operates at 24 GHz).

Modern ASDEs utilize digital processing and an interrogation technique called frequency agility, which permits the transmission of up to 16 different Ku-band frequencies within a cycle. These techniques optimize target detection in the presence of ground and rain clutter and the rejection of false targets. In the digital processing, different thresholds for rejecting clutter are used on specified areas of the airport surface, such as on paved surfaces versus grassy areas. These systems are designed to detect all aircraft and vehicles having a radar cross section of 30 ft^2 (3 m^2) or greater, and will resolve two closely spaced targets when separated by 40 ft (12 m) in range or 80 ft (24 m) in azimuth at a range of 12,000 ft (3650 m). The basic coverage is the entire airport surface out to 24,000 ft (7300 m) in range and 200 ft (60 m) in altitude. Areas where coverage is not desired are ignored for detection and blanked out on the situation display.

Automation functions. The digital processing in modern ASDEs also provides the basis for automation functions that process target data to provide controllers with improved information. Principal automation features include target tracking, the application of runway safety logic functions to the tracks, and subsequent automatic alerting of the controller to dangerous situations that are determined by the logic. Advanced airport surface surveillance systems will enhance the data already provided by the ASDE, permitting more advanced automation functions. The design requirements for these systems include providing data that allows determination of target identity and target intent from flight-plan information or from data-link messages between the control tower and aircraft. *See* ELECTRONIC NAVIGATION SYSTEMS; MOVING-TARGET INDICATION; RADAR.

Robert A. Bales

Airship

A propelled and steered aerial vehicle, dependent on the displacement of air by a lighter gas for lift. An airship, or dirigible balloon, is composed primarily of a streamlined hull, usually a prolate ellipsoid which contains one or more gas cells, fixed and movable tail surfaces for control, a car or cabin for the crew or passengers, and a propulsion system.

Past designs. Two fundamentally different designs have been successfully used for past airships, the nonrigid and the rigid. A third type, the semirigid, is

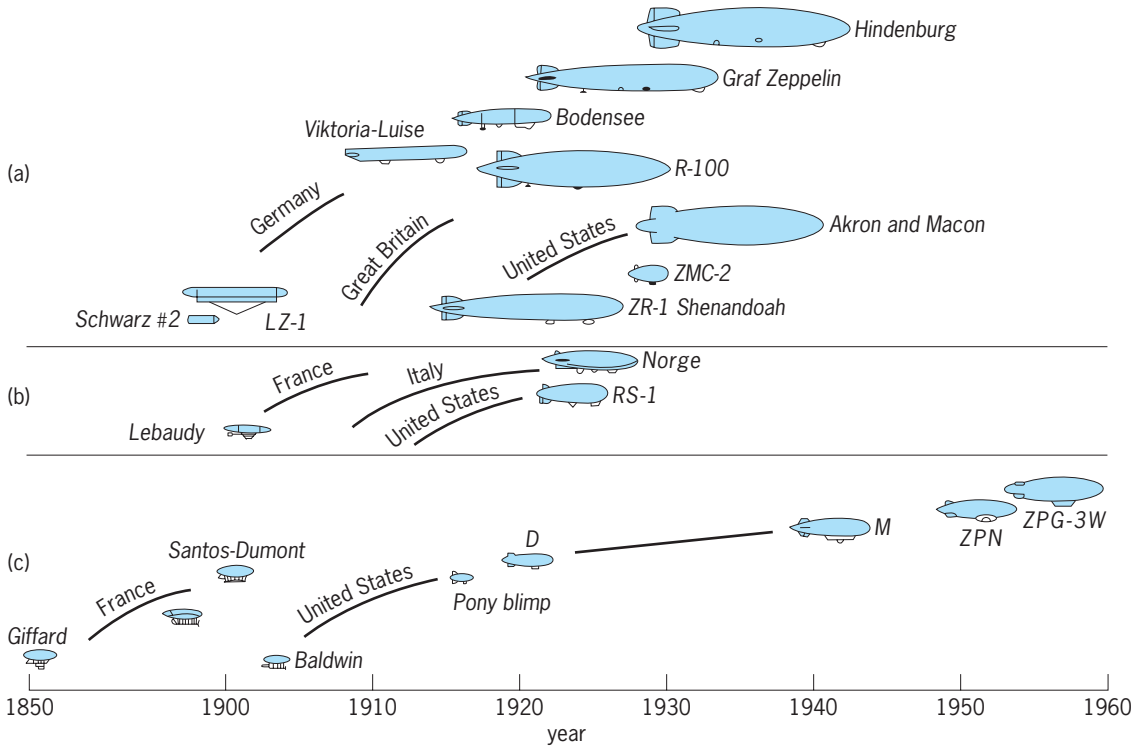


Fig. 1. History of airship development. (*a*) Rigid airships. (*b*) Semirigid airships. (*c*) Nonrigid airships.

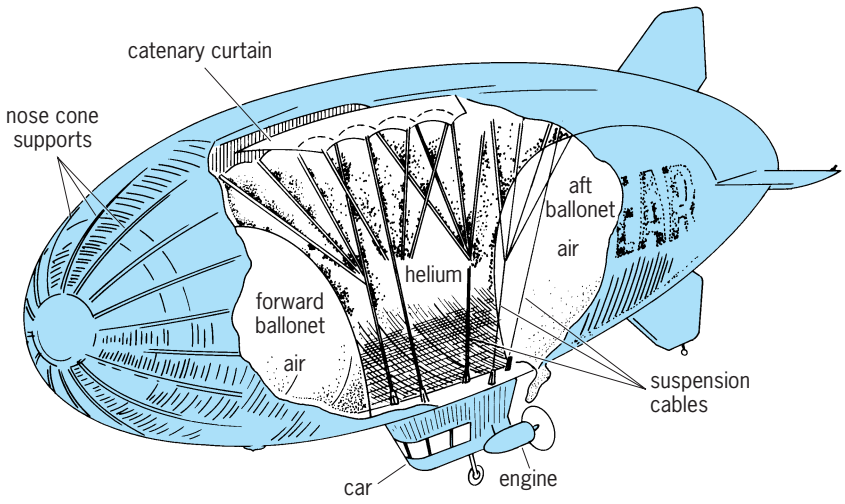


Fig. 2. Typical nonrigid airship.

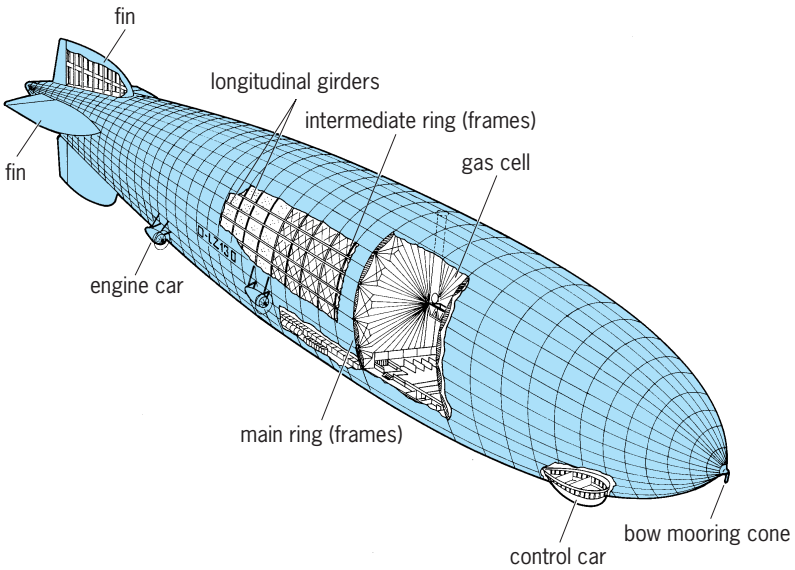


Fig. 3. Typical rigid airship.

Fig. 4. *Akron*, rigid airship of the U.S. Navy.

essentially a variant of the nonrigid type, differing by the addition of rigid keel. The principal development trends of the three types of conventional airships are depicted in **Fig. 1**.

Nonrigid. A typical nonrigid airship, or blimp (**Fig. 2**), consists of a flexible envelope, usually fabric, filled with lifting gas that is slightly pressurized. Internal air compartments (ballonets) expand and contract to maintain the pressure in the envelope as atmospheric pressure and temperature vary. Ballonet volume is controlled by ducting air from the prop wash or by electric blowers. The weights of the car structure, propulsion system, and other concentrated loads are supported by catenary systems attached to the envelope.

The nonrigid airships are historically significant for two reasons. First, a nonrigid airship was the first aircraft of any type to achieve controllable flight, in the 1850s. Second, nonrigid airships were the last type to be used on an extensive operational basis; the U.S. Navy decommissioned the last of its nonrigid airship fleet in the early 1960s. During the many years the Navy operated them, a high degree of availability and reliability was achieved. Most of these nonrigid airships were built by Goodyear, and a few, based on a modified pre-World War II Goodyear design, are used today for advertising by that company. *See* BLIMP.

Rigid. The other major type of airship is classified rigid because of its structure (**Fig. 3**). This structure was usually an aluminum ring-and-girder frame. An outer covering was attached to the frame to provide a suitable aerodynamic surface. Several gas cells were arrayed longitudinally within the frame. These cells were free to expand and contract, thereby allowing for pressure and temperature variations. Thus, despite their nearly identical outward appearance, rigid and nonrigid airships were significantly different in their construction and operation.

The rigid airship was developed primarily by the Zeppelin Company of Germany, and, in fact, rigid airships became known as zeppelins. Even the small percentage of rigid airships not built by this company was based, for the most part, on the Zeppelin concept. The rigid airships of the Zeppelin Company recorded some historical firsts in air transportation, including inaugurating the first passenger air service. The culmination of zeppelin development were the *Graf Zeppelin* and *Hindenburg* airships—unquestionably outstanding engineering achievements in their day. All of the rigid airships produced in the United States were for military purposes; none was in operation at the outbreak of World War II.

The last rigid airships built and operated in the United States were the *Akron* (**Fig. 4**) and *Macon*, built in 1931 and 1933. These airships were in many respects typical of the large rigid type but also had some unique design features. They contained just under 7,000,000 ft^3 (200,000 m^3) of helium, were a little over 800 ft (240 m) long, and had a gross lifting capability of about 450,000 lb (200,000 kg). Among the unique features of these airships were their eight engines. The propellers could be swiveled

in any vertical or fore and aft direction, which made the *Akron* and *Macon* two of the first aircraft to use vectored thrust propulsion. The panels on the side of the airship that looked like windows were actually condensers which collected water vapor out of the engine exhaust to replace weight on board the airship as fuel was burned. The internal rigid structure of the *Akron* and *Macon* was made of aluminum alloy. In fact, the use of aluminum alloy in aerospace structures was pioneered by rigid airships. The role of the *Akron* and *Macon* was to serve as flying aircraft carriers and scout vehicles. They carried five small airplanes on board to extend their range of surveillance, protect the airship, and serve as scouts for the surface fleet.

Commercial operations. The only significant past commercial airship operations were those of the Zeppelin Company and its subsidiary DELAG. This organization began carrying passengers on flights within Germany in 1910, and from 1929 to 1937 provided a unique transatlantic air service with the *Graf Zeppelin* and *Hindenburg*. Prior to the burning of the *Hindenburg* in 1937, there had been no fatal accidents in commercial service. However, none of these commercial operations can be considered a financial success, and most were subsidized by the German government.

Modern vehicle concepts. **Figure 5** shows a representative sample of the many modern airship vehicle concepts which have been proposed. The fully buoyant conventional concepts are modern versions of the classical, fully buoyant, rigid and nonrigid airship concepts. Fully buoyant means all the lift is provided by displacement. These airships would make extensive use of modern aircraft structural materials, propulsion systems, control systems, and electronics.

The other four concepts in Fig. 5 are partially buoyant, or hybrid, designs in which the buoyant lift is substantially augmented by aerodynamic or propulsive lift. Thus, these vehicles are partly heavier than air and partly lighter than air. Two of the hybrids are short takeoff and landing (STOL) vehicles, and two have vertical takeoff and landing (VTOL) capability. Generally speaking, the partially buoyant concepts would have higher cruise speed, higher fuel consumption, and higher structural weight as compared with the classical concepts. It will therefore depend on the specific applications as to which is the best concept. An important advantage of the hybrids is that they promise to alleviate the costly ground-handling requirement of past airship designs. *See* SHORT TAKEOFF AND LANDING (STOL); VERTICAL TAKEOFF AND LANDING (VTOL).

The Dynairship is representative of lifting-body concepts in which the hull is shaped to be more aerodynamically efficient than the conventional airship shape. This particular concept has a delta platform. The vehicle is flown at positive angle of attack in cruise to generate aerodynamic lift. The Megalifter concept is a typical example of a winged airship. The Helipsoid is intended to be a compromise between the ellipsoidal and deltoid shapes. It will have a better surface-area/volume ratio than the Dynairship, at the expense of the degraded stability and control characteristics. Finally, the Heli-stat, which is representative of the rotary-wing hybrids, combines an airship hull with helicopters or helicopter rotor systems. *See* BALLOON.

Mark R. Ardema

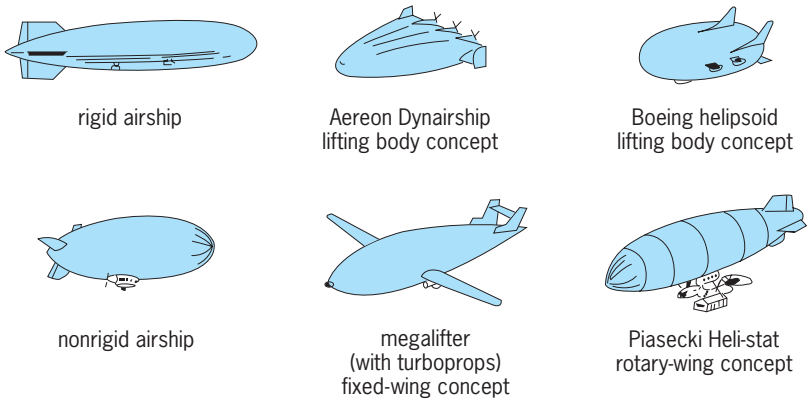
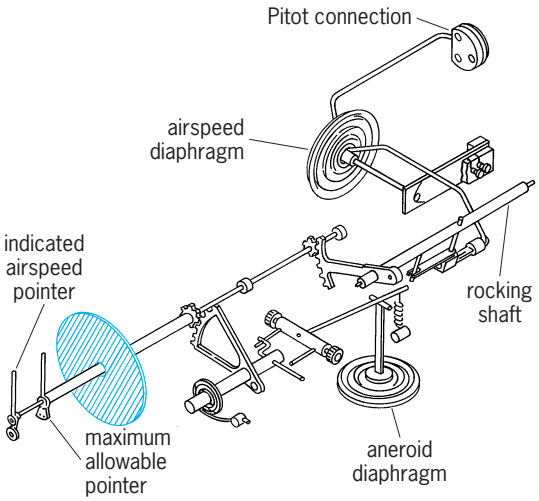


Fig. 5. Modern airship concepts.

Airspeed indicator

A device that computes and displays speed of an aircraft relative to the air mass in which the aircraft is flying. Three common types are the indicated airspeed meter, true airspeed indicator, and Machmeter.

The commonest type is the indicated airspeed meter, which measures differential pressure between the total ram pressure from the Pitot system and the total static pressure; it then converts this difference into units of speed (miles per hour or knots) under standard conditions, 29.9212 in. Hg (101.325 kilopascals) absolute and 59°F (15°C; see **illus.**). Although the indicated values are incorrect above zero altitude, the relationship to the aircraft handling remains essentially unchanged, thus providing a measure of the flyability of the aircraft. These instruments



Components of indicated airspeed meter.

are frequently supplemented with an auxiliary mechanism, driving a second pointer which computes and indicates a value known as the maximum allowable speed for the type of aircraft in which the instrument is installed.

True airspeed indicators are similar to indicated airspeed meters but include a more complex mechanism that also senses both the absolute pressure and temperature, and compensates for the change of density of the air mass, thus obtaining true airspeed. This indication is of value in computing course information; hence the true airspeed indicator is a navigation instrument whereas the indicated airspeed indicator is a flight instrument.

For those aircraft that reach higher speeds (transonic and supersonic), the ratio of the actual speed to the local speed of sound is used. Devices that compute this value are known as Machmeters. Relative Mach number is computed by determining the ratio of total ram pressure divided by total static pressure and converting the ratio into Mach number, according to the applicable formulas. Certain instruments combine airspeed, Mach number, and maximum allowable speed in one indicator to provide a central speed-data display. *See* MACH NUMBER; PITOT TUBE.

James W. Angus

Aistopoda

An order of extremely elongate, limbless fossil amphibians in the subclass Lepospondyli from Permo-Carboniferous rocks of North America and the British Isles. The order includes three families: Lethiscidae (*Lethiscus*), Ophiderpetontidae (*Coloraderpeton*, *Ophiderpeton*), and Phlegethontiidae (*Aornerpeton, Phlegethontia, Sillerpeton*). All genera are monotypic with the exception of *Ophiderpeton* (six species) and *Phlegethontia* (possibly three species). In seeming contrast to their derived state, aistopods are among the very oldest of all known fossil tetrapods.

The skulls of aistopods are fenestrated and exhibit a reduced number of bony elements. This fenestration is extreme in phlegethontiids, whose skulls may have been kinetic.

The vertebral centra are holospondylous (single-pieced), hourglass-shaped, and fused to their neural arches. Vertebrae can exceed 200 in number and frequently bear foramina for passage of spinal nerves. Ribs typically bear a unique process, which gives them a K-shape in some species.

Lethiscids and ophiderpetontids possess a ventral armor of fusiform, bony elements arranged in a chevron pattern. Such armor in phlegethontiids consists of more slender, threadlike elements.

Most aistopods were presumably aquatic, although rib specializations, and the more gracile proportions of phlegethontiids, suggest a rather snake-like, terrestrial habit. *See* AMPHIBIA; LEPOSPONDYLI.

Carl F. Wellstead

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988.

Albedo

A term referring to the reflecting properties of a surface. White surfaces have albedos close to 1; black surfaces have albedos close to 0.

Bond albedo and normal reflectance. Several types of albedos are in common use. The Bond albedo (A_B) determines the energy balance of a planet or satellite and is defined as the fraction of the total incident solar energy that the planet or satellite reflects to space. The "normal albedo" of a surface, more properly called the normal reflectance (r_n), is a measure of the relative brightness of the surface when viewed and illuminated vertically. Such measurements are referred to a perfectly white Lambert surface—a surface which absorbs no light and scatters the incident energy isotropically—usually approximated by magnesium oxide (MgO), barium sulfate ($BaSO_4$), or some other bright material. *See* PHOTOMETRY; PLANET.

Bond albedos for solar system objects range from 0.9 for Saturn's icy satellite Enceladus and Neptune's Triton to values as low as 0.01–0.02 for dark objects such as the satellites of Mars (**Table 1**). Cloud-shrouded Venus has the highest Bond albedo of any planet (0.76). The value for Earth is typically 0.35, but varies with the extent of cloud and snow cover. The Bond albedo is defined over all wavelengths, and its

TABLE 1. Bond albedos and visual geometric albedos of selected solar system objects

Object	Bond albedo (A_B)	Visual geometric albedo (p_v)
Mercury	0.12	0.14
Venus	0.76	0.59
Earth	0.35	0.37
Mars	0.24	0.15
Jupiter	0.34	0.45
Saturn	0.34	0.46
Uranus	0.34	0.48
Neptune	0.28	0.50
Pluto	0.5(?)	0.61
Moon	0.12	0.14
Phobos (M1)	0.02	0.05
Deimos (M2)	0.02	0.06
Io (J1)	0.56	0.63
Europa (J2)	0.58	0.68
Ganymede (J3)	0.38	0.43
Callisto (J4)	0.13	0.17
Amalthea (J5)	0.02	0.06
Mimas (S1)	0.60	0.75
Enceladus (S2)	0.90	1.20
Tethys (S3)	0.60	0.80
Dione (S4)	0.45	0.55
Rhea (S5)	0.45	0.65
Titan (S6)	0.20	0.20
Ariel (U1)	0.21	0.39
Umbriel (U2)	0.10	0.21
Titania (U3)	0.15	0.27
Oberon (U4)	0.12	0.23
Miranda (U5)	0.18	0.32
Triton (N1)	0.90	0.75
Nereid (N2)	0.07	0.14
Proteus (N8)	0.03	0.06
Ceres	0.03	0.06
Vesta	0.12	0.23
Comet Halley (nucleus)	0.01	0.03

TABLE 2. Normal reflectances of materials*

Material	Albedo
Lampblack	0.02
Charcoal	0.04
Carbonaceous meteorites	0.05
Volcanic cinders	0.06
Basalt	0.10
Iron meteorites	0.18
Chondritic meteorites	0.29
Granite	0.35
Olivine	0.40
Quartz	0.54
Pumice	0.57
Snow	0.70
Sulfur	0.85
Magnesium oxide	1.00

* Powders; for wavelengths near 0.5 micrometer.

value therefore depends on the spectrum of the incident radiation. For objects in the outer solar system not yet visited by spacecraft (such as Pluto), the values of A_B in Table 1 are estimates derived indirectly, since for these bodies it is impossible to measure the scattered radiation in all directions from Earth.

Normal reflectances of some common materials are listed in **Table 2**. The normal reflectances of many materials are strongly dependent on wavelength, a fact that is commonly used in planetary science to infer the composition of surfaces remotely. While the Bond albedo cannot exceed unity, the normal reflectance of a surface can do so if the material is more backscattering at opposition than the reference surface.

Geometric albedo. In the case of solar system objects, a third type of albedo, the geometric albedo (p), is commonly defined. It is the ratio of incident sunlight reflected in the backscattering direction (zero phase angle or opposition) by the object, to that which would be reflected by a circular disk of the same size but covered with a perfectly white Lambert surface. Objects like the Moon, covered with dark, highly textured surfaces, show uniformly bright disks at opposition (no limb darkening); for these, $p = r_n$. At the other extreme, a planet covered with a Lambert-like visible surface (frost or bright cloud) will be limb-darkened as the cosine of the incidence angle at opposition, and $p = {}^2/_3 r_n$. Table 1 lists the visual geometric albedo p_v, which is the geometric albedo at a wavelength of 550 nanometers.

Phase integral. For solar system objects, the ratio A_B/p is called the phase integral and is denoted by q. Here p is the value of the geometric albedo averaged over the incident spectrum, and does not generally equal p_v given in Table 1. Values of q range from near 1.5 for cloud-covered objects (1.4 for Saturn and 1.3 for Venus) to 0.2 for the very dark and rugged satellites of Mars: Phobos and Deimos. The value for the Moon is about 0.6.

Other albedo types. The quantities AB, p, and q quantify the reflection of light averaged over an entire hemisphere of a planet. They are most appropriately used when details of the distribution of the reflectance on a planet's surface are beyond the resolution of the observations. It is possible to define other types of albedos based on the geometry and nature (diffuse or collimated) of the incident beam and scattered beams. For such bidirectional albedos or reflectances, the angles of incidence and scattering as well as the angle between these two directions (phase angle) must be specified. Finally, in determining the energy balance of surfaces, one is often interested in the analog for a flat surface of the Bond albedo, that is, the fraction of the incident energy that the surface reflects. For a plane parallel beam incident at an angle i, this quantity, $A(i)$, is generally close to the value A_B, but can be significantly larger than A_B for some surfaces, for large values of i.

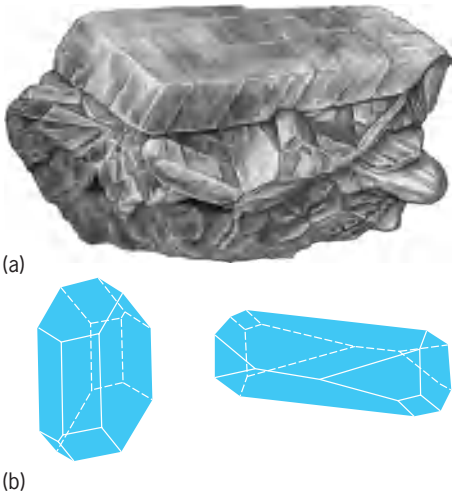
Role in planetary astronomy. The role that these albedos play in planetary astronomy is illustrated by the discovery of a distant new planet in the solar system. The distance of the planet from the Earth and Sun is accurately known from its motion; determination of its orbit requires only a small number of measurements of its position relative to the stars over a few months. The total brightness of the planet is the product of its area and (geometric) albedo. For a distant or small object whose size cannot be resolved with even the largest telescopes, an albedo may be assumed, allowing its size to be estimated from its brightness alone. What is known about the planet can be improved if thermal radiation from the planet is detected at infrared wavelengths. The quantity of incident solar energy absorbed by a planet is $(1 - A_B)$, that is, energy that is not reflected is absorbed. In this way, the Bond albedo determines the temperature of a planet. The absorbed sunlight heats the planet until it reaches an equilibrium temperature where the total energy it radiates at infrared wavelengths equals the total solar energy absorbed. The infrared brightness depends on the area of the planet and its temperature, so if both the infrared (thermal emission) and visual brightness (reflected sunlight) are measured, both the size and the albedo of a distant unresolved planet can be determined unambiguously. Using this technique, the sizes and albedos of thousands of asteroids and other objects that populate the solar system are now known. *See* HEAT RADIATION; INFRARED ASTRONOMY; RADIOMETRY.

Joseph Veverka; Jay Goguen

Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin, *The New Solar System*, 4th ed., Sky Publishing, Cambridge, MA, 1999; B. Hapke, *Theory of Reflectance and Emittance Spectroscopy*, Cambridge University Press, 1993; M. Harwit, *Astrophysical Concepts*, 4th ed., Springer, 2006.

Albite

A sodium-rich plagioclase feldspar mineral whose composition extends over the range $Ab_{100}An_0$ to $Ab_{90}An_{10}$, where Ab (= albite) is $NaAlSi_3O_8$ and An (= anorthite) is $CaAl_2Si_2O_8$ (see **illus.**). Albite occurs in crustal igneous rocks as a major component of pegmatites and granites, in association with quartz, mica (usually muscovite), and potassium



Albite. (*a*) Crystals, Amelia Court House, Virginia (*specimen from Department of Geology, Bryn Mawr College*). (*b*) Crystal habits (*after D. S. Hurlbut, Jr., Dana's Manual of Mineralogy, 17th ed., Wiley, 1959*)

feldspar (orthoclase or microcline). Sodium and potassium feldspars usually occur as distinct mineral grains, sizes varying from millimeter to meter scale. However, they are frequently intergrown, having exsolved from a single phase at high temperatures. If the intergrowth is visually observable in a hand specimen, the composite material is known as macroperthite; if visible only in a microscope, microperthite; and if submicroscopic in scale, cryptoperthite. In metamorphic rocks albite is found in granitic gneisses, and it may be the principal component of arkose, a feldspar-dominant, sedimentary rock. Cleavelandite, a platy variety, is sometimes found in lithium-rich pegmatites. *See* ARKOSE; GNEISS; PERTHITE.

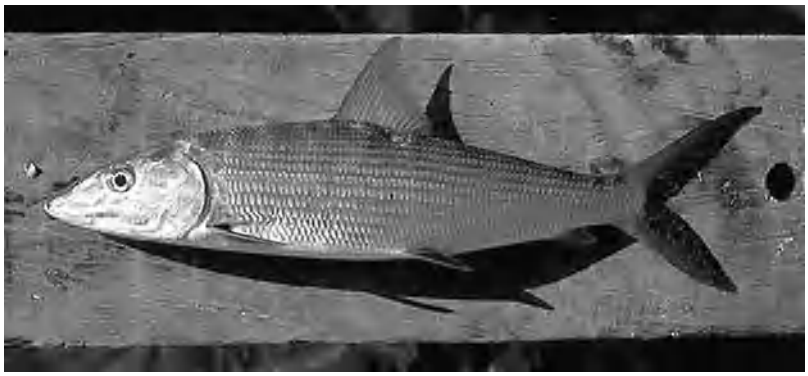
Its hardness is 6 on the Mohs scale, specific gravity 2.62, and melting point 1118°C (2044°F). If observed at temperatures above 985°C (1805°F), albite has monoclinic symmetry, provided that the distribution of aluminum (Al) and silicon (Si) among the tetrahedral sites of its framework structure is random (disordered). Below 958°C (1805°F) the symmetry becomes triclinic as the $[AlSi_3O_8]_\infty^{1-}$ tetrahedral framework collapses around the small sodium atom, primarily through the mechanism of intertetrahedral angle bending. With annealing at lower temperatures, Al and Si become completely ordered, forming what is called low albite. If the An component exceeds 2–3 mole %, this feldspar exsolves into two distinct, submicroscopic phases, one of nearly pure albite (An_{0-1}) composition, and the other of composition between An_{15} and An_{25}; the resultant lamellar intergrowth of these two phases is called peristerite, with optical interference colors resembling those on the neck feathers of a pigeon. Such material, especially that from Hybla, Ontario, Canada, is valued as a semiprecious gemstone and may be cut and polished into cabochons. Albite is used in ceramic materials, as a source of soda (Na_2O) and alumina (Al_2O_3) in glasses, and in certain, relatively soft abrasives. *See* FELDSPAR; GEM; IGNEOUS ROCKS; PEGMATITE.

Paul H. Ribbe

Albuliformes

An order of actinopterygian fishes in the subdivision Elopomorpha, along with Elopiformes, Anguilliformes, and Saccopharyngiformes, all of which have a leptocephalous larval stage. Albuliformes, comprising three families, eight genera, and 30 species, are distinguished from other Elopomorpha by an open mandibular sensory canal in the dentary and angular bones. Two suborders are recognized, Albuloidei and Notacanthoidei, the latter of which was recognized previously as the order Notacanthiformes. *See* ACTINOPTERYGII; ANGUILLIFORMES; ELOPIFORMES; NOTACANTHOIDEI.

Albuloidei, consisting of the family Albulidae, are characterized by a fusiform body; the tip of the snout overhanging an inferior mouth; a mouth bordered by premaxillae; a gular plate having a thin median splint or being entirely absent; an infraorbital lateral-line canal extending onto the premaxillae; and caudal fin rays supported by six hypural bones. Albulidae contains two subfamilies, Albulinae with one genus (*Albula*) and three species and Pterothrissinae with one genus (*Istieus*) and two species. *Albula* (see **illustration**) differs from *Istieus* (formerly *Pterothrissus*) in having 16–21 dorsal fin rays versus 55–65; a small gular plate versus no gular plate; and the maxillae toothless versus maxillae toothed. *Albula* occurs in tropical seas, rarely entering brackish and fresh water, and attains a length of 105 cm (41 in.).



***Albula vulpes*. (*Courtesy of C. Q. Jessen*)**

Istieus occurs in the eastern Atlantic from Mauritania to Namibia and in the western Pacific from China and Japan, and probably does not exceed 40 cm (16 in.) in length. *Albula* can tolerate low oxygen tension by "inhaling" air into its lunglike air bladder. It is benthic in habitat and feeds on invertebrates. *Istieus* is also benthic but usually occurs in deeper waters of continental slopes. Herbert Boschung

Bibliography. J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; D. G. Smith, Albulidae (pp. 683–684), Halosauridae (685–687), Notacanthidae (688–689), in K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic*, FAO Species Identification Guide for Fishery Purposes, vol. 2, FAO, Rome, 2002.

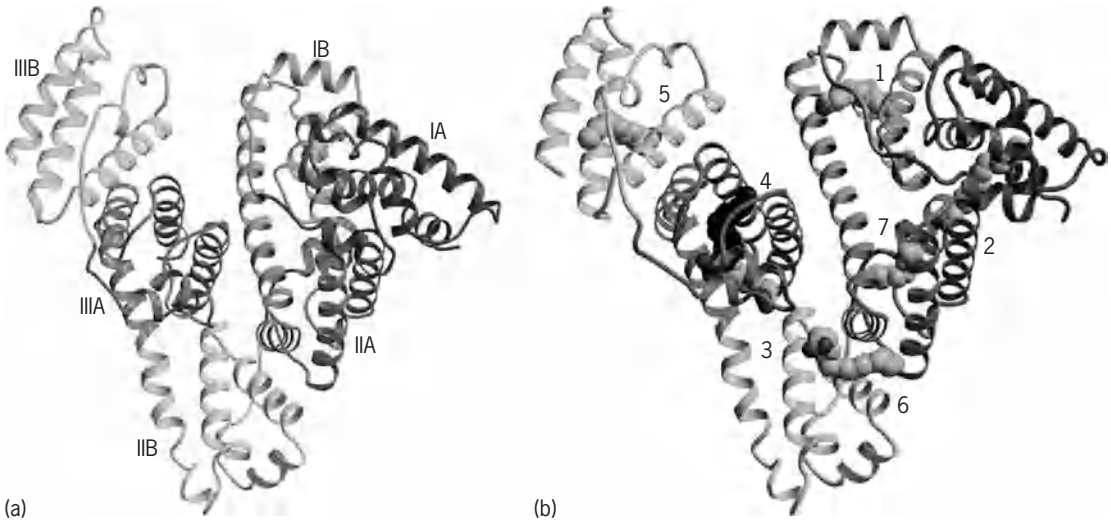
Albumin

A plasma protein produced by the liver that maintains fluid balance in the blood and transports fatty acids in the plasma and interstitial fluid. It is the most abundant protein in human serum, and one of the first discovered and earliest studied proteins. Serum albumin was precipitated from urine in 1500 by Paracelsus, and was crystallized in 1894 by A. Gürber. Probably no other protein has been studied as extensively as serum albumin, and our knowledge of its structure and interactions with its ligands has come from many researchers, using a great variety of experimental approaches. Its ability to bind many different ligands, most of which are hydrophobic anions, and several molecules of the same ligand (fatty acid) is well documented. Fatty acids, bilirubin (an orange-yellow bile pigment formed in the breakdown of red blood cells), and hematin (a blue to blackish-brown compound formed in the decomposition of hemoglobin) represent the endogenous ligands of albumin with highest affinity. *See* PROTEIN.

Structure. A major breakthrough in albumin research came in the early 1990s, when x-ray diffraction studies of crystalline human albumin provided the first accurate images of this remarkable protein. The high-resolution structure revealed an overall heart shape, both in the defatted form and with bound fatty acids (see **illustration**). The protein comprises three domains, each with two subdomains, as predicted by many studies. The fatty acids are bound throughout the protein in structurally distinct sites with similar features: the carboxyl group of each fatty acid forms a salt bridge or a hydrogen bond with basic and polar amino acid side chains, and the hydrocarbon chain of the fatty acid fits into a hydrophobic cavity between helices.

Physiological function. In human plasma, albumin is the major buffer for pH variations and is essential for maintaining the plasma colloidal osmotic pressure, causing fluid to remain in the blood instead of leaking out into the tissues. Albumin is the main transport vehicle for fatty acids for utilization in various tissues. As albumin circulates through the vascular system, fatty acids desorb rapidly from the protein in spite of their high affinity for the binding sites (see illustration) and diffuse through the adjacent cell membranes to enter cells. Albumin also binds fatty acids that are released into the plasma from cells such as fat cells for transport to other tissues, where the fatty acids serve as fuel or substrates for lipid synthesis. *See* BLOOD; LIPID METABOLISM.

Role in disease. Low plasma levels of albumin are associated with several diseases, such as diabetes, liver disease, and cardiovascular disease. Undetectably low or very low plasma levels of albumin are found in an extremely rare inherited disorder known as analbuminemia. There is inadequate compensation for albumin's role in maintaining plasma colloidal osmotic pressure, which is very low in affected individuals; however, the plasma lipoproteins



Human serum albumin. (*a*) Domain structure of the defatted protein and (*b*) location of eight myristic acid (a 14-carbon fatty acid) ligands in different binding sites. In *a*, the secondary structure of the protein is shown schematically, and the domains are designated I, II, III. The subdomains are A and B. Fatty acids with different chain lengths are present in the plasma, and the binding affinity increases with increasing chain length of the fatty acid, up to a length of 20 carbons. In *b*, showing the binding of the myristic acid, the atom type is differently shaded to show carbon and oxygen. Myristic acid in site 4 is darker to distinguish it from myristic acid in site 3.

take over the role of transport of fatty acids and bilirubin. Analbuminemia is not a lethal disorder, although edema, fatigue, and lipodystrophy (a disturbance of fat metabolism in which the subcutaneous fat disappears over some regions of the body but is unaffected in others) are common complaints. The high concentration of albumin in the general population, therefore, is not essential for life but appears to be important for optimal health.

Drug transport. Human serum allumin (HSA) has two principal drug-binding sites (sites I and II in the classical literature), which have been localized in the crystal in subdomains IIA and IIIA (see illustration). Site I binds bulky heterocyclic molecules with a delocalized negative charge in the center of a predominantly hydrophobic structure, including warfarin, phenylbutazone, thyroxine, benzodiazepines, antidiabetic agents, and tryptophan. Site II binds aromatic carboxylic acids that are ionized at physiological pH and contain a negative charge at one end of the molecule, including propofol, diazepam, ibuprofen, acetylsalicylic acid (aspirin), bilirubin, and thyroxine. Solubilization of these hydrophobic compounds by albumin contributes to a more homogeneous distribution of the drug in the body, reduces the volume of distribution, lowers the rate of clearance, and increases the plasma half-life of the drug. *See* ASPIRIN; BILIRUBIN; HETEROCYCLIC COMPOUNDS; THYROXINE.

James A. Hamilton

Bibliography. J. R. Brown and P. Shockley, Serum albumin: Structure and characterization of its ligand binding sites, Wiley, New York, 1982; J. K. Choi et al., Interactions of very long-chain saturated fatty acids with serum albumin, *J. Lipid Res.*, 43:1000–1010, 2002; S. Curry et al., Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites, *Nat. Struc. Biol.*, 5:827–835, 1998; C. Daniels, N. Noy, and D. Zakim, Rates of hydration of fatty acids bound to unilamellar vesicles of phosphatidylcholine or to albumin, *Biochemistry*, 24:3286–3292, 1985; J. A. Hamilton, Fatty acid transport: Difficult or easy, *J. Lipid. Res.*, 39:467–481, 1998; X. M. He and D. C. Carter, Atomic structure and chemistry of human serum albumin, *Nature*, 358:209–215, 1992 [published erratum appears in *Nature*, 364(6435):362, July 22, 1993]; T. Peters, Jr., *All About Albumin*, Academic Press, San Diego, 1996; A. A. Spector, Fatty acid binding to plasma albumin, *J. Lipid Res.* 16:165–179, 1995.

Alcohol

A member of a class of organic compounds composed of carbon, hydrogen, and oxygen. They can be considered as hydroxyl derivatives of hydrocarbons produced by the replacement of one or more hydrogens by one or more hydroxyl (—OH) groups.

Classification. Alcohols may be mono−, di−, tri−, or polyhydric, depending upon the number of hydroxyl groups they possess. They are classified as primary (RCH_2OH), secondary (R_2CHOH), or tertiary (R_3COH), depending on the number of hydrogen atoms attached to the carbon atom bearing the hydroxyl group. Alcohols can also be characterized by the molecular configuration of the hydrocarbon portion (aliphatic, cyclic, heterocyclic, or unsaturated). There are two systems in use for alcohol nomenclature, the common naming system and the IUPAC (International Union of Pure and Applied Chemistry) naming system. The common name is sometimes associated with the natural source of the alcohol or with the hydrocarbon portion (for example, methyl alcohol, ethyl alcohol). The IUPAC method is a systematic procedure with agreed-upon rules. The name of the alcohol is derived from the parent hydrocarbon which corresponds to the longest carbon chain in the alcohol. The final "e" in the hydrocarbon name is dropped and replaced with "ol"; and a number before the name indicates the position of the hydroxyl. Examples of these two systems are given in the **table**.

Reactions. Oxidation of primary alcohols produces aldehydes (RCHO) and carboxylic acids (RCO_2H); oxidation of secondary alcohols yields ketones ($RCOR'$). Dehydration of alcohols produces alkenes and ethers (ROR). Reaction of alcohols with carboxylic acids results in the formation of esters ($ROCOR'$), a reaction of great industrial importance. The hydroxyl group of an alcohol is readily replaced by halogens or pseudohalogens. *See* ALDEHYDE; ALKENE; CARBOXYLIC ACID; ESTER; ETHER; KETONE.

Uses. Industrially, the monohydric aliphatic alcohols are classified according to their uses as the lower alcohols (1–5 carbon atoms), the plasticizer-range alcohols (6–11 carbon atoms), and the detergent-range alcohols (12 or more carbon atoms). The lower alcohols are employed as solvents, extractants, and antifreezes. Esters of the lower alcohols are employed extensively as solvents for lacquers, paints, varnishes, inks, and adhesives. The plasticizer-range alcohols find their primary use in the form of esters as plasticizers and also as lubricants in high-speed applications such as jet engines. The detergent-range alcohols are used in the form of sulfate and ethoxysulfate esters in detergents and surfactants.

Ethanol and methanol are volatile, combustible substances, but their use as motor fuel is limited by economic considerations. In the United States, ethanol made by fermentation of agricultural products is blended with gasoline to produce a motor fuel called gasohol, but this constitutes an insignificant fraction of the energy supply. In Brazil, where sugarcane is a major agricultural product, the large-scale production of ethanol by fermentation has been promoted by the government, and ethanol is a significant and rapidly growing component of the automotive fuel supply. *See* ETHYL ALCOHOL; METHANOL.

Sources. Alcohols are derived either from natural-product processing, such as the fermentation of carbohydrates and the reductive cleavage of natural fats and oils, or by chemical synthesis based on the hydrocarbons derived from petroleum or the synthesis gas from coal.

Alcohols and their formulas

Name		
Common	IUPAC	Formula
Methyl alcohol	Methanol	CH_3OH
Ethyl alcohol	Ethanol	CH_3CH_2OH
n-Propyl alcohol	1-Propanol	$CH_3CH_2CH_2OH$
Isopropyl alcohol	2-Propanol	$(CH_3)_2CHO$
n-Butyl alcohol	1-Butanol	$CH_3(CH_2)_2CH_2OH$
sec-Butyl alcohol	2-Butanol	$CH_3CH_2CHOHCH_3$
tert-Butyl alcohol	2-Methyl-2-propanol	$(CH_3)_3COH$
Isobutyl alcohol	2-Methyl-1-propanol	$(CH_3)_2CHCH_2OH$
n-Amyl alcohol	1-Pentanol	$CH_3(CH_2)_3CH_2OH$
n-Hexyl alcohol	1-Hexanol	$CH_3(CH_2)_4CH_2OH$
Allyl alcohol	2-Propen-1-ol	$CH_2{=}CHCH_2OH$
Crotyl alcohol	2-Buten-1-ol	$CH_3CH{=}CHCH_2OH$
Ethylene glycol	1,2-Ethanediol	$HOCH_2CH_2OH$
Propylene glycol	1,2-Propanediol	$CH_3CHOHCH_2OH$
Trimethylene glycol	1,3-Propanediol	$HOCH_2CH_3CH_2OH$
Glycerol	1,2,3-Propanetriol	$CH_2OHCHOHCH_2OH$

The fermentation of sugars and starches (carbohydrates) to produce alcoholic beverages has been employed at least since history has been recorded. The industrial fermentation process is the biological transformation of a carbohydrate by a highly specialized strain of yeast to produce the desired product, such as ethanol, or 1-butanol and acetone. Fermentation is no longer the major source of 1-butanol, but still accounts for all potable ethanol and a large proportion of the ethanol used industrially worldwide. As sources of hydrocarbons based on petroleum continue to be depleted, fermentation processes based on renewable raw materials are likely to become more important. *See* DISTILLED SPIRITS; FERMENTATION; INDUSTRIAL MICROBIOLOGY.

Syntheses. The major processes employed industrially to produce aliphatic monohydric alcohols are the reduction of synthesis gas, the hydration and oxonation of hydrocarbons, the condensation of aldehydes followed by reduction, the Ziegler process, and the reduction of animal fats and vegetable oils.

Reduction. Methanol, the simplest alcohol, is produced almost entirely by the catalytic reduction of synthesis gas (carbon monoxide and hydrogen), as shown in reaction (1).

$$\underset{\text{Carbon monoxide}}{CO} + \underset{\text{Hydrogen}}{2H_2} \longrightarrow \underset{\text{Methanol}}{CH_3OH} \qquad (1)$$

Hydration. Treatment of lower unsaturated hydrocarbons with water in the presence of an acidic catalyst is employed commercially to produce ethanol and 2-propanol, as shown in reactions (2) and (3).

$$\underset{\text{Ethylene}}{CH_2{=}CH_2} + H_2O \longrightarrow \underset{\text{Ethanol}}{CH_3CH_2OH} \qquad (2)$$

$$\underset{\text{Propylene}}{CH_3CH{=}CH_2} + \underset{\text{Water}}{H_2O} \longrightarrow \underset{\text{2-Propanol}}{CH_3\underset{\displaystyle OH}{\underset{|}{C}}HCH_3} \qquad (3)$$

Most of the synthetic ethanol in the United States is made by this method.

Oxonation. A modification of the Fischer-Tropsch process, known as the oxo reaction, involves the treatment of unsaturated hydrocarbons with synthesis gas at high temperatures under high pressure in the presence of a cobalt catalyst; it is a major source of monohydric aliphatic alcohols ranging from 1-propanol to the tridecanols. A mixture of isomeric alcohols is generally obtained by this process, shown in reaction (4).

$$\underset{\text{Ethylene}}{CH_2{=}CH_2} + \underset{\text{Carbon monoxide}}{CO} + \underset{\text{Hydrogen}}{H_2} \longrightarrow \underset{\text{Propionaldehyde}}{CH_3CH_2CHO} \xrightarrow{H_2} \underset{\text{1-Propanol}}{CH_3CH_2CH_2OH} \qquad (4)$$

See FISCHER-TROPSCH PROCESS.

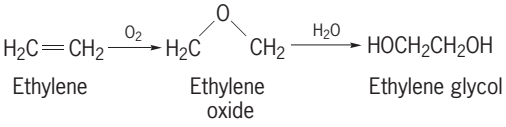
Condensation. Some of the higher aliphatic alcohols, such as 2-ethylhexanol, are synthesized by the condensation of aldehydes followed by dehydration and reduction. This base-catalyzed condensation, known as the aldol reaction, is shown in reaction (5).

$$\underset{n\text{-Butyraldehyde}}{2CH_3CH_2CHO} \xrightarrow{\text{base}} \underset{\text{Butyraldol}}{CH_3CH_2CH_2\overset{\displaystyle OH}{\overset{|}{C}}H\underset{\displaystyle CH_2CH_3}{\underset{|}{C}}HCHO} \xrightarrow[(2)\ H_2]{(1)\ -H_2O} \underset{\text{2-Ethylhexanol}}{CH_3CH_2CH_2CH_2\underset{\displaystyle CH_2CH_2}{\underset{|}{C}}HCH_2OH} \qquad (5)$$

Ziegler process. Formerly, the major source of detergent range alcohols was hydrogenolysis of animal fats and vegetable oils, but the Ziegler process, using aluminum alkyls as the catalyst and ethylene from the cracking of petroleum as the feedstock, has become the major source of industrial supply. In this sequence of reactions, a trialkyl aluminum such as

triethyl aluminum [$Al(C_2H_5)_3$] is alkylated with ethylene to give the desired higher trialkyl aluminum [AlR_3], which is then oxidized to the trialkoxy aluminum [$Al(OR)_3$], which is hydrolyzed to give a mixture of primary alcohols in the desired molecular weight range. *See* STEREOSPECIFIC CATALYST.

Ethylene glycol and propylene glycol are the most industrially significant polyhydric alcohols. They are made from ethylene, as shown in reaction (6), and



(6)

propylene, by means of their intermediate oxides. The oxides are synthesized by catalytic oxidation of the olefins. *See* ESTER; PHENOL; POLYOL. Paul E. Fanta

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, vol. 1, 4th ed., 1991.

Alcohol fuel

Any alcohol burned as a fuel. Generally, the term alcohol refers to ethanol, the first organic chemical produced by humans. If available, any alcohol may be used as a fuel, but ethanol and methanol are the only ones which are sufficiently inexpensive. Alcohols are useful fuels, even though the oxygen atom in any alcohol molecule reduces its heating value, because the molecular structure increases the combustion efficiency. *See* COMBUSTION; GASOLINE.

Ethanol. Alcohol (ethanol) was used as a motor fuel, either alone or in mixtures with gasoline (gasohol), in many countries during the early 1930s, and it became an important alternative fuel for automobiles again in the 1970s and 1980s. It burns well in engines designed for gasoline and has a high octane rating. However, because of the high cost of the raw materials required, or of their conversion, it costs much more than gasoline. Gasohol is produced by dissolving 5–15% absolute water-free alcohol in gasoline. The 5% water fraction in 95% alcohol causes phase separation because of its insolubility, making 95% alcohol unsuitable for use in gasohol. This is first seen as cloudiness, then as small droplets of a water-alcohol solution, and finally as two distinct layers. Such phase separation also occurs because alcohol is hygroscopic; that is, it tends to absorb water from the air and cause dilution, either before or after mixing with gasoline.

Gasoline, alcohol, and water have individual solubility effects on the metals, coatings, and plastic parts of a car's carburetor and other parts of its fuel system. Usually, gasohol has about the same effects on these materials as gasoline; and in the absence of absorbed water, gasohol causes no difficulties in usual servicing and operation of automobiles. Because alcohol increases the octane rating, and gives other improvements in combustion, gasohol usually performs better and gives an increase in octane number higher than would be expected on the basis of the small fraction of alcohol in the fuel. *See* OCTANE NUMBER.

Specially modified engines and parts in contact with water-free (absolute or neat) alcohol account also for the considerably greater efficiency obtained in its use. Neat alcohol is not used to any extent as a fuel in the United States, but is widely used as a 95% impure grade in Brazil in cars fitted with special parts.

Ethanol has many advantages, but it does not compete in most countries with other liquid fuels on an economic basis for two reasons. First, raw materials are much more valuable for food or other uses, if there are no political constraints or artificial pricing. Second, basically, alcohol always requires more energy to produce than it will deliver in any use. Thus it will always be inefficient and expensive from a thermal standpoint. Nevertheless, alcohol will find use as fuel for a number of reasons: (1) In many countries there may be no other indigenous raw material from which to make motor fuels. (2) The plants for its product can be quite simple, small, and inexpensive compared to the very large, highly capital-intensive plants required for making other fluid fuels. (3) Technology is simple, well known, and established worldwide. (4) Highly labor-intensive programs are required with simple and familiar rural duties for most of the labor force. *See* ETHYL ALCOHOL.

Methanol. Carbon in any material can be converted to methanol. Natural gas during the 1970s and 1980s was favored as a raw material, and its energy may be delivered at a much lower cost if the gas is converted to methanol than if it is piped directly for 4000–5000 mi (6400–8000 km), even less and with much greater safety than if the gas is converted to liquefied natural gas, shipped for 2000–3000 mi (3200–4800 km), and stored at its cryogenic temperature. *See* LIQUEFIED NATURAL GAS (LNG); NATURAL GAS.

As gas prices increase, otherwise unsalable solid fuels will be used increasingly as feedstocks for methanol. Low-grade coal, lignite, or peat may be too far from market or have too much sulfur, chemically bound water, or ash to be viable fuels. Plants built alongside deposits of such fossil fuels may pipe or ship liquid methanol to market.

Methanol has been synthesized in the United States since 1925, first from synthesis gas (also known as syngas, a mixture of CO and H_2) manufactured from coal and, more recently, almost entirely from natural gas. While the basic chemistry is still the same, the many improvements in processing, flow sheets, catalysts, and construction materials now allow single units to produce up to 1.5×10^6 gal/day (5.7×10^6 liters/day) of fuel-grade methanol. Such so-called world-size plants are necessary if methanol is to be economical for fuel use. One such unit using, for example, low-grade lignite, will fuel automobiles for as many miles as the gasoline from 55,000 barrels (8700 m^3) of crude oil a day. *See* LIGNITE.

In producing synthesis gas, the feedstock (for example, the fossil fuel), after any necessary preparation, is gasified to hydrogen, carbon monoxide, and carbon dioxide. The gasifier should operate at as high a pressure as possible to minimize the cost of compression of the synthesis gas produced up to the operating conditions of the converter: 100–300 atm ($1–3 \times 10^7$ pascals) pressure and 390–750°F (200–400°C). Correspondingly large is the air separation plant to supply 5 tons (4.5 metric tons) per minute of oxygen for the partial oxidation of a coal feedstock, and to give the oxygen in the large volume of methanol produced. The steps of air separation, syngas generation, balancing of reaction to produce the desired amount of hydrogen, and the separation of CO_2 and H_2S (then sulfur) from the synthesis gas require a vast complex of massive equipment. Purified synthesis gas passes through catalyst chambers and heat exchangers which, by correct interrelation and proper control of temperatures and space velocities, give 97% methanol, 1% water, and 1–2% higher alcohols. Modern processing (Wentworth) delivers this fuel-grade methanol without any refining being necessary for immediate use as motor fuel and with a higher energy content of the original fuel and a much lower plant cost and operating cost than earlier processes used to make the chemical grade.

Methanol for fuel use has a heating value of about 10,000 Btu/lb (23 megajoules/kg). Fifteen gallons (0.57 m^3) of methanol give 1×10^6 Btu (1×10^9 joules). To make it requires an input of from 1.5 to 2 times as much thermal energy based on the higher heating value of a solid fuel, or somewhat less if it is made from petroleum gas or liquid feedstocks. Thus the overall energy conversion from solid fuels is in the range of 50 to 65%, and the lower efficiencies in this range are for poorer coals or lignites.

Methanol is the most versatile and cheapest liquid fuel which can be made. It is also less flammable than gasoline; accidental fires are extinguished with water instead of being spread as flaming films. When it combusts, it yields neither particulates (soot) nor sulfur oxides, and yields lower quantities of nitrogen oxides than any other fuel. In an accident at sea, there is no fire or oil slick because the methanol dissolves in water. When produced from natural gas or solid fuel, it can always be regasified to give substitute natural gas (SNG) with only a small loss of energy.

Diesel engines in trucks, railroad locomotives, and ships may use a small amount of their regular diesel fuel for start-up. Then methanol may be used for at least 90% of the total fuel.

Steam may be generated in almost any furnace which can be fired by oil or gas, by changing the nozzle which feeds fuel to the combustion chamber so as to use methanol, which gives no emission problems. However, methanol's greatest potential use in electric power plants is by internal combustion in gas turbines, where it is a superior fuel. It may be used most efficiently in a combined cycle with a steam boiler and turbine on the low-temperature side. Compared with a coal-fired, steam-powered plant with its modern accessories, methanol gives a capital cost reduced by almost 50% and a thermal efficiency over a third greater (46% versus 32%).

Methanol has been preferred in automobile engines, for example, in racing cars where efficiency, performance, and safety are more important than cost. With its high octane number (110), 10% can be added to improve the performance of unleaded gasoline, along with other beneficial additives amounting to 30% or more.

Methanol is used much better neat or mixed with just a few percent of common liquids as additives to eliminate problems related to cold start, materials of construction, lubricity, and so on. Lean mixtures may be burned at compression ratios of between 12 and 16 to 1, to give 50 to 80% higher thermal efficiencies than gasoline, even though the heating value of methanol is only about one-half. Even when expensive chemical-grade methanol is used in automobiles with relatively insignificant changes, the mileage costs are less than when gasoline is used as a fuel. When methanol from a world-size fuel-grade plant is used, the fuel cost per mile will be less than half. Pollution from emissions is nonexistent. *See* METHANOL.

Donald F. Othmer

Bibliography. D. F. Othmer, *Chem. Eng*. (London), January 1981; J. H. Perry, *Methanol: Bridge to a Renewable Energy Future*, 1990.

Alcoholism

The continuous or excessive use of alcohol (ethanol) with associated pathologic results. Alcoholism is characterized by constant or periodic intoxication, although the pattern of consumption varies markedly. Individuals admitted for the first time to an alcoholism treatment center typically have been consuming 3–4 oz (80–100 g) of pure alcohol per day, corresponding to seven to nine drinks or bottles of beer or glasses of wine. Studies have shown that problem drinking in these populations starts at about 2 oz/day (60 g/day), that is, four to five drinks per day, and that these are consumed in rapid succession, leading to intoxication on three or more days per week. Individuals who consume these levels of alcohol have a greater-than-average risk of developing alcoholic liver cirrhosis. However, the levels should not be taken as absolute, since they can vary greatly in different individuals, according to body weight and other factors.

The symptoms and consequences associated with severe alcohol consumption also vary greatly; that is, in some individuals only a few may be present. These may consist of the development of physical dependence manifested as a state of physical discomfort or hyperexcitability (tremors or shakes) that is reduced by continued consumption; the development of tolerance to the effects of alcohol, which leads individuals to increase their consumption; accidents while intoxicated; blackouts, characterized by loss of memory of events while intoxicated; work problems, including dismissal; loss of friends and

family association; marital problems, including divorce; financial losses, including bankruptcy or continual unemployment. Medical problems can include gastric ulcers, pancreatitis, liver disease, and brain atrophy. The last is often associated with cognitive deficiencies, as shown by the inability to comprehend relatively simple instructions or to memorize a series of numbers. *See* COGNITION.

Individuals seeking an early treatment for their alcohol problems have very good probabilities of recovery. The lesser the number of presenting problems described above, the better the chances of favorable outcome, and so an early identification of problem drinking by family, friends, employers, or physicians becomes very important. Employee assistance programs have become an important factor in identification and referral and rehabilitation of individuals with alcohol problems in the United States, Canada, and many other countries. The types of intervention vary greatly, progressing from self-monitoring techniques, to intensive outpatient and inpatient programs, to Alcoholics Anonymous groups.

Absorption of alcohol. Alcohol is absorbed most rapidly from solutions of 15–30% (30–60 proof) and less rapidly from beverages containing below 10% and over 30%. This is to be expected at the lower concentrations, since the rate of absorption depends on the concentration gradient across the mucosal surface. At higher concentrations, ethanol abolishes the rhythmic opening of the pylorus, thus preventing the passage to the intestine, where absorption is faster. The presence of food in the stomach is also known to delay gastric emptying and thus to slow absorption. Once in the bloodstream, alcohol is distributed evenly in all tissues, according to their water content. As a rule of thumb, for a 150-lb (68-kg) individual two standard drinks (1.5 oz of a distilled beverage or 13.6 g per standard drink) will yield blood alcohol levels of about 0.06 g per 100 mL of blood. It was previously believed that all ethanol consumed was absorbed into the bloodstream. However, studies have shown that a small part of the ethanol ingested is degraded directly in the stomach without entering the blood. *See* DISTILLED SPIRITS; MALT BEVERAGE; WINE.

Effects on the nervous system. The exact mechanisms of the pharmacological actions of alcohol are not known. Alcohol can act as a stimulant at lower doses, and as a depressant at higher doses. Even at very low doses, alcohol can impair the sensitivity to odors and taste. Also, low doses are known to alter motor coordination and time and space perception, important aspects of car driving (about 50% of all fatal traffic accidents are caused by intoxicated drivers). Some effects are already seen at levels of 0.05%. Pain sensitivity is diminished with moderate doses. In some individuals, alcohol is known to diminish feelings of self-criticism and to inhibit fear and anxiety, effects which are probably related to an alcohol-induced sociability. These effects act, no doubt, as psychological reinforcers for the use of alcoholic beverages.

It is generally accepted that alcohol affects the nerve cell by preventing the production and propagation of electric impulses along a network consisting of axons and synapses. The brain functions much as an electronic system in which one nerve cell, acting as a current generator, communicates information to many other cells, which in turn receive impulses from many other areas. Some impulses are enhanced, others are blunted. Memory and conditioning appear to play an important role in integrating the impulses which are finally expressed as behaviors. Studies in the United States and England have shown that when alcohol becomes dissolved in the membrane of the cells, it fluidizes or disorganizes the membrane, which in turn leads to changes in the physical and biochemical characteristics of the latter. Chronic exposure to alcohol alters the composition of the membrane and its rigidity, so that alcohol becomes less of a disorganizing agent. The new membrane composition also appears to partly exclude the alcohol molecules from dissolving in it. These changes are likely to be related to tolerance, when more alcohol is required to produce the same effect. Studies have shown that learning and conditioning also play a role in the latter phenomenon. In general, the knowledge gained from these studies is that the state of organization (fluidity–rigidity) of the membrane is correlated with the state of excitability of the brain; the increased rigidity of the membrane that follows chronic alcohol consumption might account for the alcohol withdrawal hyperexcitability (physical dependence) and the fact that continued ingestion of alcohol is required to alleviate the physical dependence. Studies in rodents and monkeys have addressed the question of what factors can lead to alcohol consumption in these animals, which normally reject alcohol. It has been found that small amounts of alcohol-derived molecules, such as the tetrahydroisoquinolines, when injected into the brain, produce a marked increase in the preference for alcohol of the animals. These studies raise the possibility that these substances could be formed in greater amounts in individuals who consume alcohol heavily.

Studies have shown that lipid molecules in the membranes of brain cells of animals fed alcohol for a prolonged time can confer alcohol resistance to brain cell membranes of animals that have never received alcohol. Not only brain lipids are modified but also some brain proteins. Thus, it is clear that brain chemistry is modified by chronic alcohol exposure. This effect remains for days or weeks after alcohol administration has been discontinued. Brain cells communicate with each other and with cells involved in sensation by means of chemical messengers, or neurotransmitters, released by these cells. Once a neurotransmitter is externally recognized by a cell, the information is passed to the interior of the cell by second messengers, a system that involves G-proteins. The continued presence of alcohol makes the nerve cells produce less of these proteins, and as a consequence an external chemical message is not well communicated to the interior of the cell.

Studies using white blood cells of alcoholics as sensors of this deficiency have shown that as much as 75% of the external message is not transmitted to the interior of the cell. These changes are of importance in that they may explain the reduction in cognitive efficiency, in relation to learning and to processing external information, which starts at levels of alcohol consumption that are considered moderate social drinking.

A major finding in the mid-1980s was that some of the effects of alcohol can be quickly reversed by new experimental drugs. Studies have shown that alcohol enhances the actions of an inhibitory brain neurotransmitter referred to as gamma-aminobutyric acid (GABA). Benzodiazepines, such as diazepam, are anxiety-reducing and sedative drugs which also enhance the effects of GABA. These effects can be reduced by experimental antagonist molecules, which interact in the brain in the same regions where GABA is found. *See* CELL MEMBRANES; NERVOUS SYSTEM (VERTEBRATE); SYNAPTIC TRANSMISSION.

Effects on the liver. The liver is responsible for about 80% of the metabolism of alcohol. In the liver, alcohol is first oxidized to acetaldehyde and then to acetate, which is metabolized in many tissues, including the brain, heart, and muscles. A 150-lb (68-kg) person metabolizes approximately 0.4 oz (10 g) of pure alcohol per hour (about 1 oz of a distilled beverage per hour) or, if alcohol is continuously present in the bloodstream, about 8–10 oz (190–240 g) of pure alcohol per day, equivalent to 1300–1600 calories per day. Since alcoholic beverages contain negligible levels of essential nutrients, these calories are called empty calories. Many alcoholics show malnutrition due to the fact that an important part of their caloric intake is alcohol. Alcohol also impairs the absorption and the metabolism of some essential nutrients. *See* ABSORPTION (BIOLOGY); LIVER; MALNUTRITION.

In the presence of alcohol, about 80% of oxygen consumed by the liver is devoted to the metabolism of alcohol; as a consequence, other substances such as fats, normally oxidized by the liver, are not metabolized, leading to fat accumulation in the liver. Continued consumption of alcohol has been reported to increase the capacity of the liver to oxidize alcohol, such that some alcoholics can metabolize up to 16 oz (400 g) of pure alcohol per day. Since all the hepatic metabolism of alcohol, by any known route, requires oxygen, there is also an increased uptake of oxygen by the liver. *See* LIPID METABOLISM.

Alcoholic liver disease and liver cirrhosis rank among the 10 leading causes of mortality in the United States and Canada. It was initially believed that cirrhosis in the alcoholic was exclusively of nutritional origin. However, baboons fed 50% of their caloric intake as alcohol develop cirrhosis in 1–4 years, even when the alcohol is given with a nutritious diet. Baboons fed the same diet, but with carbohydrates given instead of alcohol, did not develop cirrhosis. It is not known, however, if nutrient supplementation would have reduced the damage in its initial stages. Some investigators have proposed that in humans alcoholic liver disease is produced by increased utilization of oxygen by the liver, when the extra utilization is not matched by an adequate supply of oxygen to the organ. Studies in Canada have shown that the antithyroid drug propylthiouracil, which reduces the increased oxygen demand induced by alcohol in the liver, greatly protects against the mortality of alcoholic liver disease. In a 2-year treatment study, deaths in patients treated with propylthiouracil were 50–60% lower than in patients given placebo capsules.

The mortality due to liver disease can be drastically reduced if the individuals abstain from alcohol. Since it is unlikely for liver scar tissue in cirrhosis to reverse to normal, studies suggest that complications accompanying cirrhosis may be more damaging than the cirrhosis itself.

Alcoholic liver disease is characterized by two conditions: failure of the liver to detoxify noxious substances and to produce essential products, and increased resistance to blood flow through the liver. The latter condition (portal hypertension) results in the opening of collateral vessels (esophageal varices) as an alternative for the passage of portal blood. The expanded vessels can burst, leading to internal bleeding. Also, noxious substances bypass the cleansing action of the liver. This condition is responsible for about one-half of the deaths in alcoholic liver disease. Studies have demonstrated that chronic alcohol consumption results in enlargement of the liver cells, which leads to compression of the minute intrahepatic vessels that feed the liver cells. In turn, compression of these vessels increases the resistance to blood passage and to portal hypertension. This effect is quickly reversible upon abstinence from alcohol, and may be responsible for the reduction in mortality from liver disease in abstinent alcoholics. *See* CIRRHOSIS.

Effects on sexual and reproductive functions. Alcohol consumption produces a striking reduction in the hormone testosterone in males. This appears to be due to an effect in brain areas which are responsible for specific signals for the maintenance of normal testosterone levels and to a metabolic effect in the testicular cells that produce testosterone. In addition, alcoholics have increased rates of testosterone destruction by the liver. Feminization signs are often seen in male alcoholics, while impotence and testicular atrophy are also found frequently.

Studies in France and the United States in the early 1970s were the first to report a distinct pattern of physical and behavioral anomalies in children of alcoholic women who drank heavily during pregnancies. These anomalies which characterize the fetal alcohol syndrome include facial and cranial deformities such as drooping eyelids (ptosis), small eyes (microophthalmia), and underdeveloped midface (midfacial hypoplasia). A 10-year follow-up of these children in the 1980s indicated that all were borderline for intelligence, while half of them presented IQs of 20 to 60, in the retarded range. The prevalence of this condition in the general population has not been fully established, but it ranges between 0.5 and 3 births in 1000. In the United States the risk of fetal alcohol

syndrome is about seven times higher in the black race than in the white population. Studies so far do not indicate that there is a safe level of habitual alcohol consumption, below which there is no effect on the unborn; and the role of specific drinking patterns, including a single bout, needs to be determined. In animals with short gestation periods, a single massive intoxication results in damage to the newborn. Social drinking during pregnancy has been associated with lower scores on tests of spoken English and verbal comprehension in the children. *See* FETAL ALCOHOL SYNDROME.

Early identification of alcohol abuse. Several laboratory tests are used to detect high levels of alcohol consumption. Among these are an elevated serum gamma-glutamyl transferase and the presence of enlarged red cells. In addition, rib and vertebral fractures, as seen on routine x-rays, are found 15 times more frequently in alcoholics than in normal populations. Combinations of these tests allow diagnosis of heavy alcohol consumption with a good degree of accuracy. A simple litmus-paper-like test to detect alcohol in saliva, serum, and urine consists of a strip of paper with special reagents which, when impregnated by any of the fluids, indicates the amount of alcohol present by changing color. Another test to determine intoxication involves the eyealyzer, which in 15 seconds senses alcohol vapors emanating from the lacrimal fluid normally covering the eye surface. This noninvasive device is 95% accurate and is not affected by residual alcohol in the mouth, as is the breathalyzer, and it can be used in persons that are unconscious since the eyelid can be held open.

Genetic factors. There is abundant evidence that tendency to alcoholism can be of familial origin, due to environmental, cultural, and genetic factors. A Swedish study demonstrated that identical twins are twice as likely to have a common alcoholic problem as fraternal twins. In an American-Danish study, it was shown that children of alcoholic parents are more likely to develop alcoholism (18%) than children of nonalcoholic parents (4%) when both groups of children were adopted by nonrelatives within 6 weeks of birth.

In a Swedish study, similar findings were observed where male adoptees whose biological fathers were severely alcoholic had a 20% occurrence of alcohol abuse, as compared with 6% for adoptee sons of nonalcoholic fathers. All adoptions in this study had occurred with the first 3 years of life and in most cases in the first few months. A subsequent Swedish-American study revealed the existence of two types of genetic predisposition to alcoholism. The more common type, mileu-limited alcoholism (type I), is associated with mild manifestations and adult-onset alcohol abuse in one of the biological parents. In this group, a significant contribution to alcoholism predisposition was low socioeconomic status. A second type, male-limited alcoholism (type II), accounts for about 25% of all male alcoholics and is unaffected by the environment. This type of alcoholism, often with teenage onset, was associated with severe alcoholism in the biological father but not in the mother. The biological father also tended to have a record of more serious criminality and an earlier onset of alcoholism. Alcohol abuse was nine times more frequent in adoptees of type II alcoholic fathers, regardless of the environment into which they were adopted. Brain electrical activity in response to external stimuli is markedly different in the male children of type II alcoholics, at an age as early as 12 years. It should be noted that about 60–80% of sons of alcoholics do not become alcoholic and thus environmental factors have a marked influence, for example in type I alcoholism. In identical twins, where genetic factors are constant, alcoholism occurs in both individuals in only 25% of the cases.

Studies from Germany and Japan demonstrate that some genetic factors can protect against the development of alcoholism. About one-half of the population of Japan lack a specific aldehyde dehydrogenase to rapidly metabolize the toxic acetaldehyde produced in the liver in the first reaction of alcohol oxidation. As a consequence, blood acetaldehyde levels build up, leading to facial flushing and to unpleasant reactions which may include nausea. These individuals appear not to develop alcoholism, as shown by the finding that an extremely low proportion of individuals lacking this specific aldehyde dehydrogenase seek treatment of alcoholism.

A number of studies also showed that special genetic strains of rodents can be developed which either prefer alcohol to water or reject alcohol. Some strains of alcohol-preferring rats drink to intoxication, develop physical dependence manifested as hyperexcitability upon withdrawal from alcohol, and have been shown to work (conduct specific tasks) to receive alcohol. Other genetic strains of mice can be shown to differ markedly in sensitivity to the effects of ethanol in that after receiving similar doses of ethanol, one strain (LS or long-sleep) sleeps five times longer than another strain (SS or short-sleep). These differences have been found to result from differences in the brain of these animals. *See* BEHAVIOR GENETICS.

Statistics. In the United States there are an estimated 10 million problem drinkers. Official statistics indicate that in 1986 1.2 million people were admitted for alcoholism treatment in private and government facilities. Alcohol-related deaths are estimated at 100,000 per year. Accidents and cirrhosis are the leading causes of mortality associated with excessive alcohol use, followed by hypertensive diseases, infections, and cancer. Eighty percent of all cirrhosis and about one-half of all accidental deaths, suicides, and homicides are alcohol-related. The average per-capita consumption (over 14 years of age) is 2.65 gallons (10 liters) of pure alcohol per year. Heavy drinkers, constituting about 10% of the drinking population, account for one-half of all alcohol consumed in the United States. It is estimated that alcohol creates problems for 18 million persons 18 years old or more. Both the level of consumption and the number of deaths attributed to alcohol declined during the 1980s. Women drink significantly less than men; however, there has been

an increase in drinking among women aged 35 to 64. People over age 65 consume less alcohol than younger adults and have lower prevalence of alcohol abuse.

Studies in the Scandinavian countries and Canada have shown that the price of alcoholic beverages (relative to the disposable income) is inversely related to the total amount of alcohol consumed by the population. Cirrhosis in different countries has been shown to correlate almost perfectly with the per-capita consumption, with the United States showing about one-third of the incidence found in France. Some studies have shown that the frequency of coronary heart disease is lower in countries with higher per-capita consumption of alcohol. However, others suggest that this is largely due to the fact that populations in these countries also have dietary patterns that by themselves lead to less coronary heart disease. *See* ADDICTIVE DISORDERS; BEHAVIORAL TOXICOLOGY.

Yedy Israel

Pharmacotherapy. Pharmacotherapy for alcohol rehabilitation has been gaining wider acceptance. Specific pharmacotherapies which have received the most research attention utilize naltrexone, acamprosate, and disulfiram. Other pharmacological interventions for specific indications are SSRIs, (selective serotonin reuptake inhibitors) and buspirone.

Naltrexone. Naltrexone is an opiate receptor antagonist which blocks the effects of endogenous opioids in the brain. Research from animal studies suggests that alcohol activates endogenous opioid systems and may contribute to the pleasurable effects produced by alcohol consumption. Consequently, naltrexone might reduce the reinforcing effects of alcohol consumed by people and decrease their incentive to drink. The efficacy of naltrexone treatment has been examined by several investigators in double-blind, placebo-controlled studies. These studies have demonstrated that naltrexone (50 mg per day) reduces the frequency of alcohol consumption, resumption of heavy drinking, and intensity of alcohol craving with relatively few significant side effects. Relative to placebo controls, relapse rates were reduced significantly when patients took naltrexone. Naltrexone is well tolerated; about 15% of patients suffer from nonserious side effects, primarily nausea. These findings contributed to the 1995 decision by the U.S. Food and Drug Administration (FDA) to approve naltrexone for use in alcoholism treatment.

Acamprosate. The mechanism of action of acamprosate, calcium-acetyl-homotaurinate, involves primarily the restoration of a normal *N*-methyl-D-aspartate (NMDA) receptor tone in glutamatergic systems. It decreases postsynaptic potentials in the neocortex and diminishes voluntary alcohol intake in alcohol-preferring rats. It was proposed that one of acamprosate's actions is suppressing conditioned withdrawal craving by its effects on NMDA receptors and calcium channels. Acamprosate was investigated in 20 controlled clinical trials with about 5000 patients up to early 2006. Acamprosate (1998 mg per day) lengthens time to relapse, reduces drinking days, and increases complete abstinence among alcohol-dependent patients. Adverse events tended to be mild and transient, primarily involving the gastrointestinal tract with diarrhea and abdominal discomfort in about 10% of the patients. In 2004, acamprosate was approved in the United States.

Combined treatment. It was expected that naltrexone would specifically attenuate "reward craving" (anticipating hedonic effects of alcohol), while acamprosate would diminish relief craving (anticipating unpleasant effects associated with the absence of alcohol). The combination of these two drugs would act simultaneously on two different aspects of craving and might have a more incisive effect on the risk of relapse than either treatment alone. Both drugs are well tolerated and have no marked propensity for potentially dangerous drug–drug interactions. Now preclinical and clinical data exist to demonstrate a good tolerability of combined medication and, moreover, provide some hints for a significant improvement in outcome regarding the duration of abstinence following alcohol withdrawal.

Disulfiram. Disulfiram is a drug which causes an inhibition of the enzyme aldehyde dehydrogenase. Aldehyde dehydrogenase breaks down the first metabolite of alcohol, acetaldehyde. After taking disulfiram for 3 or more days, a person who drinks alcohol usually will experience an increase in acetaldehyde blood levels. This rise will produce nausea, vomiting, tachycardia, difficulty in breathing, and changes in blood pressure leading to hypotension. These effects typically last up to 3 h. When the drug is prescribed at the standard dose (250 mg per day), dangerous side effects are minimized, although the side-effect profile of disulfiram must be considered carefully for each patient.

Given its pharmacological effects, disulfiram acts as a deterrent to future alcohol consumption by making intake an aversive experience. Alcohol use cannot be resumed safely until 6–7 days after the last disulfiram dose. Despite the compelling rationale for its potential effectiveness, studies of disulfiram's efficacy are equivocal, providing only modest evidence for reducing drinking frequency without significantly enhancing abstinence rates. While some professionals endorse the effectiveness of disulfiram treatment, others advocate more controlled clinical trials to determine its relative costs and benefits before it is used routinely in alcoholism treatment.

Several new drug therapies for alcoholism rehabilitation are under investigation. Buspirone, a nonbenzodiazepine antianxiety agent, may decrease anxiety symptoms associated with a protracted alcohol withdrawal syndrome, thus reducing alcohol relapse potential. SSRIs also seem to reduce alcohol intake, especially in clinically depressed patients suffering from a co-morbid alcohol dependence. These medications require further investigation to determine their effectiveness as pharmacotherapeutic agent, in the treatment of alcoholism.

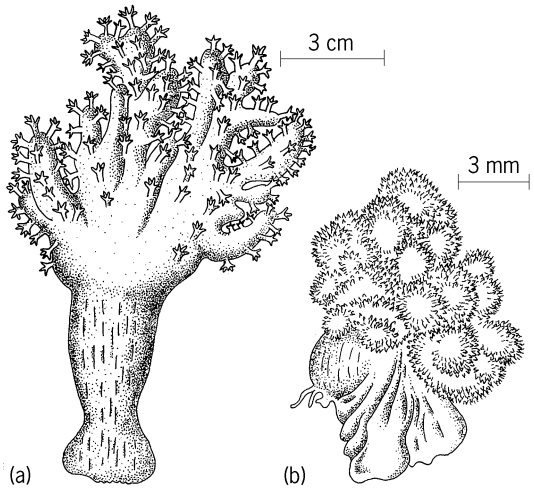
Pharmacotherapy approaches for alcohol rehabilitation are beginning to show promise as clinical interventions assisting patients in their efforts to recover from alcoholism. Concurrent behavioral and

supportive psychotherapeutic treatments, self-help program participation, and supervision of prescribed medication regimens may further optimize pharmacotherapy outcomes. Determining the best combination of pharmacotherapy approaches, psychotherapeutic interventions, and patient characteristics remains a challenge to the alcoholism treatment field. Falk Kiefer; Steve Martino

Bibliography. H. Bogleites and B. Kissin (eds.), *The Genetics of Alcoholism*, 1994; Combine Study Group, Testing combined pharmacotherapies and behavioral interventions for alcohol dependence (the COMBINE study): A pilot feasibility study, *Alcohol Clin. Exp. Res.*, 27(7):1123–1131, 2003; J. C. Garbutt et al., Pharmacological treatment of alcohol dependence: A review of the evidence, *JAMA*, 281:1318–1325, 1999; D. B. Goldstein, *Pharmacology of Alcohol*, 1983; D. W. Goodwin, *Alcoholism*, 2d ed., 1994; J. C. Hughes and C. C. Cook, The efficacy of disulfiram: A review of outcome studies, *Addiction*, 92:381–395, 1997; F. Kiefer et al., Comparing and combining naltrexone and acamprosate in relapse prevention of alcoholism: A double-blind, placebo-controlled study, *Arch. Gen. Psychiat.*, 60:92–99, 2003; R. Z. Litten and J. P. Allen, Pharmacotherapies for alcoholism: Promising agents and clinical issues, *Alcoholism Clin. Exp. Res.*, 15:620–633, 1991; C. P. Rivers, *Alcohol and Human Behavior: Theory, Research, and Practice*, 1994; J. R. Volpicelli et al., Effect of naltrexone on alcohol "high" in alcoholics, *Amer. J. Psychiat.*, 152:613–615, 1995.

Alcyonacea

An order of the cnidarian subclass Alcyonaria (Octocorallia). Alcyonacea, the soft corals (see **illus.**), are littoral anthozoans, which form massive or dendriform colonies with yellowish, brown, or olive colors. Most attach to some solid substratum; however, some remain free in sandy or muddy places. The only skeletal structures are small, elongated, spindle-shaped or rodlike, warty sclerites which are scattered over the mesoglea. The colony is supple and leathery. The polyp body is embedded in the coenenchyme, from which retractile anthocodia protrude in *Alcyonium*. In *Xenia* and *Heteroxenia*, anthocodia are nonretractile. The polyp base is protected by many sclerites and is termed a calyx.



Alcyonaceans. (*a*) *Alcyonium palmatum*. (*b*) *Dendronephytha* sp.

The aciniform gonads develop in the six ventral mesenteries and hang inside the gastrovascular cavity. *Anthomastus*, *Sarcophyton*, and *Lobophytum* are dimorphic and the siphonozooid is fertile only in *Anthomastus*. New polyps arise asexually also, from the solenial network. *See* OCTOCORALLIA (ALCYONARIA); CNIDARIA. Kenji Atoda

Aldebaran

A cool red giant star, prominently located in the constellation Taurus. With its red color, it appropriately represents the eye of the Bull. At a distance of 18.5 parsecs (60 light-years), Aldebaran, or α Tauri, is among the nearest (and brightest) giant stars to the Sun. It appears to be situated near the middle of the Hyades, but Aldebaran is actually less than half the distance of this nearby star cluster. The star is an example of a K-type giant, a very common type of evolved star that derives its energy from the thermonuclear burning of helium in a core surrounded by a thin, hydrogen-burning shell. Its spectral type of K5III corresponds to an effective temperature of 6700°F (4000 K) and a radius of about 40 times the Sun. It is nearly 150 times more luminous than the Sun and, as is typical for K giants, its brightness varies by a modest amount. Aldebaran is accompanied in a long-period binary-star system by a cool dwarf companion star some 100,000 times fainter than the giant. *See* BINARY STAR; GIANT STAR; HYADES; SPECTRAL TYPE; STELLAR EVOLUTION; TAURUS; VARIABLE STAR.

Because it is among the nearest and brightest members of its class, Aldebaran has been studied in considerable detail. A thin and very cool shell of dust surrounds the star out to 500 stellar radii. The dust grains originate in the outer atmosphere of Aldebaran and are then gently blown away by radiation pressure. Aldebaran's angular diameter of about 20 milli-arc-seconds (1/180,000 degree) has been measured by several interferometric methods, including a measurement with a two-telescope interferometer that determined the diameter to an accuracy of less than 1%. Observations from the *Hubble Space Telescope* detected the emission from ionized atoms in the chromosphere of Aldebaran, where the turbulent gases appear to be churning at speeds of more than 12 mi/s (20 km/s). *See* STAR.

Harold A. McAlister

Bibliography. A. Frankoi, D. Morrison, and S. C. Wolff, *Voyages Through the Universe*, 3d ed., Brooks/Cole, 2004; J. B. Kaler, *The Hundred Greatest Stars*, Copernicus Books, 2002; J. M. Pasachoff and A. Filippenko, *Astronomy in the New*

Millennium, 3d ed., Brooks/Cole, 2007; A. Quirrenbach et al., Angular diameter measurements of cool giant stars in strong TiO bands and in the continuum, *Astrophys. J.,* 406:215–219, 1993.

Aldehyde

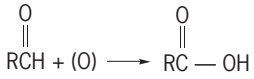
One of a class of organic chemical compounds represented by the general formula RCHO. Formaldehyde, the simplest aldehyde, has the formula HCHO, where R is hydrogen. For all other aldehydes, R is a hydrocarbon radical which may be substituted with other groups such as halogen or hydroxyl (see **table**). Aldehydes are closely related to ketones, since both types of compounds contain the carbonyl group

In the case of ketones, the carbonyl group is attached to two hydrocarbon radicals. Because of their high chemical reactivity, aldehydes are important intermediates for the manufacture of resins, plasticizers, solvents, dyes, and pharmaceuticals. *See* KETONE.

Physical properties. At room temperature formaldehyde is a colorless gas, which is handled commercially in the form of a 37% aqueous solution or as its polymer, paraformaldehyde $(CH_2O)_x$. The other low-molecular-weight aldehydes are colorless liquids having characteristic, somewhat acrid odors. The unsaturated aldehydes acrolein and crotonaldehyde are powerful lacrimators.

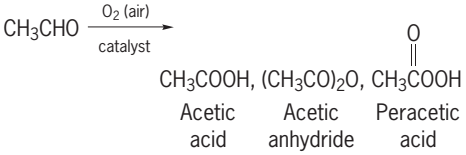
Chemical properties. The important reactions of aldehydes include oxidation, reduction, aldol condensation, Cannizzaro reaction, and reactions with compounds containing nitrogen.

Oxidation. Aldehydes are oxidized to carboxylic acids by a wide variety of oxidizing agents, such as air, oxygen, hydrogen peroxide, permanganate, peracids, silver oxide, and nitric acid, as in reaction (1).



(1)

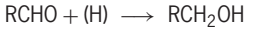
The catalytic oxidation of acetaldehyde is used for the production of three industrially important compounds, as in reaction (2), where the product



(2)

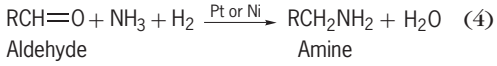
depends on the conditions of the reaction.

Reduction. Aldehydes are reduced to primary alcohols cheaply by reaction with hydrogen in the presence of a platinum or nickel catalyst, or in the laboratory by treatment with aluminum isopropoxide or oxide or with complex metal hydrides such as lithium aluminum hydride ($LiAlH_4$) or sodium borohydride ($NaBH_4$), as in reaction (3).



(3)

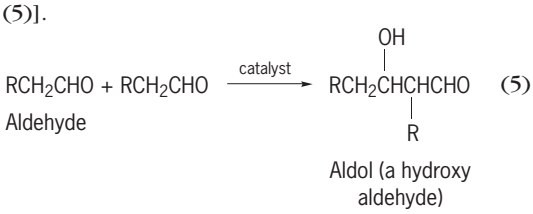
Aldehydes react with hydrogen in the presence of ammonia and a catalyst to form amines. This process [reaction (4)], known as reductive amination, is used



(4)

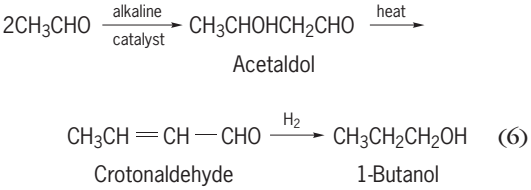
commercially to produce many important amines.

Aldol condensation. Two molecules of an aldehyde react in the presence of alkaline, acidic, or amine catalysts to form a hydroxyaldehyde. This reaction proceeds only when the carbon adjacent to the carbonyl group has at least one hydrogen atom [reaction (5)].



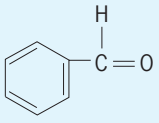
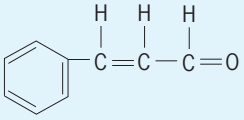
(5)

Aldols are readily dehydrated to form unsaturated aldehydes, and the latter can be hydrogenated to form saturated, primary alcohols. This reaction sequence is the basis of the commercial preparation of crotonaldehyde and 1-butanol (*n*-butyl alcohol) from acetaldehyde [reaction (6)].



(6)

Aldehydes and their formulas

Compound	Formula
Formaldehyde	HCHO
Acetaldehyde	CH_3CHO
Acetaldol	$CH_3CHOHCH_2CHO$
Propionaldehyde	C_2H_5CHO
n-Butyraldyde	$CH_3(CH_2)_2CHO$
Isobutyraldehyde	$(CH_3)_2CHCHO$
Acrolein	$CH_2{=}CHCHO$
Crotonaldehyde	$CH_3CH{=}CHCHO$
Chloral	CCl_3CHO
Chloral hydrate	$CCl_3CH(OH)_2$
Benzaldehyde	
Cinnamaldehyde	

Cannizzaro reaction. On treatment with strong alkali, aldehydes which lack a hydrogen attached to the carbon adjacent to the carbonyl group undergo a mutual oxidation and reduction, yielding one molecule of an alcohol and one molecule of the carboxylic acid salt. The reaction of benzaldehyde with aqueous potassium hydroxide is typical [reaction (7)].

$$2C_6H_5CHO + KOH \xrightarrow{H_2O} C_6H_5CH_2OH + C_6H_5COOK \quad (7)$$

Acetaldehyde and formaldehyde in the presence of calcium hydroxide react by an aldol condensation followed by a "crossed" Cannizzaro reaction to give the industrially useful polyhydroxy compound pentaerythritol [reaction (8)].

$$CH_3CHO + 4CH_2O + Ca(OH)_2 + H_2O \longrightarrow$$

Formaldehyde

$$HOCH_2-\underset{\displaystyle CH_2OH}{\overset{\displaystyle CH_2OH}{\overset{|}{\underset{|}{C}}}}-CH_2OH + (HCOO)_2Ca \quad (8)$$

Pentaerythritol — Calcium formate

Reaction with ammonia and amines. The addition of ammonia to the carbonyl group of an aldehyde is reversible, and the resulting aminoalcohol usually cannot be isolated [reaction (9)].

$$RCHO + NH_3 \rightleftharpoons \underset{\displaystyle NH_2}{\overset{|}{RCHOH}} \quad (9)$$

However, aromatic aldehydes such as benzaldehyde react with primary amines such as aniline with elimination of a molecule of water to give stable and well-characterized imines, also known as Schiff bases [reaction (10)].

$$C_6H_5CHO + N_2NC_6H_5 \longrightarrow C_6H_5CH{=}NC_6H_5 + H_2O \quad (10)$$

Oximes, phenylhydrazones, and semicarbazones. Aldehydes react readily with hydroxylamine to eliminate water and form oximes [reaction (11)].

$$\underset{\text{Aldehyde}}{RCHO} + \underset{\text{Hydroxylamine}}{H_2NOH} \longrightarrow \underset{\text{Oxime}}{RCH{=}NOH} + H_2O \quad (11)$$

Phenylhydrazine and semicarbazide react in a similar manner to produce phenylhydrazones and semicarbazones, respectively [reactions (12) and (13)].

$$\underset{\text{Aldehyde}}{RCHO} + \underset{\text{Phenylhydrazine}}{C_6H_5NHNH_2} \longrightarrow RCH{=}NNHC_6H_5 + H_2O \quad (12)$$

Phenylhydrazone

$$\underset{\text{Aldehyde}}{RCHO} + \underset{\text{Semicarbazide}}{H_2NCONHNH_2} \longrightarrow RCH{=}NNHCONH_2 + H_2O \quad (13)$$

Semicarbazone

Since these aldehyde derivatives are usually crystalline compounds with sharp melting points, they are frequently employed for the characterization and recognition of aldehydes. *See* HYDRAZINE; OXIME.

Cyanohydrin formation. Aldehydes combine with hydrogen cyanide to form hydroxynitriles, known as cyanohydrins, as shown in reaction (14) for benzaldehyde. This particular cyanohydrin, known as mandelonitrile, occurs in nature in the form of a glycoside derivative, amygdalin, which is responsible for the characteristic flavor of bitter almonds and the seeds of peaches and apricots.

$$C_6H_5CHO + HCN \longrightarrow \underset{\displaystyle OH}{\overset{|}{C_6H_5CHCN}} \quad (14)$$

Hemiacetal-acetal formation. Alcohols react with aldehydes in the presence of acidic catalysts to form hemiacetals [reaction (15)].

$$RCHO + CH_3OH \rightleftharpoons \underset{\displaystyle OH}{\overset{\displaystyle OCH_3}{\overset{/}{\underset{\backslash}{RCH}}}} \quad (15)$$

Aldehyde — Methanol — Hemiacetal

Hemiacetals may react further with alcohols in the presence of acidic catalysts, but not basic catalysts, to form acetals [reaction (16)].

$$\underset{\displaystyle OH}{\overset{\displaystyle OCH_3}{\overset{/}{\underset{\backslash}{RCH}}}} + CH_3OH \longrightarrow \underset{\displaystyle OCH_3}{\overset{\displaystyle OCH_3}{\overset{/}{\underset{\backslash}{RCH}}}} + H_2O \quad (16)$$

Hemiacetal — Methanol — Acetal

Hemiacetals are generally too unstable to permit isolation, but acetals are very stable under basic conditions, and they are therefore useful as "protecting groups," preventing reaction of the carbonyl group while transformations are carried out elsewhere in the molecule.

Condensation with active methylene compounds. Aldehydes condense with a variety of active methylene compounds in the presence of a base with elimination of a molecule of water to give unsaturated acids or acid derivatives. The reactions of benzaldehyde shown in reactions (17)–(19) are typical.

$$C_6H_5CHO + \underset{\text{Malonic acid}}{CH_2(COOH)_2} \longrightarrow C_6H_5CH{=}C(COOH)_2 \quad (17)$$

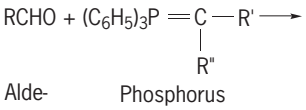
Benzylidenemalonic acid

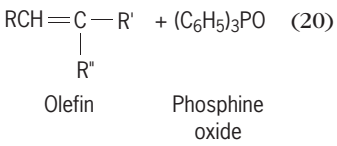
$$C_6H_5CHO + \underset{\text{Acetic anhydride}}{(CH_3CO)O} \longrightarrow \underset{\text{Cinnamic acid}}{C_6H_5CH{=}CHCOOH} \quad (18)$$

$$C_6H_5CHO + \underset{\text{Ethyl acetate}}{CH_3CO_2C_2H_5} \longrightarrow C_6H_5CH{=}CHCO_2C_2H_5 \quad (19)$$

Ethyl cinnamte

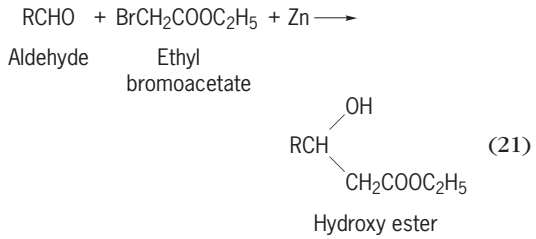
Wittig reaction. Aldehydes react with phosphorus ylides to form olefins, as shown in reaction (20). The



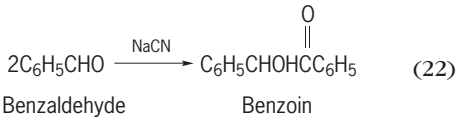


aldehyde may be aliphatic or aromatic. It may contain double or triple bonds and functional groups such as hydroxyl, alkoxyl, nitro, halo, acetal, and ester. Thus the reaction is very general, and the position of the double bond is always certain.

Reformatsky reaction. Aldehydes react with bromo esters and zinc to form hydroxy esters [reaction (21)].



Benzoin condensation. Aromatic aldehydes undergo bimolecular condensation in the presence of alkali cyanide catalyst [reaction (22)].

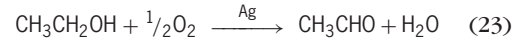


Tests for aldehydes. Tollens' reagent, a solution of silver nitrate in ammoniacal sodium hydroxide, oxidizes aldehydes, with the formation of a mirror of metallic silver. Schiff's reagent (fuchsin aldehyde reagent) reacts with aldehydes to give a violet-purple color. Fehling's solution (a blue, alkaline cupric tartrate complex) is especially valuable for the characterization of reducing sugars. A positive test is indicated by the formation of a red precipitate of cuprous oxide. In the infrared absorption spectrum, the aldehyde group has a characteristic strong band at 1660–1740 cm^{-1} due to the carbonyl stretching vibration accompanied by two weak bands at 2700–2900 cm^{-1} due to the aldehydic carbon-hydrogen stretching. In the proton magnetic resonance spectrum the aldehydic protons have a unique absorption region at 0.0–0.6 τ.

Synthesis. Because of the importance of aldehydes as chemical intermediates, many industrial and laboratory syntheses have been developed. The more important of these methods are illustrated by the following examples.

Catalytic dehydrogenation of primary alcohols. Formaldehyde is produced on a large scale industrially by the catalytic dehydrogenation of methanol. Acetaldehyde is produced similarly by the dehydrogenation of ethanol over a copper catalyst at 250–300°C (480–570°F) with formation of hydrogen as a by-product.

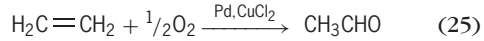
Oxidation of primary alcohols. Acetaldehyde is produced on a large scale industrially by passing a mixture of ethanol, oxygen, and steam over silver gauze at 480°C (896°F) to give an overall yield of 85–90% of the aldehyde [reaction (23)].



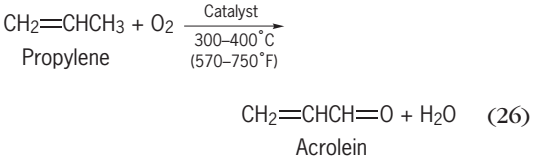
On a laboratory scale, oxidizing agents such as chromic acid or manganese dioxide have been used to produce aldehydes from primary alcohols, as in reaction (24). The reaction must be carefully con-

trolled to avoid oxidation of the aldehyde to the carboxylic acid.

Oxidation of olefins. Ethylene is oxidized directly to acetaldehyde by the Wacker process, employing a palladium–cupric chloride catalyst as shown in reaction (25).

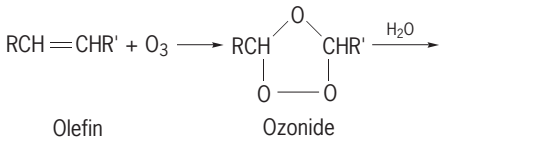


Acrolein, prepared industrially by the oxidation of propylene over a copper oxide catalyst at 350°C (660°F), as in reaction (26), is used in the manufac-

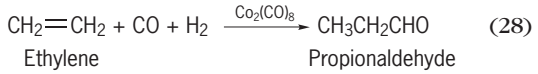


ture of glycerol and acrylic acid. *See* GLYCEROL.

In the laboratory, olefins are treated with ozone to form an ozonide which decomposes with water to form aldehydes and hydrogen peroxide [reaction (27)].



Hydroformylation of olefins. Olefins may be hydroformylated by reaction with carbon monoxide and hydrogen in the presence of a catalyst, usually cobalt carbonyl, at 180°C (360°F) and 6000 lb/in.^{2} (41 megapascals) as shown in reaction (28). This reaction is



also known as the oxo process.

Chlorination of a methyl group attached to a benzene ring. The chlorination of toluene in the presence of strong light (which serves as a catalyst) proceeds stepwise

to form benzyl chloride and benzal chloride at 135–175°C (275–347°F), by reactions (29) and (30), respectively.

$$\underset{\text{Toluene}}{C_6H_5CH_3} + Cl_2 \longrightarrow \underset{\substack{\text{Benzyl}\\\text{chloride}}}{C_6H_5CH_2Cl} + HCl \qquad (29)$$

$$C_6H_5CH_2Cl + Cl_2 \longrightarrow \underset{\substack{\text{Benzal}\\\text{chloride}}}{C_6H_5CHCl_2} + HCl \qquad (30)$$

Hydrolysis of benzal chloride gives benzaldehyde, as in reaction (31). This is one industrial process for

$$C_6H_5CHCl_2 + H_2O \longrightarrow \underset{\substack{\text{Benzal-}\\\text{dehyde}}}{C_6H_5CHO} + 2HCl \qquad (31)$$

the production of benzaldehyde.

Oxidation of a methyl group attached to a benzene ring. Certain substituted benzaldehydes, such as *p*-nitrobenzaldehyde, are readily prepared by oxidation of the corresponding toluene derivatives with chromic acid. Acetic anhydride is used to esterify the aldehyde groups and prevent further oxidation. Subsequent hydrolysis of the diacetate produces the aldehyde [reaction (32)].

$$\underset{\substack{\text{}p\text{–Nitro-}\\\text{toluene}}}{O_2NC_6H_4CH_3} + O_2 + 2\underset{\substack{\text{Acetic}\\\text{anhydride}}}{(CH_3CO)_2O} \longrightarrow$$

$$O_2NC_6H_4CH(OOCCH_3)_2$$

Diacetate

$$\downarrow H_2O$$

$$\underset{\substack{\text{}p\text{–Nitro-}\\\text{benzaldehyde}}}{O_2NC_6H_4CHO} + 2\underset{\substack{\text{Acetic}\\\text{acid}}}{HOOCH_3} \qquad (32)$$

One commercial synthesis of benzaldehyde involves the passing of a mixture of toluene and air over a heated metallic oxide catalyst at temperatures above 500°C (930°F) [reaction (33)].

$$\underset{\text{Toluene}}{C_6H_5CH_3} + O_2 \xrightarrow{\text{Catalyst}} \underset{\substack{\text{Benzalde-}\\\text{hyde}}}{C_6H_5CHO} + HCl \qquad (33)$$

Reduction of acid chlorides. Acid chlorides react with hydrogen in the presence of a specially prepared palladium catalyst to form aldehydes and hydrogen chloride, as in reaction (34). This reaction, known

$$\underset{\substack{\text{Acid}\\\text{chloride}}}{RCOCl} + H_2 \xrightarrow{Pd} \underset{\substack{\text{Alde-}\\\text{hyde}}}{RCHO} + HCl \qquad (34)$$

as the Rosenmund synthesis, is a useful laboratory method, since the acid chlorides can be prepared from the corresponding carboxylic acids.

Reaction of Grignard reagents. Aldehydes are formed by the reaction of alkylmagnesium halides (Grignard reagents) with ethyl orthoformate or diethyl formamide [reactions (35) and (36)].

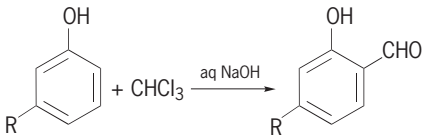
$$RMgX + HC(OCH_2CH_3)_3 \longrightarrow$$

$$RCH(OCH_2CH_3)_2 \xrightarrow{H_3O^-} RCHO \qquad (35)$$

$$RMgX + HC(=O)N(CH_3)_2 \longrightarrow$$

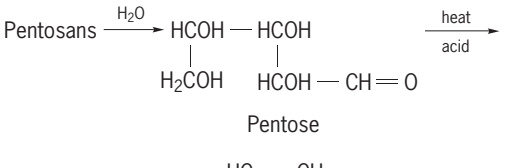
$$RCHN(CH_3)_2 \text{ (OMgX)} \xrightarrow{H_3O^+} RCHO \qquad (36)$$

Reimer-Tiemann synthesis. When phenols are treated with chloroform and strong alkali, *o*-hydroxybenzaldehydes (salicylaldehydes) are formed [reaction (37)].

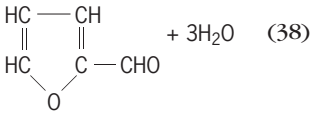


Phenols — *o*–Hydroxybenzaldehydes (salicylaldehydes) (37)

Furfural production. Cornstalks, corncobs, grain hulls, and similar farm wastes produce furfural upon heating with dilute sulfuric acid. This reaction (38), employed on a large scale industrially, proceeds from pentosans found in the agricultural residues.



Pentose



(38)

Furfural

Chloral production. Chloral is an important intermediate for the production of DDT. It is formed by chlorination of anhydrous ethanol [reaction (39)].

$$\underset{\text{Ethanol}}{2CH_3CH_2OH} + \underset{\text{Chlorine}}{8Cl_2} \longrightarrow$$

$$\underset{\text{Chloral}}{2Cl_3CCHO} + \underset{\substack{\text{Hydrogen}\\\text{chloride}}}{10HCl} \qquad (39)$$

See CARBOHYDRATE; FORMALDEHYDE. Paul E. Fanta

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; T. W. G. Solomons, C. B. Fryhle, and T. G. Solomons, *Organic Chemistry,* 7th ed., 1999.

Alder

Any of the deciduous shrubs and trees of the genus *Alnus* in the birch family (Betulaceae). There are about 30 species, 10 in the United States, widespread in cool north temperate regions, and southward in the Andes of South America. They have a smooth gray bark, elliptical or ovate saw-toothed leaves in three rows, male flowers in long catkins mostly in early spring, and clusters of several dry hard ellipsoid blackish fruits 0.5–1 in. (1.2–2.5 cm) long. The fruits are conelike and present throughout the year. Alders are common in wet soils, such as stream borders. They often form thickets even beyond treelines and are pioneers on bare areas such as landslides, roadsides, and after fire or logging. As in legumes, the roots bear swellings or nodules containing nitrogen-fixing bacteria which enrich the soil. *See* FAGALES; LEGUME; NITROGEN FIXATION.

Red alder (*Alnus rubra*) fruit and leaf.

Red alder (*Alnus rubra*; see **illus.**) is a small to large tree of the Pacific Northwest along the coast from southeast Alaska south to central California, and local in mountains of northern Idaho. It is the most common and most important hardwood in that region of conifers, occurring in pure stands and mixed with other species. This fast-growing, short-lived tree becomes 100–120 ft (30–36 m) tall on good sites. The wood is reddish or yellowish brown (whitish when freshly cut), moderately lightweight, moderately soft, and fine-textured. Principal uses are pulpwood, furniture, and cabinetwork. *See* FOREST AND FORESTRY; TREE.

Elbert L. Little, Jr.

Aldosterone

The principal mineralocorticoid (a steroid hormone that controls electrolyte and fluid balance) in humans. An adrenocortical steroid, aldosterone was first isolated in 1953 by S. A. Simpson and J. F. Tait in London. This hormone is synthesized from cholesterol in a process involving four different enzymes in the zona glomerulosa (outer zone) of the adrenal cortex. Aldosterone contains an equilibrium mixture of the aldehyde (see **illustration**) and the hemiacetal, the equilibrium favoring the latter. Normal serum concentrations range from 5 to 25 mg/dL. *See* ANDRENAL CORTEX; CHOLESTEROL; HORMONE; STEROID.

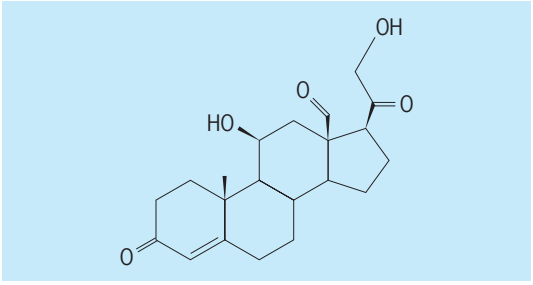
Biosynthesis. Direct stimulators of aldosterone biosynthesis include angiotensin II and III, ACTH, alpha-MSH, prolactin, vasopressin, potassium, hydrogen, ammonium, serotonin, histamine, and selected prostaglandins. The two major stimuli for aldosterone secretion are angiotensin II and potassium. Aldosterone synthesis is primarily regulated by the renin-angiotensin system. This system is activated by low sodium balance or decreased blood flow to the kidneys, where renin is produced. Renin is a proteolytic enzyme that converts the inactive angiotensinogen globulin to angiotensin I, which is then converted to angiotensin II by the angiotensin converting enzyme. *See* KIDNEY; THIRST AND SODIUM APPETITE.

Physiologic activity. Aldosterone acts primarily at the renal distal convoluted tubule promoting the excretion of potassium and retention of sodium, thereby influencing extracellular volume homeostasis and blood pressure. *See* POTASSIUM; SODIUM.

Alterations in Na^+/K^+ flux are seen between 30 and 90 minutes after aldosterone administration. This delayed response in activity is a nongenomic effect and represents binding to cytosolic steroid receptors, translocation to the nucleus, interaction with DNA, and finally genomic transcription and translation of effector proteins. Aldosterone stimulates the synthesis of epithelial sodium channels at the apical epithelial membrane of the epithelial cells in the kidney. Aldosterone increases the activity of the Na^+/K^+-ATPase. Mineralocorticoid receptors (the cytosolic steroid receptors) are present in epithelial cells in the kidney, bladder, gastrointestinal tract, sweat and salivary glands, smooth muscle, vascular endothelium, brain, and myocytes (muscle cells).

Increased aldosterone secretion contributes to some forms of arterial hypertension, disorders of sodium retention and edema, and syndromes of potassium wasting (observed when intake of potassium is less than output in the urine). Chronic aldosterone excess in the presence of high salt intake causes cardiac fibrosis in experimental animals. Patients with chronic heart failure had a significant reduction in morbidity and mortality when given a mineralocorticoid receptor antagonist, spironolactone, in addition to traditional therapy.

There are rapid (nongenomic) effects of aldosterone in smooth muscle, skeletal muscle, colonic



Molecular structure of aldosterone, 11β,21-dihydroxy-3,20-dioxo-pregn-4en-18al.

epithelial cells, and myocardial cells. These nongenomic effects have been linked to the development of increased systemic vascular resistance and therefore might participate in human hypertension and cardiovascular disease. *See* CARDIOVASCULAR SYSTEM; HYPERTENSION. Maria I. New; Alejandro Diaz

Bibliography. J. W. Funder, Aldosterone action, *Annu. Rev. Physiol.*, 55:115–130, 1993; D. N. Orth and W. J. Kovacs, The adrenal cortex, in J. D. Wilson et al. (eds.), *William Textbook of Endocrinology*, 9th ed., W. B. Saunders, Philadelphia, pp. 517–664, 1998; J. S. Willians and G. H. Willians, 50th aniversary of aldosterone, *J. Clin. Endocrinol. Metab.*, 88:2364–2372, 2003.

Alfalfa

The world's most valuable forage legume, *Medicago sativa*, also known also as lucerne, and less often as purple medic, medica, snail clover, median herb, Burgundy hay or clover, Chilean clover, and Burgoens Hoy. It is often referred to as the queen of forages. Alfalfa is produced worldwide on more than 32,000,000 hectares (80,000,000 acres). Seven countries—United States, Russia, Argentina, France, Italy, Canada, and China—account for 87% of the production area (see **table**). *See* LEGUME FORAGES; ROSALES.

Estimated hectarage in selected countries growing alfalfa

Continent and country	Year	Hectarages*
Europe		
Austria	1982	12,630
Bulgaria	1982	399,000
Czechoslovakia	1983	200,000
East Germany	1982	190,000
France	1983	566,000
Greece	1980	198,700
Hungary	1982	337,500
Italy	1982	1,300,000
Poland	1981–1983	258,000
Romania	1981	400,000
Soviet Union (European)	1971	3,375,000
Spain	1981	332,600
Switzerland	1983	6,000
West Germany	1983	31,000
Yugoslavia	1984	337,000
North America		
Canada	1981	2,544,300
Mexico	1982	245,000
United States	1981	10,559,025
South America		
Argentina	1981	7,500,000
Chile	1983	60,000
Peru	1981	120,000
Asia		
Iran	1977	270,000
Soviet Union (Siberian)	1971	1,125,000
Turkey	1969	73,700
Africa		
Republic of South Africa	1985	300,000
Oceania		
Australia	1981–1982	111,500
New Zealand	1984	101,200

*1 hectare = 2.5 acres.

Description. Alfalfa is a herbaceous perennial legume. The seed, 0.04–0.08 in. (1–2 mm) long and wide, and 0.04 in. (1 mm) thick, is kidney-shaped and yellow to olive-green or brown. Plants produce a deep taproot that may have several branches. Two to twenty-five stems are borne from a single crown. Trifoliolate leaves are produced alternately on the stem. The flower is a simple raceme with several complete and perfect florets (**Fig. 1**), which may be various shades of purple, variegated, cream, yellow, and white. Pods are usually spirally coiled. Alfalfa is considered to be cross-pollinated. Under production conditions, pollination is by honeybees (*Apis mellifera*), leafcutter bees (*Megachile*), and alkali bees (*Nomia*), depending on production location.

Origin and domestication. Alfalfa originated in the Near East, in the area extending from Turkey to Iran and north into the Caucasus. Some scientists believe that alfalfa had two distinct centers of origin. The first center was in the mountainous region of Transcaucasia, and Asia Minor and the adjoining areas of northwestern Iran. The second area was in Central Asia. Domestication of alfalfa probably corresponded with the domestication of the horse. The oldest records trace to Turkey between 1400 and 1200 B.C., where alfalfa was fed to domestic animals during the winter and was regarded as a highly nutritious feed. The exact path of movement of alfalfa into the rest of the world is difficult to document. However, maritime trade was well developed in the eastern Mediterranean prior to 1400 B.C. and could have contributed to the spread of the crop. The Romans are known to have had alfalfa as early as the second century B.C. and are given credit for the development of advanced production practices. During this same period of time a Chinese expedition passing into Turkistan to obtain Iranian horses also obtained seed of alfalfa. With the fall of the Roman empire, alfalfa almost disappeared from Europe. During the eighteenth century alfalfa was taken from Europe to New Zealand, Australia, and the New World. There are two primary avenues of introduction of alfalfa into the United States. Spanish missionaries probably brought alfalfa from Chile and Mexico to Texas, New Mexico, Arizona, and California. The Chilean types were nonhardy, and thus climatic conditions prevented their movement into the northern United States and Canada. The second avenue of introduction was into Minnesota in 1857. A immigrant named Wendelin Grimm brought a few pounds of alfalfa seed (*M. media* = *M. sativa* × *M. falcata*) from Germany. Although his initial plantings were not very successful, by harvesting the seed from surviving plants he eventually developed a hardy type that became known as Grimm alfalfa.

Establishment, management, and harvesting. Ideally, alfalfa should be planted in deep, well-drained, medium-textured soil with a pH of about 6.8. Because alfalfa is grown from a single planting for 3 or more years, land preparation is extremely important to successful establishment and production. In areas where irrigation is practiced, special attention must be paid to slope, levee construction, and drainage.

Soils should be able to supply the following kilograms of nutrients per megagram of hay produced (1 kg/Mg = 2 lb/ton): potassium, 40; calcium, 36; magnesium, 7; phosphorus, 6; sulfur, 5; and iron, chlorine, manganese, boron, zinc, copper, molybdenum, less than 0.5. In addition, the crop requires 123 lb (56 kg) of nitrogen, which is largely derived by nitrogen fixation, by *Rhizobium meliloti*. Seed should be inoculated with these bacteria, if there is any question of their presence in the soil.

Planting rates vary between 16 and 45 kg/ha (14 and 40 lb/acre) depending on soil type and planting method. The higher rates are more common on heavy soils and when seeding is done by aircraft. Optimum soil temperatures for germination and seeding development are between 68 and 86°F (20 and 30°C). Dormant cultivars have lower optimum temperatures than nondormant cultivars. Forage should be harvested at one-tenth bloom or with 0.8–1.2 in. (2–3 cm) of regrowth (**Fig. 2**). In northern production areas, one to three harvests are taken annually, whereas in areas such as the Imperial Valley of California up to ten harvests may be taken in a year. *See* NITROGEN CYCLE.

Genetics and plant breeding. Alfalfa is an autotetraploid with a basic genomic number of 8 ($2n = 4x = 32$). However, diploid, hexaploid, and $n = 7$ species are represented in the genus. The frequency of quadrivalents is low, and so the frequency of double reduction is near zero. In the United States, prior to 1940 most of the improvement of alfalfa was by importation of germplasm adapted to production areas. Since the discovery of bacterial wilt disease (caused by *Clavibacter michiganese* ssp. *insidiosum*) considerable effort has been placed on the genetic improvement of the crop, with particular emphasis on the development of multiple pest and disease resistance. Phenotypic recurrent selection is the most common breeding method used for the improvement of alfalfa, but mass selection and various forms of genotypic recurrent selection are used. Cultivars released before 1960 were combinations of one to three germplasm sources, obtained from such areas as India, Chile, Peru, Africa, and Belgium, and primarily developed through public breeding programs. Subsequent to 1960, emphasis in public programs shifted to the development of breeding methods and germplasm. Most cultivars released since 1960 are populations (synthetics), proprietary, and combinations of several germplasm sources. A few commercial hybrids have been developed, and there is currently considerable interest in complementary strain crossing to produce new cultivars. *See* BREEDING (PLANT).

Seed production and standards. Seed production is most efficient in hot and dry environments such as the San Joaquin Valley of California. Seed fields are usually planted on beds at 2–4 kg/ha (1.8–3.6 lb/acre). Certified seed is produced under strict standards, which are monitored by state member organizations of the International Crop Improvement Association. These standards include the geographic area for seed production, the number of generations and number of years of seed increase, isolation, freedom from volunteer plants and weeds, viability, and mechanical purity.

Larry R. Teuber

Fig. 1. Alfalfa stems at flowering stage. (*Iowa State University Photo Service*)

Diseases. Alfalfa is susceptible to at least 75 diseases caused by fungi, bacteria, mycoplasma-like

Fig. 2. Alfalfa of wide crown type at late bud or early bloom stage. (*Iowa State University Photo Service*)

Fig. 3. Damaging effect of bacterial wilt of alfalfa (left), contrasted with healthy plant (right).

organisms, viruses, and nematodes. Most of these occur sporadically, and rarely in epidemic proportions; however, several are responsible for appreciable losses in forage yield and seed production. The best way to reduce disease losses in alfalfa is to grow locally adapted disease-resistant varieties.

Root and crown diseases. Bacterial wilt is one of the best-known and most damaging diseases of alfalfa (**Fig. 3**). The bacteria enter the plants principally through wounds in the roots and through cut ends of newly mowed stems. They are spread in the field by surface water and by tillage and harvesting machinery. Bacterial wilt causes stunting and reduced stand longevity. Resistant varieties have been developed for all alfalfa-growing regions of the United States.

Fig. 4. Fruiting structures of the Verticillium wilt fungus. Conidia (spores) are borne singly at the tips of conidiophores. (*From J. H. Graham et al., A Compendium of Alfalfa Diseases, American Phytopathological Society, 1979*)

Verticillium wilt (**Fig. 4**), first identified in the United States in 1976, has been a damaging disease in Europe since the 1930s. It occurs in the northwestern United States and adjacent areas in Canada. Fusarium wilt, another fungus disease, occurs worldwide but is most severe in warm regions.

Phytophthora root rot, caused by a water mold, occurs worldwide. It is favored by waterlogged soils resulting from excessive irrigation, abundant rainfall, or inadequate soil drainage. Resistant varieties are available.

Other root and crown diseases include Sclerotinia crown rot, active during cool moist weather; Stagonospora root rot and leaf spot; and Fusarium and Rhizoctonia crown and root rots, more active during warm periods. Frequently, crown and root rots cannot be ascribed to a single pathogen, so the term disease complex is used to characterize the pathogens in combinations with various interacting agents (**Fig. 5**). These agents include stress factors such as winter damage, improper cultural and management practices, and root insects. No variety of alfalfa resistant to this complex of disorders has been developed so far.

Stem and leaf diseases. Many microorganisms attack the stems and foliage of alfalfa. Some, such as the fungi that cause anthracnose and spring black stem, spread into the crown and upper part of the taproot, weakening or killing the plant. Anthracnose, favored by warm humid weather, is associated with the summer decline of alfalfa in the eastern United States. Resistance to anthracnose has been incorporated into some varieties; however, many of these varieties are susceptible to a second race of the anthracnose fungus which was discovered in 1978. Spring black stem occurs in most alfalfa-growing areas of the world and is particularly damaging in cooler humid regions. The fungus can attack any part of the plant, but the shiny black streaks on stems and irregularly shaped black spots on leaves are most conspicuous. Summer black stem is another disease of stems and leaves; its reddish to smoky-brown lesions appear when temperatures increase during summer.

Several leaf diseases occur widely and cause losses through defoliation and a general weakening of the plant. Common leaf spot, yellow leaf blotch, Leptosphaerulina leaf spot, and Stemphylium leaf spot may develop so abundantly that they cause serious loss of leaves. Forage quality and yield are lowered. Some varieties have low levels of resistance. Early harvest may help reduce defoliation as well as spread of leaf and stem diseases.

Some less prevalent but occasionally locally important foliar diseases are downy mildew, leaf rust, and bacterial leaf spot.

Virus diseases. The alfalfa mosaic virus (AMV) complex consists of many strains that differ in virulence

Fig. 5. Crown rot of alfalfa. It may be caused by several fungi in combination with stress factors such as winter injury. (*From J. H. Graham et al., A Compendium of Alfalfa Diseases, American Phytopathological Society, 1979*)

and symptoms. The classic symptom is yellow mottling of leaves. Other symptoms are stunting, contortions of leaves and petioles, and, in some instances, death of plants. However, most AMV-infected plants in an alfalfa stand never show symptoms. Exact information on disease losses is lacking. Because the virus is seedborne, seed produced in areas of low virus incidence should be used. Other viruses of alfalfa include alfalfa enation, transient streak, and alfalfa latent viruses. *See* PLANT VIRUSES AND VIROIDS.

Mycoplasmalike diseases. Witches'-broom, long believed to be a leafhopper-transmitted virus, is now thought to be caused by a mycoplasmalike microorganism. The disease is of minor importance in the United States but is widespread in parts of semiarid Australia.

Nematodal diseases. Plant-parasitic nematodes of alfalfa not only damage plants directly but also aid in the transmission and entry of other pathogens. The alfalfa stem nematode is most damaging, especially in the western United States. Infected crown buds and stems are swollen and stunted and may be killed. Reuse of irrigation water is a major pathway for dissemination. Some resistant varieties have been developed. The root-knot nematodes (*Meloidogyne* spp.) and the root-lesion nematodes (*Pratylenchus* spp.) are widespread and attack many crop plants and weeds. There is usually less damage to the alfalfa crop than to more susceptible crops that follow in a rotation.

Noninfectious diseases. Alfalfa frequently shows disease symptoms not associated with pathogenic or parasitic organisms. These noninfectious or abiotic diseases are caused by the lack or excess of plant nutrients, extreme temperatures, air pollutants, pesticides, harvesting machinery, and other factors. Symptoms caused by these agents can be confused with those caused by infectious (biotic) agents. In addition, abiotic agents often influence the severity of diseases caused by biotic agents. *See* PLANT PATHOLOGY.

J. H. Graham

Bibliography. D. K. Barnes et al., Alfalfa germplasm in the United States: Genetic vulnerability, use improvement and maintenance, *USDA Tech. Bull.* 1571, 1977; C. H. Hanson (ed.), *Alfalfa Science and Technology*, American Society of Agronomy, 1972; A. A. Hanson, R. R. Hill, and D. K. Barnes (eds.), *Alfalfa and Alfalfa Improvement*, American Society of Agronomy, 1988; D. Smith, *Forage Management in the North*, 1975.

Alfvén waves

Propagating oscillations in electrically conducting fluids or gases in which a magnetic field is present.

Magnetohydrodynamics deals with the effects of magnetic fields on fluids and gases which are efficient conductors of electricity. Molten metals are generally good conductors of electricity, and they exhibit magnetohydrodynamic phenomena. Gases can be efficient conductors of electricity if they become ionized. Ionization can occur at high temperatures or through the ionizing effects of high-energy (usually ultraviolet) photons. A gas which consists of free electrons and ions is called a plasma. Most gases in space are plasmas, and magnetohydrodynamic phenomena are expected to play a fundamental role in the behavior of matter in the cosmos. *See* MAGNETISM; PLASMA (PHYSICS).

Waves are a particularly important aspect of magnetohydrodynamics. They transport energy and momentum from place to place and may, therefore, play essential roles in the heating and acceleration of cosmical and laboratory plasmas. A wave is a propagating oscillation. If waves are present, a given parcel of the fluid undergoes oscillations about an equilibrium position. The parcel oscillates because there are restoring forces which tend to return it to its equilibrium position.

In an ordinary gas, the only restoring force comes from the thermal pressure of the gas. This leads to one wave mode: the sound wave. If a magnetic field is present, there are two additional restoring forces: the tension associated with magnetic field lines, and the pressure associated with the energy density of the magnetic field. These two restoring forces lead to two additional wave modes. Thus there are three magnetohydrodynamic wave modes. However, each restoring force does not necessarily have a unique wave mode associated with it. Put another way, each wave mode can involve more than one restoring force. Thus the usual sound wave, which involves only the thermal pressure, does not appear as a mode in magnetohydrodynamics. *See* SOUND; WAVE MOTION IN FLUIDS.

The three modes have different propagation speeds, and are named fast mode (F), slow mode (S), and intermediate mode (I). The intermediate mode is sometimes called the Alfvén wave, but some

scientists refer to all three magnetohydrodynamic modes as Alfvén waves. The intermediate mode is also called the shear wave. Some scientists give the name magnetosonic mode to the fast mode.

Basic equations. The magnetohydrodynamic wave modes are analyzed by using the magnetohydrodynamic equations for the motion of a conducting fluid in a magnetic field, combined with Maxwell's equations and Ohm's law. *See* MAXWELL'S EQUATIONS.

In studies of waves, the most important of the magnetohydrodynamic equations is called the momentum equation. It expresses how fluid parcels accelerate in response to the various forces acting on the fluid. The essential forces come from the thermal pressure in the fluid, and from the tension and pressure forces associated with the magnetic field. The magnetic field can exert forces on the fluid only when an electric current flows in the fluid; thus magnetohydrodynamic waves can exist only in an electrically conducting fluid. *See* FLUID FLOW.

Two other equations express the conservation of mass of the fluid and the conservation of entropy. However, entropy is conserved only when viscosity, heat conduction, and electrical resistivity are ignored. *See* CONSERVATION OF MASS; ENTROPY.

The fourth equation is Ohm's law, which expresses the relationship between the electric current and the electric field in the fluid. If the fluid has no electrical resistivity, any nonzero electric field would lead to an infinite current, which is physically unacceptable. Thus the electric field "felt" by any fluid parcel must be zero. However, detailed analysis of the motions of electrons and ions reveals that this statement is valid only if the wave frequency is much less than the cyclotron frequency of the plasma ions; thus magnetohydrodynamic waves are very low-frequency waves. *See* OHM'S LAW; PARTICLE ACCELERATOR; RELATIVISTIC ELECTRODYNAMICS.

It is possible to combine Ohm's law with Faraday's law of induction. The resultant equation is called the magnetohydrodynamic induction equation, which is the mathematical statement of the "frozen-in" theorem. This theorem states that magnetic field lines can be thought of as being frozen into the fluid, with the proviso that the fluid is always allowed to slip freely along the field lines. The coupling between the fluid and the magnetic field is somewhat like the motion of a phonograph needle in the grooves of a record: the needle is free to slip along the grooves (the field lines) while otherwise being constrained from crossing the grooves. It is this coupling between the fluid and the magnetic field which makes magnetohydrodynamic waves possible. The oscillating magnetic field lines cause oscillations of the fluid parcels, while the fluid provides a mass loading on the magnetic field lines. This mass loading has the effect of slowing down the waves, so that they propagate at speeds much less than the speed of light (which is the propagation speed of waves in a vacuum). *See* ELECTROMAGNETIC RADIATION; FARADAY'S LAW OF INDUCTION; LIGHT.

The final equation is Ampère's law, which states that an electric current always produces a magnetic field. *See* AMPÈRE'S LAW; BIOT-SAVART LAW.

Thus there are six equations for six unknowns (the pressure, density, and velocity of the fluid; the electric current; and the electric and magnetic fields). Six equations are in principle sufficient to determine six unknowns.

Linearization of equations. Unfortunately, the six equations are too difficult to be of much use because some of them are nonlinear; that is, they contain products of the quantities for which a solution is sought. Nonlinear magnetohydrodynamics is still only in its infancy, and only a few specialized solutions are known. In order to get solvable equations, scientists accept the limitation of dealing with small-amplitude waves and linearize the equations, so that products of the unknowns are removed. Fortunately, much can still be learned from this procedure. The linearization consists of four steps: (1) All quantities are split into two parts: a background part (denoted by the subscript 0) which is constant in space and time, and a wave part (denoted by the prefix δ) which oscillates in space and time; for example, $\mathbf{B} = \mathbf{B}_0 + \delta\mathbf{B}$ and $\mathbf{v} = \mathbf{v}_0 + \delta\mathbf{v}$, where $\mathbf{B}$ denotes the vector magnetic field and $\mathbf{v}$ denotes fluid flow velocity. (2) These definitions are inserted into the six fundamental equations. (3) It is assumed that the waves are of such small amplitude that products of the wave parts can be ignored. (4) It is assumed that there is no background flow. (This last step simplifies the mathematics without loss of generality.)

The resulting equations have solutions which are harmonic in time and space, that is, which vary as $\exp(i\mathbf{k} \cdot \mathbf{r} - i\omega t)$. Here $\mathbf{r}$, is spatial position, ω is (angular) frequency, and $\mathbf{k}$ is called the wave vector. The magnitude of the wave vector represents how rapidly the wave oscillates in space, and its direction represents the direction of propagation. The magnitude of the angular frequency represents how rapidly the wave oscillates in time. Since the equations are linear, more general solutions can be constructed by adding together an arbitrary number of harmonic solutions. *See* HARMONIC MOTION.

Dispersion relation. The linearized equations have solutions only when ω and $\mathbf{k}$ are related to one another in a special way. The equation which expresses this relationship between ω and $\mathbf{k}$ is called the dispersion relation. It is a fundamental relationship in the theory of waves. *See* DISPERSION RELATIONS.

The dispersion relation for magnetohydrodynamic waves is given by Eq. (1). The quantity M_0^2 is given by Eq. (2). The quantity $k_\parallel$ is the component of $\mathbf{k}$

$$(\omega^2 - k_\parallel^2 v_A^2)(k_\perp^2 + M_0^2) = 0 \tag{1}$$

$$M_0^2 = \frac{(k_\parallel^2 c_s^2 - \omega^2)(k_\parallel^2 v_A^2 - \omega^2)}{(c_s^2 + v_A^2)(k_\parallel^2 c_T^2 - \omega^2)} \tag{2}$$

parallel to $\mathbf{B}_0$ while $k_\perp$ is the component perpendicular to $\mathbf{B}_0$. Equation (2) contains several velocities which are fundamental velocities in magnetohydrodynamics. The quantity v_A is called the Alfvén speed. The Alfvén speed is equivalent to the speed at which a wave travels along a taut string, except that the tension of the string is replaced by the tension in

the magnetic field lines. The quantity c_s is the usual speed of sound in a gas, and c_T is called the tube speed, given by Eq. (3).

$$c_T^2 = c_s^2 v_A^2/(c_s^2 + v_A^2) \tag{3}$$

Equation (1) is a cubic equation for ω^2. The three roots represent the three magnetohydrodynamic wave modes. For each of the three values of ω^2, there are two values of ω (one positive and one negative) which correspond to waves propagating in opposite directions.

A particularly useful quantity is the propagation speed, or phase speed, of the wave, given by Eq. (4).

$$v_{\text{ph}} = \omega/k \tag{4}$$

This quantity can be obtained by solving Eq. (1). The three modes are designated fast, intermediate, and slow, according to their phase speeds (**Fig. 1**). The phase speed depends on the angle θ between the wave propagation direction and the direction of $\mathbf{B}_0$. Thus θ satisfies Eq. (5). The graphs of phase speed

$$\tan\theta = k_\perp/k_\parallel \tag{5}$$

change character depending on whether c_s/v_A is less than or greater than 1. *See* PHASE VELOCITY.

All three modes can propagate along the magnetic field, while only the fast mode can propagate across the magnetic field. For propagation along the field ($\theta = 0$), two of the modes have $v_{\text{ph}} = v_A$, while the third has $v_{\text{ph}} = c_s$; the latter mode is essentially a sound wave unaffected by the magnetic field, while the other two modes are intrinsically magnetic. The intermediate mode has $v_{\text{ph}} = v_A \cos\theta$; this mode is always magnetic in character, and thus the sound speed does not affect its phase speed. In all other cases the fast and slow modes have a mixed character, and all three restoring forces combine in a complicated way. The essential properties of the modes are summarized below.

Intermediate mode. The velocity and magnetic field fluctuations in this mode are perpendicular to the plane containing $\mathbf{k}$ and $\mathbf{B}_0$. This means that the motions are pure shears. There is no compression of the plasma. This is why the sound speed does not appear in the dispersion relation. The tension in the magnetic field lines is the only restoring force involved in the propagation of the wave. This mode is therefore closely analogous to the propagation of waves on a string.

The dispersion relation is given by Eq. (6). The

$$\omega^2 = k_\parallel^2 v_A^2 \tag{6}$$

quantity $k_\perp$ does not appear because the motions on neighboring field lines do not communicate with one another. This is a consequence of the fact that the motions are shears, so that neighboring field lines never bump together.

The fluctuating parts of the fluid flow velocity and the magnetic field, $\delta\mathbf{v}$ and $\delta\mathbf{B}$, are closely related (**Fig. 2**) and in fact satisfy Eq. (7). (The symbol sgn means the algebraic sign of the quantity in parentheses.) Equation (7) has been used to identify the presence of the intermediate mode in the solar wind. There, spacecraft-borne instruments have measured the magnetic field and plasma velocity directly. The fluctuations have been found to obey Eq. (7) fairly closely. It appears that intermediate waves are rather copious in the solar wind. Further investigations have revealed that the waves usually propagate away from the Sun. The Sun may thus be an emitter of these waves. However, the waves that are found in the solar wind also exhibit many properties of turbulence; thus, nonlinear effects, such as interactions between different waves, are important. *See* SOLAR WIND; SUN.

$$\delta\mathbf{v}/v_A = -\text{sgn}(\mathbf{k}\cdot\mathbf{B}_0)\delta\mathbf{B}/\mathbf{B}_0 \tag{7}$$

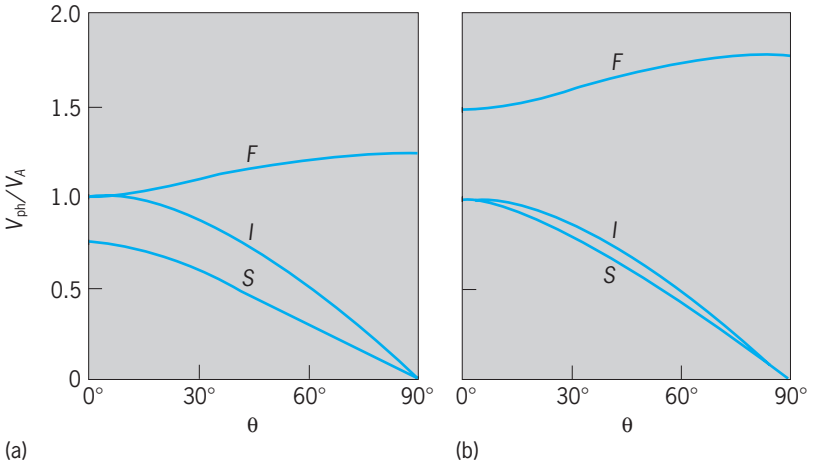


Fig. 1. Phase speed as a function of propagation direction for the fast (F), slow (S), and intermediate (I) modes. The ordinate is phase velocity V_{ph}, normalized to the Alfvén speed V_A. The abscissa is the angle θ between the wave propagation of the direction of the magnetic field B_0. (*a*) $C_s/V_A = 0.75$. (*b*) $C_s/V_A = 1.5$.

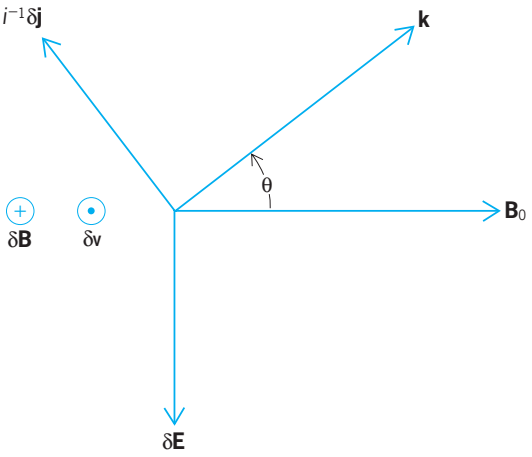


Fig. 2. Vector relationship between the various fluctuating wave quantities for the intermediate mode. Symbols next to $\delta\mathbf{B}$ and $\delta\mathbf{v}$ indicate that these quantities are pointing into and out of the paper, respectively. The electric current $\delta\mathbf{j}$ must be multiplied by i^{-1}(or $1/\sqrt{-1}$) because the current is 90° out of phase with the other quantities.

Waves propagate energy. Since the intermediate mode does not involve the thermal properties of the plasma, its energy flux density is given by the time average of the electromagnetic Poynting flux. For the intermediate mode, the wave energy flux density is always parallel or antiparallel to $\mathbf{B}_0$. Thus the intermediate mode always channels energy along the magnetic field. In fact, the energy flux density of the wave consists of the wave's kinetic energy density plus the wave's magnetic energy density, propagating along the magnetic field at speed v_A. *See* POYNTING'S VECTOR.

Because these waves channel energy along magnetic fields, they may be responsible for the observed fact that cosmical plasmas are strongly heated in the presence of magnetic fields. For example, in the solar wind the wave energy is observed to decrease at greater distances from the Sun, and the amount of wave energy that is lost is about the same as the heat energy that is gained by the solar-wind plasma. Thus, the solar wind is an example of a cosmical plasma that is heated by waves. Waves may also accelerate the solar wind through an effect which is closely analogous to the radiation pressure exerted by light waves. They may also heat the solar atmosphere. Results from the *Solar and Heliospheric Observatory (SOHO)* indicate that protons and heavier ions are strongly heated close to the Sun, in the region where the solar wind is formed. The heating is strongest in the directions perpendicular to $\mathbf{B}_0$. The high-frequency intermediate mode (also called the ion-cyclotron wave) is capable of producing this kind of heating. However, it has not yet been proven that the waves are present at the Sun. It is even less certain whether waves heat the atmospheres of other stars, or other astrophysical objects. Intermediate waves have also been proposed as a means of heating plasmas in controlled thermonuclear fusion devices. *See* NUCLEAR FUSION.

Fast mode. The fast mode is difficult to analyze. However, in many cosmical and laboratory plasmas, v_A^2 is much greater than C_S^2. This is the strong-magnetic-field case. In that case the fast mode is more easily understood.

From Eq. (1), the fast-mode dispersion relation is approximately given by Eq. (8). The direction of $\mathbf{k}$

$$\omega^2/k^2 \approx v_A^2 \tag{8}$$

does not appear in Eq. (8). Thus the waves propagate isotropically.

Fast waves are compressive, and the magnetic field strength fluctuates as well (**Fig. 3**). Thus fast waves are governed by the two restoring forces associated with the tension and pressure in the magnetic field.

Because the magnetic field is strong, the wave energy flux density is mainly due to the time average of the Poynting flux. For this mode, it is evident that the energy flux is along $\mathbf{k}$. Thus the fast mode can propagate energy across the magnetic field.

Slow mode. Like the fast mode, the slow mode is difficult to study in general, and the discussion will again be confined to strong magnetic fields, so that v_A^2 is much greater than C_S^2. The slow mode in a strong field is equivalent to sound waves which are guided along the strong magnetic field lines. The strong magnetic field lines can be thought of as a set of rigid pipes which allow free fluid motion along the pipes, but which restrict motion in the other two directions. The motions on the individual pipes are not coupled together, and thus the slow mode is analogous to the sound waves on a set of independent organ pipes. The slow mode channels energy along the magnetic field. Because the sound speed is small, by assumption, the slow mode transmits energy less effectively than the fast or intermediate modes.

Nonlinear effects. Only small-amplitude waves have been considered. Real waves have finite amplitude, and nonlinear effects can sometimes be important. One such effect is the tendency of waves to steepen, ultimately forming magnetohydrodynamic shock waves and magnetohydrodynamic discontinuities. There is an abundance of magnetohydrodynamic discontinuities in the solar wind. Nonlinear studies have shown these are the result of the nonlinear steepening of the intermediate waves. Nonlinear studies have also shown that the intermediate mode will naturally reach a state where the magnetic field can have large fluctuations of its direction, but maintain a constant field strength. Thus the tip of the magnetic field vector moves on the surface of a sphere. This state of spherical polarization is in fact commonly observed in the solar wind. *See* NONLINEAR ACOUSTICS; SHOCK WAVE.

It is also possible that waves can degenerate into turbulence. There are indications that this too happens in the solar wind. The observed power spectra show that fluctuations are present with a very broad range of frequencies and wave-vector magnitudes, which is a property of turbulence. At greater distances from the Sun, the outward-propagating waves are less dominant; they are apparently getting scrambled because of turbulent processes. *See* TURBULENT FLOW.

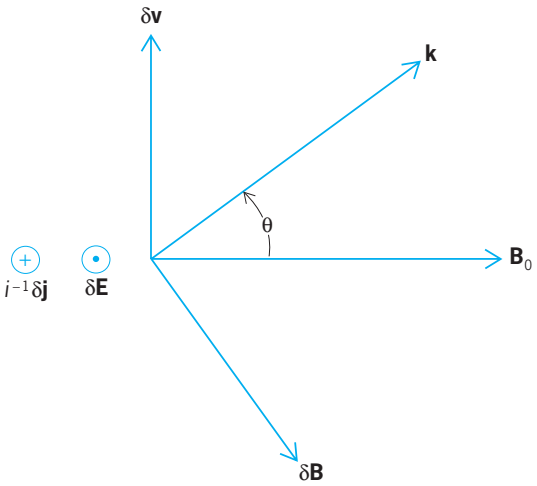


Fig. 3. Vector relationships between the various fluctuating wave quantities for the fast mode in a strong magnetic field, where v_A^2 is much greater than c_A^2.

Dissipation. Viscosity, heat conduction, and electrical resistivity have all been ignored in the above discussion. These effects will all tend to dissipate the wave energy and heat the plasma. *See* CONDUCTION (HEAT); ELECTRICAL RESISTIVITY; SOUND ABSORPTION; VISCOSITY.

In some plasmas, the electrons and ions almost never collide with one another. In such nearly collisionless plasmas, the fast and slow modes can dissipate by the process of Landau damping. Shock formation and turbulence also lead to wave dissipation.

Surface waves. Only waves in a spatially uniform background have been considered. While the analysis of magnetohydrodynamic waves in a nonuniform background is complicated, it is possible to consider an extreme limit, in which the background is uniform except at certain surfaces where it changes discontinuously. Surfaces can support magnetohydrodynamic waves, which are in some respects similar to waves on the surface of a lake. These waves may play important roles in heating cosmical and laboratory plasmas. *See* MAGNETOHYDRODYNAMICS; WAVE MOTION IN LIQUIDS. Joseph V. Hollweg

Bibliography. R. L. Carovillano and J. M. Forbes (eds.), *Solar-Terrestrial Physics*, 1983; F. F. Chen, *Introduction to Plasma Physics*, 1984; J. V. Hollweg, *Computer Phys. Rep.*, 12:205–232, 1990; E. R. Priest, *Solar Magnetohydrodynamics*, 1982; R. Schwenn and E. Marsch (eds.), *Physics of the Inner Heliosphere*, vol. 2, 1991.

Algae

An informal assemblage of predominantly aquatic organisms that carry out oxygen-evolving photosynthesis but lack specialized water-conducting and food-conducting tissues. They may be either prokaryotic (lacking an organized nucleus) and therefore members of the kingdom Monera, or eukaryotic (with an organized nucleus) and therefore members of the kingdom Plantae, constituting with fungi the subkingdom Thallobionta. They differ from the next most advanced group of plants, Bryophyta, by their lack of multicellular sex organs sheathed with sterile cells and by their failure to retain an embryo within the female organ. Many colorless organisms are referable to the algae on the basis of their similarity to photosynthetic forms with respect to structure, life history, cell wall composition, and storage products. The study of algae is called algology (from the Latin *alga*, meaning sea wrack) or phycology (from the Greek *phykos*, seaweed). *See* BRYOPHYTA; PLANT KINGDOM; THALLOBIONTA.

General form and structure. Algae range from unicells 1–2 micrometers in diameter to huge thalli [for example, kelps often 100 ft (30 m) long] with functionally and structurally distinctive tissues and organs. Unicells may be solitary or colonial, attached or free-living, with or without a protective cover, and motile or nonmotile. Colonies may be irregular or with a distinctive pattern, the latter type being flagellate or nonmotile. Multicellular algae form packets, branched or unbranched filaments, sheets one or two cells thick, or complex thalli, some with organs resembling roots, stems, and leaves (as in the brown algal orders Fucales and Laminariales). Coenocytic algae, in which the protoplast is not divided into cells, range from microscopic spheres to thalli 33 ft (10 m) long with a complex structure of intertwined siphons (as in the green algal order Bryopsidales).

Classification. Sixteen major phyletic lines (classes) are distinguished on the basis of differences in pigmentation, storage products, cell wall composition, flagellation of motile cells, and structure of such organelles as the nucleus, chloroplast, pyrenoid, and eyespot. These classes are interrelated to varying degrees, the interrelationships being expressed by the arrangement of classes into divisions (the next-higher category). Among phycologists there is far greater agreement on the number of major phyletic lines than on their arrangement into divisions.

Superkingdom Prokaryotae
- Kingdom Monera
  - Division Cyanophycota (= Cyanophyta, Cyanochloronta)
    - Class Cyanophyceae, blue-green algae
  - Division Prochlorophycota (= Prochlorophyta)
    - Class Prochlorophyceae

Superkingdom Eukaryotae
- Kingdom Plantae
  - Subkingdom Thallobionta
    - Division Rhodophycota (= Rhodophyta, Rhodophycophyta)
      - Class Rhodophyceae, red algae
    - Division Chromophycota (= Chromophyta)
      - Class: Chrysophyceae, golden or golden-brown algae
        - Prymnesiophyceae (= Haptophyceae)
        - Xanthophyceae (= Tribophyceae), yellow-green algae
        - Eustigmatophyceae
        - Bacillariophyceae, diatoms
        - Dinophyceae, dinoflagellates
        - Phaeophyceae, brown algae
        - Raphidophyceae, chloromonads
        - Cryptophyceae, cryptomonads
    - Division Euglenophycota (= Euglenophyta, Euglenophycophyta)
      - Class Euglenophyceae
    - Division Chlorophycota (= Chlorophyta, Chlorophycophyta)
      - Class: Chlorophyceae, green algae
        - Charophyceae, charophytes
        - Prasinophyceae

Placing more taxonomic importance on motility than on photosynthesis, zoologists traditionally have considered flagellate unicellular and colonial algae as protozoa, assigning each phyletic line the rank of order. *See* BACILLARIOPHYCEAE; CHAROPHYCEAE; CHLOROPHYCEAE; CHRYSOPHYCEAE; CRYPTOPHY-

CEAE; CYANOPHYCEAE; DINOPHYCEAE; EUGLENOPHYCEAE; EUKARYOTAE; EUSTIGMATOPHYCEAE; PHAEOPHYCEAE; PRASINOPHYCEAE; PROCHLOROPHYCEAE; PROKARYOTAE; PROTOZOA; PRYMNESIOPHYCEAE; RAPHIDOPHYCEAE; RHODOPHYCEAE; THALLOBIONTA; XANTHOPHYCEAE.

Cell covering. Although some unicellular algae are naked or sheathed by mucilage or scales, most are invested with a covering (wall, pellicle, or lorica) of diverse composition and construction. These coverings consist of at least one layer of polysaccharide (cellulose, alginate, agar, carrageenan, mannan, or xylan), protein, or peptidoglycan that may be impregnated or encrusted with calcium carbonate, iron, manganese, or silica. They are often perforated and externally ornamented (**Fig. 1***a, d, i, j,* and *k*). Diatoms have a complex wall composed almost entirely of silica. In multicellular and coenocytic algae,

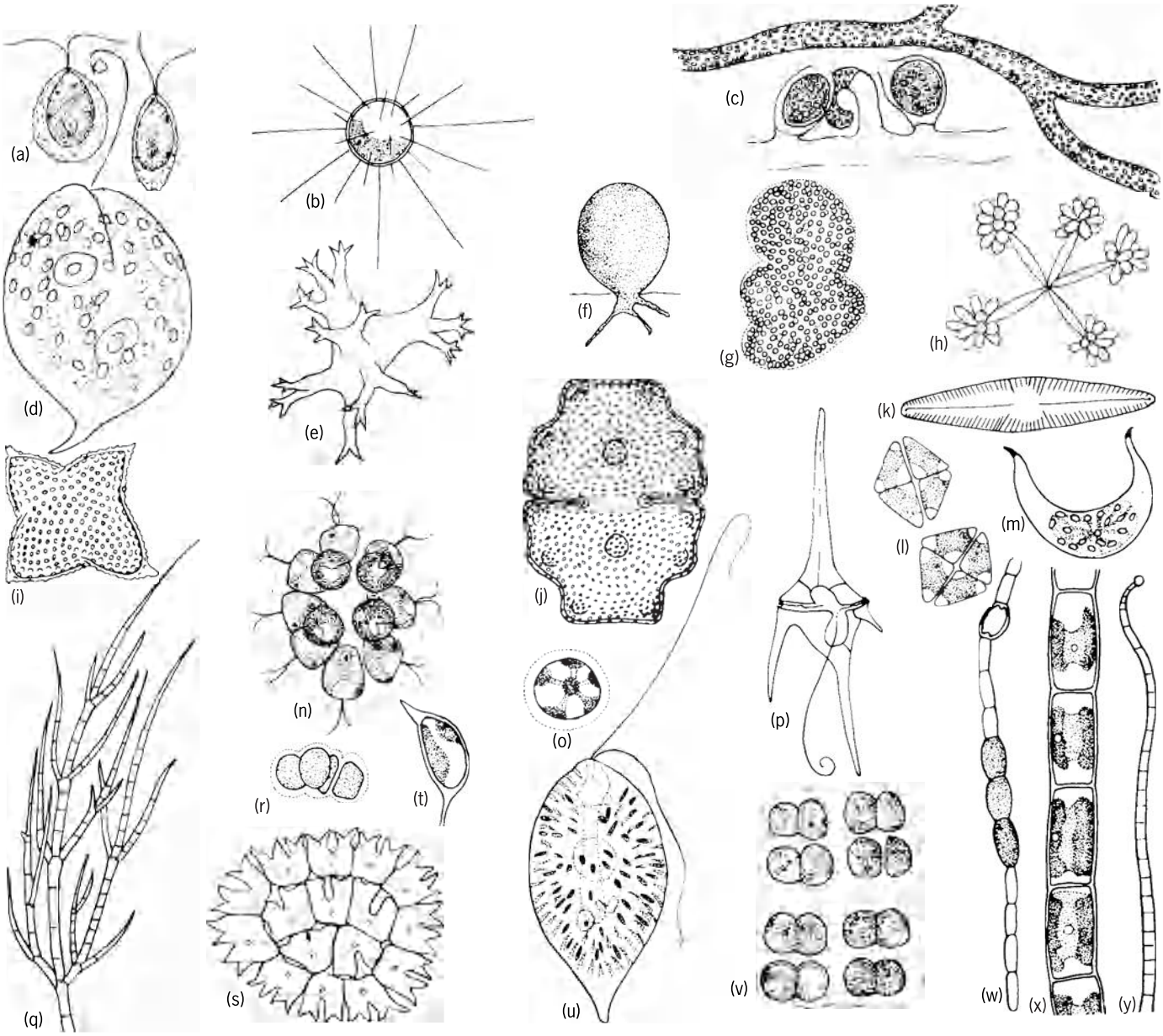


Fig. 1. Representative fresh-water algae. ***(a) Phacotus*** **(Chlorophyceae), motile unicell with a lorica;** ***(b) Golenkinia*** **(Chlorophyceae), nonmotile planktonic unicell;** ***(c) Vaucheria*** **(Xanthophyceae), siphonous coenocyte;** ***(d) Phacus*** **(Euglenophyceae), uniflagellate unicell;** ***(e) Tetraedron*** **(Chlorophyceae), nonmotile planktonic unicell;** ***(f) Botrydium*** **(Xanthophyceae), terrestrial spheroidal coenocyte;** ***(g) Microcystis*** **(Cyanophyceae), coccoid colony;** ***(h) Actidesmium*** **(Chlorophyceae), irregular colony;** ***(i) Tetragoniella*** **(Xanthophyceae), nonmotile unicell;** ***(j) Euastrum*** **(Chlorophyceae), desmid showing semicells;** ***(k) Navicula*** **(Bacillariophyceae), unicell with siliceous wall;** ***(l) Crucigenia*** **(Chlorophyceae), regular nonmotile colony;** ***(m) Cystodinium*** **(Dinophyceae), nonmotile cystlike dinoflagellate:** ***(n) Pandorina*** **(Chlorophyceae), regular motile colony;** ***(o) Asterococcus*** **(Chlorophyceae), unicell in mucilaginous sheath;** ***(p) Ceratium*** **(Dinophyceae), motile dinoflagellate;** ***(q) Stigeoclonium*** **(Chlorophyceae), branched filament;** ***(r) Chroococcus*** **(Cyanophyceae), unicell, solitary or in irregular colony;** ***(s) Pediastrum*** **(Chlorophyceae), regular nonmotile colony;** ***(t) Characium*** **(Chlorophyceae), epiphytic unicell;** ***(u) Gonyostomum*** **(Raphidophyceae), biflagellate unicell;** ***(v) Merismopedia*** **(Cyanophyceae), platelike colony;** ***(w) Wollea*** **(Cyanophyceae), one trichome from a gelatinous colony;** ***(x) Ulothrix*** **(Chlorophyceae), unbranched filament;** ***(y) Oscillatoria*** **(Cyanophyceae), trichome lacking sheath.**

most reproductive cells are naked, but vegetative cells have walls whose composition varies from class to class. *See* CELL WALLS (PLANT).

Cytoplasmic structure. Prokaryotic algae lack membrane-bounded organelles. Eukaryotic algae have an intracellular architecture comparable to that of higher plants but more varied. Among cell structures unique to algae are contractile vacuoles in some freshwater unicells, gas vacuoles in some planktonic blue-green algae, ejectile organelles in dinoflagellates and cryptophytes, and eyespots in motile unicells and reproductive cells of many classes. Nuclei of different taxonomic groups vary in the degree of nuclear breakdown during division, the presence of histone around the deoxyribonucleic acid (DNA), details of spindle formation and elongation, and chromosome behavior. Chromosome numbers vary from $n = 2$ in some red and green algae to $n \geq 300$ in some dinoflagellates. The dinoflagellate nucleus is in some respects intermediate between the chromatin region of prokaryotes and the nucleus of eukaryotes and is termed mesokaryotic. Some algal cells characteristically are multinucleate, while others are uninucleate. Chloroplasts, which always originate by division of preexisting chloroplasts, have the form of plates, ribbons, disks, networks, spirals, or stars and may be positioned centrally or along the cell wall. Photosynthetic membranes (thylakoids) are arranged in distinctive patterns and contain pigments diagnostic of individual classes. Chloroplast DNA may be dispersed throughout the organelle or (in Chromophycota) distributed in a ring. Pyrenoids—proteinaceous bodies around which carbohydrate is stored—occur in the chloroplasts of at least some species of all classes. They have been shown to contain the same carboxylating enzyme that is present in higher plants with C_3 metabolism. *See* CELL (BIOLOGY); CELL PLASTIDS; CHROMOSOME; DEOXYRIBONUCLEIC ACID (DNA); PHOTOSYNTHESIS; PLANT CELL.

Movement. In all classes of algae except Prochlorophyceae, there are cells that are capable of movement. The slow, gliding movement of certain blue-green algae, diatoms, and reproductive cells of red algae presumably results from extracellular secretion of mucilage. Ameboid movement, involving pseudopodia, is found in certain Chrysophyceae and Xanthophyceae. An undulatory or peristaltic movement occurs in some Euglenophyceae.

The fastest movement is produced by flagella, which are borne by unicellular algae and reproductive cells of multicellular algae representing all classes except Cyanophyceae, Prochlorophyceae, and Rhodophyceae. Most motile cells have two flagella, inserted apically or laterally, but some have three, four, eight, or a crown of flagella. Those that appear to be uniflagellate probably have a second, nonemergent flagellum. The basic structural feature of a flagellum—two central microtubules surrounded by nine doublets—is constant throughout the algae (as in all other plants and animals) except in male gametes of diatoms, in which the central pair is absent. Flagella may be decorated with hairs, spines, or scales, the details being of diagnostic value at various taxonomic levels. Details of the flagellar basal region are extremely variable and are considered to be of phylogenetic significance. Biflagellate cells may be isokont (with identical flagella) or heterokont (with morphologically different flagella). Flagellate cells with eyespots are phototactic, moving into areas of optimal light intensity.

Internal movement also occurs in algae in the form of cytoplasmic streaming and light-induced orientation of chloroplasts. *See* CELL MOTILITY; CILIA AND FLAGELLA.

Cell division and growth. Cell division is the process by which unicellular algae multiply and multicellular algae increase in size and complexity. Even within a single class, such as the green algae, details of cell division vary with respect to involvement and interaction of nuclei, microtubules, vesicles, and plasmalemma in wall formation. Random cell division results in diffuse growth, as in a filament or sheet. Thalli with differentiated organs, however, require the localization of cell division in discrete areas or tissues (meristems), which may be apical, basal, marginal, or intercalary. Growth in coenocytic algae entails expansion or elongation of the wall and repeated division of the nucleus and other organelles, but without formation of internal walls. *See* CELL DIVISION.

Reproduction and life histories. Sexual reproduction is unknown in prokaryotic algae and in three classes of eukaryotic unicells (Eustigmatophyceae, Cryptophyceae, and Euglenophyceae), in which the production of new individuals is by binary fission. In sexual reproduction, which is found in all remaining classes, the members of a copulating pair of gametes may be morphologically indistinguishable (isogamous), morphologically distinguishable but with both gametes motile (anisogamous), or differentiated into a motile sperm and a relatively large nonmotile egg (oogamous). Gametes may be formed in undifferentiated cells or in special organs (gametangia), male (antheridia) and female (oogonia). (The highly modified gametangia of red algae are called spermatangia and carpogonia, respectively.) When gametangia are multicellular, every cell becomes fertile, sterile protective cells being absent. Sex hormones (attractants and determinants) have been demonstrated in a few brown and green algae. Sexual reproduction may be replaced or supplemented by asexual reproduction, in which special cells (spores) capable of developing directly into a new alga are formed in undifferentiated cells or in distinctive organs (sporangia). Spores may be motile or nonmotile, and may germinate immediately upon release or may secrete a thick wall to protect themselves during a period unfavorable to germination. Many multicellular algae reproduce vegetatively by fragmentation or by the formation of propagules. *See* REPRODUCTION (PLANT).

Life histories range from binary fission in some unicells to an elaborate sequence of four multicellular somatic phases in many red algae, involving

highly specialized reproductive systems. The different life history patterns are defined in terms of the position of syngamy and meiosis in relation to somatic phases. The three basic types of life history described below may be characterized as having zygotic meiosis, sporic meiosis, and gametic meiosis, respectively. In exceptional cases, meiosis regularly occurs in a somatic tissue, resulting in chimeralike thalli and a complicated life history.

In fresh-water algae, sexual reproduction is generally correlated with the onset of adverse environmental conditions (drying up of the habitat in summer or freezing in winter). The zygote resulting from gametic union secretes a thick, protective wall, transforming itself into a resting stage (zygospore). At the return of favorable growing conditions, the zygote germinates by undergoing meiosis and giving rise usually to four haploid spores, each of which produces a new thallus. The habitat of marine algae, by contrast, undergoes seasonal changes more gradually, and the zygote usually develops immediately into a diploid somatic phase. Adverse conditions of light and temperature are endured, not by thick-walled spores, but by filamentous or crustose stages that receive protection from predation and physical stress by nestling in crevices.

In green and brown marine algae, the diploid somatic phase is autonomous, but in red algae, which have two successive diploid phases, only the second is autonomous, the first remaining an integral part of the parent gamete-producing plant (gametophyte). The autonomous diploid plant (sporophyte) in turn produces spores by meiosis. The haploid spores (meiospores) then give rise directly to unisexual (dioecious) or bisexual (monoecious) gametophytes, which produce gametes to complete the cycle. This life history, which involves an alternation or sequence of somatic phases correlated with changes in ploidy level, is often inappropriately called alternation of generations. The haploid and diploid somatic phases may be similar to one another (isomorphic) or dissimilar in varying degrees (heteromorphic). Extreme cases are presented by kelps, in which the sporophyte often reaches a length of 100 ft (30 m) while the gametophytes are microscopic.

A few algae (Fucales, some Bryopsidales) have a life history in which there is only one somatic phase (a bisexual plant) or two expressions (male and female) of one somatic phase, which are diploid. The formation of gametes entails meiosis, and the zygote develops directly into a new diploid thallus.

Culture studies supplemented by field studies have revealed that some algae are extremely plastic with regard to life history, being able to bypass one or another phase by recycling (repeating), apomeiosis (short-circuiting of sexual reproduction), pedogenesis, parthenogenesis, and vegetative multiplication, and to produce special microscopic (usually filamentous) stages in response to seasonal, long-term, or sporadic changes in the environment. Some species of benthic marine algae have different life histories in different parts of their geographic range, determined probably by temperature and light regimes.

Metabolism. Most algae are autotrophic, obtaining energy and carbon through photosynthesis. All photosynthetic algae liberate oxygen and use chlorophyll *a* as the primary photosynthetic pigment. Secondary (accessory) photosynthetic pigments, which capture light energy and transfer it to chlorophyll *a*, include chlorophyll *b* (Prochlorophyceae, Euglenophyceae, Chlorophycota), chlorophyll *c* (Chromophycota), fucoxanthin among other xanthophylls (Chromophycota), and phycobiliproteins (Cyanophyceae, Rhodophyceae, Cryptophyceae). Other carotenoids, especially β-carotene, protect the photosynthetic pigments from oxidative bleaching. Except for different complements of accessory pigments (resulting in different action spectra), photosynthesis in algae is identical to that in higher plants. Carbon is predominantly fixed through the C_3 pathway. *See* CAROTENOID; CHLOROPHYLL.

Obligately or facultatively heterotrophic algae occur in most classes. These algae are parasites (on other algae, higher plants, and a broad spectrum of animals including humans) or free-living osmotrophs or phagotrophs.

The chemical products of photosynthesis in most algae are stored as one of two polymers of glucose—starch similar to that in higher plants, in which the glucose units have α-1,4-linkages, or laminaran, chrysolaminaran, and paramylon, with β-1,3-linkages. Other storage products (sugars, alcohols, and lipids) are less common, but may be abundant in certain taxonomic groups. Phosphorus in the form of polyphosphate and nitrogen in the form of proteins or peptides are stored by some algae.

The source of carbon for most photosynthetic algae is carbon dioxide (CO_2), but some can use bicarbonate. Many photosynthetic algae are also able to use organic substances (such as hexose sugars and fatty acids) and thus can grow in the dark or in the absence of CO_2. Colorless algae obtain both energy and carbon from a wide variety of organic compounds in a process called oxidative assimilation.

Most algae obtain nitrogen for protein synthesis from nitrate or ammonium, but some unicells can use organic sources such as urea or amino acids. Some blue-green algae—with few exceptions only those with heterocysts—can meet their requirements by fixing atmospheric nitrogen. *See* NITROGEN FIXATION.

Algae, like higher plants, require a certain complement of elements for successful growth and reproduction. In addition to the elements required by higher plants, at least some algae require vanadium, silicon, iodine, or sodium. Unlike higher plants, many algae need an external source of certain vitamins (vitamin B_{12}, thiamin, and biotin, singly or in combination). These vitamins are usually available in natural waters as a result of the metabolic activity of bacteria, but they may be depleted during phytoplankton blooms. *See* PLANT METABOLISM.

Numerous substances are liberated into water by living algae, often with marked ecological effects.

These extracellular products include simple sugars and sugar alcohols, wall polysaccharides, glycolic acid, phenolic substances, and aromatic compounds. Some secreted substances inhibit the growth of other algae and even that of the secreting alga. Some are toxic to fishes and terrestrial animals that drink the water. Nitrogen-fixing blue-green algae release a large proportion of their fixed nitrogen into the water. As part of the plankton tug-of-war, some algae secrete vitamins while others secrete vitamin-inactivating substances. *See* ALLELOPATHY.

Occurrence. Algae are predominantly aquatic, inhabiting fresh, brackish, and marine waters without respect to size or degree of permanence of the habitat. They may be planktonic (free-floating or motile) or benthic (attached). Benthic marine algae are commonly called seaweeds. Substrates include rocks (outcrops, boulders, cobbles, pebbles), plants (including other algae), animals, boat bottoms, piers, debris, and less frequently sand and mud. Some species occur on a wide variety of living organisms, suggesting that the hosts are providing only space. Many species, however, have a restricted range of hosts and have been shown to be (or are suspected of being) at least partially parasitic. Certain algae parasitize fishes while various unicells are characteristic symbionts of protozoa, coelenterates, mollusks, and flatworms. All reef-building corals contain dinoflagellates, without which their calcification ability is greatly reduced. Different phases in a life history may have different substrate preferences. The depth at which algae can grow is determined by the availability of light (water transparency). In the ocean, transparency is affected by latitude (angle of incident sunlight), distance from terrestrial runoff, and physical nature of the bottom. In exceptionally clear areas, such as the Mediterranean and off the east coast of Florida, algae grow at depths of at least 330 ft (100 m). *See* PHYTOPLANKTON.

Many fresh-water algae have become adapted to a nonaquatic habitat, living on moist soil, masonry and wooden structures, and trees. A few parasitize higher plants (expecially in the tropics), producing diseases in such crops as tea, coffee, and citrus. There are several specialized habitats. In deserts, algae (chiefly blue-greens) live on the undersurface and in cracks of rocks. Representatives of several classes (mostly unicells) grow on snow and ice, which they tint green or red depending upon the amount of protective reddish-orange pigment present in the cells. Thermophilic algae (again, chiefly blue-greens) live in hot springs at temperatures up to 163°F (73°C), forming a calcareous deposit known as tufa. *See* TUFA.

One of the most remarkable adaptations of certain algae (blue-greens and greens) is their coevolution with fungi to form a compound organism, the lichen. Most lichens are subaerial, being especially common on the bark and dead twigs of trees and on rocks exposed to moisture-laden winds, but a few are marine. *See* LICHENS.

Geographic distribution. Fresh-water algae, which are distributed by spores or fragments borne by the wind or by birds, tend to be widespread if not cosmopolitan, their distribution being limited by the availability of suitable habitats. Certain species, however, are characteristic of one or another general climatic zone, such as cold-temperate regions or the tropics. Marine algae, which are spread chiefly by water-borne propagules or reproductive cells, often have distinctive geographic patterns. Along temperate shores, individual species tend to have fairly limited ranges, the end points being determined primarily by water temperature. Thus, the marine floras of Japan, Pacific North America, South Africa, southern Australia, and New Zealand all have a high proportion of endemic species. Distributional patterns of marine algae are correlated with patterns of general oceanic circulation, the main currents not only providing a means and determining the direction of transport of propagules but also controlling the temperature. Major floristic provinces include the North Pacific (Japan eastward through Baja California, Mexico, established by the Kuroshio Current System), the North Atlantic (Cape Hatteras, North Carolina, northeastward through Atlantic Europe, established by the Gulf Stream System), the tropical Atlantic (established by the Atlantic Equatorial Current System), the Arctic (related to the Arctic Ocean), the Antarctic/Subantarctic (related to the Antarctic Convergence and the circumpolar current), and the Indo-Pacific (transcending modern current systems and related to ancient positions of land masses). The marine algal flora at any given latitude on the Pacific coast of America is at least 90% different from that at the same latitude on the Atlantic coast. *See* POPULATION DISPERSAL; POPULATION DISPERSION.

Many taxonomic groups are widely distributed, but others are characteristic of particular climatic zones or geographic areas. The large brown seaweeds called kelps are mostly confined to cold waters of the Northern Hemisphere, although certain genera occur only in the Southern Hemisphere. *Sargassum*, also a brown alga, is extremely common in warm waters everywhere. Coenocytic green algae, many of which have calcareous thalli, are characteristic of all tropical areas except the eastern Pacific. Crustose coralline algae are common in all oceans, but individual genera have distribution patterns related to climatic zones and hemispheres. There is a large pantropical floristic element comprising red, brown, and green seaweeds. *See* PLANT GEOGRAPHY.

Vertical distribution. On coasts that are subjected to tidal fluctuation and offer the protection of regular fog or cloud cover, an intertidal belt of seaweeds may be found. When this belt is at least 3 ft (1 m) in breadth, it is often subdivided into horizontal bands that are determined by the ability of particular species to withstand desiccation and wave exposure and by the availability of suitable substrate. Intertidal zonation of seaweeds is correlated with that of invertebrates. Much less is known about patterns of vertical subtidal zonation, where the determining factors appear to be water movement (surge and currents), light intensity, substrate, and predation. Thus, some species grow only at relatively great depths or

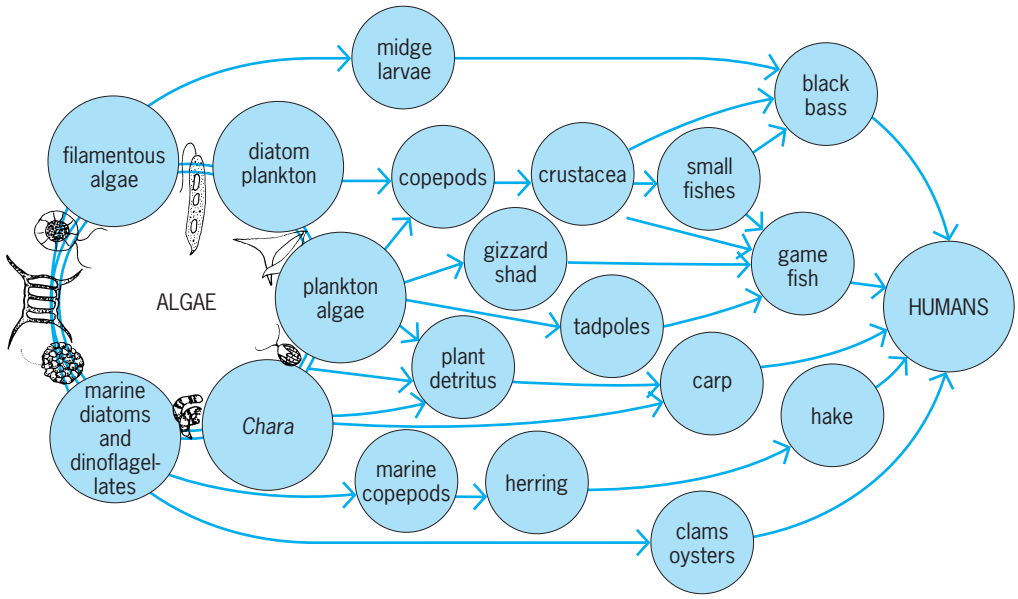


Fig. 2. Food chains from algae to humans. (*After H. C. Sampson et al., eds., Textbook of Botany, 3d ed., Harper, 1966*)

only on vertical rock faces or only on mollusk shells on quiet sandy bottoms. Algae growing on shores of lakes with wave action (such as Lake Michigan) exhibit small-scale vertical zonation.

Economic importance. Planktonic algae, as the primary producers in oceans and lakes, support the entire aquatic trophic pyramid and thus are the basis of the fisheries industry (**Fig. 2**). Concomitantly, their production of oxygen counteracts its uptake in animal respiration. The ability of certain planktonic algae to assimilate organic nutrients makes them important in the treatment of sewage. *See* FOOD WEB; SEWAGE TREATMENT.

Numerous red, brown, and green seaweeds as well as a few species of fresh-water algae are consumed by the peoples of eastern Asia (especially Japan), Indonesia, Polynesia, and the North Atlantic. *Porphyra* (nori), *Laminaria* (kombu), and *Monostroma* (aonori) are produced by intensive mariculture in Japan and China. Large brown seaweeds may be chopped and added to poultry and livestock feed or applied whole as fertilizer for crop plants. The purified cell-wall polysaccharides of brown and red algae (alginate, agar, carrageenan) are used as gelling, suspending, and emulsifying agents in numerous industries, including food processing, cosmetics, pharmaceuticals, textiles, paper, and printing, as well as in dentistry, medicine, and biological research. Some seaweeds have specific medicinal properties, such as effectiveness against worms. The blue-green alga *Spirulina*, a traditional food in parts of Mexico and central Africa, is being grown commercially and marketed as a high-protein dietary supplement. Nitrogen-fixing blue-green algae growing in rice paddies provide significant nitrogen enrichment. Petroleum is generally believed to result from bacterial degradation of organic matter derived primarily from planktonic algae. Diatomaceous earth and diatomite (fossil diatom deposits) have many industrial applications. Unicellular algae, because of the ease with which they can be grown in pure culture and thus provide quantities sufficient for experimentation, have been used traditionally in certain biochemical, biophysical, and physiological studies. *Chlamydomonas, Chlorella, Ankistrodesmus*, and *Euglena* have thus been the subject of thousands of investigations, yielding fundamental biological information. *See* AGAR; ALGINATE.

On the negative side, algae can be a nuisance by imparting tastes and odors to drinking water, clogging filters, and making swimming pools, lakes, and beaches unattractive. Sudden growths (blooms) of planktonic algae can produce toxins of varying potency. In small bodies of fresh water, the toxin (usually from blue-green algae) can kill fishes and livestock that drink the water. In the ocean, toxins produced by dinoflagellate blooms (red tides) can kill fishes and render shellfish poisonous to humans. Ciguatera poisoning in tropical areas has been traced to toxin produced by benthic dinoflagellates and accumulated in reef-dwelling fishes. Aquarium fishes may die from an infection by a parasitic dinoflagellate, while humans may suffer from protothecosis, a gastrointestinal disease caused by a colorless chlorophycean unicell (*Prototheca*). In warm humid regions, leaves of crop plants (tea, coffee, citrus) may be badly damaged by a parasitic green alga. A few marine algae have spread far beyond their original home and become noxious weeds, damaging shellfish plantings, clogging the propellers of small boats, and contributing a large biomass to odoriferous beach wrack. *See* EUTROPHICATION; TOXIN.

Fossil algae. At least half of the classes of algae are represented in the fossil record, usually abundantly, in the form of siliceous, calcareous, or organic remains, impressions, or indications. Blue-green algae were among the first inhabitants of the Earth, appearing in rocks at least as old as 2.3 billion years.

Their predominance in shallow Precambrian seas is indicated by the extensive development of stromatolites, laminated reeflike structures produced by algal mats that entrap detrital sediments and sometimes deposit calcium carbonate. Modern stromatolites (as at Shark Bay, Western Australia) implicate blue-green algae, although bacteria may have been at least partly responsible for some of the ancient formations. *See* STROMATOLITE.

All three classes of seaweeds (reds, browns, and greens) were well established by the close of the Precambrian, 600 million years ago (mya). Shallow-water reef deposits were formed in various geological periods by calcareous red and green seaweeds, and the process continues today. The chlorophycean order Dasycladales, in which the entire thallus may be calcified, has left a rich record of more than 150 genera dating from the Precambrian. The calcified zygospores (gyrogonites) of Charophyceae are found in fresh-water deposits as old as the Devonian (400 mya). The relatively low diversity among living dasyclads and charophytes clearly indicates that they are relicts. Among red algae, coralline algae date from the Devonian, while two frequently encountered groups (Solenoporaceae and Gymnocodiaceae) are known only as fossils. *See* GEOLOGIC TIME SCALE.

By far the greatest number of fossil taxa belong to classes whose members are wholly or in large part planktonic. Siliceous frustules of diatoms and endoskeletons of silicoflagellates, calcareous scales of coccolithophorids, and highly resistant organic cysts of dinoflagellates contribute slowly but steadily to sediments blanketing ocean floors, as they have for tens of millions of years. Cores obtained in the Deep Sea Drilling Project have revealed an astounding chronology of the appearance, rise, decline, and extinction of a succession of species and genera. From this chronology, much can be deduced about the climate, hydrography, and ecology of particular geological periods. More than 350 genera of dinoflagellates and an even greater number of genera of coccolithophorids (Prymnesiophyceae) have been described from strata younger than the Permian (about 250 mya). Centric diatoms appear in the Cretaceous (about 100 mya), with pennate diatoms apparently evolving in the Paleocene (65 mya). Because of their abundance, planktonic algae have been important rock builders in shallow seas. Mesozoic coccolithophorids resulted in the accumulation of extensive chalks and limestones (such as the white cliffs of Dover) and there are massive Tertiary deposits of diatoms (diatomaceous earth or diatomite). *See* PALEOBOTANY.

Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae*: *Structure and Reproduction*, 1978; V. J. Chapman, *Seaweeds and Their Uses*, 1970; J. D. Dodge, *The Fine Structure of Algal Cells*, 1973; E. Gantt (ed.), *Handbook of Phycological Methods*: *Developmental and Cytological Methods*, 1980; J. A. Hellebust and J. S. Craigie (ed.), *Handbook of Phycological Methods*: *Physiological and Biochemical Methods*, 1978; C. S. Lobban and W. J. Wynne (ed.), *The Biology of Seaweeds*, 1981; F. E. Round, *The Ecology of Algae*, 1981; J. R. Stein (ed.), *Handbook of Phycological Methods*: *Culture Methods and Growth Measurements*, 1973; W. D. P. Stewart (ed.), *Algal Physiology and Biochemistry*, 1974; H. Tappan, *The Paleobiology of Plant Protists*, 1980; O. R. Zaborsky (ed.), *CRC Handbook of Biosolar Resources*, vol. 1, pt. 1, 1982.

Algebra

The branch of mathematics dealing with the solution of equations. These equations involve unknowns, or variables, along with fixed numbers from a specified system. The origins of algebra were based on the need to develop equations that modeled real-world problems. From this came a very extensive theory based on the need to find the values that can be successfully used in the equations.

Number systems. Classical algebra is conducted in one of several number systems. The most basic is the natural numbers, consisting of the counting numbers: 1, 2, 3, 4.... The natural numbers along with their negatives and 0 form the set of integers: ..., −2, −1, 0, 1, 2,.... Any number which can be expressed as a quotient a/b of two integers a and b is called a rational number. If the decimal expansion of a number either is repeating or terminates in a string of zeros, it is always possible to find integers a and b such that the quotient a/b gives the original number. Numbers where the decimal expansion cannot be written as a quotient of integers are the so-called irrational numbers. Early mathematicians did not recognize the existence of such numbers until geometric considerations showed that they must exist as the lengths of sides of right triangles.

The sets of rational numbers and irrational numbers have no elements in common, and together they make up the real-number system. It is within the system of real numbers that algebraic tasks are most often performed. However, there is a still larger set of numbers, known as the complex numbers, which serve much more completely in solving equations. The complex numbers consist of numbers of the form $a + bi$, where i is a symbol representing $\sqrt{-1}$. The symbol i is commonly called the imaginary root of −1, and the numbers of the form bi are known as the set of imaginary numbers. In a complex number $a + bi$, the real number a is called the real part while the real number b is the imaginary part. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

Real numbers are often represented on a number line (**Fig. 1***a*). In this method of depiction, there is a one-to-one correspondence between the points on the line and the real numbers. There is no such representation of the complex numbers since the number line implies an ordering notion which does not easily extend beyond the real numbers. However, the complex numbers can be represented on a

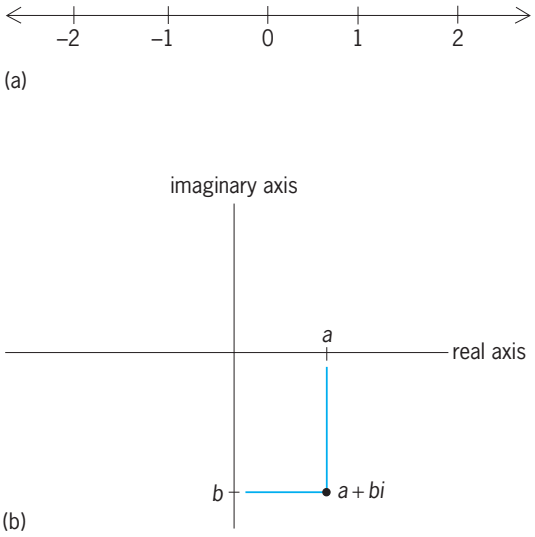


Fig. 1. Representations of number systems. (*a*) Of real numbers on a number line. (*b*) Of complex numbers on a two-dimensional plane.

two-dimensional plane (Fig. 1*b*), by plotting the real part on the horizontal axis and the imaginary part on the vertical axis.

Operations. Within any number system there is a set of operations which can be performed on some or all of the members of the system. Operations take a pair of numbers in the system and produce a single new number from this pair. The most standard operations are addition (+), multiplication (·), subtraction (−), and division (÷). A set is closed under an operation if the application of the operation to any pair of numbers in the set results in another number in the same set. For example, the set of natural numbers is closed under + but not under − (since, for example, $1 - 2 = -1$ is not a natural number). The rational numbers, the real numbers, and the complex numbers are closed under all four of the basic operations (except in the case where division by 0 may occur). *See* ARITHMETIC.

Other operations arise from the four basic ones. Exponentiation is derived from the process of repeatedly multiplying a number by itself. If n is a positive integer, the symbol a^n indicates that the number a is multiplied by itself n times. If $n > 0$ is an integer, then a^{-n} refers to the quotient $1/a^n$. Fractional exponents can also be considered by defining $a^{1/n}$ to be the number b such that $b^n = a$. This notion is further extended by letting $a^{m/n} = (a^{1/n})^m$. For real numbers, some fractional exponents may not exist. For example, $(-1)^{1/2}$ is not a real number. The complex numbers, however, are closed with respect to exponentiation.

Algebraic expressions. In algebra, symbols are frequently used to designate unknown values. For example, letters such as x and y can stand for any one of a set of numbers satisfying certain conditions. These symbols are combined with numbers and the basic operations to form algebraic expressions such as (1).

$$3x^2 + 5y^3x - 6 \qquad \frac{(x^5 - 1)^{3xy}}{y^2 + 2^x} \tag{1}$$

The symbols are usually called variables, while the numbers (or constants) multiplied by them are the coefficients of the expression. The coefficients are all assumed to come from a designated number system. Algebraic expressions take on specific values when each of the variables involved is assigned a numerical value.

Algebraic equations. A statement is a relationship between several algebraic expressions. This relationship can be equality (=) or an inequality (< or >), but in any case it puts restrictions on the values of the variables. For example, each of expressions (2)

$$x^2 + 3xy = y - 3 \qquad 2xy > y^5 - 1 \tag{2}$$

represents a relationship between the variables x and y, while Eq. (3) relates the three variables, x, y, and z.

$$4z - 2xy = 5 + yz^2 \tag{3}$$

If, when specific values from an appropriate number system are substituted for the variables in an equation, the resulting statement is true, then these numbers are said to be a solution to the equation. The solutions of a particular equation are dependent on the number system in use. For instance, the equation $x^4 = 2$ has no solutions in the set of rational numbers, two solutions ($x = 2^{1/4}$ and $x = -2^{1/4}$) in the real numbers, and four solutions in the complex field.

Algebraic equations are used to describe many phenomena in the real world. Their application runs through almost all disciplines in the sciences, social sciences, and business. Once an equation is established as a good model of some situation, it is necessary to determine the solutions to the equation, as these represent instances when the conditions described by the equation are actually met.

One of the most basic kinds of equations is the linear equation. If such an equation involves the variables $x_0, x_1, \ldots, x_n$, then it can be written in the form of Eq. (4), where $a_0, a_1, \ldots, a_n, b$ are coefficients

$$a_0x_0 + a_1x_1 + \cdots + a_nx_n = b \tag{4}$$

from some number system.

In the simplest form, where only two variables, x and y, are involved, a linear equation represents a straight line in two dimensions. The equation of such a line can be written in the form $y = mx + b$, where m is the slope of the line (the rate of change of the y variable with respect to the x variable) and b is the y intercept, or the y value corresponding to $x = 0$.

Another important algebraic equation is the quadratic equation, which has the standard form $y = ax^2 + bx + c$ for coefficients a, b, c. The values of x that produce the value $y = 0$ are found through the quadratic formula, Eq. (5). These x values are known

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{5}$$

as roots. In the complex number system, there are always either one or two roots to a quadratic equation. However, in the real number system, the number of roots of a particular quadratic equation is

given by the discriminant, $b^2 - 4ac$. There are two real roots to a quadratic equation if the discriminant is positive, one real root if it is 0, and no real roots if it is negative. For example, the equation $y = 3x^2 - 2x + 1$ has the discriminant given by Eq. (6), and

$$b^2 - 4ac = (-2)^2 - 4(3)(1) = -8 \qquad (6)$$

hence has no real roots. The complex roots are x_1 and x_2 given by Eqs. (7). The ability to always find

$$x_1 = \frac{-(-2) + \sqrt{(-2)^2 - 4(3)(1)}}{2(3)} = \frac{1 + \sqrt{2}i}{3} \qquad (7a)$$

$$x_2 = \frac{-(-2) - \sqrt{(-2)^2 - 4(3)(1)}}{2(3)} = \frac{1 - \sqrt{2}i}{3} \qquad (7b)$$

roots of quadratic equations (as well as much more complicated ones) in the complex numbers is an indicator of the usefulness of this number system.

Functions. Some of the most fundamental algebraic equations are those involving two variables. When one of these variables, say y, can be expressed in terms of the other variable, say x, in such a way that each value of x produces exactly one value of y, then the relationship described by the equation is known as a function. More general relationships which may or may not produce unique values of y from each value of x are called relations. In a function such as has been described, the variable x is called the independent variable while the resulting variable y is called the dependent variable. The possible values that can be used by the independent variable make up the domain of the function (or, more generally, the relation). The resulting values of the dependent variable compose the range of the function.

It is most common and quite practical to write a variable y which is dependent on a variable x in the form $y = f(x)$. This notation, known as function notation, indicates that every y value is determined by an x value. An equation such as $y = x^2 + 3$ would be written in functional notation as $f(x) = x^2 + 3$. If $x = a$, then the corresponding value of y is given by $y = f(a)$. In the particular function $f(x) = x^2 + 3$, the value of y associated with $x = -1$ is $y = f(-1) = (-1)^2 + 3 = 4$.

Suppose an equation involves the variables x and y. It is said that the ordered pair (a, b) is part of the function (or relation) relating x and y whenever the value $x = a$ produces the value $y = b$. The ordered pair $(-1, 4)$ is part of the function $f(x) = x^2 + 3$, since $f(-1) = 4$. Equations involving three variables, say x, y, z, produce ordered triples (a, b, c) describing corresponding values of the variables. This process can readily be extended so that an equation with n variables would give rise to a set of ordered "n-tuples" where the terms for the corresponding variables can be substituted successfully into the equation.

Graphs. Relations, and more specifically functions, are frequently represented effectively with graphs. If an equation involves the variables x and y, then a two-dimensional coordinate system can be used to depict the values that correspond in the equation. If the ordered pair (a, b) is given by the equation, then it can be graphically depicted (**Fig. 2***a*). All of the ordered pairs given by the equation combine to give a complete graphic representation of the relationship described by the equation (Fig. 2*b*).

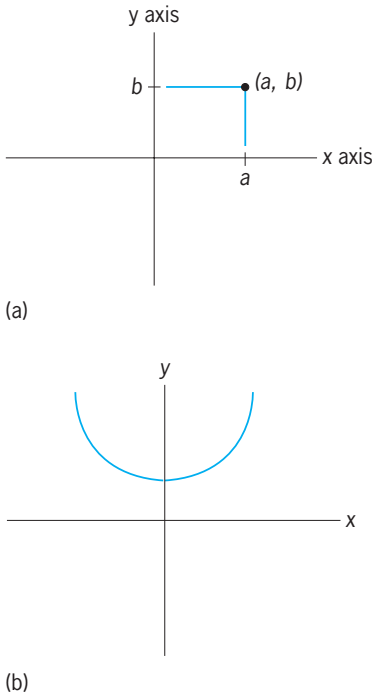


Fig. 2. Representation of relations by graphs. (*a*) Graphic depiction of an ordered pair (*a, b*) given by an equation. (*b*) Graph of the function $f(x) = x^2 + 3$.

Functions are readily distinguished among more general relations by their graphical representations. Since each x value from the domain of a function produces only one y value in the range, and since all of the points on a vertical line have the same x value, it must be the case that each vertical line intersects the graph of a function in at most one place. For example, from the vertical line test it is readily seen that the relation described in Eq. (8) is a function

$$y = 2x^2 - 4x + 5 \qquad (8)$$

(**Fig. 3***a*) while the relation described by Eq. (9) is not (Fig. 3*b*).

$$y^2 = x - 1 \qquad (9)$$

One of the most important properties of a function is given by the places where the dependent variable takes on the values 0. These correspond to the points where the graph crosses, or touches, the horizontal or x axis. The values of x for which $f(x) = 0$ are called zeros (or roots) of the function. *See* ANALYTIC GEOMETRY; GRAPHIC METHODS.

Polynomials. There are many special functions which are commonly used to describe real situations. One of the most basic is the polynomial, which is a function of the form (10). Each coefficient a_0,

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \qquad (10)$$

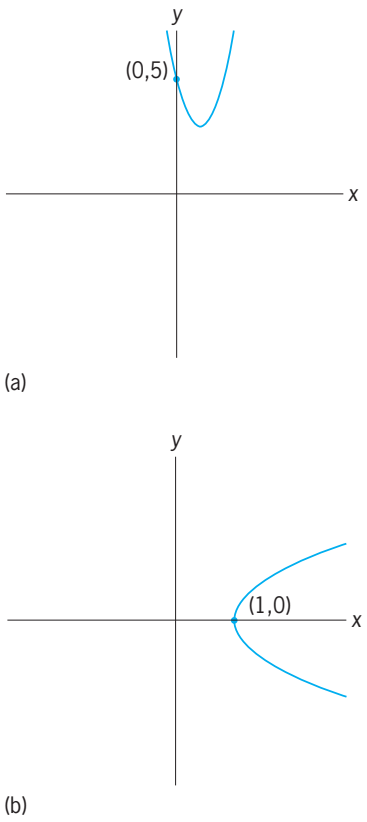


Fig. 3. Graphs of the relations described by the equations (*a*) $y = 2x^2 - 4x + 5$ and (*b*) $y^2 = x - 1$. The vertical-line test shows that the first relation is a function while the second is not.

$a_1, \ldots, a_n$ is a real number, and the exponents of the variables are nonnegative integers. The largest exponent of x is called the degree of the polynomial. When graphed, a polynomial of degree n will reverse directions no more than $n - 1$ times. For example, the graph of the degree-4 polynomial in Eq. (11) [**Fig. 4**] changes direction three times.

$$f(x) = 3x^4 - 4x^3 - 12x^2 + 8 \qquad (11)$$

A polynomial of degree n will have no more than n roots in the real number system. Sometimes it may have fewer. For example, the function $f(x) = x^2 + 1$ has no real roots, while the function $f(x) = x^2 - 1$ has the roots $x = -1$ and $x = 1$. The function $f(x) = x^2$ has the single root $x = 0$. Whenever $x = a$ is a root of a polynomial $f(x)$ of degree n, then the polynomial can be factored as $f(x) = (x - a)\, g\,(x)$, where $g\,(x)$ is a polynomial of degree $n - 1$. If the term $(x - a)$ can be successively factored from a polynomial k times, then $x = a$ is said to be a root of multiplicity k. The equation $y = (x - 1)(x - 1)$ has only one root, $x = 1$, but this root is of multiplicity 2.

It is in the consideration of roots of an equation that the complex number system becomes particularly important. If a polynomial is formed with coefficients from the complex number system, then the sum of the multiplicities of the roots (from the complex numbers) of the polynomial coincides with the degree of the polynomial. Consequently, a complex number polynomial $f(x)$ can be factored completely as $f(x) = (x - a_1)(x - a_2) \ldots (x - a_n)$, where $a_1, a_2, \ldots, a_n$ are the complex number roots and n is the degree of the polynomial. This is not always possible for real polynomials. For example, the polynomial $f(x) = x^2 + 1$ cannot be factored further by using polynomials with real coefficients. It is called irreducible over the real numbers. As a complex polynomial, it can be factored as $x^2 + 1 = (x + i) \cdot (x - i)$.

Factoring of polynomials. Even in the real numbers, a polynomial often may be factored into other polynomials of lesser degree. This allows for analysis of the polynomial by consideration of the simpler factors. Some of the most common and useful factoring patterns are given in Eqs. (12).

$$x^2 - a^2 = (x - a)(x + a) \qquad (12a)$$

$$x^2 + 2ax + a^2 = (x + a)(x + a) \qquad (12b)$$

$$x^2 - 2ax + a^2 = (x - a)(x - a) \qquad (12c)$$

$$x^3 + 3ax^2 + 3a^2x + a^3 = (x + a)(x + a)(x + a) \qquad (12d)$$

$$x^3 + 3ax^2 - 3a^2x - a^3 = (x - a)(x - a)(x - a) \qquad (12e)$$

$$x^3 - a^3 = (x - a)(x^2 + ax + a^2) \qquad (12f)$$

$$x^3 + a^3 = (x + a)(x^2 - ax + a^2) \qquad (12g)$$

Rational functions. Functions which are of the form $f(x) = g(x)/h(x)$, where $g(x)$ and $h(x)$ are polynomials, are called rational functions. They frequently appear in instances where comparisons of two polynomials are necessary. A rational function is defined only at points where the denominator is nonzero. The roots of a rational function agree with

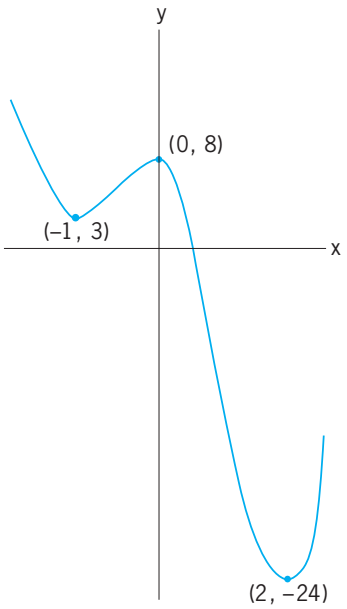


Fig. 4. Graph of the degree-4 polynomial $f(x) = 3x^4 - 4x^3 - 12x^2 + 8$, which changes direction three times.

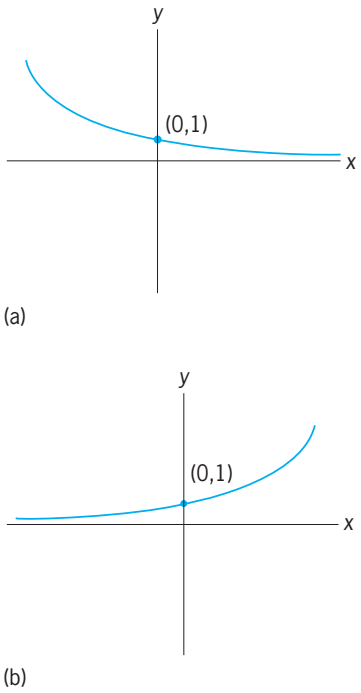


Fig. 5. Graphs of exponential functions. (*a*) Decreasing graph of $y = (\frac{1}{2})^x$. (*b*) Increasing graph of $y = 2^x$.

the roots of the polynomial that is found in the numerator.

Exponential functions. In general, an exponential function is one which has the form $f(x) = b^x$ for some constant b in the number system. The graphs of these functions take on one of two basic types. The graph is decreasing if $0 < b < 1$ (**Fig. 5***a*), and it is increasing if $b > 1$ (Fig. 5*b*). The case of $b = 1$ is trivial, since all values of the function are 1.

Exponential functions are of special importance in describing a situation where the rate of growth is proportional to the amount present. Among the situations that can be described by exponential functions are those involving the calculation of interest, radioactive decay, and population growth.

Calculation with exponential functions is based on the basic equalities given in Eqs. (13).

$$b^{x+y} = b^x b^y \tag{13a}$$

$$b^{x-y} = \frac{b^x}{b^y} \tag{13b}$$

$$(ab)^x = a^x b^x \tag{13c}$$

$$(b^x)^y = b^{xy} \tag{13d}$$

Logarithmic functions. Another function that plays an especially important role in algebra is the logarithm. In general, if $b > 1$, the logarithm to the base b of a value x is denoted by Eq. (14). The logarithm

$$y = \log_b x \tag{14}$$

function is defined in terms of exponentials by reversing the roles of the x and y variables. It follows that if $y = \log_b$, then y is the exponent which when applied to b produces x, as in Eqs. (15). The logarithm function, like the exponential function, has a characteristic graph (**Fig. 6**).

$$y = \log_b x \leftrightarrow x = b^y \tag{15}$$

Basic rules that are necessary for calculation with logarithms are given in Eqs. (16). For example, since

$$\log_b(xy) = \log_b x + \log_b y \tag{16a}$$

$$\log_b\left(\frac{x}{y}\right) = \log_b x - \log_b y \tag{16b}$$

$$\log_b x^r = r(\log_b x) \tag{16c}$$

$$\log_b b^x = x \tag{16d}$$

$$b^{\log_b x} = x \tag{16e}$$

logarithms are exponents, the exponent for the product in Eq. (16*a*) is the sum of the exponents for the factors. *See* LOGARITHM.

Systems of linear equations. Frequently, a situation is described by a system of linear equations which must all be satisfied simultaneously. For example, the system of Eqs. (17) involves three variables and three equations.

$$\begin{aligned} 2x + 3y - 4z &= 2 \\ 4x - 5y - z &= 0 \\ -3x + y - 2z &= 1 \end{aligned} \tag{17}$$

More general systems involve n variables and m equations and take the form of Eqs. (18). A solution

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{18}$$

to the system is a set of numbers $c_i, c_2, \ldots, c_n$ which satisfies each equation.

The equations in a linear system can be multiplied by constants, where a single number is multiplied by each term in the sum. They can also be added together, where the coefficients of the corresponding variables are summed. A linear combination of one or more equations in the system is formed by multiplying each equation by a constant and then

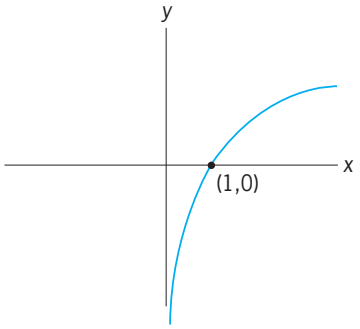


Fig. 6. Graph of the logarithm function $y = \log_b x$.

summing the resulting equations. When one equation in a system is a linear combination of one or more of the other equations, the system is called dependent. The solutions to a dependent system are the same as the solutions to the system formed by deleting the equation that is a linear combination of other equations in the system. In order to find solutions to a system, it is frequently expedient to first try to eliminate equations that are linear combinations of others in the system. *See* LINEAR SYSTEMS OF EQUATIONS.

Wayne B. Powell

Bibliography. D. T. Christy, *College Algebra*, 2d ed., 1993; C. H. Edwards and D. Penney, *Calculus and Analytic Geometry*, 4th ed., 1994.

Algebraic geometry

The study of zero sets of polynomial equations. Examples are the parabola $y - x^2 = 0$, thought of as sitting in the (x, y)-plane, and the locus of all points (t, t^2, t^3, t^4), which is defined by Eqs. (1), in the

$$y^2 = xz \qquad yx = wz \qquad wy = x^2 \tag{1}$$

coordinate space with coordinates w, x, y, z. Another interesting example is an elliptic curve, typically defined by an equation like Eq. (2), where A and B are

$$y^2 = x(x - A)(x - B) \tag{2}$$

constants. Objects such as these are called affine algebraic sets. They exist in affine n-space, denoted $\mathbf{A}^n$, which is defined to be the coordinate space with coordinates $x_1, \ldots, x_n$. The coordinate ring $k[V]$ of an affine algebraic set V is the set of functions on V obtained by restriction from polynomial functions on the ambient affine space. These functions can be added, subtracted, and multiplied (that is, they form a ring). For example, the coordinate ring of the parabola has in it functions y and x satisfying $y = x^2$. This relation can be used to eliminate all references to y, so that the coordinate ring of the parabola is identified with the ring of polynomials in x. *See* ANALYTIC GEOMETRY; POLYNOMIAL SYSTEMS OF EQUATIONS; RING THEORY.

The possibility of studying a question geometrically via the zero loci of polynomials or algebraically via the algebra of the coordinate ring gives the subject much of its power and flavor. This has led to an amazing growth in applications to other disciplines. For example, the integers $\ldots, -2, -1, 0, 1, 2, \ldots$ form a ring that is algebraically similar to the coordinate ring of the affine line, and algebrogeometric methods have come to play a central role in number theory. In studying the path space of strings and the partition function, modern physicists have made the moduli space of curves a central object of research. Finally, a number of differential equations in engineering and physics can best be studied algebrogeometrically, although their solutions are not algebraic functions. *See* DIFFERENTIAL EQUATION; NUMBER THEORY; SUPERSTRING THEORY.

Basic numerical invariants. An algebraic set V is irreducible if $fg = 0$ in $k[V]$ implies that either f is 0 or g is 0. For example, the algebraic set defined by $xy = 0$ is not irreducible. An irreducible algebraic set is called an algebraic variety.

Dimension. The procedure of defining the dimension of an algebraic variety V begins with the intuition that affine n-space has dimension n. Geometrically, if V is defined by the vanishing of several polynomials in n variables, so that V is contained in affine n-space, the coordinates on $\mathbf{A}^n$ are changed in what is said to be a general way and the projection given by Eqs. (3) is considered. It turns out that for

$$\mathbf{A}^n \to \mathbf{A}^d \qquad (x_1, \ldots, x_n) \to (x_1, \ldots, x_d) \tag{3}$$

one value of $d \leq n$, this mapping when restricted to V will be surjective with finite fibers, that is, there will exist at least one and at most a finite number of points of V of the form $(a_1, \ldots, a_d, \ldots)$ for given $a_1, \ldots, a_d$. This value of d is the dimension of V. Algebraically, if $k[V]$ denotes the coordinate ring of V, the function field $k(V)$ is considered, consisting of fractions f/g with f and g in $k[V]$ and g not zero. [The word "field" refers to the fact that in $k(V)$ it is possible to add, subtract, multiply, and divide.] There will exist, in $k(V)$, d elements $z_1, \ldots, z_d$ satisfying no polynomial relations among themselves, but having the property that, for any w in $k(V)$, there is a nonzero polynomial $P(x_1, \ldots, x_{d+1})$ with $P(z_1, \ldots, z_d, w) = 0$. *See* FIELD THEORY (MATHEMATICS).

Degree. The degree of an algebraic set is a central notion of enormous depth and power. Intuitively, if V is a variety of dimension d in $\mathbf{A}^n$, the degree of V is the number of points of intersection of V with $\mathbf{A}^{n-d}$. There are a number of pitfalls here. So far, there has been no specification of where the functions take values. If the coordinates are real numbers, then the parabola $y - x^2 = 0$ meets the line $y = 1$ in two points, but it meets the line $y = -1$ in zero points (**Fig. 1**). This problem is solved by taking coordinates in the field $\mathbf{C}$ of complex numbers. (More generally, it is possible to work over any algebraically closed

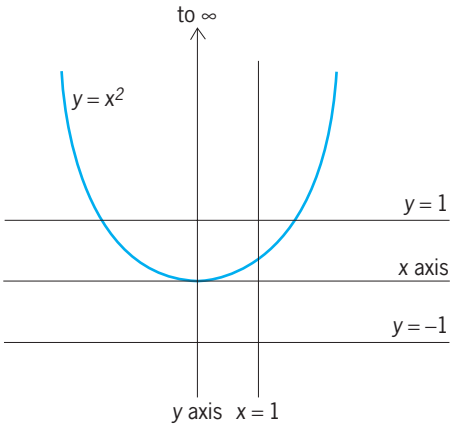


Fig. 1. Parabola in affine 2-space, $\mathbf{A}^2$. While it would seem that lines in $\mathbf{A}^2$ can meet the parabola in 0, 1, or 2 points, with the introduction of algebraically closed fields of coordinates and projective varieties, all such intersections have two points, counting multiplicity.

field. A field F is said to be algebraically closed if any polynomial with coefficients in F has a root in F. The complex numbers have this property.) It can also be argued that the parabola meets the line $y = 0$ in just one point, but this is a point of tangency and therefore must be counted twice. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

A more serious difficulty is how to arrange for the parabola to meet the lines $x =$ constant in two points. A glance at Fig. 1 suggests that one point of intersection should be thought of as lying at ∞ (infinity). The affine (x, y) plane containing the parabola can be imagined positioned as the locus $Z = 1$ inside $\mathbf{A}^3$ with coordinates X, Y, Z. The set of all lines through the origin in $\mathbf{A}^3$ meeting the parabola (**Fig. 2**) is contained in the locus $X^2 - YZ = 0$, but there is one line, defined by $Z = X = 0$, lying in this locus but not meeting the parabola. The lines through the origin meeting the line $x = c$ lie in the plane $X - cZ = 0$ in $\mathbf{A}^3$, and the common zeros, satisfying Eq. (4),

$$X - cZ = 0 = X^2 - YZ \tag{4}$$

consist of two lines, one being the line that joins (0,0,0) with $(c, c^2, 1)$ and the other being the line $Z = X = 0$. This motivates defining projective n-space, $\mathbf{P}^n$, to be the set of all lines through the origin in $\mathbf{A}^{n+1}$. For example, the Riemann sphere $\mathbf{P}^1$ is equivalent to $\mathbf{A}^1$ together with the point ∞ at infinity. Projective algebraic sets in $\mathbf{P}^n$ are defined by the vanishing of homogeneous polynomials in the coordinates on $\mathbf{A}^{n+1}$. For many purposes, it is preferable to work with projective algebraic sets, which can be thought of as completions or compactifications of the affine algebraic sets defined above. In the above example, the collection of lines through the origin lying on the homogeneous conic $X^2 - YZ = 0$, which comprises the lines that meet the parabola $x^2 = y$ together with the line $X = Z = 0$, is the projective completion of the parabola. If V is a projective algebraic variety of dimension n in $\mathbf{P}^N$ and L is a linear space of dimension $N - n$ such that the intersection of V and L is finite, then the number of points of intersection of V and L, counted with multiplicity, is independent of L and is called the degree of V. *See* PROJECTIVE GEOMETRY.

Classification. The problem of classifying projective algebraic varieties typically leads to a finite

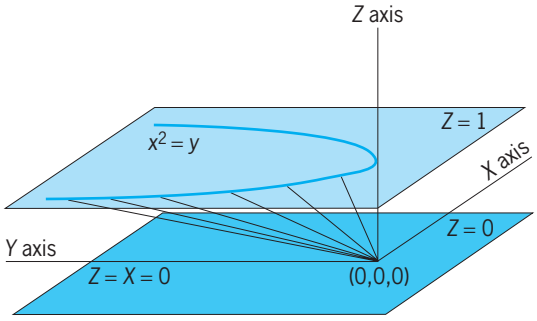


Fig. 2. Projective completion of the parabola of Fig. 1, consisting of lines through the origin lying on the homogeneous conic $X^2 - YZ = 0$, including lines meeting the parabola $x^2 = y$ together with the line $Z = X = 0$.

Fig. 3. Riemann surfaces of low genus. (*a*) Genus 0. (*b*) Genus 1. (*c*) Genus 2.

number of discrete invariants together with a continuous parameter space, the moduli space of the problem, which is itself an algebraic variety. The classification of projective varieties of dimension 1 (algebraic curves) and dimension 2 (algebraic surfaces) will be considered.

Algebraic curves. To help visualize the classification of algebraic curves, the coordinates will be chosen to be complex numbers, and the curves will be assumed to be nonsingular, that is, to have a well-defined tangent at each point. In this case, the algebraic curves are topological surfaces (frequently referred to as Riemann surfaces), and, by topology, they can all be realized as (hollow) doughnuts with $g = 0, 1, 2, \ldots$ holes (**Fig. 3**). The number of holes is the genus of the curve and is the only discrete invariant.

Algebraically, a polynomial in two variables $P(x, y)$ can be written as in Eq. (5). After projective comple-

$$a_0(x)y^n + a_1(x)y^{n-1} + \cdots + a_n(x) = 0 \tag{5}$$

tion and elimination of singular points (by a process called normalization), the locus of zeros of P yields an algebraic curve C. For a general value of x, there are n solutions for y, and therefore C can be visualized as an n-sheeted covering of the projective line $\mathbf{P}^1$. The function y has poles on C, but it is single-valued and satisfies the given polynomial relation. Simple, but by no means trivial, examples are the elliptic and hyperelliptic equation (6). Values of x (possibly including

$$y^2 = f(x) \tag{6}$$

$x = \infty$) for which there are fewer than n solutions in y are called branch points. For a general polynomial $P(x, y)$ there will be $n - 1$ solutions in y over branch points. In this case, the number b of branch points is even, and the genus of C is given by the Hurwitz formula (7).

$$g = 1 - n + (b/2) \tag{7}$$

See CONFORMAL MAPPING; TOPOLOGY.

Curves of a given genus form a continuous family parametrized by the moduli space $\mathbf{M}_g$. The only curve of genus zero is the Riemann sphere, and therefore $\mathbf{M}_g$ is a point. Curves of genus 1 (elliptic curves) are complex tori, which can be viewed as quotients of the complex plane modulo a lattice L_z (called the period lattice) consisting of elements of the form $m + nz$ for fixed complex number z and all integers m and n. Scaling the lattice by a complex factor does not change the quotient curve, and therefore, if a, b, c, and d are integers satisfying Eq. (8) and

w is given by Eq. (9), then L_w and L_z give the same curve. [Indeed, L_z is related to L_w by Eq. (10).] Al-

$$ad - bc = \pm 1 \tag{8}$$

$$w = (az + b)(cz + d)^{-1} \tag{9}$$

$$L_z = (cz + d)L_w \tag{10}$$

gebraic functions on $\mathbf{M}_1$ are meromorphic functions $f(z)$ defined for $z = x + y\sqrt{-1}$ and $y \neq 0$, satisfying $f(z) = f(w)$ with any w as above, and not growing too rapidly as $y \to \infty$. These coincide with rational functions $P[j(z)]/Q[j(z)]$ in the classical j-function given by Eq. (11). In fact, $\mathbf{M}_1 = \mathbf{A}^1$ with parameter

$$j(z) = q^{-1}(1 + 744q + 196884q^2 + \cdots) \qquad q = e^{2\pi iz} \tag{11}$$

j. For $g > 1$, $\mathbf{M}_g$ is an algebraic variety of dimension $3g - 3$. Although not complete, it may be completed to a projective variety by admitting points corresponding to certain singular curves (stable curves).

A divisor D on a curve C is a formal sum $n_1p_1 + n_2p_2 + \cdots + n_rp_r$ with n_i an integer and each p_i a point of C. The degree of D equals $n_1 + n_2 + \cdots + n_r$. To an algebraic function f on C is associated a divisor (f) of degree 0 by associating to each singular point its multiplicity as a zero or pole. By definition, the jacobian $J(C)$ is the group of divisors of degree 0 modulo the subgroup of divisors (f) associated to functions. The jacobian $J(C)$ is a g-dimensional complex torus, whose period lattice is given by integrating a basis of g algebraic differential one-forms over closed loops. In fact, $J(C)$ is a projective algebraic variety (called a principally polarized abelian variety), and the classical Torelli theorem states that the association $C \to J(C)$ embeds $\mathbf{M}_g$ into the Siegel moduli space $\mathbf{A}_g$ of principally polarized abelian varieties of dimension g. The Schottky problem of identifying the image of $\mathbf{M}_g$ in $\mathbf{A}_g$ has been solved.

Algebraic surfaces. The classification of projective varieties of dimension 2 (algebraic surfaces) hinges on a careful analysis of line bundles and linear systems. To map a variety V to $\mathbf{P}^n$ means to associate to each point of V a line through the origin in $\mathbf{A}^{n+1}$, but a modern algebraic geometer would take a dual view and think of the space of linear functions on the line. The collection of these dual lines associated to points of V forms a line bundle. A linear function f on $\mathbf{A}^{n+1}$ induces an element f_v in the line over v in V, and hence a section of the line bundle. The set of v where the f_v vanishes is a Cartier divisor on V. A two-dimensional subspace of linear functions gives rise to a linearly varying family of Cartier divisors called a pencil on V.

This correspondence between line bundles, linear systems, and Cartier divisors lies at the heart of the classification of algebraic surfaces. The key difficulty in calculating the dimension of the space of all sections of a given line bundle is usually overcome via the Riemann-Roch theorem. The space of sections of line bundle L is denoted $H^0(V, L)$, but there are also defined higher cohomology groups $H^1(V, L)$ and $H^2(V, L)$. These are finite-dimensional vector spaces. The Riemann-Roch theorem calculates the Euler-Poincaré characteristic given by Eq. (12).

$$\chi(L) = \text{dimension of } H^0(V, L) - \text{dimension of } H^1(V, L) + \text{dimension of } H^2(V, L) \tag{12}$$

Among the interesting kinds of surfaces are (1) rational surfaces, whose function field $k(V)$ is generated by two independent transcendental elements ($\mathbf{P}^2$ is a rational surface); (2) ruled surfaces, with a one-parameter family of curves of genus 0; (3) elliptic surfaces, with a pencil of elliptic curves; and (4) K3 surfaces, a family of surfaces including nonsingular surfaces of degree 4 in $\mathbf{P}^3$. These K3 surfaces are of interest to physicists as providing a plausible model for the universe.

Applications to differential equations. The linear Picard-Fuchs equations arise very naturally. If $[X_s]$ is a family of algebraic curves (or indeed of varieties of any dimension), depending on a parameter s, then the de Rham cohomology of a curve X_s can be thought of as the space of closed differential forms modulo exact forms, that is, as the space of integrands to be integrated over closed loops m_s on X_s (Fig. 3). These integrands can themselves be differentiated with respect to the parameter s. Since they form a finite-dimensional space, there will be some operator having the form given by Eq. (13)

$$D = a_0(s)(d/ds)^n + a_1(s)(d/ds)^{n-1} + \cdots + a_n(s) \tag{13}$$

[Picard-Fuchs equation] which vanishes identically. A solution may be obtained by choosing a de Rham cohomology class w_s and a loop m_s varying continuously with s. The integral in Eq. (14) satisfies Eq. (15). As s describes a loop in the parameter

$$F(s) = \int w_s \tag{14}$$

$$DF(s) = \int Dw_s = 0 \tag{15}$$

space, m_s does not necessarily return to its original position, and therefore $F(s)$ is in general a multiple-valued function of s. For example, the hypergeometric equation (16) can be studied in this way.

$$s(1 - s)(d^2f/ds^2) + [c - (a + b + 1)s](df/ds) - abf = 0 \tag{16}$$

See HYPERGEOMETRIC FUNCTIONS.

There has also been much interest in certain nonlinear partial differential equations such as the Korteweg–de Vries equation (17). The coordinate

$$\partial f/\partial t = \partial^3 f/\partial s^3 + 6f\partial f/\partial s \tag{17}$$

ring R of an affine algebraic curve is embedded inside a large ring D of differential operators of the form

$D = a_0(s)(d/ds)^n + \cdots$ mentioned above. This embedding depends on the choice of a line bundle L. When L is made to vary with respect to a second parameter t, the Korteweg–de Vries equation and related equations arise as equations of motion for a fixed element of R inside D.

Applications to number theory. Three major conjectures have been solved by using algebrogeometric techniques. The Weil conjecture concerned numbers of solutions of equations mod p. A key idea was to reinterpret solutions modulo p as fixed points of the Frobenius endomorphism given by Eq. (18). The

$$(x_1, \ldots, x_n) \to (x_1^p, \ldots, x_n^p) \tag{18}$$

Lefschetz fixed-point theorem in algebraic topology counts the number of fixed points of an endomorphism. This theorem was ported to algebraic geometry via étale cohomology and used to help solve the conjecture.

The Mordell conjecture stated that a curve of genus greater than 1 can have only a finite number of points whose coordinates are rational numbers. This implies, for example, that many polynomials in two variables, like the Fermat equation (19), have

$$x^n + y^n - 1 = 0 \qquad (n \geq 4) \tag{19}$$

only finitely many rational solutions. The proof of the Mordell conjecture involved an arithmetic generalization of the concept of degree, which was then applied to study the geometry of the Siegel moduli space.

Finally, it seems likely that the Fermat conjecture itself, which states that Eq. (19) has no rational solutions with $xy \neq 0$, has been proved. Thus it can be said that Eq. (20) has no solution in integers with n

$$a^n + b^n + c^n = 0 \tag{20}$$

prime and $abc \neq 0$. Central to the proof is the Frey elliptic curve, given by Eq. (21), which is associated

$$y^2 = x(x - a^n)(x + b^n) \tag{21}$$

to a solution of Eq. (20).

These problems illustrate the power of geometry in dealing with seemingly algebraic problems like the Mordell and Fermat conjectures which admit no plausible, purely algebraic attack. Even the study of mathematics at the school and college level, where the algebraic manipulation of formulas tends to be emphasized, can benefit from greater attention to deeper and more powerful geometric insights. *See* GEOMETRY.

Spencer J. Bloch

Bibliography. S. S. Abhyanker, *Algebraic Geometry for Scientists and Engineers*, 1992; J. Harris, *Algebraic Geometry: A First Course*, 1992; R. Hartshorne, *Algebraic Geometry,* rev. ed., 1991; I. R. Shafarevich, *Basic Algebraic Geometry,* 1977, paper, 1990.

Alginate

A major constituent (10–47% dry weight) of the cell walls of brown algae. Extracted for its suspending, emulsifying, and gelling properties, it is one of three algal polysaccharides of major economic importance, the others being agar and carrageenan. *See* AGAR; CARRAGEENAN.

Alginate is a linear (unbranched) polymer in which regions have a predominance of either D-mannuronic acid with β-1,4 linkages or L-guluronic acid with α-1,3 linkages. The ratio of the monomers varies greatly among different species, in the same species at different seasons, and in different parts of the same plant. Regions rich in mannuronic acid have a ribbonlike conformation, while those rich in guluronic acid have kinks with a marked affinity for calcium ions. Alginate gels are not formed by cooling, as with agar and carrageenan, but by the addition of calcium ions, which causes separate polymeric chains to cohere along surfaces rich in guluronic acid.

The chief sources of alginate are members of the family Fucaceae (rockweeds) and the order Laminariales (kelps), harvested from naturally occurring stands on North Atlantic and North Pacific shores. The supply has been significantly increased through mariculture, especially in China. Processing is carried out chiefly in the United Kingdom, United States, Norway, Canada, France, and China. In the extraction process, washed and macerated plants (fresh or dried) are digested with alkali to solubilize the alginate, and the filtrate is precipitated with a concentrated solution of calcium chloride. The calcium alginate is converted to alginic acid by dilute hydrochloric acid. After further purification and chemical treatment, the product is marketed as sodium, potassium, or ammonium alginate, usually in powdered form.

Because of its colloidal properties, alginate finds numerous industrial applications, especially in the food, textile, paper, printing, paint, cosmetics, and pharmaceutical industries. About half of the consumption is in the making of ice cream and other dairy products, in which alginate prevents the formation of coarse ice crystals and provides a smooth texture. As an additive to paint, it keeps the pigment in suspension and minimizes brush marks. An alginate gel is used in making dental impressions. *See* ICE CREAM; MILK; PHAEOPHYCEAE.

Paul C. Silva; Richard L. Moe

Algorithm

A well-defined procedure to solve a problem. The study of algorithms is a fundamental area of computer science. In writing a computer program to solve a problem, a programmer expresses in a computer language an algorithm that solves the problem, thereby turning the algorithm into a computer program. *See* COMPUTER PROGRAMMING.

Classic algorithms. Efficient algorithms for solving certain classic problems were discovered long before

the advent of electronic computers.

Sieve of Eratosthenes. The algorithm known as the Sieve of Eratosthenes dates back to the 3d century B.C. It solves the problem of finding all the prime numbers between 1 and a positive integer N. It works as follows: First write the numbers from 1 to N, and cross out 1. Then cross out the multiples of 2 except 2 itself. Then find the first number not yet crossed out after 2. The number is 3. Cross out the multiples of 3 except 3 itself. Then find the first number not yet crossed out after 3. The number is 5. Cross out the multiples of 5 except 5 itself. And so on. This process is repeated until no more numbers can be crossed out.

For example, the case $N = 20$, carried through the crossing out of multiples of 3, gives:

~~1~~ 2 3 ~~4~~ 5 ~~6~~ 7 ~~8~~ ~~9~~ ~~10~~ 11 ~~12~~
13 ~~14~~ ~~15~~ ~~16~~ 17 ~~18~~ 19 ~~20~~

Consideration of the numbers 5, 7, and so on will not cross out any more numbers. The numbers that have not been crossed out (2, 3, 5, 7, 11, 13, 17, 19) are the prime numbers between 1 and 20. *See* NUMBER THEORY.

Euclid's algorithm. Euclid's algorithm, detailed in his treatise *Elements*, solves the problem of finding the greatest common divisor of two positive integers. It makes use of the fact that the greatest common divisor of two numbers is the same as the greatest common divisor of the smaller of the two and their difference. It works as follows: First find the larger of the two numbers and then replace it by the difference between the two numbers. Then find the larger of the resultant two numbers and then replace it by the difference between the two numbers. And so on. When the two numbers become equal, their value is the greatest common divisor and the algorithm terminates (finishes).

The steps of the algorithm when it is used to find the greatest common divisor of 48 and 18 are:

Replace 48 by 30 (= 48 − 18)	~~48~~	18
Replace 30 by 12 (= 30 − 18)	~~30~~	18
Replace 18 by 6 (= 18 − 12)	12	~~18~~
Replace 12 by 6 (= 12 − 6)	~~12~~	6
The answer is 6	6	6

Operation. An algorithm generally takes some input (for example, the value N in the Sieve of Eratosthenes and the two given integers in Euclid's algorithm), carries out a number of effective steps in a finite amount of time, and produces some output. An effective step is an operation so basic that it is possible, at least in principle, to carry it out using pen and paper. In computer science theory, a step is considered effective if it is feasible on a Turing machine or any of its equivalents. A Turing machine is a mathematical model of a computer used in an area of study known as computability, which deals with such questions as what tasks can be algorithmically carried out and what cannot. *See* AUTOMATA THEORY; RECURSIVE FUNCTION.

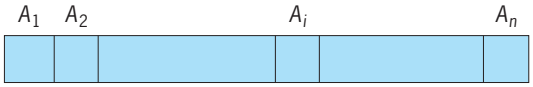


Fig. 1. An array of *n* elements.

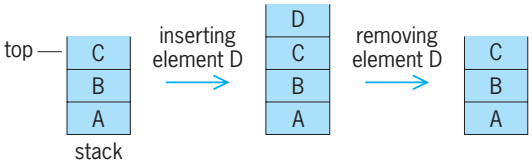


Fig. 2. Inserting and deleting an element from a stack.

Data structures. Many computer programs deal with a substantial amount of data. In such applications, it is important to organize data in appropriate structures so as to make it easier or faster to process the data. In computer programming, the development of an algorithm and the choice of appropriate data structures are closely intertwined, and a decision regarding one often depends on knowledge of the other. Thus, the study of data structures in computer science usually goes hand in hand with the study of related algorithms. Commonly used elementary data structures include records, arrays, linked lists, stacks, queues, trees, and graphs.

A record is a group of related components. For example, a record representing a date may have three components: month, day, and year. An array is a sequence of indexed (subscripted) components, A_1, $A_2, \ldots, A_n$, stored in the memory of a computer in such a manner that each component A_i can be accessed by specifying its index (subscript) i (**Fig. 1**). The components of an array can be accessed individually and independently. In contrast, a linked list is a sequence of components stored in such a manner that the components can be accessed only sequentially. That is, the first component must be accessed before the second, the second before the third, and so on. A stack is a sequence of components that allows component access, addition, and removal only at one end of the sequence called the top of the stack (**Fig. 2**). Adding an element to the top of a stack is

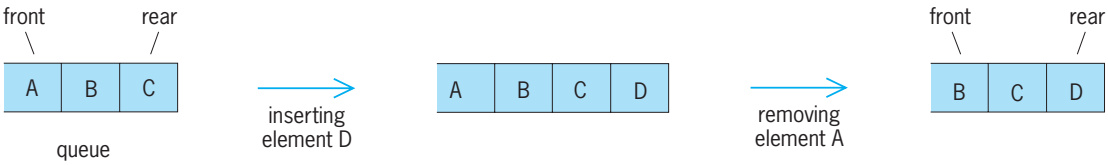


Fig. 3. Inserting and removing an element from a queue.

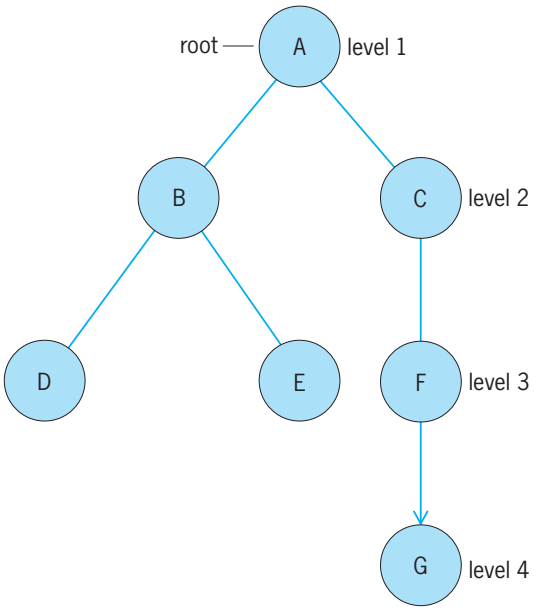


Fig. 4. An example of a tree.

generally called a push operation. Removing an element from a stack is called a pop operation. A queue is a sequence of components that allows component access and removal at one end (called the front of a queue) and addition of new components at the other end of the sequence (called the rear of a queue) [**Fig. 3**].

A tree is a data structure whose components are arranged in a hierarchy of levels. The topmost level has only one component called the root of the tree, but other levels can have many components. Each component on some level (except the root) is a child of some component on the level above. For example, each component on the second level is a child of the root, and each component on the third level is a child of some component on the second level, and so on. In summary, every component of a tree has one parent located on the level above (except the root, which does not have a parent) and possibly several children located on the level below. The number of levels in such a hierarchy is called the height of a tree. For example, the components *A*, *B*, *C*, *D*, *E*, *F*, and *G* will form a tree of height 4 if they are related in the following manner: A is the parent of *B* and *C*, *B* is the parent of *D* and *E*, *C* is the parent of *F*, and *F* is the parent of *G* (**Fig. 4**). A binary tree is a tree in which each component may have up to two children.

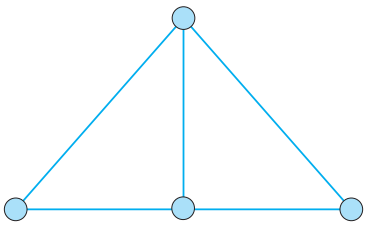


Fig. 5. A graph with four vertices and five edges.

A graph is similar to a tree in that it represents a collection of components and their relationship. However, the relationship among its components is not necessarily hierarchical, and any component can be related to any number of components. A component is called a vertex or node, and a relation between two vertices is called an edge. A convenient way to show a graph is to draw a picture in which a vertex is represented by a dot and an edge is represented by a line connecting two vertices (**Fig. 5**). If the edges have a direction assigned to them, the graph is called a directed graph. *See* GRAPH THEORY.

Elementary algorithms. Many algorithms are useful in a broad spectrum of computer applications. These elementary algorithms are widely studied and considered an essential component of computer science. They include algorithms for sorting, searching, text processing, solving graph problems, solving basic geometric problems, displaying graphics, and performing common mathematical calculations.

Sorting arranges data objects in a specific order, for example, in numerically ascending or descending orders. Internal sorting arranges data stored internally in the memory of a computer. Simple algorithms for sorting by selection, by exchange, or by insertion are easy to understand and straightforward to code. However, when the number of objects to be sorted is large, the simple algorithms are usually too slow, and a more sophisticated algorithm, such as heap sort or quick sort, can be used to attain acceptable performance. External sorting arranges data records stored on disk or tape.

Searching looks for a desired data object in a collection of data objects. Elementary searching algorithms include linear search and binary search. Linear search examines a sequence of data objects one by one. Binary search adopts a more sophisticated strategy and is faster than linear search when searching a large array. A collection of data objects that are to be frequently searched can also be stored as a tree. If such a tree is appropriately structured, searching the tree will be quite efficient.

A text string is a sequence of characters. Efficient algorithms for manipulating text strings, such as algorithms to organize text data into lines and paragraphs and to search for occurrences of a given pattern in a document, are essential in a word processing system. A source program in a high-level programming language is a text string, and text processing is a necessary task of a compiler. A compiler needs to use efficient algorithms for lexical analysis (grouping individual characters into meaningful words or symbols) and parsing (recognizing the syntactical structure of a source program). *See* SOFTWARE ENGINEERING; WORD PROCESSING.

A graph is useful for modeling a group of interconnected objects, such as a set of locations connected by routes for transportation. Graph algorithms are useful for solving those problems that deal with objects and their connections—for example, determining whether all of the locations are connected, visiting all of the locations that can be reached from a

given location, or finding the shortest path from one location to another.

Geometric objects such as line segments, circles, and polygons are fundamental in modeling many physical objects, such as a building or an automobile. Geometric algorithms are useful for solving many problems in designing and analyzing a geometric model of a physical object. Basic geometric algorithms include those for determining whether two line segments intersect, and whether a given point is within a given polygon. A related area, computer graphics, is the study of methods to represent and manipulate images, and to draw them on a computer screen or other output devices. Basic algorithms in the area include those for drawing line segments, polygons, and ovals. *See* COMPUTER GRAPHICS.

Mathematical algorithms are of wide application in science and engineering. Basic algorithms for mathematical computation include those for generating random numbers, performing operations on matrices, solving simultaneous equations, and numerical integration. Modern programming languages usually provide predefined functions for many common computations, such as random number generation, logarithm, exponentiation, and trigonometric functions.

Recursion. An algorithm for solving a problem can often be naturally expressed in terms of using the algorithm itself to solve a simpler instance of the problem. Euclid's algorithm makes use of the fact that the greatest common divisor of two different integers is the same as the greatest common divisor of the smaller number of the two and their difference. By recognizing that the problem of finding the greatest common divisor of the smaller number and the difference is a simpler instance of the given problem, the algorithm can be stated recursively: If the two given numbers are equal, their value is the answer; otherwise, (this is the recursive part) find the greatest common divisor of the smaller number and their difference. The following shows the steps of this recursive algorithm when used to find the greatest common divisor of 48 and 18.

a. Since 18 is smaller, find the greatest common divisor of 18 and 30 (= 48 − 18).
b. Since 18 is smaller, find the greatest common divisor of 18 and 12 (= 30 − 18).
c. Since 12 is smaller, find the greatest common divisor of 12 and 6 (= 18 − 12).
d. Since 6 is smaller, find the greatest common divisor of 6 and 6 (= 12 − 6).
e. The greatest common divisor is 6.

Since 6 is the greatest common divisor of 6 and 6, it is also the greatest common divisor of 6 and 12, of 18 and 12, of 18 and 30, and of the given numbers 48 and 18.

Analysis of algorithms. A problem to be solved by an algorithm generally has a natural problem size, which is usually the amount or the magnitude of data to be processed, and the number of operations that the algorithm performs depends predominantly on the problem size. A common approach to characterize the time efficiency of an algorithm is to count the number of operations that the algorithm performs in the worst case. For example, the problem of searching an array $A_1, A_2, \ldots, A_n$ for a value V has the problem size n. The linear search algorithm successively compares the desired value V with the array components $A_1, A_2, \ldots, A_n$ until it finds that a component is equal to V or until it has compared V with all of the components without finding one that is equal to V. The worst case obviously requires the algorithm to compare V with all n components. Therefore, the number of operations performed by linear search is essentially proportional to n, and linear search is said to have a running time of $O(n)$.

If the array is sorted, binary search can be used instead of linear search. Binary search first compares V with the middle component A_m, where m is $(1 + n)/2$. If A_m is equal to V, it terminates; otherwise, the search continues with at most one-half of the array components remaining to be considered. (Depending on whether V is greater than or less than A_m, either the components $A_1 \ldots A_{m-1}$ or the components $A_{m+1} \ldots A_n$ still need to be considered in the search, since the array is sorted.) A comparison of V with the middle component of the part of the array remaining to be considered either will lead to success of the search and termination of the algorithm, or again will reduce the number of elements to be searched by at least one-half of those remaining to be considered, and so forth. In summary, each comparison either leads to termination of the algorithm or reduces the number of components remaining to be considered in the search by at least one-half. In the worst case, about $\log_2 n$ comparisons are needed to reduce the number of components remaining in the search to one (by then the search is trivial). The number of operations is essentially proportional to $\log_2 n$, and the algorithm is said to have a running time of $O(\log n)$. When the problem size n is large, $\log_2 n$ is much smaller than n, indicating that binary search will be much faster than linear search when searching a large array.

An open question in the study of computational complexity concerns a class known as NP-complete problems. The class includes such problems as the traveling salesperson problem (minimizing the distance traveled in a tour of a number of cities) and finding a hamiltonian path in a graph (a path that passes through each vertex once). Although the correctness of proposed solution to these problems can be checked in polynomial time (a running time that is a polynomial function of the problem size), no polynomial-time algorithms to solve these problems have been found. (Known algorithms to solve them have an exponential running time, which makes them impractical when the problem size is large.) Furthermore, if any NP-complete problem can be solved in polynomial time, so can all others in the

class. Whether these problems can be solved in polynomial time remains an open question. *See* CRYPTOGRAPHY.

Randomized algorithms. By using a randomizer (such as a random number generator) in making some decisions in an algorithm, it is possible to design randomized algorithms that are relatively simple and fast. Since the output of a randomizer may vary from run to run, the output and the running time of a randomized algorithm may differ from run to run for the same input. There is a nonzero probability that the output of a randomized algorithm may be incorrect and execution of the algorithm may be slower than desired. A design goal of randomized algorithms is to keep such a probability small. However, there are critical systems that cannot tolerate a nonzero, albeit small, probability of incorrect output or a longer running time.

Adaptive systems. In many applications, a computer program needs to adapt to changes in its environment and continue to perform well. An approach to make a computer program adaptive is to use a self-organizing data structure, such as one that is reorganized regularly so that those components most likely to be accessed are placed where they can be most efficiently accessed. A self-modifying algorithm that adapts itself is also conceivable. For developing adaptive computer programs, biological evolution has been a source of ideas and has inspired evolutionary computation methods such as genetic algorithms. *See* GENETIC ALGORITHMS.

Parallel algorithms. Certain applications require a tremendous amount of computation to be performed in a timely fashion. An approach to save time is to develop a parallel algorithm that solves a given problem by using a number of processors simultaneously. The basic idea is to divide the given problem into subproblems and use each processor to solve a subproblem. The processors usually need to communicate among themselves so that they may cooperate. The processors may share memory, through which they can communicate, or they may be connected by communication links into some type of network such as a hypercube. *See* CONCURRENT PROCESSING; MULTIPROCESSING; SUPERCOMPUTER.

Samuel C. Hsieh

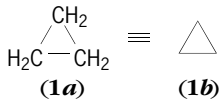
Bibliography. F. M. Carrano, P. Helman, and R. Veroff, *Data Abstraction and Problem Solving with C++*, Addison Wesley Longman, 1998; E. Horowitz, S. Sahni, and S. Rajasekaran, *Computer Algorithms/C++*, Computer Science Press, 1996; B. R. Preiss, *Data Structures and Algorithms with Object-Oriented Design Patterns in C++*, Wiley, 1999.

Alicyclic hydrocarbon

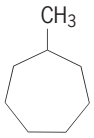
An organic compound that contains one or more closed rings of carbon atoms. The term alicyclic specifically excludes carbocyclic compounds with an array of π-electrons characteristic of aromatic rings. Compounds with one to five alicyclic rings of great variety and complexity are found in many natural products such as steroids and terpenes. By far the majority of these have six-membered rings. *See* AROMATIC HYDROCARBON; STEROID; TERPENE.

Structures and nomenclature. The bonding in cyclic hydrocarbons is much the same as that in open-chain alkanes and alkenes. An important difference, however, is the fact that the atoms in a ring are part of a closed loop. Complete freedom of rotation about a carbon-carbon bond (C—C) is not possible; the ring has faces or sides, like those of a plate.

Simple monocyclic hydrocarbons are usually represented as bond line structures; for example, cyclopropane (**1***a*) is usually represented as structure (**1***b*) and methylcycloheptane as structure (**2**). These

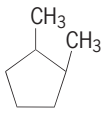


(**1***a*) (**1***b*)

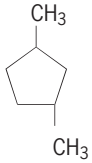


(**2**)

hydrocarbons are named by adding the prefix cyclo to the stem of the alkane corresponding to the number of atoms in the ring. When two or more substituents are attached to the ring, the relative positions and orientation must be specified: *cis* on the same side and *trans* on the other, as in *cis*-1,2-dimethylcyclopentane (**3**) and trans-1,3-dimethylcyclopentane (**4**).

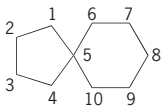


(**3**)

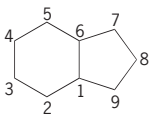


(**4**)

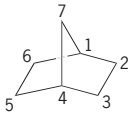
In bicyclic compounds the rings can be joined in three ways: spirocyclic, fused, and bridged, as illustrated in the structures for spiro[4.5]decane (**5**), bicyclo[4.3.0]nonane (**6**), and bicyclo[2.2.1]heptane (**7**). In each case the name indicates the total



(**5**)



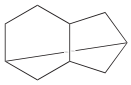
(**6**)



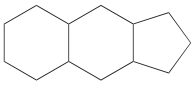
(**7**)

number of carbon atoms,and the number of atoms in each bridge. Atoms are numbered as shown.

Any of these bicyclic systems can be transformed to a tricyclic array by introduction of another bond between nonadjacent carbons, as in structure (**8**), or an additional ring, as in structure (**9**). Cyclic structure

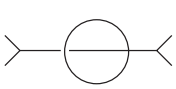


(**8**)

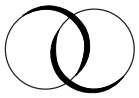


(**9**)

gives rise to the possibility of compounds made up of molecular subunits that are linked mechanically rather than chemically. In rotaxanes (**10**), bulky groups are introduced at the ends of a long chain that is threaded through a large ring ($>C_{30}$). Cyclization of the ends leads to a catenane (**11**). Several



(**10**)



(**11**)

examples of compounds with these structures have been prepared.

Ring strain and conformation. In a planar ring of n carbon atoms, the internal bond angles are given by

$$180\left(\frac{n-2}{n}\right)$$

and the angles (α) for three-, four-, and five-membered rings are 60°, 90°, and 108°, respectively. The angle between bonds of a tetrahedral carbon atom is 109°. Thus in the smaller rings, particularly cyclopropane, a significant distortion of the tetrahedral bond angle is required, and the ring is strained. (**Fig. 1**). *See* BOND ANGLE AND DISTANCE.

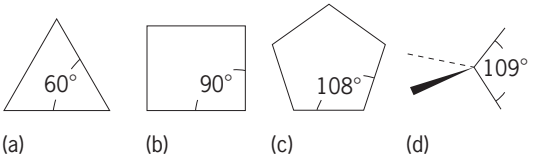


Fig. 1. Internal bond angles (α) for (*a*–*c*) planar rings and (*d*) a tetrahedral carbon atom.

Six-membered and larger rings can have normal bond angles by adopting nonplanar puckered conformations. For cyclohexane, two nonplanar arrangements are called chair and boat (**Fig. 2**). *See* CONFORMATIONAL ANALYSIS.

Angle strain is not the only consideration in the conformation of alicyclic compounds. Another factor is the interaction that occurs when bonds on adjacent carbons are aligned (eclipsing interaction; Fig. 2), as in cyclopropane or the boat form of cyclohexane. The rings in cyclobutane and cyclopentane are bent slightly to reduce these eclipsing interactions; cyclohexane exists almost entirely in a somewhat twisted chair conformation. In medium rings of 8–11 carbon atoms, another destabilizing effect is the interaction of atoms situated across the ring from each other.

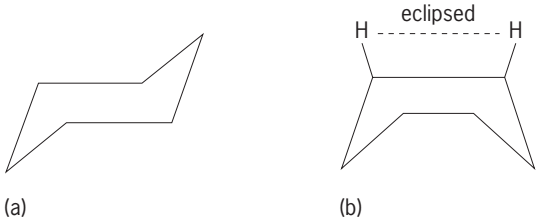


Fig. 2. Nonplanar (*a*) chair and (*b*) boat structures of cyclohexane.

The presence of a *cis* double bond in small rings increases the ring strain somewhat, but cyclobutene and even cyclopropene can be prepared. A cyclopropene ring is present in the naturally occurring sterculic acid (**12**). Alicyclic rings with a *trans* double bond or a triple bond are possible when the ring contains eight or more carbons, as in *trans*-cyclooctene (**13**).

(**12**)

(**13**)

Properties. The boiling points, melting points, and densities of cycloalkanes are all higher than those of their open-chain counterparts, reflecting the more compact structures and greater association in both liquid and solid. Geometrical constraints in the smaller rings have significant effects on reactions of cycloalkanes and derivatives. Because of ring strain, ring-opening reactions of cyclopropane, such as isomerization to propene, take place under conditions that do not affect alkanes or larger-ring cycloalkanes [reaction (1)]. Comparison of reaction rates of

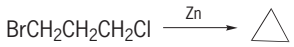
(1)

compounds with different ring size or *cis-trans* configuration has provided important insights about reaction mechanism and conformational analysis. *See* ORGANIC REACTION MECHANISM.

Ring-forming reactions. A number of useful methods have been devised that lead to alicyclic rings. These reactions are of three types: C-C bond formation between atoms in an open-chain precursor,

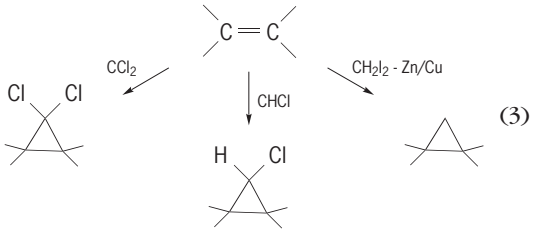
cycloaddition or cyclooligomerization, and expansion or contraction of a more readily available ring. In cycloaddition, two molecules react with formation of two bonds; in cyclooligomerization, three or more molecules combine to form three or more bonds.

Three-membered rings. Cyclopropane can be prepared by dehalogenation of 1-bromo-3-chloropropane (**14**) with zinc (Zn), as in reaction (2). A more general



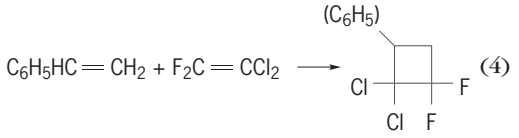
(**14**) (2)

approach is cycloaddition of a carbene or carbenoid reagent to a double bond, as in reactions (3).



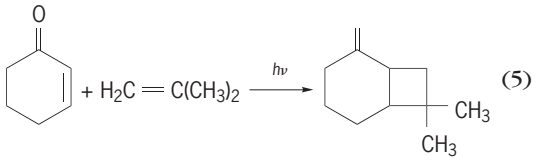
(3)

Four-membered rings. One of the important routes to cyclobutanes is cycloaddition of two double bonds. These reactions are facile with allenes, fluoroalkenes, and other alkenes in which a free-radical intermediate can be stabilized, as in reaction (4). 2+2 Cycload-



(4)

ditions (in which a two-atom molecule combines with a two-atom partner to give a four-atom ring) also occur photochemically [reaction (5)]. Four-

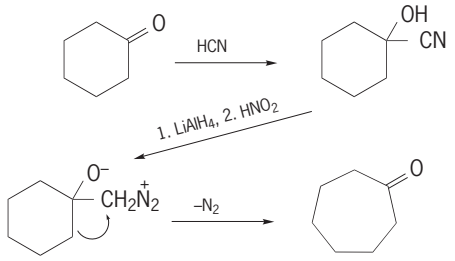


(5)

membered rings can also be obtained by ring contraction reactions of cyclopentanes.

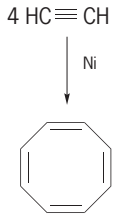
Five- and six-membered rings. Compounds with these rings are available in quantity, and there are many ways to interconvert them. Intramolecular reactions between groups separated by four or five atoms are highly favorable; an example is the cationic cyclization in the biosynthesis of terpenes and steroids. Another very general route to cyclohexane derivatives is Diels-Alder cycloaddition. *See* BIOSYNTHESIS; DIELS-ALDER REACTION.

Seven-membered rings. Cycloheptane derivatives can be obtained by ring expansion reactions such as those in reactions (6).



(6)

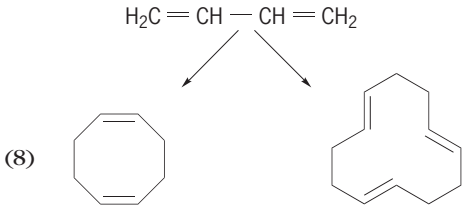
Eight-membered rings. The compound of central importance in this series is cyclooctatetraene (**15**), which is derived from four molecules of acetylene by cyclooligomerization with a nickel (Ni) catalyst [reaction (7)]. Similarly, dimerization of 1,3-dienes



(7)

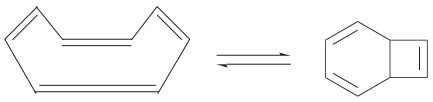
(**15**)

gives cycloocta-1,5-dienes [reaction (8)] and trimerization leads to cyclododecatrienes [reaction (9)].



(8) (9)

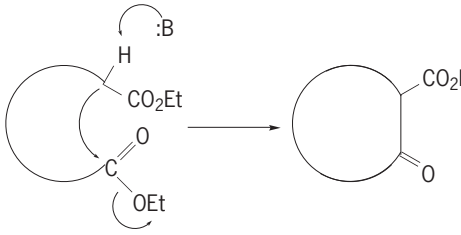
Cyclooctatetraene is a highly reactive, nonaromatic polyene that exists in a tub conformation. In several reactions, the products are those arising from the bicyclic valence tautomer (an isomer formed rapidly and reversibly by changes in valence bonds), as in reaction (10).



(10)

See TAUTOMERISM.

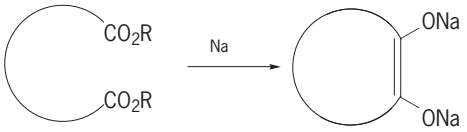
Larger rings. Ring closure between ester or nitrile groups at the ends of a chain by base-catalyzed condensation can be used to obtain rings of 14 atoms or more, as in reaction (11), where Et = C_2H_5. Very



(11)

dilute solutions must be used in order to minimize

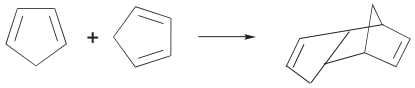
intermolecular reaction. The yields of medium rings (9–12 ring atoms) by these reactions are negligible, however, because the geometry necessary for such an enolate addition to the ester is unfavorable. A useful method for all rings is known as the acyloin condensation of a diester on molten sodium (Na), as in reaction (12), where R = an organic group. In this



(12)

process, the ends of the chain can be aligned by moving on the sodium surface.

Major compounds. Two alicyclic compounds are manufactured in large volume; both are products of the petroleum industry. Cyclopentadiene (**16**) is formed from various alkylcyclopentanes in naphtha fractions during the refining process. It is a highly reactive diene and spontaneously dimerizes in a $4+2$ cycloaddition [reaction (13)]; it is used as a copolymer in several resins.



(13)

(**16**)

Cyclohexane is produced in large quantity by hydrogenation of benzene. The principal use of cyclohexane is conversion by oxidation in air to a mixture of cyclohexanol and the ketone, which is then oxidized further to adipic acid for the manufacture of nylon. *See* ORGANIC SYNTHESIS.

James A. Moore

Bibliography. F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry: Structure and Mechanisms (Part A)*, 4th ed., 2004; M. Grossel, *Alicyclic Chemistry*, 1997; Z. Rappaport and S. Patai (eds.), *Chemistry of Alkanes and Cycloalkanes*, 1992.

Alismatales

A small order of flowering plants, division Magnoliophyta (Angiospermae), which gives its name to the subclass Alismatidae of the class Liliopsida (monocotyledons). It consists of three families (Alismataceae, Butomaceae, and Limnocharitaceae) and less than a hundred species. They are aquatic and semiaquatic herbs with a well-developed, biseriate perianth that is usually differentiated into three sepals and three petals, and with a gynoecium of several or many, more or less separate carpels. Each flower is usually subtended by a bract. *Butomus umbellatus* (flowering rush) and species of *Sagittaria* (arrowhead, family Alismataceae) of this order are sometimes cultivated as ornamentals.

The Alismatales are usually regarded as the most primitive existing group of monocotyledons, but because of their trinucleate pollen and nonendospermous seeds they cannot be considered as directly ancestral to the rest of the group. Instead they appear to form a near-basal side branch which has undergone its own sort of specialization while remaining relatively primitive in gross floral structure.

The Alismatales and some related orders have often been treated as a single order Helobiae or Helobiales, embracing most of what is here treated as the subclass Alismatidae. *See* ALISMATIDAE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM.

Arthur Cronquist

Alismatidae

A relatively primitive subclass of the class Liliopsida (monocotyledons) of the division Magnoliophyta (Angiospermae), the flowering plants, consisting of 4 orders, 16 families, and less than 500 species. Typically they are aquatic or semiaquatic, with apocarpous flowers and nonendospermous seeds. They have trinucleate pollen, and the stomates usually have two subsidiary cells. The orders Alismatales, Hydrocharitales, and Najadales are closely related among themselves and have often been treated as a single order, Helobiae or Helobiales. The Triuridales differ from the other orders in being terrestrial and mycotrophic, without chlorophyll, and in having abundant endosperm in the seeds. *See* ALISMATALES; HYDROCHARITALES; TRIURIDALES; MAGNOLIOPHYTA; NAJADALES; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

Alkali

Any compound having highly basic properties, strong acrid taste, and ability to neutralize acids. Aqueous solutions of alkalies are high in hydroxyl ions, have a pH above 7, and turn litmus paper from red to blue. Caustic alkalies include sodium hydroxide (caustic soda), the sixth-largest-volume chemical produced in the United States, and potassium hydroxide. They are extremely destructive to human tissue; external burns should be washed with large amounts of water. The milder alkalies are the carbonates of the alkali metals; these include the industrially important sodium carbonate (soda ash) and potassium carbonate (potash), as well as the carbonates of lithium, rubidium, and cesium, and the volatile ammonium hydroxide. Sodium bicarbonate is a still milder alkaline material. *See* ACID AND BASE; PH.

Soda ash and potash washed from the ashes of wood or other biomass fires were among the first manufactured chemicals, and they produced crude soap when reacted with fats. Today, the manufacture of soap and synthetic detergents still requires large amounts of alkalies. *See* DETERGENT; SOAP.

Synthetic soda ash had been made from salt since about 1840, with the Solvay process, using limestone and ammonia, predominating since 1950. Beginning about 1940, large sources of mineral deposits of impure soda ash crystallized from ancient lakes were mined and purified in increasing quantities; however, in the 1980s, production of synthetic and natural soda ash became about equal.

Until about 1900, all caustic soda was made chemically from soda ash by action with slaked lime, and this method was used for most of the soda ash manufactured until about World War II. Meanwhile, the availability of inexpensive electric power had allowed the commercial development of salt electrolysis. This process yields (at the cathode) a 10–12% sodium hydroxide solution and hydrogen gas and chlorine gas (at the anode). Since demand for chlorine increased rapidly after 1940, large quantities of its chemical equivalent, caustic soda, were being produced and sold. Subsequent industrial problems caused wide market swings involving the replacement of soda ash by caustic soda, and the marketing and pricing of soda ash, caustic soda, and chlorine.

About 50% of the caustic soda produced goes into making many chemical products, about 16% into pulp and paper, 6.5% each into aluminum, petroleum, and textiles (including rayon), with smaller percentages into soap and synthetic detergents, and cellophane. For soda ash, about 50% goes to react mainly with sand in making glass, 25% to making miscellaneous chemicals, 6.5% each to alkaline cleaners and pulp and paper, and a few percent to water treatment and other uses. *See* ALKALI METALS; ELECTROCHEMICAL PROCESS; GLASS; HYDROXIDE; PAPER; SOAP; TEXTILE CHEMISTRY.

Donald F. Othmer

Alkali emissions

Light emissions in the upper atmosphere from elemental lithium, potassium, and especially sodium. These alkali metals are present in the upper atmosphere at altitudes from about 50 to 62 mi (80 to 100 km) and are very efficient in resonant scattering of sunlight. The vertical column contents (number of atoms per square meter) of the alkali atoms are easily deduced from their respective emission intensities. First detected with ground-based spectrographs, the emissions were observed mainly at twilight since they tend to be overwhelmed by intense scattered sunlight present in the daytime. A chemiluminescent process gives rise to so-called nightglow emissions at the same wavelengths. The development of lidars (laser radars) that are tuned to the resonance lines have enabled accurate resolution of the concentrations of these elements versus altitude for any time of the day. *See* AERONOMY; CHEMILUMINESCENCE.

There is little doubt that the origin of these metals is meteoritic ablation. Rocket-borne mass spectrometers have found that meteoritic ions are prevalent above the peaks of neutral atoms with a composition similar to that found in carbonaceous chondrites, a common form of meteorites. The ratio of the concentrations of ions to neutral atoms rises rapidly with altitude above 55 mi (90 km). *See* METEORITE.

Sodium emission. Sodium (Na) is the most abundant alkali metal in meteorites. The sodium D-lines, a doublet at 589 and 589.6 nanometers, were first detected in the nightglow in the late 1920s and at twilight a decade later. The nominal peak concentration of sodium is 3×10^9 atoms m^{-3} near 55 mi (90 km) where the total gas concentration of the atmosphere is 7×10^{19} atoms (or molecules) m^{-3}. The sodium concentration declines by half at 3 mi (5 km) above and below the peak. The altitude of the maximum concentration for the sodium layer is in accord with the fact that the bulk of the micrometeorites ablate at this height, since their mean velocity is about 22 mi/s (35 km/s).

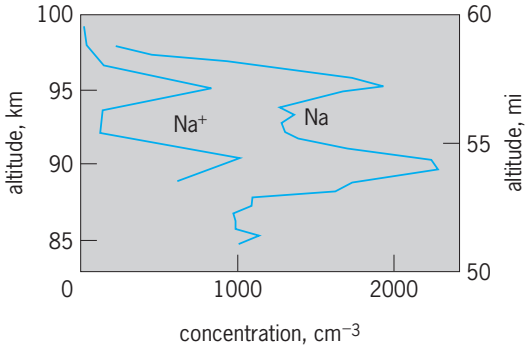


Fig. 1. Summer observations of sodium (Na) and sodium ion (Na^+) at high latitudes. (*After W. Swinder, Sodium chemistry: A brief review and two new mechanisms for sudden sodium layers, Planet. Space Sci., 40: 247–253, 1992*)

Mesospheric dynamics commonly distorts the sodium profile. Lidar measurements have provided the best evidence of these distortions and will allow scientists to achieve a better understanding of the dynamics of the mesosphere. *See* MICROMETEORITE.

The discovery of the occasional appearance of so-called sudden sodium layers (**Fig. 1**) is indicative of the complex dynamics and chemistry in the upper mesosphere. While some fresh meteor trails may conceivably contribute to such narrow layers of neutral meteor atoms, it is likely that the principal mechanism for sudden sodium layers originates with the neutralization of meteoritic ions, which can be layered by a process involving the winds and the Earth's magnetic field.

Lidar has established that the vertical column content of sodium, about 3×10^{13} atoms m^{-2} at mid-latitudes during equinox, varies very little diurnally. Twilight spectroscopic data, confirmed by the more recent lidar data, have long indicated that there is a seasonal variation in sodium, with more present in winter. The variation increases with latitude, about a factor of 2 near 20° latitude and a factor of 5 near 55° (**Fig. 2**). Data for polar latitudes are scarce. This pattern is thought to relate to the chemistry. Thus, the main depletion process for sodium, reaction (1),

$$Na + O_2 + M \rightarrow NaO_2 + M \qquad (1)$$

where M is either oxygen (O_2) or nitrogen (N_2), proceeds more slowly as temperature increases. The temperature at mesospheric altitudes is greater in winter than in summer. *See* LIDAR.

Other factors may be more or equally important, like the seasonal variation of atomic oxygen. The decline of sodium below its peak would be even

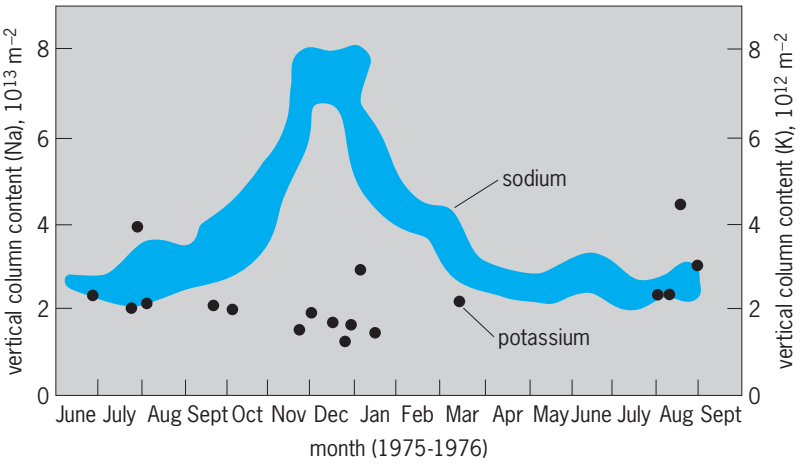


Fig. 2. Seasonal patterns of the vertical column contents of sodium and potassium at 44° latitude. (*After W. Swider, Chemistry of mesospheric potassium and its different seasonal behavior as compared to sodium, J. Geophys. Res., 92(D5):5621–5626, 1987*)

sharper if it were not for reactions (2*a*) and (2*b*).

$$NaO_2 + O \rightarrow NaO + O_2 \qquad (2a)$$

$$NaO + O \rightarrow Na + O_2 \qquad (2b)$$

Process (2*b*) can leave sodium in an excited˜state. Relaxation of the excited atoms to the ground state produces the sodium D-lines. The father of aeronomy, S. Chapman, proposed this nightglow mechanism a half century ago. Current models of the nightglow yield emission intensities compatible with observations.

Potassium emission. Potassium (K) is 15 times less abundant than sodium in meteorites. The ratio of the potassium and sodium column contents in the mesosphere ranges from 1/10 to 1/100 because, unlike sodium, the column content of potassium, 2×10^{12} m^{-2}, is fairly constant with season (Fig. 2). Once thought related to a possible additional source for sodium like sea spray, this variation may have a chemical origin. Although specific reactions may be equivalent, the particular rate coefficients may differ. For example, the rate coefficient for the formation of potassium superoxide (KO_2) is about twice that for sodium superoxide (NaO_2), and is less strongly dependent on temperature. The potassium doublet at 767 and 770 nm is estimated to have a nightglow intensity near the night sky background, 50 times smaller than the typical intensity of the sodium nightglow. The potassium nightglow has never been detected.

Lithium emission. Lithium (Li) is 35 times less abundant in meteorites than potassium, and 500 times less abundant than sodium. The meteoric source of lithium can be swamped at times by other sources, such as nuclear explosions, volcanic eruptions, and deliberate releases of lithium for aeronomic investigations of the upper atmosphere. Nevertheless, the lithium emission at 671 nm has been observed at twilight by spectrometers and at night by lidars, which indicate that the vertical content, 10^{10} m^{-2} at equinox, increases in winter like sodium. The lithium nightglow emission is undetectable. *See* ALKALI METALS; ATMOSPHERIC CHEMISTRY; MESOSPHERE.

William Swider

Bibliography. P. P. Batista et al., Characteristics of the sporadic sodium layers observed at 23°S, *J. Geophys. Res.*, 94:15349–15358, 1989; J. M. C. Plane, The chemistry of meteoric metals in the Earth's upper atmosphere, *Int. Rev. Phys. Chem.*, 10:55–106, 1991; W. Swider, Sodium chemistry: A brief review and two new mechanisms for sudden sodium layers, *Planet. Space Sci.*, 40:247–253, 1992.

Alkali metals

The elements of group 1 in the periodic table. Of the alkali metals, lithium differs most from the rest of the group, and tends to resemble the alkaline-earth metals (group 2 of the periodic table) in many ways. In this respect lithium behaves as do many other elements that are the first members of groups in the periodic table; these tend to resemble the elements in the group to the right rather than those in the same group. Francium, the heaviest of the alkali-metal

Isotopes of the alkali metals

Element	Normal at. wt	Mass no.	Radio-active	Half-life*
Lithium, Li	6.939	5	Yes	10^{-21} s
		6	No	Stable (7.5)
		7	No	Stable (92.5)
		8	Yes	0.83 s
		9	Yes	0.17 s
Sodium, Na	22.9898	20	Yes	0.23 s
		21	Yes	23.0 s
		22	Yes	2.6 y
		23	No	Stable (100)
		24	Yes	15.0 h
		25	Yes	60 s
Potassium, K	39.102	37	Yes	1.2 s
		38	Yes	7.7 m
		39	No	Stable (93.1)
		40	Yes	1.2×10^9 y
		41	No	Stable (6.9)
		42	Yes	12.4 h
		43	Yes	22 h
		44	Yes	27 m
Rubidium, Rb	85.47	81	Yes	4.7 h
		82	Yes	6.3 h
		83	Yes	80 d
		84	Yes	23 m
		85	No	Stable (72.2)
		86	Yes	19 d
		87	Yes	6.2×10^{10} y (27.8)
		88	Yes	18 m
		89	Yes	15 m
		90	Yes	2.7 m
		91	Yes	14 m
Cesium, Cs	132.905	127–132	Yes	Short
		133	No	Stable (100)
		134	Yes	3×10^6 y
		135	Yes	2.3 y
		136	Yes	1.3 d
		137	Yes	37 y
		138–145	Yes	Short
Francium, Fr		223	Yes	21 m

*Figures in parentheses indicate the percentage occurrence in nature.

elements, has no stable isotopes and exists only in radioactive form (see **table**). *See* PERIODIC TABLE.

In general, the alkali metals are soft, low-melting, reactive metals. This reactivity accounts for the fact that they are never found uncombined in nature but are always in chemical combination with other elements. This reactivity also accounts for the fact that they have no utility as structural metals (with the possible exception of lithium in alloys) and that they are used as chemical reactants in industry rather than as metals in the usual sense. The reactivity in the alkali-metal series increases in general with increase in atomic weight from lithium to cesium. *See* CESIUM; ELECTROCHEMICAL SERIES; FRANCIUM; LITHIUM; POTASSIUM; RUBIDIUM; SODIUM. Marshall Sittig

Bibliography. H. V. Borgstedt and C. K. Matthews, *Applied Chemistry of the Alkali Metals*, 1987; J. Bowser, *Inorganic Chemistry*, 1993; A. S. Kertes and C. A. Vincent (eds.), *Alkali Metal, Alkaline-Earth Metal and Ammonium Halides in Amide Solvents*, 1980.

Alkaline-earth metals

Usually calcium, strontium, and barium, the heaviest members of group 2 of the periodic table (excepting radium). Other members of the group are beryllium, magnesium, and radium, sometimes included among the alkaline-earth metals. Beryllium resembles aluminum more than any other element, and magnesium behaves more like zinc and cadmium. The gap between beryllium and magnesium and the remainder of the elements of group 2 makes it desirable to discuss these elements separately. Radium is often treated separately because of its radioactivity.

J. J. Berzelius first reduced the three alkaline-earth metals to the elementary state, but obtained them as amalgams by electrolysis. Humphry Davy in 1808 isolated the metals in the pure state by distillation of amalgams produced electrolytically. Today industrial preparation of these elements involves electrolysis of their molten chlorides or reduction of their oxides with aluminum.

The alkaline earths form a closely related group of highly metallic elements in which there is a regular gradation of properties. The metals, none of which occurs free in nature, are all harder than potassium or sodium, softer than magnesium or beryllium, and about as hard as lead. The metals are somewhat brittle, but are malleable, extrudable, and machinable. They conduct electricity well; the specific conductivity of calcium is 45% of that of silver. The oxidation potentials of the triad are as great as those of the alkali metals.

The alkaline earths exist as large divalent cations in all their compounds, in which the elements are present in the 2+ oxidation state. The metals have a gray-white luster when cut but tarnish readily in air. They burn brilliantly in air when heated, and form the metal monoxide, except for barium, which forms the peroxide. A certain amount of nitride is formed simultaneously, especially with calcium. All the metals dissolve readily in acid. Whereas calcium reacts smoothly with water to yield hydrogen, the heavier members react as violently as sodium does. All the metals are soluble in liquid ammonia, yielding strongly reducing, electrically conducting, blue solutions. The order of solubility in water for most salts is calcium > strontium > barium, except that the order is reversed for the fluorides, hydroxides, and oxalates. All three elements unite directly with hydrogen to form hydrides and with nitrogen to form nitrides, but whereas ease of formation of a nitride increases with atomic number, ease of formation of a hydride decreases.

The elements and their compounds find important industrial uses in low-melting alloys, deoxidizers, and drying agents and as cheap sources of alkalinity. *See* BARIUM; BERYLLIUM; CALCIUM; MAGNESIUM; PERIODIC TABLE; RADIUM; STRONTIUM. Reed F. Riley

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; T. R. Dickson, *Introduction to Chemistry*, 7th ed., 1995; K. Mackay, *Introduction to Modern Inorganic Chemistry*, 6th ed., 2002.

Alkaloid

A cyclic organic compound that contains nitrogen in a negative oxidation state and is of limited distribution among living organisms. Over 10,000 alkaloids of many different structural types are known; and no other class of natural products possesses such an enormous variety of structures.

Therefore, alkaloids are difficult to differentiate from other types of organic nitrogen-containing compounds.

Simple low-molecular-weight derivatives of ammonia, as well as polyamines and acyclic amides, are not considered alkaloids because they lack a cyclic structure in some part of the molecule. Amines, amine oxides, amides, and quaternary ammonium salts are included in the alkaloid group because their nitrogen is in a negative oxidation state (the oxidation state designates the positive or negative character of atoms in a molecule). Nitro and nitroso compounds are excluded as alkaloids. The almost-ubiquitous nitrogenous compounds, such as amino acids, amino sugars, peptides, proteins, nucleic acids, nucleotides, prophyrins, and vitamins, are not alkaloids. However, compounds that are exceptions to the classical-type definition (that is, a compound containing nitrogen, usually a cyclic amine, and occurring as a secondary metabolite), such as neutral alkaloids (colchicine, piperine), the β-phenyl-ethylanines, and the purine bases (caffeine, theophylline, theobromine), are accepted as alkaloids.

Nomenclature. The nomenclature of alkaloids has not been systematized. Most alkaloids bear the suffix -ine and their names are derived in various ways: from the generic name of the plant, for example hydrastine from *Hydrastis canadensis*; from the specific plant name, for example cocaine from *Erythroxylum coca*; from the common name of the drug yielding

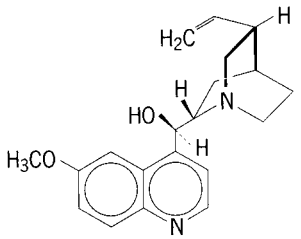
them, for example, ergotamine from ergot; or from a particular physiological activity exhibited, for example, morphine from Morpheus, Greek god of dreams. Only the pelletierine group is named for a person, P. J. Pelletier, who discovered a number of alkaloids in the first third of the nineteenth century.

Occurrence. Alkaloids are derived from both plants and animals, including marine organisms.

Plant-derived alkaloids. Alkaloids often occur as salts of plant acids such as malic, meconic, and quinic acids. Some plant alkaloids are combined with sugars, for example, solanine in potato (*Solanum tuberosum*) and tomatine in tomato (*Lycopersicum esculentum*). Others occur as amides, for example, piperine from black pepper (*Piper nigrum*), or as esters, for example, cocaine from coca leaves (*Erythroxylum coca*). Still other alkaloids occur as quaternary salts or tertiary amine oxides.

Although about 40% of all plant families contain at least one alkaloid-bearing species, only about 9% of the over 10,000 plant genera produce alkaloids. Among the angiosperms, alkaloids occur in abundance in certain dicotyledons and especially in the families Apocynaceae (dogbane, quebracho), Asteraceae (groundsel, ragwort), Berberidaceae (European barberry), Fabaceae (broom, gorse, lupine), Loganiaceae (*Strychnos* species), Menispermaceae (moonseed), Papaveraceae (poppies, chelidonium), Ranunculaceae (aconite, larkspur), Rubiaceae (cinchona bark, ipecac), Rutaceae (citrus, fagara), and Solanaceae (tobacco, deadly nightshade, tomato, potato, thorn apple). Rarely are they present in cryptogamia, gymnosperms, or monocotyledons. Among the monocotyledons, the Amaryllidaceae (amaryllis, narcissus) and Liliaceae (meadow saffron, veratrums) are families that bear alkaloids.

Examples of well-known alkaloids are morphine and codeine from the opium poppy (*Papaver somniferum*), strychnine from *Strychnos* species, quinine (**1**) from *Cinchona* bark and various *Cinchona*



(1)

species, and coniine from poison hemlock (*Conium maculatum*). Other important alkaloids are colchicine from the autumn crocus (*Colchicum autumnale*), caffeine from coffee, tea, kola, and maté, and nicotine from the tobacco plant (*Nicotiana tabacum*) and Aztec tobacco (*N. rustica*). *See* CAFFEINE; COFFEE; COLA; COLCHICINE; NICOTINE ALKALOIDS.

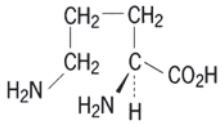
Animal-derived alkaloids. While most alkaloids have been isolated from plants, a large number have been isolated from animal sources. They occur in mammals, anurans (frogs, toads), salamanders, arthropods (ants, millipedes, ladybugs, beetles, butterflies), marine organisms, mosses, fungi, and certain bacteria. Samandarine has been isolated from the European fire salamander (*Salamandra maculosa*). Very toxic steroidal alkaloids occur in certain frogs and toads; for example, the Columbian arrow-poison frog (*Phyllobates aurotaenia*) produces a highly lethal venom containing batrachotoxin. The Canadian beaver (*Castor fiberi*) produces castoramine, and the scent gland of the musk deer (*Moschus moschiferus*) produces muscopyridine. The Southern fire ant (*Solenopsis invicta*) secretes a powerful venom containing a series of 2,6-dialkylpiperidines. The European millipede (*Glomeris marginata*) when provoked discharges a defensive secretion containing glomerine, a quinazolone. When molested, the European ladybug secretes hemolymph at the joints to afford protection against predators; this hemolymph contains precoccinelline and its oxide, coccinelline. Certain species of butterflies elaborate pheromones derived from exogenous pyrrolizidine alkaloid precursors; the pyrrolizidine alkaloid occurs in the secretions of *Lycorea ceres ceres* and *Danaus gilippus berenice*, where it elicits olfactory-receptor responses in female antennae and serves as a chemical messenger that induces mating behavior. *See* PHEROMONE.

Marine-derived alkaloids. Many alkaloids have been isolated from marine organisms, both plant and animal. Of great interest are the alkaloids occurring in certain toxic dinoflagellates, such as *Gonyaux tamarensis* in the North Atlantic and *G. cantanella* in the South Atlantic. The toxic principles include the alkaloids saxitoxin, gonyautoxin-II, and gonyautoxin-III. These paralytic alkaloids also occur in the Alaska butter clam (*Saxidomus giganteus*) and toxic mussels (*Mytilus californianus*). Tetrodotoxin, present in the ovaries and liver of the Japanese puffer fish (*Spheroides rubripes* and *S. vermicularis*), in the Taiwanese goby fish (*Gobius cringer*), and in the California newt (*Taricha torosa*), is one of the most lethal low-molecular-weight toxins known.

Isolation and purification. Since few plant species produce only a single alkaloid, there is the problem of separating complex mixtures. The alkaloid mixture is usually extracted from the finely powdered plant material by percolation with some solvent, such as methanol or aqueous ethanol. Evaporation leaves a residue, which is then dissolved in an organic solvent such as chloroform, methylene chloride, or toluene, and filtered. The solution is extracted repeatedly first with dilute sulfuric or hydrochloric acid and, after basification, with an organic solvent. These extracts when evaporated yield different mixtures of alkaloids which may be further separated by column chromatography over alumina or silica gel, vacuum liquid chromatography, countercurrent distribution, centrifugally accelerated radial thin-layer chromatography, or preparative-layer chromatography. Eventually, it is possible to obtain compounds in a high state of purity. *See* CHEMICAL SEPARATION TECHNIQUES; CHROMATOGRAPHY.

Biosynthesis. Of great interest to chemists is the mode of alkaloid synthesis in plants. The

biosynthesis of most alkaloids can be derived hypothetically, by using a few well-known chemical reactions, from such relatively simple precursors as phenylalanine, tyrosine, tryptophan, histidine, methionine, acetate units, terpene units, and amino acids such as anthranilic acid, lysine, and ornithine. A number of simple alkaloids have been synthesized from amino acid derivatives under physiological conditions: pelletierine is derived from the amino acid lysine, while cocaine and nicotine are derived from ornithine (**2**). Many alkaloids, such as

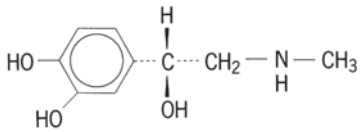


(**2**)

morphine, berberine, and emetine, incorporate a tetrahydroisoquinoline unit derived from tyrosine; and others, such as strychnine, quinine, reserpine, and camptothecin, are derived from tryptophan. The terpenoid and steroidal alkaloids, such as the diterpenoid alkaloid aconitine and the steroidal alkaloid tomatidine from the tomato, are likely derived from mevalonic acid lactone. *See* BIOSYNTHESIS; STEROID; TERPENE.

Function. The exact function of alkaloids in plants is not well understood, but they are often regarded as by-products of plant metabolism. They are sometimes considered to be reservoirs for protein synthesis; as protective agents to discourage attack by animals, insects, or microbes; as regulators of growth, metabolism, and reproduction; and as end products of detoxification of substances whose accumulation might be injurious to the plant.

Medicinals. Many alkaloids exhibit marked pharmacological activity, and some find important uses in medicine. Atropine, the optically inactive form of hyoscyamine, is used widely in medicine as an antidote to cholinesterase inhibitors such as physostigmine and insecticides of the organophosphate type; it is also used in drying cough secretions. Morphine and codeine are narcotic analgesics, and codeine is also an antitussive agent, less toxic and less habit forming than morphine. Colchicine, from the corms and seeds of the autumn crocus, is used as a gout suppressant. Caffeine, which occurs in coffee, tea, cocoa, and cola, is a central nervous system stimulant; it is used as a cardiac and respiratory stimulant and as an antidote to barbiturate and morphine poisoning. Emetine, the key alkaloid of ipecac root (*Cephaelis ipecacuanha*), is used in the treatment of amebic dysentery and other protozoal infections. Epinephrine (**3**), or adrenaline, produced in most



(**3**)

animal species by the adrenal medulla, is used as a bronchodilator and cardiac stimulant and to counter allergic reactions, anesthesia, and cardiac arrest.

Pilocarpine, from *Pilocarpus jaborandi*, in hydrochloride or nitrate form is cholinergic and is used topically in the treatment of glaucoma; it is also used orally or subcutaneously to stimulate secretion of saliva in individuals who undergo therapy with ganglionic blocking agents. Quinine, used in many areas of the world to treat malaria, is used in the United States to treat chloroquine-resistant falciparum malaria. Quinidine, a stereoisomer of quinine, is an antiarrhythmic agent used to control auricular fibrillation. Reserpine, from *Rauwolfia serpentina*, produces sedation, tranquilization, and depression of blood pressure; it is sometimes used to decrease high blood pressure and as a tranquilizer in cases of nervousness, hysteria, and stress. Taxol, from the Pacific yew tree (*Taxus brevifolia*), is used in the treatment of ovarian and breast cancer. Tubocurarine chloride (curare, from *Strychnos* and *Chondodendron* species) is employed as a skelet al muscle relaxant during surgery and to control convulsions resulting from strychnine poisoning and tetanus. Vinblastine and vincristine, from the Madagascar periwinkle plant (*Catharanthus roseus*), are potent anticancer drugs; the former is used to treat generalized Hodgkin's disease and choriocarcinoma, while the latter is used to treat acute leukemia. *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; EPINEPHRINE; MORPHINE ALKALOIDS; PHARMACOGNOSY; QUININE; STRYCHNINE ALKALOIDS.

S. William Pelletier

Bibliography. A. Brossi and G. A. Cordell (eds.), *The Alkaloids: Chemistry and Physiology*, vol. 42, 1992; G. A. Cordell, *Introduction to Alkaloids: A Biogenetic Approach*, 1981; D. R. Dalton, *The Alkaloids: The Fundamental Chemistry—A Biogenetic Approach*, 1979; K. Mothes, H. R. Shutte, and M. Luckner, *Biochemistry of Alkaloids*, 1985; S. W. Pelletier (ed.), *Alkaloids: Chemical and Biological Perspectives*, vols. 1–9, 1983–1995; S. W. Pelletier, The nature and definition of an alkaloid, *Alkaloids: Chemical and Biological Perspectives*, vol. 1, 1983; T. Robinson, *The Biochemistry of Alkaloids*, 2d ed., 1981; I. W. Southon and J. Buckingham (eds.), *Dictionary of Alkaloids*, 1989.

Alkane

An organic compound with the general formula C_nH_{2n+2}. Alkanes are open-chain (aliphatic or noncyclic) hydrocarbons with no multiple bonds or functional groups. They consist of tetrahedral carbon atoms, up to 10^5 carbons or more in length. The C-C σ bonds are formed from sp^3 orbitals, and there is free rotation around the bond axis. *See* MOLECULAR ORBITAL THEORY; STRUCTURAL CHEMISTRY.

Alkanes provide the parent names for all other aliphatic compounds in systematic nomenclature. Alkanes are designated by the ending -ane appended to a stem denoting the chain length. The straight-chain isomer is designated by the prefix *n*(normal); other isomers are named by specifying the size of the branch and its location (see **table**). The number of isomers increases enormously in larger molecules;

Nomenclature and properties of alkanes

Name	Structure	Boiling point, °C (°F)	Heat of formation (ΔH_f°), kJ
Methane	CH_4	−162 (−260)	−74.5
Ethane	CH_3CH_3	−89 (−128)	−83.45
Propane	$CH_3CH_2CH_3$	−42 (−44)	−104.6
n-Butane	$CH_3(CH_2)_2CH_3$	−0.5 (33)	−125.7
2-Methylpropane (isobutane)	$(CH_3)_2CHCH_3$	−11.7 (10.9)	−134.2
n-Pentane	$CH_3(CH_2)_3CH_3$	36.1 (97.0)	−146.87
2-Methylbutane (isopentane)	$(CH_3)_2CHCH_2CH_3$	29.9 (85.8)	−153.7
2,2-Dimethyl propane (neopentane)	$(CH_3)_4C$	9.4 (49)	−167.9
n-Hexane	$CH_3(CH_2)_4CH_3$	68.7 (156)	−167.0

thus there are 75 isomers of $C_{10}H_{22}$ and over 4 billion for $C_{30}H_{62}$.

Properties. Alkanes with four or fewer carbons are gases at atmospheric pressure. Higher *n*-alkanes are liquids or, above about 20 carbons, solids known as paraffin wax. Alkanes have densities lower than that of water and have very low water solubility; other properties depend on the degree of branching (see table). The heat of formation (ΔH°_f) is a measure of the energy content of a compound relative to the component elements in standard states. For example, comparison of heat of formation values for three alkane isomers with the molecular formula C_5H_{12} indicates that the relative energy content of the three pentane isomers decreases with increased branching; that is, the branched isomer is thermodynamically most stable. *See* PARAFFIN.

Sources. Alkanes are the major components of natural gas and petroleum, which are the only significant sources. Much smaller amounts of alkanes have been produced from coal at various times and locations, either indirectly by the Fischer-Tropsch process or by direct liquefaction. *See* COAL LIQUEFACTION; FISCHER-TROPSCH PROCESS; NATURAL GAS; PETROLEUM.

Certain long-chain alkanes and also some alicyclic hydrocarbons are very widely distributed both in sedimentary soils and shales and among living organisms. Phytane and pristane are derived from phytol, and these alkanes are of significance as biological markers in tracing diagenesis. Both insects and plants secrete alkanes, which function as moisture barriers and protective coatings. Cuticular waxes on fruit and leaves contain predominantly the *n*-alkanes with 29, 31, and 33 carbons; the preponderance of odd-carbon chains reflects the origin by loss of carbon monoxide (CO) or carbon dioxide (CO_2) from even-carbon precursors. *See* DIAGENESIS.

Individual lower alkanes can be separated from the more volatile distillate fractions of petroleum, but beyond the C_7–C_8 range the alkanes obtained are mixtures of many isomers. Compounds of a specific structure can be prepared on a laboratory scale by chemical synthesis. Various C-C bond-forming steps such as coupling or condensation are carried out to build up the desired carbon skeleton. The final step is usually removal of a functional group by some type of reduction. Several insect pheromones, for example 15,19,23-trimethylheptatriacontane ($C_{40}H_{82}$) from the female tsetse fly, have been synthesized for biological study. *See* PHEROMONE.

Chemistry. Much of the chemistry of alkanes begins at the petroleum refinery, where several reactions are carried out to adjust the hydrocarbon composition of crude oil to that needed for a constantly changing set of applications. Major reactions are (1) isomerization of straight-chain alkanes to branched compounds; (2) cracking to produce smaller molecules; (3) alkylation, for example, combination of propylene (**1**) and butane (**2**) to give 2,3-dimethylpentane (**3**), as in reaction (1); and

$$\underset{\textbf{(1)}}{CH_3CH{=}CH_2} + \underset{\textbf{(2)}}{CH_3CH_2CH_2CH_3} \longrightarrow \underset{\textbf{(3)}}{CH_3CH(CH_3)CH(CH_3)CH(CH_3)CH_2CH_3} \quad (1)$$

(4) cyclodehydrogenation (platforming), in which aromatizations occur. An important objective in some of these processes is to increase the yield of highly branched alkanes in the C_6–C_8 range needed for gasoline. *See* ALKYLATION (PETROLEUM); AROMATIZATION; CRACKING; GASOLINE.

By far the most important end use of alkanes is combustion as fuel to provide heat and electric or motive power. In most cases, complete oxidation is not achieved, and varying amounts of incompletely oxidized fragments, carbon monoxide, and elemental carbon are produced.

Controlled partial oxidation is possible if all the C-H bonds in an alkane are equivalent or if one C-H bond is significantly weaker than all the others. An example of the latter situation is isobutane, which is converted to the hydroperoxide on industrial scale for the manufacture of *t*-butyl alcohol. *See* AUTOXIDATION; COMBUSTION.

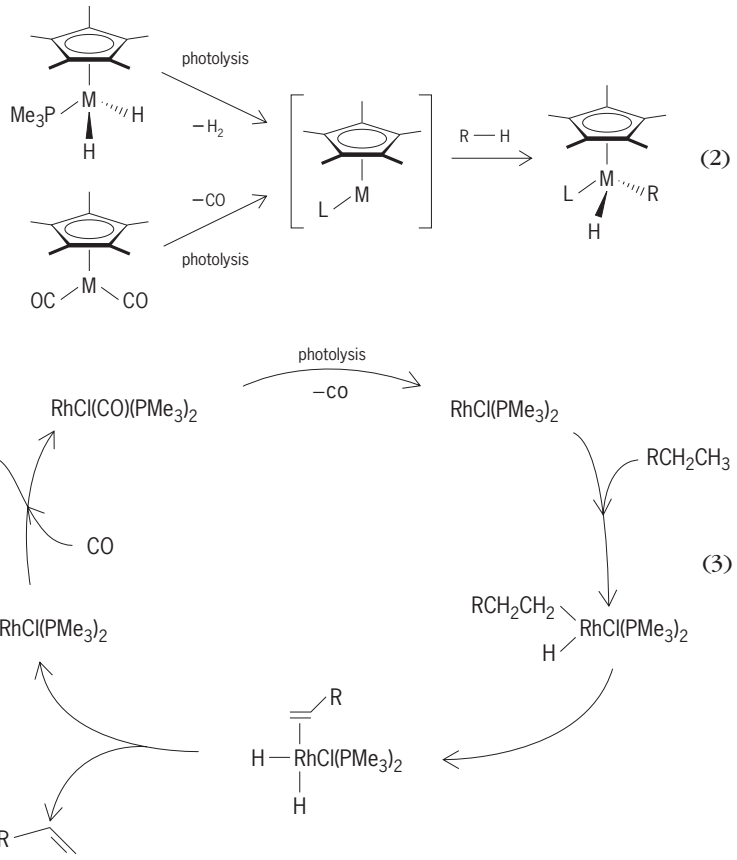
Alkanes have been referred to as paraffin hydrocarbons to indicate their low affinity or reactivity. They contain no unshared electron pairs or accessible empty bonding orbitals, and they are unaffected by many reagents that attack π bonds or other functional groups. One type of reaction that does occur is substitution by a radical chain process. Examples are chlorination and vapor-phase nitration, involving the odd-electron species Cl· and NO_2·, respectively. Neither reaction is selective; when two or more types of C-H bonds are present in the alkane, mixtures of products are usually obtained. Thus

propane gives rise to 1- and 2-chloropropanes as well as dichloro compounds. Nitration of propane leads to a mixture of 1- and 2-nitropropane, and also nitromethane and nitroethane by C-C bond cleavage. *See* HALOGENATED HYDROCARBON; HALOGENATION; NITRATION.

Since alkanes are the most abundant class of organic compounds available, much effort has been directed to the development of methods for converting them selectively to other compounds. Contributions to this objective have come from several diverse sources. Biodegradation of petroleum hydrocarbons is a well-known phenomenon and is useful in decontamination of oil leaks and spills. Closely related is utilization of the metabolic apparatus of individual microorganisms to carry out a specific reaction, such as hydroxylation of the C-H bond at the 11 position of a steroid. Many such biochemical oxidations occur by means of a cytochrome P-450 enzyme system in which an oxygen atom is transferred from dioxygen (O_2) to an alkane at the iron atom in a heme protein. Synthetic metalloporphyrins and other iron complexes have been studied as model systems for these oxidations. *See* BIODEGRADATION; CYTOCHROME; PORPHYRIN. James A. Moore

Activation by organometallic compounds. Another contribution to the activation of alkanes is from the field of organometallic chemistry. One way that C-H bonds have been cleaved involves the formation of a vacant site on a metal in a low oxidation state that contains electron-donating ligands. The vacant site provides room for the incoming hydrocarbon, and the electron-rich metal readily gives up its electrons in a process called oxidative addition. The net reaction results in the use of two electrons from the C-H bond plus two electrons from the metal to form two metal-carbon bonds, each with two electrons. The vacant site is most easily generated by irradiation of the precursor complex with ultraviolet light, a technique that makes the ligands fall off easily at ambient temperatures. If the irradiation is carried out in a pure hydrocarbon solvent, the metal oxidatively adds to the solvent, resulting in the formation of a complex with metal-hydrogen and metal-carbon bonds. The C-H bond has been completely cleaved, and further chemistry can now be envisioned with the fragments attached to the metal. Reaction (2) shows several specific metal complexes that will undergo this type of oxidative addition chemistry with saturated and unsaturated hydrocarbons such as methane, ethane, hexane, and benzene. In reaction (2), M = rhodium (Rh) or iridium (Ir); Me = methyl group (CH_3); Me_3P = trimethylphosphine; L = PMe_3 or carbonyl (C═O); and R = CH_3, C_2H_5 (ethyl), C_6H_{13} (hexyl), or C_6H_5 (phenyl). With linear alkanes such as hexane, there is a preference of cleavage of the less hindered primary (terminal) C-H bonds at the end of the molecule over the more hindered secondary (internal) C-H bonds of the molecule.

In one application based on this reactivity, a complex has been used to convert alkanes into alkenes plus hydrogen by using light energy. The complex chlorocarbonyl bis-(trimethylphosphine) rhodium



$[RhCl(CO)(PMe_3)]$ reacts with alkanes to give many thousands of turnovers of alkenes upon photolysis, as in reactions (3). Beginning with the rhodium complex $RhCl(CO)(PMe_3)_2$, the fundamental step involving activation of the C-H bond can be seen in the catalytic cycle following the loss of carbon monoxide (CO). The loss of a hydrogen atom on the carbon that is adjacent to the one bound to the metal is a common reaction in organometallic chemistry, and it leads to the production of the alkene product still bound to the rhodium. The olefin is then released from the complex. Finally, hydrogen is displaced by reaction with CO to regenerate the initial metal complex. Light energy is required in this cycle to remove CO from the metal. *See* ORGANOMETALLIC COMPOUND. William D. Jones

Bibliography. J. A. Davies et al. (eds.), *Selective Hydrocarbon Activation,* 1990; R. J. Fessenden and J. S. Fessenden, *Organic Chemistry,* 6th ed., 1998.

Alkene

A hydrocarbon that contains a carbon-carbon double bond (C═C); dienes or polyenes contain two or more double bonds. The synonym olefin is often used, especially in connection with industrial-scale production and applications. The term unsaturated distinguishes compounds such as fats that contain double bonds from saturated molecules that do not.

Structure and names. The carbon atoms joined by a double bond have three planar sp^2 orbitals and a p orbital with lobes above and below. The double

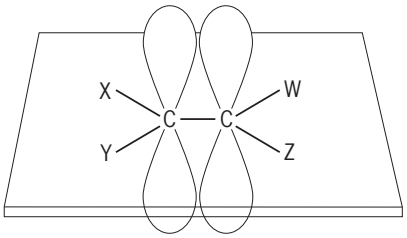
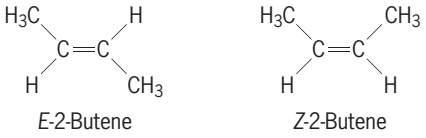


Fig. 1. Molecular orbital representation of a carbon double bond; W, X, Y, and Z represent attached groups.

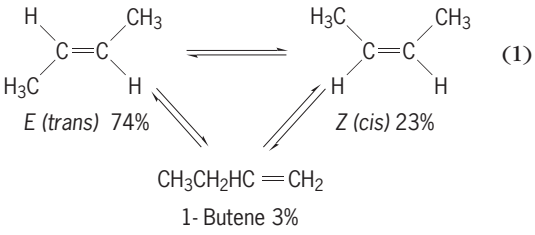
bond is a combination of a single σ bond between the nuclei and a π bond resulting from an electron pair in overlapping *p* orbitals (**Fig. 1**). As a result, the attached groups *w, x, y*, and *z* are coplanar, and rotation around the C-C axis is possible only by disrupting the π bond. A double bond is shorter (0.133 nanometer) and stronger (610 kilojoules) than a typical C-C bond (0.154 nm, 347 kJ). *See* CHEMICAL BONDING; MOLECULAR ORBITAL THEORY.

The names of alkenes are formed by changing the -ane ending of the corresponding alkane to −ene, with a number to indicate the double-bond location. Since the double bond is a barrier to rotation, an alkene such as 2-butene has two stereoisomeric forms. These geometric isomers are designated as *E* or *trans* (similar groups on opposite side) or *Z* (*cis*, similar groups on same side), as in the structures below. For the simplest alkenes (C_2–C_4) the ending



-ylene is often used. Two other common names are vinyl for the group —HC═CH_2 and allyl for the group —CH_2HC═CH_2. *See* ALKANE.

An alkene can undergo isomerization by shifts of the double bond, such as by treatment with an acid catalyst. Isomers with the double bond in different positions or geometries have different internal energies and therefore different stabilities. In most cases a *trans* isomer is favored over the *cis*, because larger groups are farther apart. Moreover, isomers in which the double bond is in an internal position on the chain are usually more stable than those with the double bond at the end. For the three straight-chain butene isomers, the equilibrium composition is shown in notation (1).



(1)

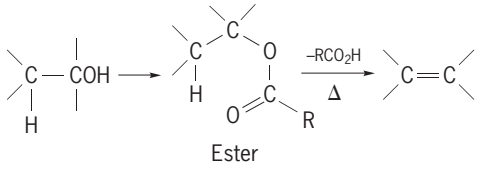
See ISOMERIZATION.

Source and preparation. Ethylene (H_2C═CH_2) and the C_3–C_5 alkenes are produced in very large volume by catalytic cracking of higher hydrocarbons or dehydrogenation of alkanes during petroleum refining. Alkenes in the $C_{10\text{-}20}$ range are important starting materials in several large-scale industrial applications. These are prepared by controlled oligomerization of ethylene with a triethylaluminum catalyst [reaction (2), where $n = 5$–10].

$$n\,HC_2{=}CH_2 \xrightarrow{(CH_3CH_2)_3Al} CH_3CH_2(CH_2CH_2)_{n-2}HC{=}CH_2 \qquad (2)$$

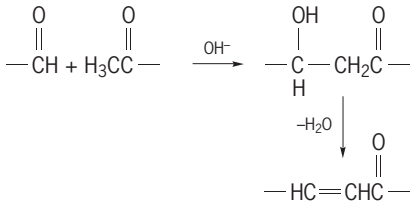
See CRACKING; DEHYDROGENATION.

Several general methods for preparing alkenes involve elimination reactions. Two long-known examples are removal of a hydrogen-halogen group (HX) from a halide with strong base, and loss of water from alcohols. Both reactions can often give more than one alkene, and acid-catalyzed dehydration may lead to rearrangement. In the latter case, conversion of the alcohol to an ester and pyrolysis may be preferable [reaction (3)]. Dehydration of a β-



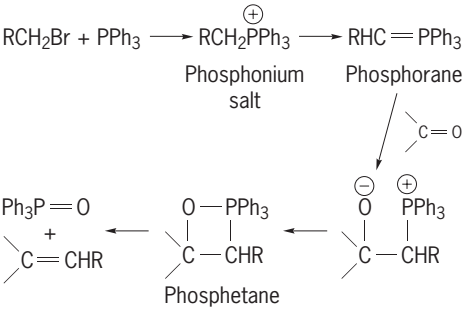
(3)

hydroxycarbonyl system is the second step in many carbonyl condensation reactions [reaction (4)].



(4)

A widely used preparative method for alkenes is the Wittig reaction. In this process a phosphorane adds to a carbonyl compound, and the double bond is formed by loss of triphenylphosphine oxide from a phosphetane [reaction (5)]. Another approach to



(5)

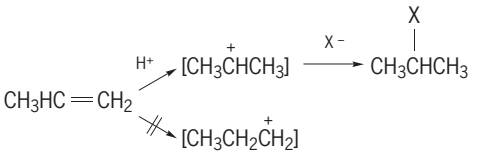
the formation of a double bond is reduction of a triple bond. *See* ALKYNE; ORGANOPHOSPHORUS COMPOUND.

Chemical activity. The double bond is a highly reactive and versatile functional group. Because of the readily accessible electrons in the π bond, alkenes react with a variety of Lewis acids, or electrophiles.

A characteristic reaction is addition, with formation of two new σ bonds in place of the π bond.

Hydrogenation. Addition of hydrogen occurs on the surface of a finely divided metal catalyst, usually nickel, palladium, or platinum. This reaction is useful for several purposes; one application is the partial hydrogenation of unsaturated vegetable oils to remove some of the double bonds and raise the melting temperature. *See* FAT AND OIL (FOOD).

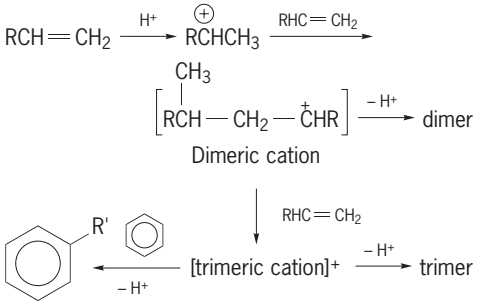
Proton acids. Several types of addition products can arise from the reaction of alkenes with acids. In a simple case hydrochloric acid (HCl) or hydrobromic acid (HBr) add to propene to give the 2-halopropane; with aqueous sulfuric acid the elements of water (H_2O) add, and 2-propanol is obtained [reaction (6),



(6)

where X= chlorine (Cl), bromine (Br), or hydroxyl (OH)]. These additions occur by way of an intermediate carbenium ion that rapidly reacts with a nucleophile such as halide ion or water. The structure of the product depends on the fact that the cation with the larger number of alkyl groups attached to the electron-deficient carbon is more stable and more easily formed. *See* REACTIVE INTERMEDIATES.

The carbenium ion formed from an alkene may undergo other reactions as well. Under suitable conditions it can react as an electrophile with another molecule of the alkene to give branched dimers, trimers, and so forth [reaction (7), where R= CH_3

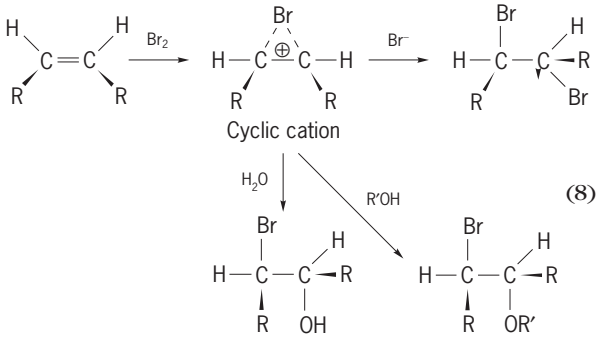


(7)

or C_2H_5 and R′ = C_9H_{18} or $C_{12}H_{24}$]. An alkene with acid can also be used to alkylate a benzene ring in a variant of the Friedel-Crafts reaction. Both processes are used in the production of alkylbenzenes for detergents and other uses. *See* FRIEDEL-CRAFTS REACTION.

Halogenation. Bromine or chlorine adds very rapidly to alkenes; bleaching of the red color of Br_2 has long been recognized as a test for a double bond. The reaction of ethylene with chlorine to give 1,2-dichloroethane is the first step in the manufacture of vinyl chloride.

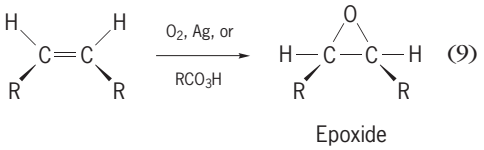
In the addition of halogen to a double bond [shown in reaction (8) for Br_2], a cyclic cation is initially formed. In the second step, bromide ion attacks this intermediate at a C-Br bond. The result is *anti* addition, with the Br atoms bonded to opposite sides of the original double bond. When the reaction is carried out in an aqueous or alcoholic solution of the halogen, water or alcohol is the nucleophile in the second step and the product is the bromo alcohol or ether.



(8)

See HALOGENATED HYDROCARBON.

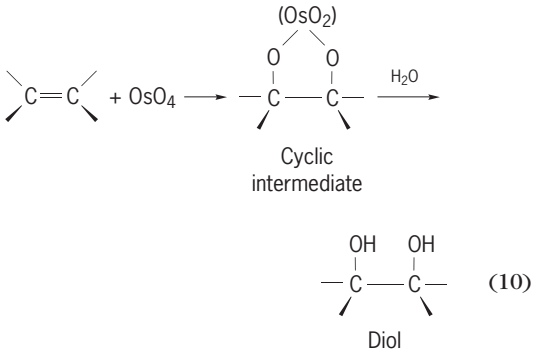
Oxidation. Addition of oxygen to a double bond gives an epoxide. This reaction is carried out on an industrial scale by passing an alkene and air over a silver (Ag) catalyst. In smaller-scale work, epoxides can be obtained by treating an alkene or other unsaturated compound with a peroxide or peracid [reaction (9)].



(9)

See EPOXIDE.

Alkenes can also be oxidized to diols. This reaction takes place when a transition-metal oxide such as permanganate (MnO_4) or osmium tetroxide adds to the double bond. A cyclic intermediate results, and this is hydrolyzed to give the diol [reaction (10)].



(10)

Hydroboration. Alkenes combine readily with borane (BH_3). The resulting alkyl boranes are highly reactive compounds and are useful and versatile reagents in organic synthesis. *See* HYDROBORATION.

Free radicals. Double bonds undergo addition reactions by attack of free radicals, X·. Some of the important radicals are obtained from HX compounds such as hydrogen bromide (HBr), hydrogen sulfide (H_2S), a thiol (HSR), and chloroform ($HCCl_3$). The process is initiated by homolysis of HX, often with air [reaction (11), where X=HBr, HSR, or $HCCl_3$]. A chain reaction then begins (propagation), with X· adding to the π bond [reaction (12)]. Subsequent reaction of

the intermediate radical with another HX molecule gives a product and a new X· radical to continue the chain [reaction (13)]. The direction of addition

Initiation:

$$HX \xrightarrow{O_2} X\cdot + HO_2\cdot \qquad (11)$$

Propagation:

$$RHC{=}CH_2 + X\cdot \longrightarrow R\dot{C}HCH_2X \qquad (12)$$

$$R\dot{C}H\,CH_2X + HX \longrightarrow RCH_2CH_2X + X\cdot \qquad (13)$$

is due to stabilization of the electron-deficient radical by alkyl substituents, just as in the case of a carbenium ion [reaction (6)]. When X = Br, it will be noted that addition by the radical mechanism [reactions (12) and (13)] occurs in the direction opposite to that in the ionic mechanism. The observation and clarification of this reversal of the addition of HBr was a milestone in the understanding of how reactions occur. *See* CHAIN REACTION (CHEMISTRY).

The free-radical intermediate in reaction (12) can react with HX as in reaction (13), but it can also react with another alkene molecule, or several molecules, before being capped. This possibility is shown in reaction scheme (14) for the free-radical addition of

$$CCl_4 \xrightarrow{h\nu} Cl\cdot + \cdot CCl_3 \xrightarrow{H_2C=CH_2} CCl_3CH_2CH_2\cdot \qquad (14)$$

$$CCl_3CH_2CH_2\cdot \xrightarrow{n-1\ H_2C=CH_2} CCl_3(CH_2CH_2)_n\cdot \xrightarrow{CCl_4} CCl_3(CH_2CH_2)_nCl$$

carbon tetrachloride (CCl_4) to ethylene. This reaction results in a mixture of products; the main products have the following yields; $n = 1$, 9%; $n = 2$, 57%; and $n = 3$, 24%. The overall process is called telomerization. *See* FREE RADICAL.

Reactions of metal complexes. Among the most important properties of alkenes is the formation of π complexes by donation of an electron pair to the coordination shell of a transition metal. This is a very general reaction, and the complexes are central to some of the most important chemistry, including industrial processes, involving alkenes. Chief among these is the production of polyethylene and polypropylene by polymerization of the alkenes with Ziegler-Natta catalysts, in which monomer units are incorporated into the chain at a coordination site on a transition metal such as titanium(III). *See* COORDINATION COMPLEXES; ORGANOMETALLIC COMPOUND; POLYMERIZATION; STEREOSPECIFIC CATALYST.

In complex formation, electron density shifts from the π bond to the metal, and the coordinated alkene is therefore susceptible to addition of nucleophiles such as an OH or NH group. Palladium (Pd) is a particularly versatile metal in several reactions. In the Wacker process for the production of acetaldehyde, ethylene is complexed with Pd^{II} [reaction (15)]. After addition of water, the complex undergoes β-elimination of enol, and palladium is thereby reduced

$$Pd^{2+} + \begin{matrix}CH_2\\ \|\\ CH_2\end{matrix} \longrightarrow Pd\!\!\begin{matrix}CH_2\\ \|\\ CH_2\end{matrix} \xrightarrow{H_2O} \begin{matrix}H-CHOH\\ |\\ Pd^+-CH_2\end{matrix}$$

$$\xrightarrow{\beta\text{-elimination}} \begin{matrix}H\\ |\\ Pd\end{matrix}\!\!\begin{matrix}CHOH\\ \|\\ CH_2\end{matrix} \longrightarrow \begin{matrix}O\\ \|\\ CH_3CH\end{matrix} \qquad (15)$$

$$\begin{matrix}H\\ |\\ Pd\end{matrix}\!\!\begin{matrix}CHOH\\ \|\\ CH_2\end{matrix} \xrightarrow{-H^+} Pd^0 \xrightarrow{\frac{1}{2}O_2,\ Cu^{2+}} Pd^{2+}$$

Enol Acetaldehyde

to Pd^0. It is reoxidized to Pd^{II} by copper ion (Cu^{2+}), which is in turn reoxidized by oxygen (air). The net reaction is thus a two-electron oxidation of ethylene, with both palladium and copper required in only catalytic amounts. Analogous reactions in which alcohol or acetic acid are used rather than water produce vinyl ether and vinyl acetate, respectively. Extension of this reaction to higher 1-alkenes provides a convenient and practical method to prepare methyl ketones [reaction (16)].

$$RHC{=}CH_2 \xrightarrow[H_2O]{Pd^{II},\ air} \begin{matrix}O\\ \|\\ RCCH_3\end{matrix} \qquad (16)$$

Methyl ketone

Another example of the utility of palladium π complexes is the substitution of an aryl (Ar) or vinyl group for a vinyl hydrogen in an alkene [reaction (17)]. In this reaction the aryl group is transferred

$$ArX + \begin{matrix}H\\ |\\ RC{=}CR_2\end{matrix} \xrightarrow{Pd^{II}} \begin{matrix}Ar\\ |\\ RC{=}CR_2\end{matrix} + HX \qquad (17)$$

from the metal to the complexed alkene, followed by β-elimination of the product.

Hydroformylation is a major industrial reaction in which an alkene is converted to an aldehyde containing an additional carbon. The alkene is complexed to cobalt (Co) in the form of $HCo(CO)_3$. Transfer of hydrogen from the metal is followed by insertion of carbon monoxide (CO) and finally reduction of a Co-CO bond [reaction (18)]. *See* HYDROFORMYLATION.

A different type of organometallic chemistry underlies a process known as olefin metathesis, in which the groups attached to the double bonds in two alkenes are scrambled. The reaction proceeds by way of metal carbene complexes of molybdenum or tungsten (W), and the reversible formation of metallacyclobutanes [reaction (19); L = CO]. In this way an alkene with a double bond in the center of a chain can be converted by reaction with ethylene to two new terminal alkenes.

Allylic compounds. The position adjacent to a double bond is termed allylic, and a group or atom at that position is strongly affected by the π bond. Thus in a radical or an ionic intermediate, the electron deficiency or excess is delocalized by conjugation with the p orbitals (**Fig. 2**). As a consequence, these species are lower in energy and more easily formed than those in a saturated molecule. For this reason, several reactions of alkenes occur by

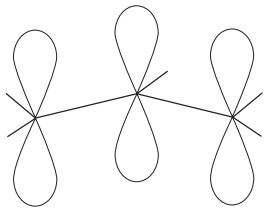


Fig. 2. Molecular orbital representation of a reactive intermediate, with delocalization by conjugation with the *p* orbitals.

substitution of the allylic position rather than addition at the double bond. Examples are the catalytic oxidation of propene to acrylic acid or acrylonitrile, or free-radical chlorination to allyl chloride [reaction (20)]. *See* DELOCALIZATION.

Reactions such as hydrolysis of an allylic type chloride often occur with rearrangement, since the allylic carbenium ion can be solvated at either end of the allylic system. An example is 1-chloro-2-butene, which gives mainly 1-buten-3-ol and a small amount of 2-buten-1-ol. The other chloride gives the same mixture of alcohols [reaction (21), where the positive charge and the π bond in the intermediate are delocalized over two carbon atoms].

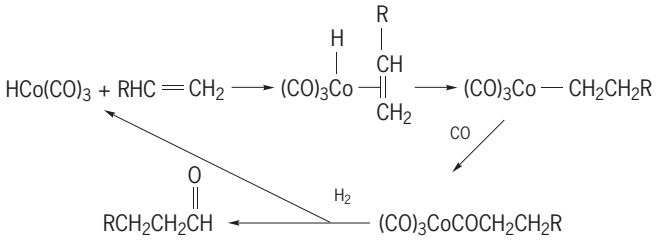
Allylic compounds readily form π complexes with palladium, nickel, and other metals. These π-allyl complexes, in which the metal is complexed to three carbons, provide selective activation of the allylic system for C-C bond formation with a nucleophilic carbon, as in reaction (22) where Nu represents the nucleophile and X = Cl, RO, or ROC = O.

Dienes. Two double bonds, for example in a pentadiene, can exist in three arrangements: cumulative or allenic, conjugated, and nonconjugated, as in 2,3-, 1,3-, and 1,4-pentadiene, respectively. The 2,3-isomer is an allene, for example, 2,3-pentadiene ($CH_3HC{=}C{=}CHCH_3$); 1,3-pentadiene ($H_2C{=}CH\text{-}CH{=}CHCH_3$) is conjugated; 1,4-pentadiene ($H_2C{=}CHCH_2CH{=}CH_2$) is nonconjugated. In 2,3-pentadiene the central carbon has *sp* hybrid orbitals and is susceptible to attack by either electrophilic or nucleophilic reagents. A disubstituted allene such as 2,3-pentadiene is a chiral molecule and exists in enantiomeric forms (**Fig. 3**).

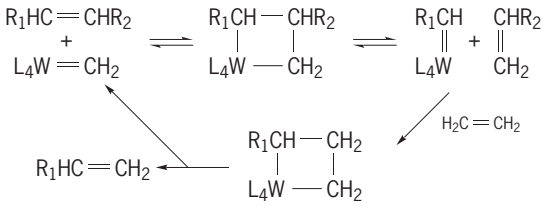
The characteristic feature of a conjugated diene is 1,4-addition. In the reaction of 1,3-pentadiene with bromine, electrophilic attack gives an allylic cation which can lead to either 1,4- or 1,2-addition [reaction (23)]. The major product generally results from the former.

Another important aspect of conjugated diene chemistry is a group of reactions that take place simply by shifts of σ and π bonds. These include ring opening of cyclohexadienes, Diels-Alder cycloadditions, and double-bond (sigmatropic) rearrangements, referred to collectively as pericyclic reactions. *See* CONJUGATION AND HYPERCONJUGATION; DIELS-ALDER REACTION; PERICYCLIC REACTION; WOODWARD-HOFFMANN RULE.

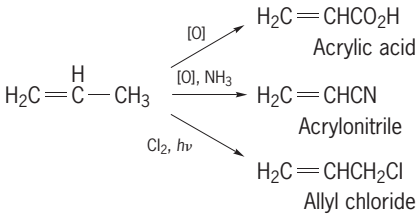
Fig. 3. Enantiomeric forms of the disubstituted allene 2,3-pentadiene. The bold dot represents the central carbon atom.



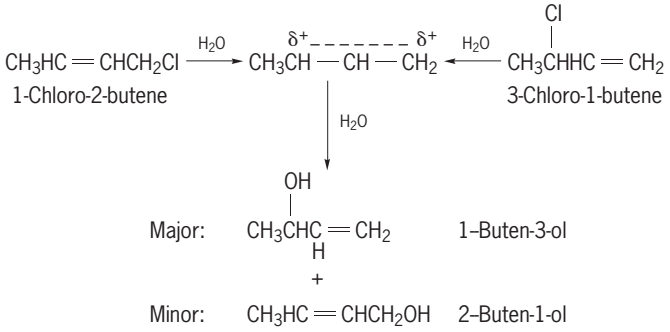
(18)



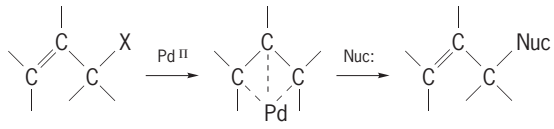
(19)



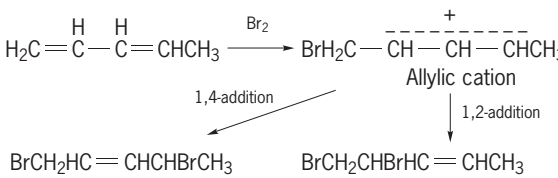
(20)



(21)



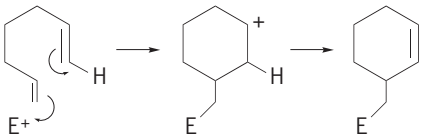
(22)



(23)

In nonconjugated dienes the double bonds can react independently, but with a 1,5- or 1,6-diene electrophilic attack at one double bond may be

accompanied by cyclization [reaction (24), where E

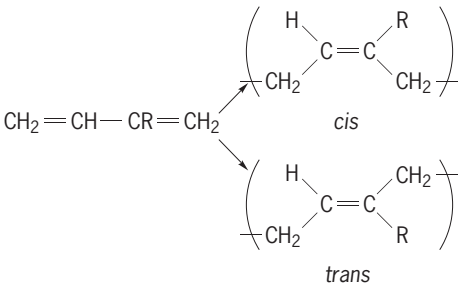


(24)

represents the electrophile]. This process frequently occurs in the biosynthesis of steroids and terpenoids.

Dienes form complexes of several kinds with transition metals. With Pd^{II} salts, butadiene forms π-allyl complexes. Complexes of iron with conjugated dienes are coordinated to four atoms. Typical compounds are linear or cyclic diene–iron tricarbonyl complexes. The elusive cyclobutadiene forms stable complexes with iron and nickel; the bonding is delocalized to achieve an 18-electron count around the metal, as in metallocenes. *See* METALLOCENES.

Metal complexes are involved in the oligomerization of dienes to cyclic dimers and trimers. Polymerization of 1,3-dienes occurs by 1,4-addition and can lead to either *cis*- or *trans*-linked chains [reaction (25),where R═CH_3, the 1,3-diene is isoprene,



(25)

and where R═Cl, the 1,3-diene is chloroprene]. The stereochemistry can be controlled by the choice of catalysts. These polymers represent the structures of elastomeric materials. *Cis*-1,4-(poly)isoprene obtained by polymerization with R_3Al-$TiCl_4$ catalysts is the basic unit of natural rubber from *Hevea* species. The *trans* polyisoprene is obtained with other R_3Al catalysts and is the repeating unit of gutta-percha. Addition of small amounts of *cis*-1,4(poly)butadiene significantly improves the wear resistance of natural rubber. Neoprene, the polymer of 2-chlorobutadiene (chloroprene), has several advantageous features. *See* RUBBER.

Polyenes. Compounds with three or more double bonds are found in numerous naturally occurring sources. The long-chain (C_{12}–C_{20}) carboxylic acids present as esters in fats and other lipids typically contain two or three nonconjugated double bonds. The most conspicuous polyenes are carotenoids, the red-yellow pigments that occur abundantly in plants and many other organisms. These compounds are typically C_{40} molecules with an extended system of conjugated *trans* double bonds. Vitamin A is a pentaene derived in the organism by cleavage of β-carotene. Reversible *cis-trans* isomerization of one double bond plays a central role in color vision. *See* CAROTENOID; VITAMIN A.

A very long polyene chain is obtained by polymerization of acetylene with an organometallic catalyst. This material can be treated to give an electrically conducting polymer. *See* ORGANIC CONDUCTOR.

James A. Moore

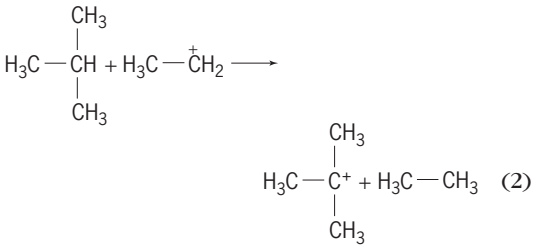
Bibliography. A. Streitwieser, C. Heathcock, and E. M. Kosower, *Introduction to Organic Chemistry*, 4th ed. 1992.

Alkylation (petroleum)

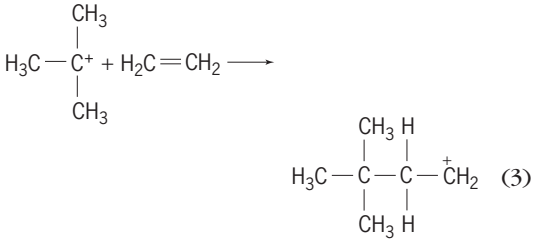
In the petroleum industry, a chemical process in which an olefin (ethylene, propylene, and so forth) and a hydrocarbon, usually 2-methylpropane, are combined to produce a higher-molecular-weight and higher-carbon-number product. The product has a higher octane rating and is used to improve the quality of gasoline-range fuels. The process was originally developed during World War II to produce high-octane aviation gasoline. Its current main application is in the production of unleaded automotive gasoline. *See* GASOLINE; OCTANE NUMBER.

Reactions. The alkylation reaction is initiated by the addition of a proton (H^+) to the olefin [reaction (1)]. The protonated olefin (carbonium reactions) then reacts with the isobutane by abstraction of a proton from the isobutane to produce the *t*-butyl carbonium reactions [reaction (2)]. Reaction of this tertiary carbonium reactions with the olefin proceeds by combination of the two species [reaction (3)] to produce a more complex six-carbon carbonium reation which yields a stabilized product by abstraction of a proton from another molecule of isobutane [reaction (4)]. The reaction progresses to more product by reaction of the *t*-butyl and carbonium reactions, as already shown in reaction (3). *See* REACTIVE INTERMEDIATES.

(1)

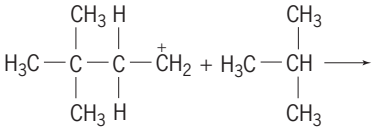


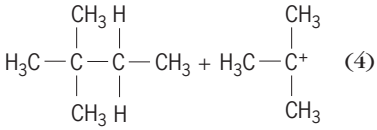
(2)



(3)

Processes. In actual refinery practice, the feedstock for the alkylation process is isobutane (or an isobutane-enriched stream) recovered from refinery





(4)

Alkylate

gases or produced by isobutane isomerization. The olefin for the feedstock is derived from the gas production of a catalytic cracker. *See* CRACKING; ISOMERIZATION.

The acid catalyst is sulfuric acid (H_2SO_4), hydrofluoric acid (HF), or aluminum chloride ($AlCl_3$). When sulfuric acid is employed, the reaction must be maintained at 35–45°F (2–8°C) to reduce unnecessary side reactions; with hydrogen fluoride and with aluminum chloride, refrigeration is not necessary, and temperatures up to 115°F (45°C) may be employed.

Sulfuric acid is used with propylene and higher-carbon-number olefins but not with ethylene because of the tendency to produce the acid-wasting ethyl hydrogen sulfate ($C_2H_5HSO_4$). Hydrogen fluoride is more generally used, and has the advantage of being more readily separated from the products because of a boiling point below that of sulfuric acid. Aluminum chloride is used to a lesser extent than both of the acids and requires injection of water to produce the hydrogen chloride promoter. *See* CATALYSIS.

Thermal alkylation is a noncatalyzed process that requires temperatures on the order of 950±25°F (about 510°C) and pressures in the range 3000–8000 lb/in.^{2} (18–55 megapascals). The reaction is believed to proceed by way of a (neutral) free-radical intermediate and is much less specific than the acid-catalyzed process. The starting hydrocarbon is usually isobutane, and the olefin is ethylene, propylene, or higher olefins such as butenes. The nature of the reactants influences the quality of the alkylate. For example, alkylation of 1-butene with isobutane gives an alkylate with a research octane rating of about 93, whereas using 2-butene as the olefin yields an alkylate having a research octane number of about 98. *See* ALKANE; ALKENE; FREE RADICAL; PETROLEUM PROCESSING AND REFINING.

James G. Speight

Bibliography. S. Glasstone, *Energy Deskbook*, 1983; J. McKetta, *Petroleum Processing Handbook*, 1992; J. G. Speight, *The Chemistry and Technology of Petroleum*, 3d ed., 1999.

Alkyne

A hydrocarbon that contains a triple carbon bond, —C≡C—. The first compound in the series is acetylene. Names of alkynes are formed by changing the -ane ending of the corresponding alkane to -yne and numbering as needed, for example, 2-pentyne (structure 1), the 2 indicating that the triple bond is at the second carbon atom of the chain. Some compounds can be named as a substituted acetylene, for example, diphenylacetylene (**2**); or the ethynyl group (—C≡CH) can be named as a substituent, for example, 1-ethynylcyclopentanol (**3**).

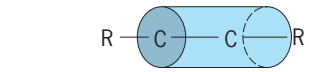
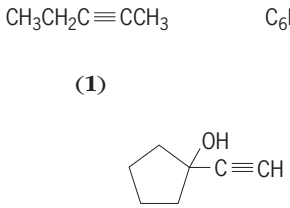


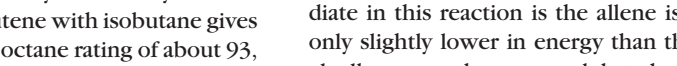
Fig. 1. Geometry of a linear arrangement of triply bonded carbon atoms and two attached groups.



See ACETYLENE; ALKANE.

Structure and properties. A triple bond consists of a σ bond and two π bonds between two *sp* carbon atoms. The carbons are held closer together (0.121 nanometer) and more strongly (836 kilojoules/mole) than those in a double bond. The π bonds are orthogonal, and the electron density is distributed in a cylindrical sheath around the C-C axis. The resulting geometry is a linear arrangement of the triply bonded carbons and the two attached atoms (**Fig. 1**). *See* MOLECULAR ORBITAL THEORY.

The energy difference between a 1-alkyne and 2-alkyne (20 kJ) is larger than that between 1- and 2-alkenes, and a 1-alkyne can be completely isomerized with a base catalyst [reaction (1)]. An intermediate in this reaction is the allene isomer, which is only slightly lower in energy than the 1-alkyne. Cycloalkynes can be prepared, but the requirement of a linear four-atom system for a triple bond places a limit on the minimum size of a cycloalkane ring. Cyclooctyne, with an eight-membered ring, can be isolated; seven- and six-membered cyclic alkynes can be trapped as transient intermediates. In cycloalkynes with medium rings (8–11 carbons) the amount of allene isomer is substantial, since only three atoms are collinear and the cyclic allene in this case is significantly more stable than the cycloalkyne. *See* ISOMERIZATION; REACTIVE INTERMEDIATES.



A distinctive property of acetylene and other 1-alkynes is the acidity of the —C≡C—H bond. Because the electrons in an *sp* carbon orbital are closer to carbon than those in *sp*^{2} or *sp*^{3} orbitals, the anion —C≡C:$^-$ is stabilized. The acidities of a few relevant compounds are given in the **table**. The terminal C-H

Comparison of acidity values of several compounds

Compound	Acidity, pK_A ($-\log K_A$)
Ethane (H_3C-CH_3)	50
Ethylene ($H_2C=CH_2$)	44
Ammonia (NH_3)	35
Acetylene ($HC\equiv CH$)	25
Ethanol (CH_3CH_2OH)	16

in acetylene (and other alkynes) is far more acidic (about 10^{20}–10^{25} times) than hydrogen in alkenes or alkanes. Hydrogen can be removed from an alkyne by reaction with sodium amide, a reaction that is conveniently carried out in liquid ammonia [NH_3; reaction (2)]. Many other bases, such as alkyllithiums, can

$$\underset{\text{Alkyne}}{RC\equiv CH} + NaNH_2 \xrightarrow{NH_3} \underset{\text{Sodium acetylide}}{RC\equiv CNa} + NH_3 \qquad (2)$$

be used to prepare the acetylide anion ($-C\equiv C:^{\ominus}$). *See* PK.

Preparation. One general method for introducing a triple bond into a carbon chain is by elimination reactions, for example, removal of HX (X = halogen) from a vinyl halide [$H_2C=CHX$] with very strong base, or removal of two HBr molecules from a dibromide [reaction (3)].

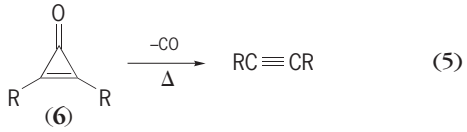
$$\begin{array}{l} -H_2C-CBr_2- \\ -HBrC-CHBr- \end{array} \xrightarrow{-HBr} -HC=CBr- \xrightarrow{-HBr} -C\equiv C- \qquad (3)$$

Another starting point for elimination is an enol ester such as a trifluoromethylsulfonate (**4**) or an enol phosphate (**5**), as in reactions (4). The latter

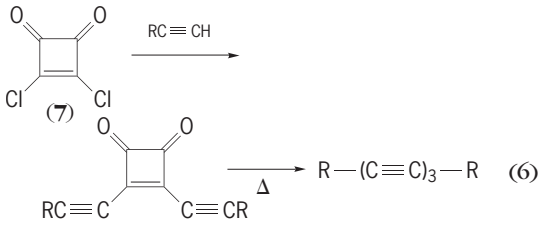
$$\underset{\text{Carbonyl compound}}{-CH_2C(=O)-} \xrightarrow{(CF_3SO_2)_2O} \underset{\textbf{(4)}}{-CH=C(OSO_2CF_3)-} \text{ or } \underset{\textbf{(5)}}{-CH=C(OPO_3H)-} \longrightarrow -C\equiv C- \qquad (4)$$

has been suggested as the precursor of the triple bonds in naturally occurring acetylenes.

Alkynes are also obtained by thermal elimination of neutral molecules, as in the preparation of cycloalkynes by oxidizing the hydrazones of 1,2-diketones. Another reaction is elimination of carbon monoxide (CO) from a cyclopropenone (**6**) [reaction (5)] or cyclobutanedione (**7**) [reaction



(**6**) (5)



(**7**) (6)

(6)]. The latter reaction can be used to build up chains containing several triple bonds.

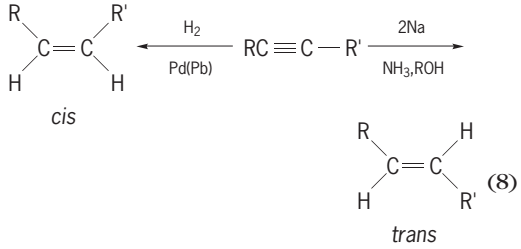
A second general approach for introducing a triple bond is using reactions involving an acetylide anion in either a substitution or addition reaction. The acetylide anion is an excellent nucleophile, similar in reactivity to cyanide ion ($N\equiv C:^{\ominus}$). Any alkyl halide or sulfonate ester that is a practical substrate for substitution can be converted to an alkyne. With a dihalide (**8**) and excess acetylide, a diacetylene (**9**) is available [reaction (7)].

$$\underset{\textbf{(8)}}{Br(CH_2)_nBr} + 2HC\equiv CNa \xrightarrow[\text{Sodium acetylide}]{} HC\equiv C(CH_2)_nC\equiv CH \qquad (7)$$

(**9**)

Addition of acetylides to aldehydes, ketones, or epoxides leads to the corresponding alkynols. An acetylide can be acylated with an acid chloride or carbonated to give the ketone or carboxylic acid, respectively.

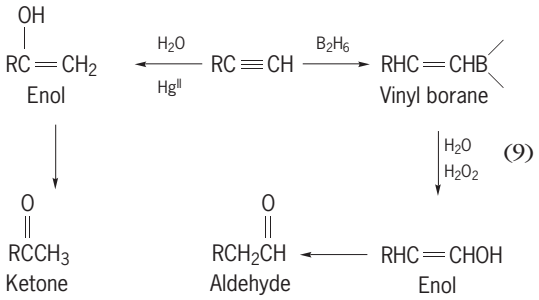
Addition reactions. In general, electrophilic additions are slower with triple bonds than double bonds because π electrons in the latter are more accessible. However, several addition reactions of alkynes are useful. Hydrogenation of a triple bond can be carried out with a palladium (Pd) catalyst that is deactivated ["poisoned" with the addition of lead (Pb)] to permit selective reduction to the *cis*-alkene. Reduction to the *trans*-alkene can be accomplished with sodium metal (Na) in ammonia (NH_3) containing alcohol [reaction (8)].



(8)

Addition of water to a terminal triple bond occurs in the presence of a mercuric (Hg^{II}) salt, leading

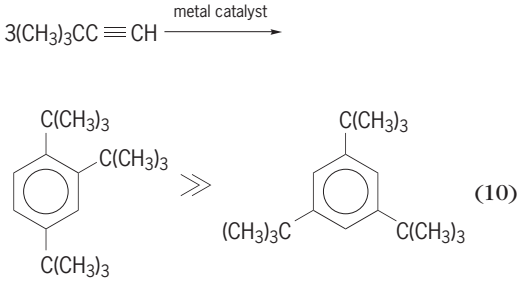
initially to an enol and thence to a ketone. Reaction of the alkyne with borane (B_2H_6) gives a vinyl borane that can be oxidized in the presence of water (H_2O) and hydrogen peroxide (H_2O_2) to the isomeric enol of the aldehyde [reactions (9)]. Dialkyl aluminum



(9)

reagents add to alkynes. The alkenylaluminum products can be converted stereoselectively to alkenes or to alkenyl halides or allylic alcohols.

Cyclization reactions. In early work benzene was observed among the products formed on heating acetylene. Substituted acetylenes are quite stable thermally, but metal-catalyzed thermal trimerization gives trisubstituted benzenes, with the 1,2,4-isomer greatly predominating over the 1,3,5-isomer [reaction (10)].

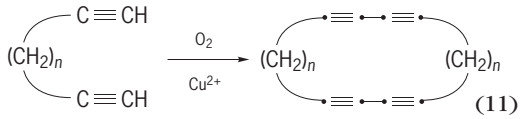


(10)

This bias for the 1,2,4 pattern indicates a mechanism that is not a simple head-to-tail cyclization, and may involve some type of metal-oriented dimeric complex. The reaction is a route to benzenes with adjacent bulky groups.

The formation of cyclooctatetraene in high yield by reaction of acetylene in the presence of a nickel (Ni^{II}) catalyst was a seminal finding in organometallic chemistry. A variety of other cyclizations of alkyne-metal complexes have been observed. In some of these reactions carbon monoxide molecules are also involved, leading to cyclopentadienones and tropones. The cocylization of an alkyne with an alkene and carbon monoxide, catalyzed by dicobalt octacarbonyl [$Co_2(CO)_8$], is a general and flexible method for cyclopentenone preparation. *See* ORGANOMETALLIC COMPOUND.

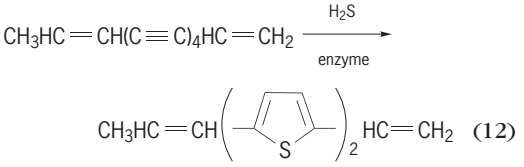
Diynes and polyynes. 1-Alkynes can be coupled by mild oxidation to the conjugated diacetylene. The reaction is carried out by air oxidation of the cuprous (Cu^{I}) acetylide or oxidation of the alkyne with cupric (Cu^{II}) ion in the presence of an amine. For coupling of two different alkynes, the reaction is carried out with the copper derivative of one alkyne and a different 1-bromoalkyne. The yields in these couplings are usually good, and a variety of functional groups, including additional triple bonds, can be present. Oxidative coupling of terminal dialkynes leads to macrocyclic polyacetylenes [reaction (11)]. This is



(11)

a very direct way to prepare large-ring compounds, although rings with different numbers of diyne units and also polymeric products are formed. *See* MACROCYCLIC COMPOUND.

Polyynes occur in many species of higher plants, particularly in the families Compositae, Umbelliferae, and Asteraceae as well as in Basidiomycetes and other fungi. Many hundreds of compounds have been isolated and characterized; nearly all of these are structurally and biogenetically related to fatty acids with unbranched C_{10}–C_{20} chains. A number of compounds are derived from a series of C_{13} hydrocarbons that contain a combination of six triple and double bonds. A common feature is the occurrence of thiophenes arising from addition of the biochemical equivalent of hydrogen sulfide (H_2S) to two adjacent triple bonds [reaction (12)]. Many other functional



(12)

groups such as hydroxyl, epoxy, furan, and pyran rings are also present.

A group of antibiotics that are potent inhibitors of tumor growth have as a common feature a macrocyclic enediyne system. The mechanism of action of these compounds involves rearrangement of the multiple bonds to a benzene diradical which then interferes with deoxyribonucleic acid (DNA) synthesis. *See* ANTIBIOTIC.

Compounds containing an extended sequence of conjugated triple bonds with the general formula —$(C{\equiv}C)_n$— have been detected in interstellar matter as well as in plant sources. Polyyne chains with up to 16 triple bonds (**Fig. 2***a*), and also molecules such

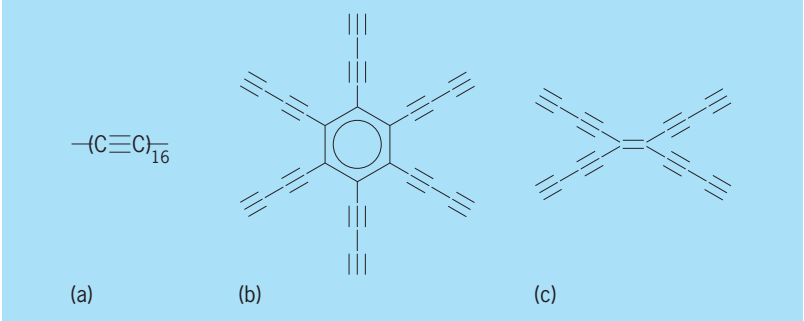


Fig. 2. Compounds containing an extended sequence of conjugated triple carbon bonds: (*a*) polyyne chain with 16 triple bonds; (*b*) hexa(butadiynyl)benzene; and (c) tetrabutadiynyl ethylene.

as hexa(butadiynyl)benzene (Fig. 2*b*) and tetrabutadiynyl ethylene (Fig. 2*c*), have been prepared. These compounds are stable with a bulky group such as trialkylsilyl at the ends of the polyyne chains; but with terminal hydrogen, they are extremely reactive and may ignite in air or explode. Compounds with extended polyyne systems are of interest as precursors of carbon rods or two-dimensional carbon networks, both of which have potential as new materials. James A. Moore

Bibliography. L. Brandsma and H. D. Verkruijsse, *Syntheses of Acetylenes, Allenes and Cumulenes*, 1981; F. Diedrich and Y. Rubin, Synthetic approaches to molecular carbon allotropes, *Angew. Chem. Int. Ed.*, 31:1101–1123, 1992; H. G. Viehe (ed.), *Chemistry of Acetylenes*, 1969.

Allantois

A fluid-filled sac- or sausagelike, extraembryonic membrane lying between the outer chorion and the inner amnion and yolk sac of the embryos of reptiles, birds, and mammals. The allantois eventually fills up the space of the extraembryonic coelom in most of these animals. It is composed of an inner layer of endoderm cells, continuous with the endoderm of the embryonic gut, or digestive tract, and an outer layer of mesoderm, continuous with the splanchnic mesoderm of the embryo. It arises as an outpouching of the ventral floor of the hindgut and dilates like a filling balloon into a large allantoic sac which spreads throughout the extraembryonic coelom. The allantois remains connected to the hindgut by a narrower allantoic stalk which runs through the umbilical cord. *See* AMNION; CHORION; GERM LAYERS.

The allantois eventually fuses with the overlying chorion to form the compound chorioallantois, which lies just below the shell membranes in reptiles and birds. The chorioallantois is supplied with an extensive network of blood vessels and serves as an important respiratory and excretory organ for gaseous interchange. The allantoic cavity also serves as a reservoir for kidney wastes in some mammals, in reptiles, and in birds. In the latter two groups the allantois assists in the absorption of albumin. In some mammals, including humans, the allantois is vestigial and may regress, yet the homologous blood vessels persist as the important umbilical arteries and veins connecting the embryo with the placenta. *See* FETAL MEMBRANE; PLACENTATION. Nelson T. Spratt, Jr.

Bibliography. B. I. Balinsky, *Introduction to Embryology,* 5th ed., 1981.

Allele

Any of a number of alternative forms of a gene. Allele is a contraction of allelomorph, a term which W. Bateson used to designate one of the alternative forms of a unit showing mendelian segregation. New alleles arise from existing ones by mutation. The diversity of alleles produced in this way is the basis for hereditary variation and evolution. *See* GENE; MENDELISM; MUTATION.

The genetic material is a coded set of instructions in chemical form (deoxyribonucleic acid) for making a living cell. Each gene is like an essential word in the code. The different alleles of a given gene determine the degree to which the specific hereditary characteristic controlled by that gene is manifested. The particular allele which causes that characteristic to be expressed in a normal fashion is often referred to as the wild-type allele. Mutations of the wild-type allele result in mutant alleles, whose functioning in the development of the organism is generally impaired relative to that of the wild-type allele. In this analogy, the wild-type allele corresponds to the correct word in the genetic code and the mutant alleles correspond to mistakes in the spelling of that word. *See* DEOXYRIBONUCLEIC ACID (DNA); GENETIC CODE.

An allele occupies a fixed position or locus in the chromosome. In the body cells of most higher organisms, including humans, there are two chromosomes of each kind and hence two alleles of each kind of gene, except for the sex chromosomes. Such organisms and their somatic cells are said to carry a diploid complement of alleles. A diploid individual is homozygous if the same allele is present twice, or heterozygous if two different alleles are present. Let *A* and *a* represent a pair of alleles of a given gene; then *A/A* and *a/a* are the genetic constitutions or genotypes of the two possible homozygotes, while *A/a* is the genotype of the heterozygote. Usually the appearance or phenotype of the *A/a* individuals resembles that of the *A/A* type; *A* is then said to be the dominant allele and *a* the recessive allele. In the case of the sex chromosomes, one sex (usually the male in most higher animals, with the exception of birds) has only one X chromosome, and the Y lacks almost all of the genes in X. The male thus carries only one dose of X-linked genes and is said to be hemizygous for alleles carried on his X chromosome. As a result, if a male inherits a recessive mutant allele such as color blindness on his X chromosome, he expresses color blindness because he lacks the wild-type allele on his Y chromosome. *See* CHROMOSOME; SEX-LINKED INHERITANCE.

The wild-type allele of a gene is symbolized by attaching a superscript plus sign to the gene symbol; for example, a^+ is the wild-type allele of the recessive mutant gene *a*. When no confusion arises, the symbol + is used instead of a^+. This means that $a/+$ is equivalent to a/a^+. In some cases a mutant is dominant to the wild-type allele, and then the mutant symbol is usually capitalized.

In a population of diploid individuals, it is possible to have more than two alleles of a given gene. The aggregate of such alleles is called a multiple allelic series. Since genes are linear sequences of hundreds or even thousands of nucleotide base pairs, the potential number of alleles of a given gene which can arise by base substitution alone is enormous. For example, a gene coding for a protein with only 100 amino acids would be considered a very small gene, and would have 300 base pairs in its coding region; yet, since each base pair can exist in four different chemical

forms, 4^{300} possible mutant alleles could arise within the coding region alone, or approximately 10^{180} alleles—a number vastly larger than the estimated number of particles in the known universe. Actually, many other types of mutant alleles of a gene do occur, including duplications, deletions, and inversions of DNA within the noncoding as well as coding regions of the gene. Indeed, many spontaneous mutations in at least some higher organisms, such as yeast, maize, and *Drosophila,* turn out to be insertions of large segments of DNA into either the noncoding or coding regions. These insertions, known as transposons, appear to be virallike elements that have infected the germ line. *See* TRANSPOSONS.

Mutations of independent origin that involve changes at the same nucleotide positions in a gene are known as homoalleles, while mutations that involve changes at different nucleotide positions are called heteroalleles. Recombination between alleles of the same gene does not occur or tends to be extremely rare in higher organisms. Also, recombination within the gene in such cases tends to be different in outward properties from ordinary crossing over and is known as gene conversion. *See* RECOMBINATION (GENETICS).

It is often difficult to distinguish operationally between a series of multiple alleles of a single gene and a cluster of closely linked genes with similar effects. Such a cluster, originally known as a pseudoallelic series, is now termed a gene complex, gene cluster, or multigene family, except that, in the latter case, the genes of the family are not necessarily closely linked. In bacteria, gene clusters are usually coordinately regulated and are termed operons. Gene cloning has verified the existence of gene clusters in higher organisms as well as in bacteria. *See* BACTERIAL GENETICS; COMPLEMENTATION (GENETICS); GENETICS; OPERON.

Edward B. Lewis

Allelopathy

The biochemical interactions among all types of plants, including microorganisms. The term is usually interpreted as the detrimental influence of one plant upon another but is used more and more, as intended originally, to encompass both detrimental and beneficial interactions.

Forms and occurrence. At least two forms of allelopathy are distinguished: (1) the production and release of an allelochemical by one species inhibiting the growth of only other adjacent species, which may confer competitive advantage for the allelopathic species; and (2) autoallelopathy, in which both the species producing the allelochemical and unrelated species are indiscriminately affected.

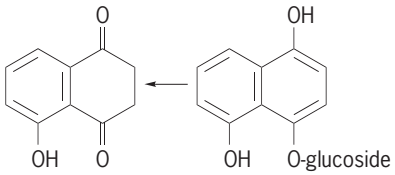
The term allelopathy, frequently restricted to interactions among higher plants, is now applied to interactions among plants from all divisions, including algae. Even interactions between plants and herbivorous insects or nematodes in which plant substances attract, repel, deter, or retard the growth of attacking insects or nematodes are considered to be allelopathic.

Interactions between soil microorganisms and plants are important in allelopathy. Fungi and bacteria may produce and release inhibitors or promoters. Through the secretion of siderophores they can immobilize iron, which may deprive harmful rhizobacteria and pathogens of this important trace element. Siderophores from pseudomonad bacteria, however, may reduce iron uptake by plants, whereas siderophores from agrobacteria may enhance it. Some bacteria enhance plant growth through fixing nitrogen, others through providing phosphorus. The activity of nitrogen-fixing bacteria may be affected by allelochemicals, and this effect in turn may influence ecological patterns. Allelochemicals, released from plant litter by microbial action, are often also removed from the rhizosphere by microorganisms. The rhizosphere must be considered the main site for allelopathic interactions. *See* NITROGEN FIXATION; RHIZOSPHERE.

Chemicals produced and secreted by microorganisms that inhibit other microorganisms are called antibiotics. In principle, this is a form of allelopathy. Vascular plants react to microbial infection as well as to abiotic elicitors by synthesizing phytoalexins, which are low-molecular-weight, broad-spectrum antimicrobial compounds, another form of allelopathy. In contrast to phytoalexins, allelochemicals are produced without the action of an elicitor. *See* ANTIBIOTIC; PHYTOALEXINS; SOIL MICROBIOLOGY; TOXIN.

Allelopathy is clearly distinguished from competition: in allelopathy a chemical is introduced by the plant into the environment, whereas in competition the plant removes or reduces such environmental components as minerals, water, space, gas exchange, and light. In the field, both allelopathy and competition usually act simultaneously, and assessing their individual influence and function is a continuing challenge. In controlled laboratory experiments with extracts or exudates from plants suspected of allelopathy, bioassays often yield significant allelopathic effects that are not manifest in nature. Soil and weather conditions, and the influence of surrounding plants and microbial interactions, carefully controlled in the laboratory, may mask allelopathic effects in the field. The unequivocal demonstration of allelopathy in the field as a result of the action of chemical compounds showing allelopathic activity in a laboratory bioassay is difficult because of the complexity and dynamic nature of the environmental interactions.

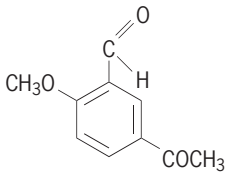
Allelopathy was first observed in ancient times: Pliny (A.D. 23–79) related the inhibition of plant growth around walnut trees. The black walnut tree *(Juglans nigra)* produces a potent allelochemical, juglone (5-hydroxynaphthoquinone; structure **1**); that inhibits growth of many annual



(**1**) (**2**)

herbs. Tomato and alfalfa under or near black walnut trees wilt and die. Alder *(Alnus glutinosa),* planted as nurse trees in black walnut orchards, decline and die from allelopathy within 8–13 years. Water leachate from leaves and extracts from the soil under a walnut tree contain juglone. Adding juglone to plant beds inhibits or kills the plants. Juglone is not, however, a general toxin for all plants; Kentucky bluegrass *(Poa pratensis)* and brambles such as wild raspberries and blackberries *(Rubus* spp.) grow under and around black walnut trees. In the cells of walnut trees, juglone is present in a safe form as 1,4,5-trihydroxynaphthalene glucoside (**2**), which is not inhibitory. Upon leaching from the leaves and bark by rain or release from the roots, the glucoside is hydrolyzed and the free hydroxynaphthalene is oxidized to the active allelochemical, juglone. The release of inhibitors, for example, hydrocyanic acid (HCN), from harmless precursors, such as cyanogenic glycosides, occurs frequently in allelopathy.

Desert plants. For plants growing in habitats with extreme climates, such as a desert, competition for the limited resources is critical, and allelopathy may have survival value. Desert shrubs are often surrounded by a bare zone; thus, all the moisture of that zone remains available to the shrub and is not shared with other plants. In the Mojave Desert of California incienso *(Encelia farinosa)* inhibits the growth of desert annuals, which will grow only around dead incienso shrubs. From the decomposing leaf litter, 5-acetyl-1-2-methoxybenzaldehyde (**3**) is released and



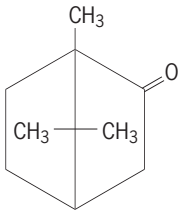
(**3**)

persists in the desert soil, functioning as an allelochemical. Incienso is apparently not affected by its own toxin. *See* DESERT.

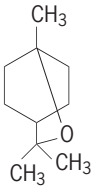
Chaparral. Important allelochemicals in chaparrals are terpenoids and water-soluble compounds, such as organic acids and phenols.

Terpenoids. The obvious and striking vegetation pattern of bare ground and stunted herbs and grasses around the shrub thickets of sagebrush *(Salvia leucophylla)* and wormwood *(Artemisia californica)* in the Californian chaparral is caused by an allelopathic effect of the volatile terpenes produced by these shrubs. *Salvia* and *Artemesia* are surrounded by a zone of bare soil 3–6 ft wide (1–2 m), followed, toward the periphery, by a zone of annual herbs and grasses with stunted growth and limited development and, finally, thriving grassland with wild oats *(Avena fatua),* brome grass *(Bromus rigidus),* and other grass species. Careful analyses have eliminated shade, nutrient deficiency, water availability, root density, slope of ground, and animal feeding and running activity as the causes for growth inhibition. While grazing by small mammals, birds, and insects—all taking advantage of the shrub for shelter—helps to keep down growth, grazing does not initiate or usually maintain the zones of inhibition. *See* HERBIVORY.

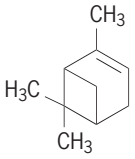
The uphill and downhill inhibition zones of *Salvia* shrubs are about equal in size, indicating that water-soluble inhibitors are unlikely in the dry climate of the chaparral. This suggests the effect of volatile compounds. Volatile terpenes from *Salvia* and *Artemisia* and other shrubs and their surrounding soil were identified and found to be inhibitory to seed germination and plant growth. Camphor (**4**) and 1,8-cineole (**5**) are the most effective inhibitors; α-pinene (**6**), β-pinene (**7**), camphene (**8**), and thujone (**9**) are



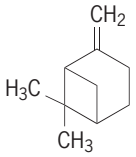
(**4**)



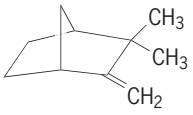
(**5**)



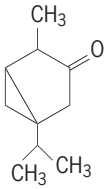
(**6**)



(**7**)



(**8**)



(**9**)

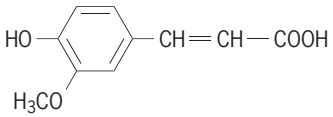
also active volatile allelochemicals. Growth inhibition is most obvious around shrubs growing in the fine-textured Zaca clay, which absorbs the terpenes and thus facilitates contact with the roots.

Spectacular fires, which sweep the chaparral approximately every 25 years, volatilize the soil-bound terpenes and destroy the shrubs. The burned sites quickly turn green with annual herbs and grasses that dominate for approximately 6 years until the shrubs grow back from surviving underground parts. With the accumulation of terpenes in the soil, the zones of bare ground and stunted growth reappear and remain apparent until the next fire.

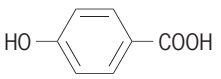
Terpenoids are common plant products and might be expected to serve as allelochemicals in other plant species. Indeed, for eucalyptus trees and sassafras, which do not inhabit the chaparral, terpenoids have been identified as effective allelochemicals; in other

plants, especially conifers, terpenoids are suspected of acting as such also. *See* TERPENE.

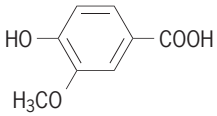
Water-soluble allelochemicals. Chamise *(Adenostema fasciculatum)* and manzanita *(Arctostaphylos glandulosa),* two common chaparral shrubs species, lack volatile terpenes, but because growth around them is also inhibited, other allelochemicals may be involved. Ferulic acid (**10**), *p*-hydroxybenzoic acid (**11**), and vanillic acid (**12**) are found in their leaves



(**10**)



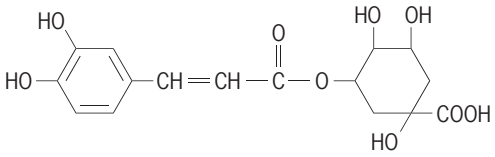
(**11**)



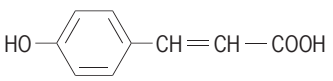
(**12**)

as well as in the surrounding soil and act as strong inhibitors of seed germination and growth. Other phenols and phenolic acids are present in the leaves but are not always detected in the soil. Rainfall in the chaparral is only moderate. The coastal fog, which prevails throughout the year, however, supplies sufficient moisture to leach the phenolic compounds into the soil, rendering them effective allelochemicals. *See* CHAPARRAL.

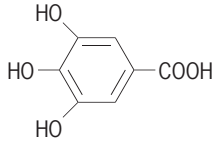
Old-field succession. Allelopathy has been studied in the four stages of old-field succession of the Oklahoma-Kansas prairie. The pioneer weed stage, lasting only 2–3 years, is characterized by sturdy weeds with low mineral needs, such as sunflower *(Helianthus annuus)*, ragweed *(Ambrosia psilostachya)*, Johnson grass *(Sorghum halepense)*, crab grass *(Digitaria sanguinalis)*, and spurge *(Euphorbia supina)*. Autoallelopathy of the pioneer weeds accelerates their own replacement by the dominant species of the second stage: namely the triple-awn grass *(Aristida oligantha)*, which lasts about 9–13 years. Tolerant of the pioneer weed-produced allelochemicals, such as chlorogenic acid (**13**), *p*-coumaric acid (**14**), gallic acid (**15**), and *p*-hydroxybenzaldehyde (**16**), e-awn grass also



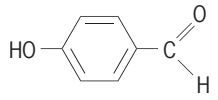
(**13**)



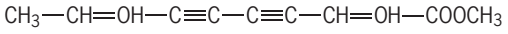
(**14**)



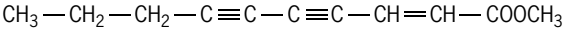
(**15**)



(**16**)



(**17**)



(**18**)

has low mineral requirements. Allelochemicals produced from triple-awn grass, such as gallic acid and tannic acid, inhibit free-living and symbiotic nitrogen-fixing bacteria, thus giving the grass a selective advantage over invading species of the third stage that require higher nitrogen concentrations. Inhibition of the nitrifying bacteria by allelochemicals produced by the invading species, in turn, controls the rather slow succession of the third and fourth stages. Gradually, ammonia in the soil increases to such a level that species of the third stage, perennial bunchgrasses—lasting some 30 years—and finally, the climax prairie, can slowly claim the field.

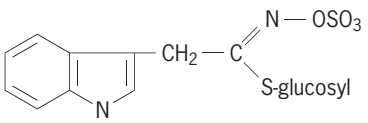
In Japanese urban wasteland, weed succession, which is similar to that of old fields, appears to be controlled by polyacetylene-type allelochemicals. Matricaria esters (**17**) and lachnophyllum ester (**18**) in the pioneer weeds, *Solidago altissima* and *Erigeron* species, have also been identified in the soil, suggesting ecological significance. *See* ECOLOGICAL SUCCESSION.

Aqueous habitats. Allelopathy is not restricted to terrestrial habitats but occurs also in fresh water and marine environments. The spikerush (*Eleocharis coloradoensis*) is an aquatic plant that has been shown to produce an allelochemical effect. Allelopathic substances in saltwater are of special interest with regard to the red tide, that is, blooms of microorganisms that discolor the water and discharge toxins. Algae-produced allelochemicals may be able to limit the growth of red tide producers. *Nannochloris*, an alga, synthesizes a cytolytic agent affecting a red tide–causing dinoflagellate. Other algae have been found to secrete fatty acids that inhibit algal growth.

Agriculture. The reduction of crop yields by weed competition is aggravated by the allelopathic effect of weeds on crops. It was already noted in 1832 that thistles *(Cirsium)* diminished the yield of oats, and spurge *(Euphorbia)* diminished the yield of flax. The growth inhibition of wheat by couch grass *(Agropyron repens)* results from impaired phosphate uptake,

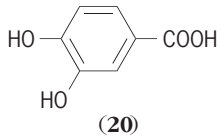
even in the presence of adequate phosphate supply, and results in severe crop losses. Leachates from Italian ryegrass *(Lolium multiflorum)*, its litter, and the soil in which it grows inhibit the growth of oat, clover, lettuce, and bromegrass. Foxtail species *(Setaria)* reduce the growth of corn by as much as 35% through allelopathy. The combined effects of foxtail allelopathy and competition can result in 90% growth reduction.

Crop plants may inhibit their own growth and reduce the yield of subsequent crops, an observation well known to farmers as soil sickness long before the investigations of allelopathy. Crop rotation cures soil sickness only when the subsequent crop is not affected by the accumulated allelochemicals of the previous crop or when they have been detoxified by soil microorganisms. Cabbage, rich in thiocyanates, inhibits grass species, tobacco, and grapes. Woad *(Isatis tinctoria)* contains high levels of glucosinolates (mustard oil glycosides), especially glucobrassicin (**19**), and its planting is there-



(**19**)

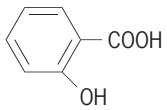
fore often avoided because of its ill effects on succeeding crops. Straw from wheat, corn, rye, rice, or sorghum decomposing in the field yields phenolic compounds, such as vanillic acid (**12**), ferulic acid (**10**), *p*-hydroxybenzoic acid (**14**), *p*-coumaric acid (**13**), *o*-hydroxyphenylacetic acid, and protocatechuic acid (**20**), all of which exert allelopathic ef-



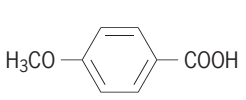
(**20**)

fects. In addition, an increase in the soil microflora in response to the crop residue generally results in nitrogen immobilization and, hence, in reduced growth of the succeeding crop. As usual in nature, competition and allelopathy occur simultaneously, and the relative effect of each on plant growth and yield is difficult to assess.

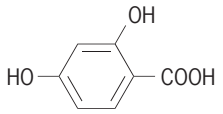
Soil microorganisms have a critical role in the allelopathy responsible for clover soil sickness. Clover roots exude toxic isoflavonoids, which do not cause allelopathy but are converted by microorganisms to salicylic acid (**21**), *p*—methoxybenzoic acid (**22**), 2,4-hydroxybenzoic acid (**23**), and other phenolic com-



(**21**)



(**22**)



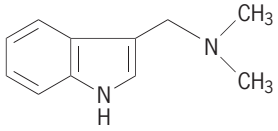
(**23**)

pounds that are responsible for clover soil sickness.

Soil microorganisms may also prevent allelopathy. In sandy soils, sorghum autoallelochemicals reduce the yield of the first and subsequent sorghum crops. In clay-rich soils, sorghum autoallelopathy does not become manifest because fungi degrade the allelochemicals before growth of the new seedlings starts.

In the complex interaction between crops and weeds and among crops lie exciting possibilities for utilizing the allelopathic properties of crop plants for weed control. Allelopathic weed control with crop plants has long been practiced by farmers and horticulturists. The challenge is twofold: to minimize the negative impact of allelochemicals on crop growth and yield and to exploit allelopathic mechanisms for pest control and crop growth regulation strategies. Allelochemicals in new crop cultivars may provide naturally occurring pesticides that can limit or suppress weeds as well as prevent insect and nematode attack and damage. The biotechnology resources for the production of herbicide-resistant crops could then be channeled into the engineering of other desired crop qualities. Allelochemicals may furnish an entirely new generation of naturally produced weed-controlling compounds, replacing synthetic herbicides and pesticides with nonaccumulating, easily degradable substances. *See* HERBICIDE.

The natural occurrence of many plant species in pure stands that apparently cannot be invaded by other species may result, in part, from allelopathic mechanisms that could be exploited to achieve weed-free crop stands. Barley, considered a smother crop, was thought to inhibit weeds through competition with its extensive root system. However, even in the absence of competition, barley inhibits weeds, such as chickweed (*Stellaria media*), shepherd's purse (*Capsella bursa-pastoris*), and tobacco, but not wheat. Gramine (**24**), an alkaloid, which occurs

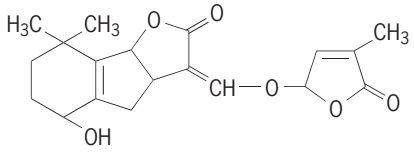


(**24**)

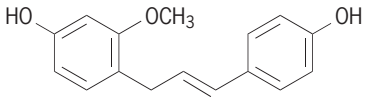
in barley and inhibits *Stellaria* growth, may function as an allelochemical. Allelochemicals in seeds, including alkaloids (for instance, in coffee beans), may protect seeds from microbial attack and destruction while they remain viable and buried in the soil for years. Agricultural practice uses cover crops, which are either killed by frost or by desiccant sprays, and the allelochemicals of the crop residues suppress weed growth. Oats, corn, sunflower, sorghum, and fescue appear promising in allelopathic weed control. *See* ALKALOID.

Not all weed-crop interactions, however, are necessarily detrimental for the crop. Corn cockle (*Agrostemma githago*), growing in wheat fields in former Yugoslavia, increases the wheat yield but suppresses its own growth. Agrostemin and gibberellin have been identified as the allelochemicals of corn cockle. Triacontanol, $CH_3(CH_2)_{28}CH_2OH$, an allelochemical from alfalfa, promotes growth of tomato, cucumber, and lettuce among other species. *See* GIBBERELLIN.

Parasitism. Infestation of vascular plants by vascular plant parasites, often treated as an oddity, is an allelopathic relationship. The seeds of the parasite witchweed (*Striga asiatica*) germinate in response to root exudate from host plants. While the active compounds have not yet been identified, strigol (**25**) induces witchweed seed germination and the formation of a haustorium (a nutrient-absorbing structure) that attaches to the host. Strigol is beneficial to the parasite by stimulating its growth and facilitating recognition of the host, but is detrimental to the plant producing it, when the parasite finds the producer and attaches itself. Purple false foxglove (*Agalinis purpurea*), which parasitizes legumes, responds to xenognosin A (**26**) with haustorium formation.



(**25**)



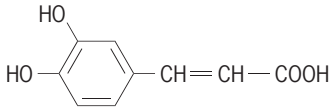
(**26**)

Forestry. As in agriculture and horticulture, allelopathy along with other ecologically important factors affects the pattern of forest growth and development and therefore has strong economic implications. While examples of tree allelopathy, such as that of the black walnut, were well known, the general significance of allelopathy for forestry has been recognized only in recent times.

Oak trees left as seed trees may impede reforestation by retarding the growth of the tree seedlings. Salicyclic acid (**21**) has been identified as one of the allelochemicals, but strong competition from oak roots and physical suppression of seed germination by oak litter also take part in retarding growth.

Bare areas under hackberry (*Celtis laevigata*) result from accumulation of the same phenolic allelochemicals that affect many other plant communities. Synergistic action of these allelochemicals, while often suspected, has been observed with hackberry toxins. Seasonality of allelopathy in a forest of sycamore (*Platanus occidentalis*), rough-leaved hackberry (*C. occidentalis*), and white oak (*Quercus alba*) may indicate that certain allelochemicals are effective upon release, whereas others may require chemical change by microorganisms before they become active. Production and release of toxins must exceed decomposition for allelopathy to occur. *Grevillia robusta*, a stately tree in the subtropical rainforest of Australia, does not form pure stands, and, not surprisingly, does not grow in monoculture beyond 10–12 years, when it declines because of strong autoallelopathy.

Bracken stands or bracken-grassland communities surrounded by forest often remain free of trees. Phenolic compounds, including caffeic acid (**27**) and



(**27**)

ferulic acid (**10**), appear to be the controlling allelochemicals. A dense ground cover of bracken, wild oat grass (*Danthonia compressa*), goldenrod (*Solidago rugosa*), and flat-topped aster (*Aster umbellatus*), does not allow a forest in the Allegheny Plateau of northwestern Pennsylvania to grow back, and half a century after clear-cutting, only a few scattered black cherry (*Prunus serotina*) and red maple (*Acer rubrum*) grow. Other woodland ferns, such as the hay-scented fern (*Dennstaedtia punctilobula*) and New York fern (*Thelypteris noveboracensis*), also contribute to the allelopathic effect.

Release of allelochemicals from bracken appears to be timed for the most effective suppression of surrounding vegetation. In the tropics, toxins are released from green bracken leaves year-round, while in southern California and in the Pacific Northwest, toxins are released from dead bracken fronds, litter, and roots when seeds germinate, either at the beginning of the rainy season or in the spring, respectively.

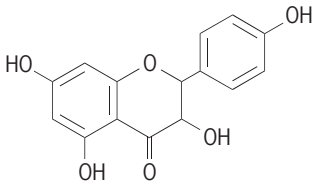
Allelopathic inhibition of forest regrowth may be indirect by affecting the mycorrhizal fungi (fungi which grow mutualistically in association with roots of vascular plants). Shrub litter, heather (*Calluna vulgaris*), and prairie grasses inhibit Norway spruce (*Picea abies*) development by preventing the establishment of mycorrhiza. *See* FOREST AND FORESTRY; MYCORRHIZAE.

Action mechanisms of allelochemicals. While numerous allelochemicals have been identified, most studies on allelopathy deal with unidentified compounds. Allelochemicals are secondary plant products and belong to a few major classes of chemical compounds: terpenoids, phenolic compounds, phenylpropane derivatives, flavonoids, long-chain fatty acids, organic cyanides, probably also alkaloids and purines, and perhaps some steroids. In the field, allelopathy involves a complex of compounds; a single specific compound does not appear to produce the observed allelochemical effects on neighboring plants. This complicates the investigation of the action mechanism of allelochemicals.

There is no single set of physiological reactions controlled by allelochemicals that result in allelopathy; the action of allelochemicals is diverse and affects a large number of physiological functions and biochemical reactions. At the whole-plant level, allelochemicals interfere with internode elongation, leaf expansion, cell division, dry-weight accumulation, as well as seed germination. Growth responses mediated by natural plant growth regulators, such as auxin or gibberellin, are often impaired. In field experiments, allelochemicals from some crop residues promote bacterial and fungal lesions on roots, enhancing diseases of the test seedlings. *See* PLANT HORMONES.

Common to many allelochemicals is their effect on membrane permeability. Ion uptake can be enhanced or reduced; certain phenolic acids inhibit the uptake of phosphate and potassium. Fescue leachates impair uptake of phosphorus and nitrogen in sweetgum. Some flavonoids, phenolic acids, and juglone inhibit membrane adenosine triphosphatase activity that is involved in the control of membrane transport. Sphagnum acidifies its habitat by releasing hydrogen ions, which may make conditions less suitable for competing plants.

Significant effects on photosynthesis by allelochemicals have been observed. Allelochemicals from pigweed and foxtail litter affect photosynthesis in corn and soybeans and, in addition, alter the biomass partitioning into the leaves compared with that of the total plant. Stomatal movement, which regulates gas exchange and water status, and net photosynthesis may also be inhibited. At the cellular and subcellular level, kaempferol (**28**), a flavonol, interferes with the



(**28**)

photosynthesis of isolated chloroplasts by inhibiting coupled electron transport and photophosphorylation (adenosine triphosphate synthesis). In isolated mitochondria, kaempferol inhibits oxidative phosphorylation. Respiration was also inhibited by juglone (**1**) and other allelochemicals. The biosynthesis and metabolism of proteins, lipids, and organic acids are also affected by allelochemicals. While it is relatively easy and straightforward to isolate and identify allelochemicals, to test them in bioassays, and to measure their physiological and biochemical reactions, in the field it is difficult to prove directly that the same compounds move from the producing plant to the neighboring plants and function as allelochemicals in the field. Manfred Ruddat

Bibliography. W. Y. Garner and J. Harvey, Jr. (eds.), *Chemical and Biological Control in Forestry*, ACS Symp. Ser. 238, 1984; M. B. Greene and P. A. Hedin (eds.), *Natural Resistance of Plants to Pests: Role of Allelochemicals*, ACS Symp. Ser. 296, 1986; J. B. Harbone, *Introduction to Ecological Biochemistry*, 4th ed., 1994; M. M. Harlin, Allelochemistry in marine macroalgae, *CRC Crit. Rev. Plant Sci.*, 5:237–249, 1987; A. R. Putnam and Chung-Shih Tang (eds.), *The Science of Allelopathy*, 1986; E. L. Rice, *Allelopathy*, 2d ed. 1984; A. C. Thompson (ed.), *The Chemistry of Allelopathy: Biochemical Interactions among Plants*, ACS Symp. Ser. 268, 1985; G. R. Waller (ed.), *Allelochemicals: Role in Agriculture and Forestry*, ACS Symp. Ser. 330, 1987.

Allergy

Altered reactivity in humans and animals to allergens (substances foreign to the body that cause allergy) induced by exposure through injection, inhalation, ingestion, or skin contact. The most common clinical manifestations of allergy are hay fever, asthma, hives, atopic (endogenous) eczema, and eczematous skin lesions caused by direct contact with allergens such as poison ivy or certain chemicals.

Allergens. A large variety of substances may cause allergies: pollens, animal proteins, molds, foods, insect venoms, foreign serum proteins, industrial chemicals, and drugs. Most natural allergens are proteins or polysaccharides of moderate molecular size (molecular weights of 10,000 to 200,000). Chemicals or drugs of lower molecular weight (haptens) have first to bind to the body's own proteins (carriers) in order to become fully effective allergens.

For the development of the hypersensitivity state underlying clinical allergies, repeated contact with the allergen is required. Duration of the sensitization period is usually dependent upon the sensitizing strength of the allergen and the intensity of exposure. Some allergens (for example, saliva, urine, and hair proteins of domestic animals) are more sensitizing than others. In most instances, repeated contact with minute amounts of allergen is required: several annual seasonal exposures to grass pollens or ragweed pollen usually occur before an overt manifestation of hay fever. On the other hand, allergy to cow milk proteins in infants can develop within a few weeks. When previous contacts with allergens have not been apparent (for example, antibiotics in food), an allergy may become clinically manifest even upon the first conscious encounter with the offending substance.

Besides the intrinsic sensitizing properties of allergens, individual predisposition of the allergic person to become sensitized also plays an important role. Clinical manifestations, such as hay fever, allergic asthma, and atopic (endogenous) dermatitis, occur more frequently in some families: the inheritance of a capacity to develop these forms of allergy has been called atopy or the atopic state. These terms indicating a disease "not at its proper place" are misnomers but are sanctioned by historical usage. In other clinical forms of allergy, genetic predisposition, though possibly present as well, is not as evident.

Mechanisms. Exposure to sensitizing allergens may induce several types of immune response,

and the diversity of immunological mechanisms involved is responsible for the various clinical forms of allergic reactions which are encountered in practice. Three principal types of immune responses are encountered: the production of IgE antibodies, IgG or IgM antibodies, and sensitized lymphocytes. *See* ANTIBODY; IMMUNOGLOBULIN.

IgE antibodies. IgE antibodies belong to a peculiar class of serum immunoglobulins which is responsible for the majority of allergies of the so-called immediate or anaphylactic type. IgE antibodies are present in blood in very small amounts (normally less than 0.2 microgram per milliliter), but have the capacity to bind strongly to the membrane of tissue mast cells and blood basophils. These cells are the major effector components in an immediate allergic reaction. They contain a wide variety of preformed inflammatory mediators, such as histamine and serotonin. They also have the capacity to form, upon interaction with allergen, further substances causing inflammatory changes in tissue. The latter include increased vascular permeability that results in swelling and is manifested by redness and wheals as observed in hives. Among such mediators formed after cell stimulation, leukotrienes, the platelet-activating factor (PAF), and several other substances which attract other blood cells, such as eosinophils, to the site of reaction have gained the most attention. Mast cells and basophils may be compared to powder kegs which will be fired upon the interaction of allergen with its corresponding cell surface–bound IgE antibody. Understandably, such a reaction proceeds quite rapidly in a sensitized individual following renewed contact with allergen. Within minutes, hay fever patients develop symptoms after inhaling grass pollen, as do individuals allergic to bee venom after being stung. Generally, individuals possessing a genetic predisposition to atopic diseases are extremely prone to develop IgE antibodies to a variety of inhaled or ingested allergens. The IgE antibodies may also bind, although with lesser affinity, to some mononuclear cells and in particular to Langerhans cells in the skin. This phenomenon may be involved in the mechanism of some eczematous skin lesions, as seen in the frequent allergic condition termed atopic dermatitis. *See* HISTAMINE; SEROTONIN.

IgG or IgM antibodies. The production of IgG of IgM antibodies requires, as a rule, more massive exposure to allergens than that leading to the formation of IgE antibodies. Upon renewed contact with sufficient amounts of allergen, immune complexes may form with serum antibodies and be deposited in various tissues where they cause an acute inflammatory reaction. Preferential sites for such reactions, depending upon the mode of entry and distribution of the allergen, are alveolar walls of the lungs, glomeruli of kidneys, synovial membranes of joints, and walls of small blood vessels of the skin. Immune complex deposition below the basement membrane of blood vessels induces a cascade activation of complement (a multicomponent system of interacting blood proteins) which generates further active mediators of inflammation. The most important include the fragments of complement factors C3 and C5 (denominated C3a and C5a or anaphylatoxins) which are chemotactic (attracting) for white blood cells, especially polymorphonuclear leukocytes, attracting them to infiltrate the site of the reaction. Once arrived there, these cells proceed to destroy excess allergen and immune complexes, and by various mechanisms enhance the inflammatory reaction. The process takes several hours, and reactions of this type (Arthus reactions) become clinically apparent and peak only 6 to 12 h following contact with allergen. *See* IMMUNE COMPLEX DISEASE; INFLAMMATION.

Sensitized lymphocytes. Sensitized lymphocytes are responsible for the clinical manifestations of so-called delayed hypersensitivity. If an individual has been exposed to allergen and has developed sensitized lymphocytes capable of reacting with it, renewed exposure will trigger the lymphocytes to produce an array of glycoproteins called lymphokines. Lymphokines are also mediators capable of causing considerable inflammation in the surrounding tissues. Although produced at first only by the relatively small proportion of lymphocytes which specifically recognize and interact with allergen, lymphokines have amplifying effects. They stimulate, indiscriminately, other cell types, such as the macrophages and monocytes, attract these cells to the site of the reaction, and cause the formation of an inflammatory infiltrate. In the skin, such a reaction manifests itself by redness, swelling, and the formation of an inflammatory papule. A typical example is the inflammatory nodular reaction experienced upon injection in the skin of protein allergens produced by tubercle bacilli (tuberculin reaction). Such a reaction occurs only in people who were once infected with tubercle bacilli, and who have developed an immune response to these microorganisms.

Skin reactions occurring upon contact withbreak sensitizing agents such as poison ivy or industrial chemicals also are caused by the same immune mechanism. Since production of lymphokines by lymphocytes following stimulation by allergens takes several hours and since the reaction becomes clinically apparent only when a minimal number of infiltrating cells (mostly monocytes) have accumulated at the reaction site, the time elapsed between allergen contact and clinical reaction is usually 24 to 72 h. Such reactions are therefore termed delayed reactions, in contrast to the immediate (anaphylactic) reactions which occur within minutes after contact with allergen C. *See* HYPERSENSITIVITY.

Clinical forms. Among the clinical manifestations of allergy are allergic rhinitis, bronchial asthma, food allergy, occupational allergy, skin allergy, and anaphylactic shock.

Allergic rhinitis. The most frequent manifestation of allergy is undoubtedly allergic rhinitis. Allergic rhinitis was originally misnamed hay fever because it was believed that hay causes the disorder, and "fever" was a term loosely applied to many ailments, even when not accompanied by fever. However, in addition to nasal symptoms (watery nasal discharge, sneezing, runny eyes, and itchy nose, throat, and roof of the

mouth), the individual may experience fatigue, irritability, and loss of appetite.

In industrialized countries, the frequency of allergic rhinitis has been estimated to be 5 to 8% of the population and appears to be increasing. Allergic rhinitis may be seasonal, as the traditional hay fever, or perennial, occurring year-round. The seasonal form is usually caused by pollens from grasses, weeds, or trees. Related symptoms appear only during that time of the year when the pollens to which the individual is sensitive occur in the air. When allergens are present all the time, allergic rhinitis may occur year-round. Frequent causes are allergens present in the house or workplace, such as dust, molds, and animal danders.

Chronic or recurrent rhinitis, however, is not always related to allergy; in many instances it may be caused by an abnormal irritability of the nose (so-called intrinsic or vasomotor rhinitis) attributable to still ill-defined factors, such as hormonal factors or stress.

Bronchial asthma. The allergic disease which by its frequency and severity causes the most problems is bronchial asthma. Although an attack of asthma may be easily described, since it involves acutely occurring shortness of breath and wheezing, a general accurate definition, encompassing all possible causes and mechanisms of the disease, is still not available. Constriction of tiny bronchi in the lung airways, accompanied by an augmentation of bronchial secretions and the swelling of bronchial mucosa, contributes to obstruct the free passage of air through the lungs, producing wheezing. However, wheezing is not always indicative of asthma. Furthermore, in numerous instances, the cause of asthma is not related to allergy against some airborne allergens. It is customary to classify asthma as extrinsic, that is, due to contact with an external allergen, or intrinsic, that is, due to endogenous causes which are still poorly defined, such as stress or bronchial hyperreactivity owing to hormonal and other factors. However, even if the causes of disease may be very dissimilar, the mechanisms and particularly the mediators responsible for allergic inflammation in the lung may be the same, leading to almost indistinguishable clinical pictures.

The same airborne allergens which cause allergic rhinitis may also be responsible for seasonal or year-round asthma. Indeed, allergic rhinitis and asthma are often associated in the same individual. In addition, ingestion of allergens in foods and inhalation of drugs by aerosol or of sensitizing industrial chemicals at the workplace are frequent causes of allergic asthma.

Allergic asthma is considered a major public health problem owing to its severity, and to the possibility of death during an acute and prolonged asthma attack (status asthmaticus), and also because of the possible evolution to severe chronic respiratory insufficiency with invalidism in neglected cases. Available statistics indicate that in most countries 2–4% of the population is affected by asthma. In view of the large variety of possible causes of asthma, different mechanisms may be operating. In classical cases of allergic asthma due to airborne allergens, the presence of IgE antibodies probably plays a major role, but it is not excluded that other forms of immune reactions also occur. *See* ASTHMA.

Food allergy. This may occur as a reaction to natural products (for example, celery, milk, and egg proteins) as well as to chemicals or preservatives added to foods (for example, dyestuffs and antioxidants). Allergy to foods may manifest itself at the primary site of contact with allergen, namely in the gastrointestinal tract, causing symptoms such as diarrhea, abdominal pain, nausea, and vomiting. However, it may also take more general forms once the allergen has been absorbed, causing symptoms in the upper airways (rhinitis and asthma) and in the skin (hives), and, in cases of extreme hypersensitivity, generalized anaphylactic shock.

In infants from allergic families, the development of allergy to cow milk is quite frequent; strict breast feeding during the first months of life is therefore increasingly advocated. The breast-feeding mother should abstain from ingesting large quantities of potential allergens such as milk and eggs, since the food allergens can pass into the mother's milk and sensitize the child. After a few months, the susceptibility of infants to sensitization by ingestion appears to decrease markedly.

Occupational allergies. Such allergies become increasingly frequent in industrialized countries. Chemicals used in plastic and pharmaceutical industries or in metal refining and construction work are frequent offenders. However, natural organic materials, such as baker's flour, animal products, castor bean, silk, and furs, may be involved as well. Here, too, the most frequent manifestations involve the upper respiratory tract and the skin.

A peculiar form of occupational lung disease, hypersensitivity pneumonitis, is characterized by repeated episodes of cough, fever, chills, and shortness of breath, occurring 4–10 h after massive exposure to some airborne allergen. In contrast to allergic asthma, the lesions do not occur at the level of the bronchi but of the alveolar walls in the lung. The inflammatory reaction taking place there may have severe consequences when occurring repeatedly, since it leads to progressive fibrosis of the lung, impairment of gas exchanges with blood, and respiratory insufficiency. The mechanisms of this type of allergic reaction are not entirely elucidated, but it is known that IgE antibodies are not involved. IgG antibodies are frequently present and possibly cause the formation of immune complexes with allergen. However, delayed hypersensitivity mechanisms are probably at play as well. Among the most frequent offending allergens are molds sometimes found in hay (causing "farmer's lung") and in various other locations, such as air conditioners, cellars, and logs used by woodworkers.

Skin allergy. This condition may take several forms. Atopic dermatitis is a chronic superficial itching

inflammation of the skin (eczema) which occurs in 0.5–0.7% of the population and is often of long duration. Atopic dermatitis usually occurs early in life together with cow milk allergy (infantile eczema). Its evolution is unpredictable, and the disease may seriously affect psychological health and working ability of the individual. IgE antibodies frequently are notably increased. This disease exhibits the same genetic predisposition ("atopy") as allergic rhinitis and asthma.

Allergic contact dermatitis is characterized by a red rash, swelling, intense itching, and sometimes blisters at the site of contact with the allergen, which may be of plant origin (for example, poison ivy or turpentine) or may be an industrial chemical. Lesions are caused by sensitized lymphocytes and belong to the delayed hypersensitivity type of allergy.

Urticaria (hives) is a very common manifestation of skin allergy. When evoked by sensitization to some allergen, IgE antibodies are present. Foods, drugs such as the penicillins, and, more exceptionally, natural inhalation allergens such as pollens are the most frequently involved. However, utricaria, resulting from the release of inflammatory mediators (particularly histamine) by skin mast cells, is even more frequently caused by nonallergic mechanisms. Some foods (for example, crustaceans or strawberries) contain substances which, when ingested in sufficient quantities, can directly trigger mast cells without prior sensitization and formation of IgE antibodies.

Anaphylaxis. The most feared allergic reaction is an acute event known as anaphylactic shock. This is an IgE-mediated reaction with sudden onset: symptoms develop within minutes after exposure to the triggering allergen. The victim may experience a feeling of great anxiety; symptoms include swelling and skin redness often accompanied by hives, vomiting, abdominal cramps and diarrhea, life-threatening breathing difficulties due to swelling of the throat, and severe bronchospasm. A drastic fall in blood pressure may cause fatal shock within a few minutes. Insect stings and drugs such as the penicillins are now the most frequent causes of anaphylactic shock; in earlier times the therapeutic injection of animal serum proteins (for example, horse antitetanus serum) was often responsible. Some acute reactions, called anaphylactoid reactions, may cause similar symptoms but are not due to IgE-mediated sensitization. Such reactions are encountered mostly in people intolerant to some drugs, such as aspirin, or to contrast media used in diagnostic radiology. *See* ANAPHYLAXIS.

Diagnosis. Diagnosis of allergic diseases encompasses several facets. The large variety of clinical forms of allergic reactions and the multitude of allergens possibly involved frequently require prolonged observation and in-depth investigation of the individual's life habits. Since many clinical manifestations of allergy are mimicked by nonallergic mechanisms, it is usually necessary to use additional diagnostic procedures to ascertain whether the person has developed an immune response toward the incriminated allergen. Such procedures primarily consist of skin tests, in which a small amount of allergen is applied on or injected into the skin. If the individual is sensitized, a local immediate reaction ensues, taking the form of a wheal (for IgE-mediated reactions), or swelling and redness occur after several hours (for delayed hypersensitivity reactions). The blood may also be analyzed for IgE and IgG antibodies by serological assays, and sensitized lymphocytes are investigated by culturing them with the allergen.

Since the discovery of the responsible allergens markedly influences therapy and facilitates prediction of the allergy's outcome, it is important to achieve as precise a diagnosis as possible. Most of the tests described above indicate whether the individual is sensitized to a given allergen, but not whether the allergen is in fact still causing the disease. Since in most cases the hypersensitive state persists for many years, it may well happen that sensitization is detected for an allergen to which the individual is no longer exposed and which therefore no longer causes symptoms. In such cases, exposition tests, consisting of close observation of the individual after deliberate exposure to the putative allergen, may yield useful information.

Therapy. Therapy of allergic reactions may be performed at several levels. Whenever feasible, the most efficient treatment, following identification of the offending allergen, remains elimination of allergen from the person's environment and avoidance of further exposure. This treatment is essential for allergies caused by most household and workplace allergens.

Inflammatory reactions and corresponding symptoms caused by allergic mediators may be relieved by several drugs acting as pharmacological antagonists to these mediators, for example, antihistaminics or steroids. Other drugs act at a more central level and influence the ease with which mast cells or basophils release their mediators; this is apparently the case for chromones. Since the state of hypersensitivity usually remains present over many years, especially when contact with allergen is periodically renewed (as is the case with pollens), such an approach to allergy therapy may be only palliative.

Attempts to influence the hypersensitive state itself in IgE-mediated allergy are more decisive. Repeated injections of increasing doses of allergen progressively diminish the degree of hypersensitivity. This procedure, which must usually be pursued over several years, is referred to as hyposensitization or immunotherapy. *See* IMMUNOLOGY; IMMUNOTHERAPY.

A. L. de Weck

Bibliography. R. Davies and S. Ollier, *Allergy: The Facts*, 1989; H. F. Krause, *Otolaryngic Allergy and Immunology,* 1989; L. M. Lichtenstein and A. S. Fauci, *Current Therapy in Allergy, Immunology, and Rheumatology*, 5th ed., 1996; G. Melillo and J. A. Nadel (eds.), *Respiratory Allergy, Advances in Clinical Immunology and Pulmonary Medicine*, 1993; D. Reinhardt and E. Schmidt (eds.), *Food Allergy,* 1988; I. M. Roitt, *Essential Immunology,* 1991.

Alligator

A large aquatic reptile of the family Alligatoridae. Common usage generally restricts the name to the two species of the genus *Alligator*. The family also includes three genera (five species) of caiman. With the crocodiles and gharial (also spelled gavial), these are survivors of archosaurian stock and are considered close to the evolutionary line which gave rise to birds.

Alligator species. The two living species have a disjunct subtropical and temperate zone distribution. The American alligator (*A. mississippiensis*) ranges throughout the southeastern United States from coastal North Carolina (historically from southeastern Virginia) to the Rio Grande in Texas, and north into southeastern Oklahoma and southern Arkansas (see **illustration**).

Poaching and unregulated hunting for the valuable hide decimated the alligator populations until the animal was placed on the U.S. Endangered Species List in 1967. The species has responded well to protection and has become abundant in many areas in its range, particularly in Florida and parts of Georgia, Louisiana, and Texas, where it is now a common sight in the freshwater habitats, including swamps, marshes, lakes, rivers, and even roadside ditches.

The second species is the Chinese alligator (*A. sinensis*), restricted to the region of the Yangtze River valley, where it inhabits burrows in the floodplains and riverbanks. This species is considered "critically endangered" by international conservation organizations. It is believed that less than 200 individuals remain in the wild.

The American alligator is by far the larger of the two species, exceeding 15 ft (4.5 m). The average length of *A. sinensis* is 4–5 ft (1.2–1.5 m).

Caiman species. There are five species of caimans, once often sold as "baby alligators." Caimans differ from alligators in technical details of their internal anatomy and scale characteristics.

The black caiman (*Melanosuchus niger*) of the Amazon Basin of South America strongly resembles the American alligator in superficial appearance and may reach a length approaching 16 ft (5 m). The spectacled caiman (*Caiman crocodilus*), with its several subspecies, is sold in the pet trade. It is the most widely distributed of the caimans, ranging from southwestern Mexico to Paraguay and Bolivia. Spectacled caimans reach a length of 8 ft (2.5 m) and can be distinguished from alligators by the presence of a ridge across the snout between their eyes, the nose bridge of the "spectacles."

The broad-nosed caiman (*C. latirostris*) occurs from southern Brazil to Paraguay and northern Argentina and reaches a length of 9 ft (3 m). The two species of smooth-fronted caiman, *Paleosuchus trigonatus* and *P. palpebrosus*, are the smallest of the family, reaching a maximum length of 7 ft (2 m). They occur over most of northern South America, south to Peru, Bolivia, and southeastern Brazil.

Anatomical features. Alligators, including caimans, are generally distinguished from crocodiles by their broader, rounded, and more massive snout, the arrangement of their teeth, and other technical anatomical details. As in all crocodilians, the teeth are conical and sharp, equivalent in shape (homodont), and replaced throughout life. Moreover, as in all crocodilians, alligators possess a functionally four-chambered heart, a secondary palate, a septum or "diaphragm" that separates the lung and peritoneal regions, a compressed tail, webbed feet, and other adaptations for an aquatic existence.

Nesting. All alligators build a conical nest of available materials near the water's edge, where they deposit 15–60 leathery-shelled eggs. Incubation takes 30–60 days; females of at least some species may guard the nest, assist in the hatching process, and remain with the young for as long as 3 years. In some cases, females carry young in their mouth to nearby water.

Habitat and behaviors. Alligators are generalized carnivores, feeding on any invertebrates, fishes, reptiles and amphibians, birds, or mammals that can be caught and overpowered. They play a dominant role in the energy and nutrient cycling in their habitats, and through their construction of "gator holes," which often retain water in times of drought, are considered important to the survival of many others species.

Alligators have also been found to have a complex social system that relies heavily on communication. Visual, tactile, auditory, and even subauditory communications have been described in alligator populations. Males roar during the breeding season to attract females, and females bellow and growl to communicate receptiveness to males. Using trunk muscles and low-frequency sound, alligators create rapid vibrations on the water surface to communicate. The vibrations coupled with the low-frequency sound can be heard at great distances (perhaps up to a kilometer). Males use a diversity of sounds as a threat warning, including a "cough" and hiss. Juveniles employ distress calls when they feel threatened. These calls attract the attention of the nearby female. *See* ARCHOSAURIA; CARDIOVASCULAR SYSTEM; CROCODILE; CROCODYLIA; DENTITION; REPTILIA.

W. Ben Cash; Howard W. Campbell

Half-submerged American alligator (*Alligator mississippiensis*). (*Photo courtesy of W. Ben Cash*)

Bibliography. S. Grenard and W. Loutsenhizer, *Handbook of Alligators and Crocodiles*, Krieger,

Melbourne, FL, 1995; S. B. Hedges and L. L. Poling, A molecular phylogeny of reptiles, *Science*, 283:998–1001, 1999; D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, New York, 2001; F. H. Pough et al., *Herpetology*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 2004; K. A. Vliet, Social dynamics of the American alligator (*Alligator mississippiensis*), *Amer. Zool.*, 29:1019–1031, 1989; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

Allometry

The study of changes in the characteristics of organisms with body size. Characteristics such as body parts or timing of reproductive events do not necessarily change in direct proportion to body size, and the ways in which they change relative to body size can often provide insights into organisms, construction and behavior. Large organisms are often not merely magnified small ones.

Variables. Many characteristics, ranging from brain size and heart rate to life span and population density, change consistently with body size. These relationships normally fit a simple power function given by Eq. (1), where y is the variable under study,

$$y = km^b \tag{1}$$

m is body mass, k is the allometric constant, and b is the allometric exponent. The use of logarithms makes the equation easier to visualize—the exponent becomes the slope of a straight line when the logarithm of the variable (y) is plotted against the logarithm of body mass (m) [Eq. (2)]. Unless the allo-

$$\log(y) = \log(k) + b\log(m) \tag{2}$$

metric exponent (b) equals 1, the ratio of y/m varies with m. The terms isometry, positive allometry, and negative allometry are sometimes used for $b = 1$, $b > 1$, and $b < 1$, respectively. If the variable of interest is the mass of an organ, under isometry organ mass is a fixed proportion of body mass; under positive allometry larger organisms have disproportionately large organs; and under negative allometry larger organisms have disproportionately small organs. *See* LOGARITHM.

Proportions. The mass of the brain as a proportion of body mass from adults of different mammalian species is an example of negative allometry. Proportions vary among related species of different sizes (that is, interspecific allometry); for example, the brain is about 5% of body mass in a mouse but only 0.05% in a whale (**Fig. 1**). Proportions often vary within a species as well. Individuals change shape during growth, a phenomenon known as ontogenetic or growth allometry; for example, among humans the head is 25% of a baby's length but only 15% of an adult's height. Within species, small and large adults of the same sex may be isometric with respect to some bodily proportions, but there may also be static intraspecific allometry. The different types of allometry (**Fig. 2**) often show a character scaling in different ways, reflecting different underlying causes.

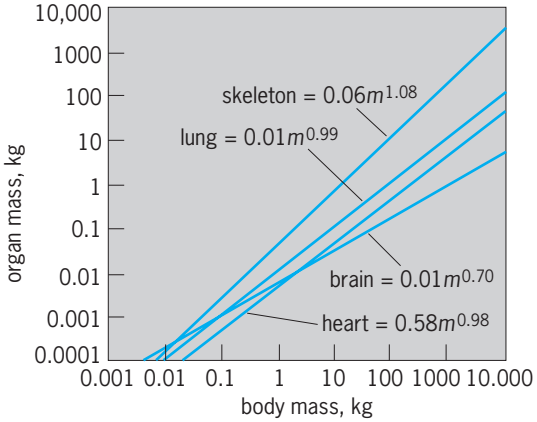


Fig. 1. Allometric lines for the mass of some organs across mammal species. (These are best-fit lines: individual species may lie above or below them.) Note that the skeleton is disproportionately heavy in large species (positive allometry), whereas the brain is disproportionately light (negative isometry). Heart and lungs scale approximately isometrically.

Even if small and large organisms are the same shape, size is significant because surface area and volume scale differently with body length. Consider what happens to a cube when the length of each side is doubled: the surface area increases fourfold, the volume (and hence mass) eightfold. The same relations hold for any shape: the surface area is proportional to the length squared and the mass is proportional to the length cubed. It follows, therefore, that larger organisms have less surface area per unit mass. This relationship has profound consequences, helping to explain, for instance, why shrews and hummingbirds are the smallest endotherms—any organism much smaller would lose heat too quickly through its relatively large surface. Size is also

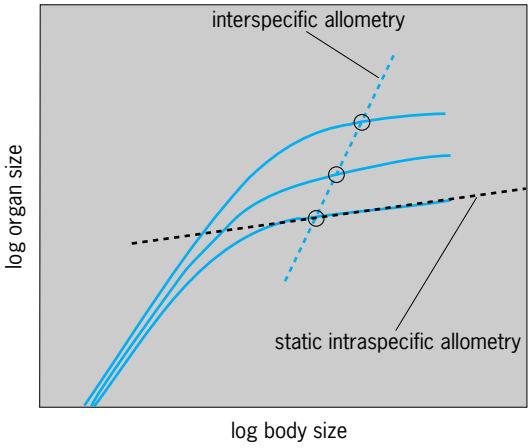


Fig. 2. Three different types of allometry. Each curve shows the growth allometry of organ size in a different species. The adults of each species show static intraspecific allometry. The circles indicate a particular stage of the life history (for example, sexual maturity). The line of interspecific allometry links points from different species.

significant for rates; for example, longer limbs can take longer strides, so larger animals need fewer strides to cover a given distance. Across mammal species, limb length scales as $m^{0.35}$, so the frequency of strides of two animals running at the same speed will differ according to their differences in $m^{0.35}$.

Allometric lines. In themselves, allometric lines are rough and probably simplified descriptions of the combined effects of the many (usually unknown) selective forces correlated with size; it still must be explained why the exponent takes the value that it does. For instance, in the skeletons of geometrically similar land animals, mass is proportional to length cubed while the cross-sectional area of any given bone, and hence its strength, scales as the square of its length. Beyond some critical size, these animals would collapse under their own weight. In fact, skeletal mass is proportional to $m^{1.08}$ for mammals and birds because bones face static (compressing) stress when the animal is stationary and elastic (buckling) stress during locomotion. Mechanical principles can be used to ascertain whether skeletons are adapted to prevent fracture by compression or by buckling. *See* BONE.

The leg bones of a quadruped with a vertical posture are like tall, thin, cylinders. To withstand compression equally at all body weights—a condition known as static stress similarity—their cross-sectional area must increase as m^1 and their length scales as $m^{1/3}$, so bone mass should scale as $(m^1)(m^{1/3}) = m^{4/3}$. But bone mass is proportional to $m^{1.08}$, so skeletons do not show static stress similarity.

Bones may instead be adapted to withstand buckling equally at all body sizes—the condition known as elastic stress similarity. The limb bones of hoofed mammals fit the model of a column, but those of other mammals do not, perhaps because their limbs are less pillarlike and so face different stresses. Elastic stress similarity is also found further afield. Limb bones of bipedal theropod dinosaurs fit the theory. Tree trunks and limbs are also thin columns faced with buckling stresses: tree trunks retain elastic stress similarity as they grow, and branch thickness scales with length as predicted. *See* STRESS AND STRAIN.

Many allometries are less well understood. Across mammal species, for example, small mammals clearly get enough oxygen for their maximal needs, so large species would seem to have more lung than they need.

Deviations from the line. Allometric lines are often used as baselines for comparison: if it is assumed that a line shows how organ mass has to change with body mass in order to keep doing the same job, organisms above or below the line have larger or smaller organs than expected for their body size. The question then arises as to why particular groups deviate from the line. Finding other variables that are correlated with this body-size-independent variation allow for powerful tests of hypotheses of adaptation. *See* ADAPTATION (BIOLOGY).

For example, across mammal species the scaling of brain mass (proportional to $m^{0.7}$) is similar to the scaling of metabolic rate (proportional to $m^{0.75}$). The similarity prompted the suggestion that brain size of newborn mammals (which is linearly related to adult brain size) scales as $m^{0.75}$ because it is somehow constrained by the mother's metabolic rate. However, another hypothesis also predicts an exponent near 0.7; surface area scales as $m^{2/3}$, as discussed earlier. If processing requirements are set by the amount of peripheral sensory input, brain size might scale as $m^{2/3}$ too. Choosing between these hypotheses without very good estimates of the exponent seems impossible. However, deviations from the line provide a critical test. The metabolic rate argument predicts that species with fast metabolisms for their size will also have relatively large brains; however, data do not support this prediction. Once deviations are understood, they can be used to make inferences. For instance, eye-socket size scales allometrically among extant primate species, but nocturnal species lie above the line. This knowledge leads to the inference of the activity patterns of species known only from fossils.

Using allometry in this way assumes that the part under study has the same function throughout the group being considered. If function differs, so may the scaling. For example, wing length scales as $m^{1/3}$ across most bird species. Hummingbird wings, however, must support the bird during hovering as well as more conventional flight, and hovering requires relatively larger wings, so hummingbird wing length scales as $m^{2/3}$.

Models. Allometry is perhaps most powerful when known allometric relations are coupled with theory to produce new hypotheses that can be tested. For example, mammal species with large bodies grow more slowly, mature later, have fewer but larger young after longer gestation periods, and live longer than do small species; these traits remain correlated with each other among deviations from the allometric lines. Allometric models predict not only the allometric exponents but also the correlations among the deviations. The models even provide a framework for asking questions vital to understanding scaling, such as what factors affect the evolution of body size itself. *See* BIOPHYSICS. Andy Purvis; Paul H. Harvey

Bibliography. R. McN. Alexander, *Optima for Animals*, 1982; E. L. Charnov, *Life History of Invariants*, 1993; P. H. Harvey and J. R. Krebs, Comparing brains, *Science*, 249:140–146, 1990; P. H. Harvey and M. D. Pagel, *The Comparative Method in Evolutionary Biology*, 1991; T. A. McMahon and J. T. Bonner, *On Size and Life*, 1983; K. Schmidt-Nielsen, *Scaling: Why Is Animal Size So Important?*, 1984.

Allosteric enzyme

Any one of the special bacterial enzymes involved in regulatory functions. End-product inhibition is a bacterial control mechanism whereby the end product of a biosynthetic pathway can react with the first enzyme of the pathway and prevent its activity. This end-product inhibition is a device through which

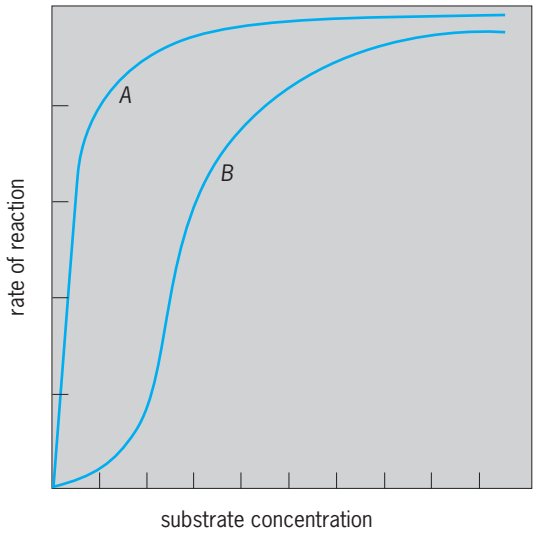


Fig. 1. Rate of reaction of an enzyme in relation to substrate concentration. Curve A obtained with ordinary enzymes, curve B obtained with allosteric enzymes.

a cell conserves its economy by shutting off the synthesis of building blocks when too many are present. *See* BACTERIAL PHYSIOLOGY AND METABOLISM.

Kinetics. Enzymes are protein substances which act as biological catalysts in converting one substance, the substrate, to another, the product. To understand the nature of the special allosteric enzymes, they must first be compared with ordinary enzymes as to their kinetics, that is, the way they work with increasing amounts of substrate. With ordinary enzymes the rate of reaction increases as the amount of substrate is increased until eventually a saturation level is obtained and no further increase occurs. The curve obtained is described as a hyperbola (**Fig. 1**). From this curve it can be said that the substrate molecules tend to occupy more and more binding sites until all the sites are in use. The allosteric enzymes, when similarly plotted, instead of showing a hyperbolic shape, often show a sigmoid, or S-shaped, curve. This means that there is a cooperative effect between substrate molecules so that, when the first molecules are bound to the enzyme, the subsequent ones are bound more readily. A clue to this concept of cooperative substrate interaction was obtained by the study of how the hemoglobin of the blood reacts with oxygen. Here, too, the curve describing the saturation of hemoglobin with oxygen traces a sigmoid shape.

Inhibition. Another characteristic of enzymes is that their action can be prevented by inhibitors. In ordinary enzymes the inhibitors are substances which closely resemble the structure of the substrate. Because of this structural similarity, they can jam the site that the substrate normally occupies. They can thus be likened to a false coin which fits the slot of the machine but does not operate it. The surprising feature about the inhibitors of allosteric enzymes is their lack of structural similarity with the substrate. Hence, they cannot be expected to work by competing with the substrate for a common slot. Instead, they apparently have their own special site of attachment, and when they become attached there, they somehow alter the site where the substrate fits. The substrate kinetics of the inhibited reaction supports this concept. Furthermore, S-shaped curves are also obtained when the inhibition rates are plotted against amounts of the inhibitor. This means that there is cooperative interaction between inhibitor molecules in the same way as there is cooperative interaction between substrate molecules.

Activation. The initial idea of allosteric inhibition was used as an explanation for the negative feedback control shown by the regulatory enzymes. There are also times when the control must have a positive effect, and these are brought about by allosteric activators. Consider, for example, two parallel production lines of metabolic activity leading to two different end products that must be built into a common structural unit. If there is too much of one of the building blocks, then the regulatory mechanisms can react in two ways. In one case the end product can shut off its own production by negative feedback control, and in the other case it can speed up the other parallel line by positive control. In this way the activity of allosteric enzymes may be either decreased by inhibitors or increased by activators. Again, if the substrate curves are studied with and without the activator, it is found that the allosteric activator converts the S-shaped curve into the ordinary hyperbolic type.

Desensitization. It is possible to treat allosteric enzymes in various ways so that they lose their sensitivity to the allosteric inhibitor. The enzymes, though desensitized with respect to inhibition, retain their activity with the substrate and, indeed, this activity is often increased to the same extent as is obtained with an activator. The S-shaped kinetics is changed to the normal variety. The agents that can cause this desensitization include heat, acid, heavy metals such as mercury, and agents which react with sulfhydryl (SH) groups on the proteins. These agents are known to be able to separate complex proteins into subunits. Allosteric enzymes are therefore considered to be complex proteins made up of identical monomers. The single unit, the monomer, is active with normal hyperbolic kinetics and is not affected by the regulatory molecules (inhibitors and activators). The multiple structure, several units combined, is the form that shows the properties of allosterism, S-shaped kinetics, and alteration by the regulators.

Allosteric model. The above considerations have led to the development of a model to explain the action of allosteric enzymes (**Fig. 2**). It suggests that the enzyme molecule is a complex consisting of identical subunits, each of which has a site for binding the substrate and another for binding the regulatory substance. These subunits interact in such a way that two conformational forms may develop. One form is in a relaxed condition (R state) and has affinity for the substrate and activator; the other form is in a constrained or taut condition (T state) and has affinity for the inhibitor. The forms exist in a state of equilibrium, but this balance can be readily tipped

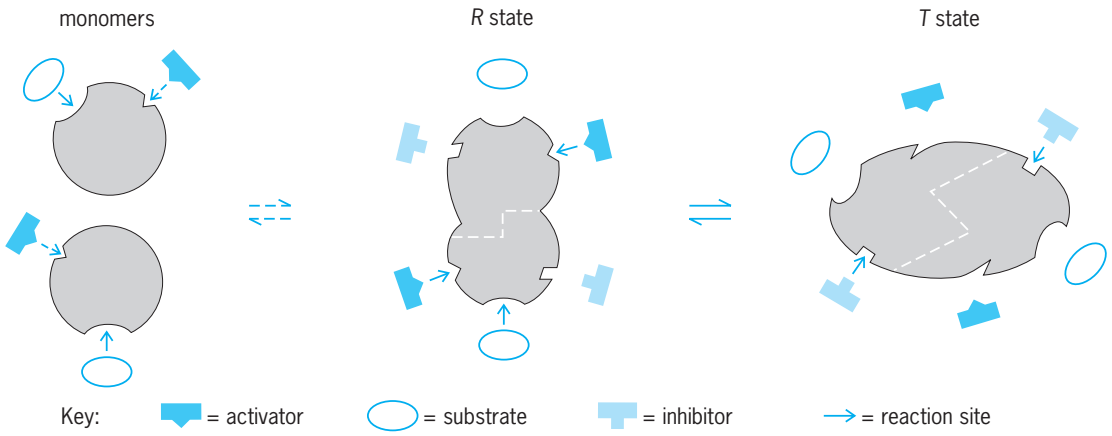


Fig. 2. Model of allosteric enzyme composed of two subunit monomers. Regulatory effects are obtained by shifting back and forth between the two states, relaxed (R state) and taut (T state). In the R state it reacts with substrate and activators, and in the T state it reacts with inhibitors. (*After J. Monod, J. Wyman, and J.-P. Changeux, On the nature of allosteric transitions: A plausible model, J. Mol. Biol., 12:88, 1965*)

by binding one of the reactants. The substrate and activator are bound by the relaxed form; when this happens, the balance is tipped in favor of that state. Conversely, the inhibitor will throw the balance toward the constrained state. The balance is thus tipped one way or the other, depending on the relative concentrations of substrate and inhibitor. Since the two states require subunit interaction for their maintenance, it can be seen why dissociation of the subunits leads to a simple monomeric enzyme which no longer exhibits allosteric effects. The model also shows how the binding sites may interact in either a cooperative or antagonistic manner. *See* ENZYME.

Joseph S. Gots

Bibliography. J.-P. Changeux, Control of biochemical reactions, *Sci. Amer.*, 212(4):36–45, 1965; B. D. Davis, *Microbiology*, 3d ed., 1980; J. Monod, J. Wyman, and J.-P. Changeux, On the nature of allosteric transitions: A plausible model, *J. Mol. Biol.*, 12:88, 1965.

Allotheria

One of the four subclasses of Mammalia, containing a single order, the Multituberculata. The Allotheria first appeared in the Late Jurassic and survived well into the Cenozoic, a period of at least 100,000,000 years. Fossils are known from North America, Europe, and Asia.

The diagnostic features of the subclass are in the dentition, which is very specialized. There is a pair of enlarged incisors above and below; reduced lateral incisors may persist in the upper jaw. Canines are absent, leaving a diastema between incisors and cheek teeth. Premolars are variable and often reduced. The lower molars have five or more cusps in two parallel longitudinal rows, the upper molars two or three parallel rows of cusps; hence the molars are multituberculate. *See* DENTITION; MAMMALIA; MULTITUBERCULATA.

D. Dwight Davis; Frederick S. Szalay

Allowance

An intentional difference in sizes of two mating parts. With running or sliding fits, in which mating parts move relative to each other, allowance is a clearance, usually for a film of oil. In this sense, allowance is the space "allowed" for motion between parts. To avoid binding between parts, a minimum clearance is critical for a running fit.

With force or shrink fits, in which mating parts, once assembled, are fixed in position relative to each other, allowance is an interference of metal; that is, a portion of metal in one part tends to occupy the same space as the adjacent portion of metal in the mating part. In this sense, allowance is the interference "allowed" to produce pressure between parts. To avoid breaking the external part, a maximum interference is critical for a drive, force, or shrink fit.

Fits are classed as running or sliding fits for parts that move freely against each other. Location fits provide accurate orientation of mating parts. Force fits require appreciable assembly pressure and produce more or less permanent assembly of parts. *See* FORCE FIT; LOCATION FIT; PRESS FIT; RUNNING FIT; SHRINK FIT; TOLERANCE.

Paul H. Black

Alloy

A metal product containing two or more elements as a solid solution, as an intermetallic compound, or as a mixture of metallic phases. This article will describe alloys on the basis of their technical applications. Alloys may also be categorized and described on the basis of compositional groups. *See* INTERMETALLIC COMPOUNDS; METAL; SOLID SOLUTION.

Except for native copper and gold, the first metals of technological importance were alloys. Bronze, an alloy of copper and tin, is appreciably harder than copper. This quality made bronze so important an alloy that it left a permanent imprint on the

civilization of several millennia ago now known as the Bronze Age. Today the tens of thousands of alloys involve almost every metallic element of the periodic table. *See* BRONZE; PERIODIC TABLE; PREHISTORIC TECHNOLOGY.

Alloys are used because they have specific properties or production characteristics that are more attractive than those of the pure, elemental metals. For example, some alloys possess high strength, others have low melting points, others are refractory with high melting temperatures, some are especially resistant to corrosion, and others have desirable magnetic, thermal, or electrical properties. These characteristics arise from both the internal and the electronic structure of the alloy. For a discussion of the physical structures of alloys and the interatomic forces in alloys. *See* ALLOY STRUCTURES.

Bearing alloys. These alloys are used for metals that encounter sliding contact under pressure with another surface; the steel of a rotating shaft is a common example. Most bearing alloys contain particles of a hard intermetallic compound that resist wear. These particles, however, are embedded in a matrix of softer material which adjusts to the hard particles so that the shaft is uniformly loaded over the total surface. The most familiar bearing alloy is babbitt metal, which contains 83–91% tin (Sn); the remainder is made up of equal parts of antimony (Sb) and copper (Cu), which form hard particles of the compounds SbSn and CuSn in a soft tin matrix (**Fig. 1**). Other bearing alloys are based on cadmium (Cd), copper, or silver (Ag). For example, an alloy of 70% copper and 30% lead (Pb) is used extensively for heavily loaded bearings. Bearings made by powder metallurgy techniques are widely used. These techniques are valuable because they permit the combination of materials which are incompatible as liquids, for example, bronze and graphite. Powder techniques also permit controlled porosity within the bearings so that they can be saturated with oil before being used, the so-called oilless bearings. *See* ANTIFRICTION BEARING; COPPER ALLOYS; POWDER METALLURGY; WEAR.

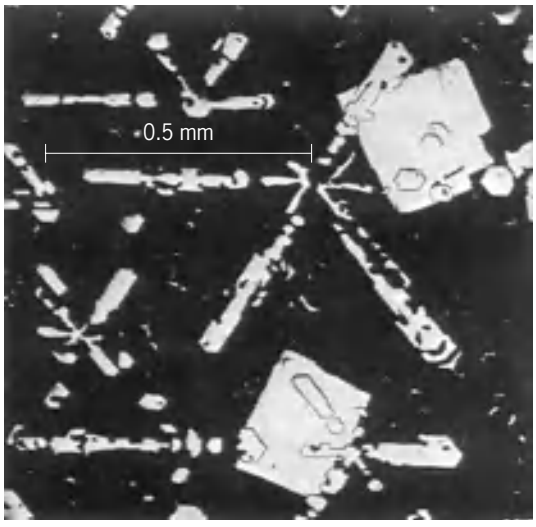


Fig. 1. Babbitt metal (84% Sn–7% Cu–9% Sb). The white areas are crystals of CuSn (star dendrites) and SbSn (nearly square); these resist wear. The dark matrix is a soft, tin-rich alloy, which adjusts to the hard particles to give uniform loading. (*From R. M. Brick et al., Structure and Properties of Alloys, 3d ed., McGraw-Hill, 1965*)

Corrosion-resisting alloys. Certain alloys resist corrosion because they are noble metals. Among these alloys are the precious-metal alloys, which will be discussed separately. Other alloys resist corrosion because a protective film develops on the metal surface. This passive film is an oxide which separates the metal from the corrosive environment. Stainless steels and aluminum alloys exemplify metals with this type of protection. Stainless steels are iron alloys containing more than 12% chromium (Cr). Steels with 18% Cr and 8% nickel (Ni) are the best known and possess a high degree of resistance to many corrosive environments. Aluminum (Al) alloys gain their corrosion-deterring characteristics by the formation of a very thin surface layer of aluminum oxide (Al_2O_3), which is inert to many environmental liquids. This layer is intentionally thickened in commercial anodizing processes to give a more permanent Al_2O_3 coating. Monel, an alloy of approximately 70% nickel and 30% copper, is a well-known corrosion-resisting alloy which also has high strength. Another nickel-base alloy is Inconel, which contains 14% chromium and 6% iron (Fe). The bronzes, alloys of copper and tin, also may be considered to be corrosion-resisting. *See* ALUMINUM ALLOYS; CORROSION; IRON ALLOYS; NICKEL ALLOYS; STAINLESS STEEL.

Dental alloys. Amalgams are predominantly alloys of silver and mercury (Hg), but they may contain minor amounts of tin, copper, and zinc (Zn) for hardening purposes, for example, 33% silver, 52% mercury, 12% tin, 2% copper, and less than 1% zinc. Liquid mercury is added to a powder of a precursor alloy of the other metals. After being compacted, the mercury diffuses into the silver-base metal to give a completely solid alloy. Gold-base dental alloys are preferred over pure gold (Au) because gold is relatively soft. The most common dental gold alloy contains gold (80–90%), silver (3–12%), and copper (2–4%). For higher strengths and hardnesses, palladium and platinum (up to 3%) are added, and the copper and silver are increased so that the gold content drops to 60–70%. Vitallium [an alloy of cobalt (65%), chromium (5%), molybdenum (3%), and nickel (3%)] and other corrosion-resistant alloys are used for bridgework and special applications. *See* AMALGAM; DENTISTRY; GOLD ALLOYS; SILVER ALLOYS.

Die-casting alloys. These alloys have melting temperatures low enough so that in the liquid form they can be injected under pressure into steel dies. Such castings are used for automotive parts and for office and household appliances which have moderately complex shapes. This processing procedure eliminates the need for expensive machining and forming operations. Most die castings are made from zinc-base or aluminum-base alloys. Magnesium-base alloys also find some application when weight reduction is paramount. Low-melting alloys of lead and tin are not common because they lack the necessary

strength for the above applications. A common zinc-base alloy contains approximately 4% aluminum and up to 1% copper. These additions provide a second phase in the metal to give added strength. The alloy must be free of even minor amounts (less than 100 ppm) of impurities such as lead, cadmium, or tin, because impurities increase the rate of corrosion. Common aluminum-base alloys contain 5–12% silicon (Si), which introduces hard-silicon particles into the tough aluminum matrix. Unlike zinc-base alloys, aluminum-base alloys cannot be electroplated; however, they may be burnished or coated with enamel or lacquer. Advances in high-temperature die-mold materials have focused attention on the die-casting of copper-base and iron-base alloys. However, the high casting temperatures introduce costly production requirements, which must be justified on the basis of reduced machining costs. *See* METAL CASTING.

Eutectic alloys. In certain alloy systems, a liquid of a fixed composition freezes to form a mixture of two basically different solids or phases. An alloy that undergoes this type of solidification process is called a eutectic alloy. A typical eutectic alloy is formed by combining 28.1% of copper with 71.9% of silver. A homogeneous liquid of this composition on slow cooling freezes to form a mixture of particles of nearly pure copper embedded in a matrix (background) of nearly pure silver.

The advantageous mechanical properties inherent in composite materials such as plywood composed of sheets or lamellae of wood bonded together and fiber glass in which glass fibers are used to reinforce a plastic matrix have been known for many years. Attention is being given to eutectic alloys because they are basically natural composite materials. This is particularly true when they are directionally solidified so as to yield structures with parallel plates of the two phases (lamellar structure) or long fibers of one phase embedded in the other phase (fibrous structure). Directionally solidified eutectic alloys are being given serious consideration for use in fabricating jet engine turbine blades. For this purpose eutectic alloys that freeze to form tantalum carbide (TaC) fibers in a matrix of a cobalt-rich alloy have been heavily studied. *See* EUTECTICS; METAL MATRIX COMPOSITE.

Fusible alloys. These alloys generally have melting temperatures below that of tin (450°F or 232°C), and in some cases as low as 120°F (50°C). Using eutectic compositions of metals such as lead, cadmium, bismuth, tin, antimony, and indium achieves these low melting temperatures. These alloys are used for many purposes, for example, in fusible elements in automatic sprinklers, forming and stretching dies, filler for thin-walled tubing that is being bent, and anchoring dies, punches, and parts being machined. Alloys rich in bismuth (Bi) were formerly used for type metal because these low-melting metals exhibited a slight expansion on solidification, thus replicating the font perfectly for printing and publication.

High-temperature alloys. Energy conversion is more efficient at high temperatures than at low; thus the need in power-generating plants, jet engines, and gas turbines for metals which have high strengths at high temperatures is obvious. In addition to having strength, these alloys must resist oxidation by fuel-air mixtures and by steam vapor. At temperatures up to about 1380°F (750°C), the austenitic stainless steels (18% Cr–8% Ni) serve well. An additional 180°F (100°C) may be realized if the steels also contain 3% molybdenum. Both nickel-base and cobalt-base alloys, commonly categorized as superalloys, may serve useful functions up to 2000°F (1100°C). Nichrome, a nickel-base alloy containing 12–15% chromium and 25% iron, is a fairly simple superalloy. More sophisticated alloys invariably contain five, six, or more components; for example, an alloy called René-41 contains approximately 19% chromium, 1.5% aluminum, 3% titanium (Ti), 11% cobalt (Co), 10% molybdenum, 3% iron, 0.1% carbon (C), 0.005% boron (B), and the balance nickel. Other alloys are equally complex. The major contributor to strength in these alloys is the solution-precipitate phase of Ni_3 (Ti,Al), γ'. It provides strength because it is coherent with the nickel-rich γ phase. Cobalt-base superalloys may be even more complex and generally contain carbon which combines with the tungsten and chromium to produce carbides that serve as the strengthening agent (**Fig. 2**). In general, the cobalt-base superalloys are more resistant to oxidation than the nickel-base alloys are, but they are not as strong. Molybdenum-base alloys have exceptionally high strength at high temperatures, but their brittleness at lower temperatures and their poor oxidation resistance at high temperatures have limited their use. However, coatings permit the use of such alloys in an oxidizing atmosphere, and they are finding increased application. A group of materials called cermets, which are mixtures of metals and compounds

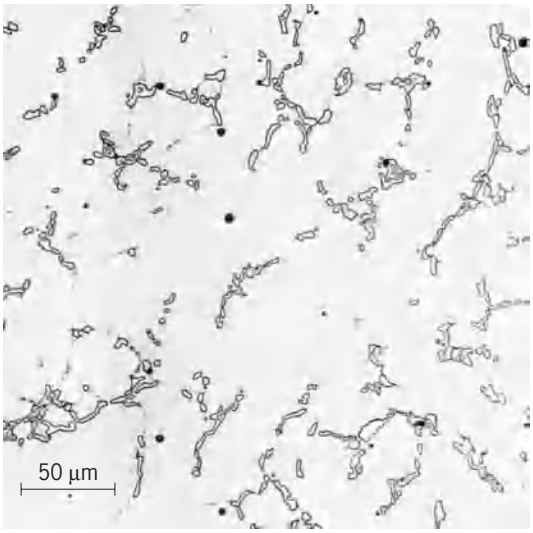


Fig. 2. Cobalt-base superalloy HS31. The dispersed phase is a carbide, $(CoCrW)_6C$, which strengthens the metal. (*Haynes Stellite Co.*)

such as oxides and carbides, have high strength at high temperatures, and although their ductility is low, they have been found to be usable. One of the better-known cermets consists of a mixture of titanium carbide and nickel, the nickel acting as a binder or cement for the carbide. *See* CERMET.

Joining alloys. Metals are bonded by three principal procedures: welding, brazing, and soldering. Welded joints melt the contact region of the adjacent metal; thus the filler material is chosen to approximate the composition of the parts being joined. Brazing and soldering alloys are chosen to provide filler metal with an appreciably lower melting point than that of the joined parts. Typically, brazing alloys melt above 750°F (400°C), whereas solders melt at lower temperatures. A 57% copper–42% zinc–1% tin brass is a general-purpose alloy for brazing steel and many nonferrous metals. A silicon-aluminum eutectic alloy is used for brazing aluminum, and an aluminum-containing magnesium eutectic alloy brazes magnesium parts. The most common solders are based on lead-tin alloys. The prevalent 60% tin–40% lead solder is eutectic in composition and is used extensively for electrical circuit production, in which temperature limitations are critical. A 35% tin–65% lead alloy has a range of solidification and is thus preferred as a wiping solder by plumbers. *See* BRAZING; SOLDERING; WELDING AND CUTTING OF METALS.

Light-metal alloys. Aluminum and magnesium (Mg), with densities of 1.6 and 1.01 oz/in.^{3} (2.7 and 1.75 g/cm^3), respectively, are the bases for most of the light-metal alloys. Titanium (2.6 oz/in.^{3} or 4.5 g/cm^3) may also be regarded as a light-metal alloy if comparisons are made with metals such as steel and copper. Aluminum and magnesium must be hardened to receive extensive application. Age-hardening processes are used for this purpose. Typical alloys are 90% aluminum–10% magnesium, 95% aluminum–5% copper, and 90% magnesium–10% aluminum. Ternary (three-element) and more complex alloys are very important light-metal alloys because of their better properties. The aluminum-zinc-magnesium system of alloys, used extensively in aircraft applications, is a prime example of one such alloy system. *See* ALUMINUM; MAGNESIUM; TITANIUM.

Low-expansion alloys. This group of alloys includes Invar (64% iron–36% nickel), the dimensions of which do not vary over the atmospheric temperature range. It has special applications in watches and other temperature-sensitive devices. Glass-to-metal seals for electronic and related devices require a matching of the thermal-expansion characteristics of the two materials. Kovar (54% iron–29% nickel–17% cobalt) is widely used because its expansion is low enough to match that of glass. *See* THERMAL EXPANSION.

Magnetic alloys. Soft and hard magnetic materials involve two distinct categories of alloys. The former consists of materials used for magnetic cores of transformers and motors, and must be magnetized and demagnetized easily. For alternating-current applications, silicon-ferrite is commonly used. This is an alloy of iron containing as much as 5% silicon. The silicon has little influence on the magnetic properties of the iron, but it increases the electric resistance appreciably and thereby decreases the core loss by induced currents. A higher magnetic permeability, and therefore greater transformer efficiency, is achieved if these silicon steels are grain-oriented so that the crystal axes are closely aligned with the magnetic field. Permalloy (78.5% nickel–21.5% iron) and some comparable cobalt-base alloys have very high permeabilities at low field strengths, and thus are used in the communications industry. Ceramic ferrites, although not strictly alloys, are widely used in high-frequency applications because of their low electrical conductivity and negligible induced energy losses in the magnetic field. Permanent or hard magnets may be made from steels which are mechanically hardened, either by deformation or by quenching. Some precipitation-hardening, iron-base alloys are widely used for magnets. Typical of these are the Alnicos, for example, Alnico-4 (55% iron–28% nickel–12% aluminum–5% cobalt). Since these alloys cannot be forged, they must be produced in the form of castings. Hard magnets are being produced from alloys of cobalt and the rare-earth type of metals. The compound RCo_5, where R is samarium (Sm), lanthanum (La), cerium (Ce), and so on, has extremely high coercivity.

Precious-metal alloys. In addition to their use in coins and jewelry, precious metals such as silver, gold, and the heavier platinum metals are used extensively in electrical devices in which contact resistances must remain low, in catalytic applications to aid chemical reactions, and in temperature-measuring devices such as resistance thermometers and thermocouples. The unit of alloy impurity is commonly expressed in karats, when each karat is 1/24 part. The most common precious-metal alloy is sterling silver (92.5% silver, with the remainder being unspecified, but usually copper). The copper is very beneficial in that it makes the alloy harder and stronger than pure silver. Yellow gold is an Au-Ag-Cu alloy with approximately a 2:1:1 ratio. White gold is an alloy which ranges from 10 to 18 karats, the remainder being additions of nickel, silver, or zinc, which change the color from yellow to white. The alloy 87% platinum (Pt)–13% rhodium (Rh), when joined with pure platinum, provides a widely used thermocouple for temperature measurements in the 1830–3000°F (1000–1650°C) temperature range. *See* GOLD; SILVER.

Shape memory alloys. These alloys have a very interesting and desirable property. In a typical case, a metallic object of a given shape is cooled from a given temperature T_1 to a lower temperature T_2 where it is deformed so as to change its shape. Upon reheating from T_2 to T_1, the shape change accomplished at T_2 is recovered so that the object returns to its original configuration. This thermoelastic property of the shape memory alloys is associated with the fact that they undergo a martensitic phase transformation (that is, a reversible change in crystal structure

that does not involve diffusion) when they are cooled or heated between T_1 and T_2.

For a number of years the shape memory materials were essentially scientific curiosities. Among the first alloys shown to possess these properties was one of gold alloyed with 47.5% cadmium. Considerable attention has been given to an alloy of nickel and titanium known as nitinal. The interest in shape memory alloys has increased because it has been realized that these alloys are capable of being employed in a number of useful applications. One example is for thermostats; another is for couplings on hydraulic lines or electrical circuits. The thermoelastic properties can also be used, at least in principle, to construct heat engines that will operate over a small temperature differential and will thus be of interest in the area of energy conversion. *See* SHAPE MEMORY ALLOYS.

Thermocouple alloys. These include Chromel, containing 90% nickel and 10% chromium, and Alumel, containing 94% nickel, 2% aluminum, 3% chromium, and 1% silicon. These two alloys together form the widely used Chromel-Alumel thermocouple, which can measure temperatures up to 2200°F (1204°C). Another common thermocouple alloy is constantan, consisting of 45% nickel and 55% copper. It is used to form iron-constantan and copper-constantan couples, used at lower temperatures. For precise temperature measurements and for measuring temperatures up to 3000°F (1650°C), thermocouples are used in which one metal is platinum and the other metal is platinum plus either 10 or 13% rhodium. *See* LASER ALLOYING; STEEL; THERMOCOUPLE. Lawrence H. Van Vlack; Robert E. Reed-Hill

Prosthetic alloys. As discussed here, prosthetic alloys are alloys used in internal prostheses, that is, surgical implants such as artificial hips and knees. External prostheses are devices that are worn by patients outside the body; alloy selection criteria are different from those for internal prostheses. In the United States, surgeons use about 250,000 artificial hips and knees and about 30,000 dental implants per year.

Alloy selection criteria for surgical implants can be stringent primarily because of biomechanical and chemical aspects of the service environment. Mechanically, an implant's properties and shape must meet anticipated functional demands; for example, hip joint replacements are routinely subjected to cyclic forces that can be several times body weight. Therefore, intrinsic mechanical properties of an alloy, for example, elastic modulus, yield strength, fatigue strength, ultimate tensile strength, and wear resistance, must all be considered. Likewise, because the pH and ionic conditions within a living organism define a relatively hostile corrosion environment for metals, corrosion properties are an important consideration. Corrosion must be avoided not only because of alloy deterioration but also because of the possible physiological effects of harmful or even cytotoxic corrosion products that may be released into the body. (Study of the biological effects of biomaterials is a broad subject in itself, often referred to as biocompatibility.) The corrosion resistance of all modern alloys stems primarily from strongly adherent and passivating surface oxides, such as titanium oxide (TiO_2) on titanium-based alloys and chromium oxide (Cr_2O_3) on cobalt-based alloys.

The most widely used prosthetic alloys therefore include high-strength, corrosion-resistant ferrous, cobalt-based, or titanium-based alloys: for example, cold-worked stainless steel; cast Vitallium; a wrought alloy of cobalt, nickel, chromium, molybdenum, and titanium; titanium alloyed with aluminum and vanadium; and commercial-purity titanium. Specifications for nominal alloy compositions are designated by the American Society for Testing and Materials (ASTM).

Prosthetic alloys have a range of properties. Some are easier than others to fabricate into the complicated shapes dictated by anatomical constraints. Fabrication techniques include investment casting (solidifying molten metal in a mold), forging (forming metal by deformation), machining (forming by machine-shop processes, including computer-aided design and manufacturing), and hot isostatic pressing (compacting fine powders of alloy into desired shapes under heat and pressure). Cobalt-based alloys are difficult to machine and are therefore usually made by casting or hot isostatic pressing. Some newer implant designs are porous-coated, that is, they are made from the standard ASTM alloys but are coated with alloy beads or mesh applied to the surface by sintering or other methods. The rationale for such coatings is implant fixation by bone ingrowth.

Some alloys are modified by nitriding or ion-implantation of surface layers of enhanced surface properties. A key point is that prosthetic alloys of identical composition can differ substantially in terms of structure and properties, depending on fabrication history. For example, the fatigue strength approximately triples for hot isostatically pressing versus as-cast Co-Cr-Mo alloy, primarily because of a much smaller grain size in the microstructure of the former.

No single alloy is vastly superior to all others; existing prosthetic alloys have all been used in successful and, indeed, unsuccessful implant designs. Alloy selection is only one determinant of performance of the implanted device. *See* METAL, MECHANICAL PROPERTIES OF; PROSTHESIS. John Brunski

Superconducting alloys. Superconductors are materials that have zero resistance to the flow of electric current at low temperatures. There are more than 6000 elements, alloys, and compounds that are known superconductors. This remarkable property of zero resistance offers unprecedented technological advances such as the generation of intense magnetic fields. Realization of these new technologies requires development of specifically designed superconducting alloys and composite conductors. An alloy of niobium and titanium (NbTi) has a great number of applications in superconductivity; it becomes superconducting at 9.5 K (critical superconducting temperature, T_c). This alloy is preferred because of its ductility and its ability to carry large

amounts of current at high magnetic fields, represented by $J_c(H)$ [where J_c is the critical current and H is a given magnetic field], and still retain its superconducting properties. Brittle compounds with intrinsically superior superconducting properties are also being developed for magnet applications. The most promising of these are compounds of niobium and tin (Nb_3Sn), vanadium and gallium (V_3Ga), niobium and germanium (Nb_3Ge), and niobium and aluminum (Nb_3Al) which have higher T_c (15 to 23 K) and higher $J_c(H)$ than NbTi.

Superconducting materials possess other unique properties such as magnetic flux quantization and magnetic-field-modulated supercurrent flow between two slightly separated superconductors. These properties form the basis for electronic applications of superconductivity such as high-speed computers or ultrasensitive magnetometers. Development of these applications began using lead or niobium (T_c of 7 K and 9 K) in bulk form, but the emphasis then was transferred to materials deposited in thin-film form. Lead-indium (PbIn) and lead-gold (PbAu) alloys are more desirable than pure lead films, as they are more stable. Improved vacuum deposition systems eventually led to the use of pure niobium films as they, in turn, were more stable than lead alloy films. Advances in thin-film synthesis techniques led to the use of the refractory compound niobium nitride (NbN) in electronic applications. This compound is very stable and possesses a higher T_c (15 K) than either lead or niobium. *See* REFRACTORY.

Novel high-temperature superconducting materials may have revolutionary impact on superconductivity and its applications. These materials are ceramic, copper oxide–based materials that contain at least four and as many as six elements. Typical examples are yttrium-barium-copper-oxygen (T_c 93 K); bismuth-strontium-calcium-copper-oxygen (T_c 110 K); and thallium-barium-calcium-copper-oxygen (T_c 125 K). These materials become superconducting at such high temperatures that refrigeration is simpler, more dependable, and less expensive. Much research and development has been done to improve the technologically important properties such as $J_c(H)$, chemical and mechanical stability, and device-compatible processing procedures. It is anticipated that the new compounds will have a significant impact in the growing field of superconductivity. *See* CERAMICS; SUPERCONDUCTIVITY. D. Gubser

Bibliography. G. S. Brady, H. R. Clauser, and J. A. Vaccari, *Materials Handbook*, 15th ed., 2002; W. D. Callister, *Materials Science and Engineering: An Introduction*, 6th ed., 2002; J. Frick (ed.), *Woldman's Engineering Alloys*, 9th ed., 2001; J. A. Helsen and H. J. Breme, *Metals as Biomaterials*, 1998; D. P. Henkel and A. Pense, *Structures and Properties of Engineering Materials*, 5th ed., 2001; J. Mordike and P. Haasen, *Physical Metallurgy*, 3d ed., 1996; C. T. Lynch (ed.), *Handbook of Materials Sciences*, vol. 2, 1975; L. H. Van Vlack, *Elements of Material Science and Engineering*, 6th ed., 1989.

Alloy structures

Metals in actual commercial use are almost exclusively alloys, and not pure metals, since it is possible for the designer to realize an extensive variety of physical properties in the product by varying the metallic composition of the alloy. As a case in point, commercially pure or cast iron is very brittle because of the small amount of carbon impurity always present, while the steels are much more ductile, with greater strength and better corrosion properties. In general, the highly purified single crystal of a metal is very soft and malleable, with high electrical conductivity, while the alloy is usually harder and may have a much lower conductivity. The conductivity will vary with the degree of order of the alloy, and the hardness will vary with the particular heat treatment used. For commercial applications of alloys and other information *see* ALLOY

The basic knowledge of structural properties of alloys is still in large part empirical, and indeed, it will probably never be possible to derive formulas which will predict which metals to mix in a certain proportion and with a certain heat treatment to yield a specified property or set of properties. However, a set of rules exists which describes the qualitative behavior of certain groups of alloys. These rules are statements concerning the relative sizes of constituent atoms for alloy formation, and concerning what kinds of phases to expect in terms of the valence of the constituent atoms. The rules were discovered in a strictly empirical way, and for the most part, the present theoretical understanding of alloys consists of rudimentary theories which describe how the rules arise from the basic principles of physics.

Prior to a discussion of the rules proposed by W. Hume-Rothery concerning the binary substitutional alloys, some information concerning alloy composition diagrams will be presented. Robb M. Thomson

Phase Diagrams

Alloys are classified as binary alloys, composed of two components; as ternary alloys, composed of three components; or as multicomponent alloys. Most commercial alloys are multicomponent. The composition of an alloy is described by giving the percentage (either by weight or by atoms) of each element in it.

A phase diagram is a graphical description of the kinds and amounts of the phases that can be expected in an alloy as a function of its composition, temperature, and pressure when it has reached thermodynamic equilibrium. The phases may be liquid, vapor, or solid with various ordered and disordered crystal structures. A phase diagram does not provide information about how rapidly equilibrium can be reached; when a phase diagram is determined experimentally, it is necessary not only to find out what phases are present but also to assure that the alloy is in the stable equilibrium state. Equilibrium is reached when the Gibbs free energy of the system has reached its lowest possible value. The thermodynamic principle of minimum Gibbs energy imposes

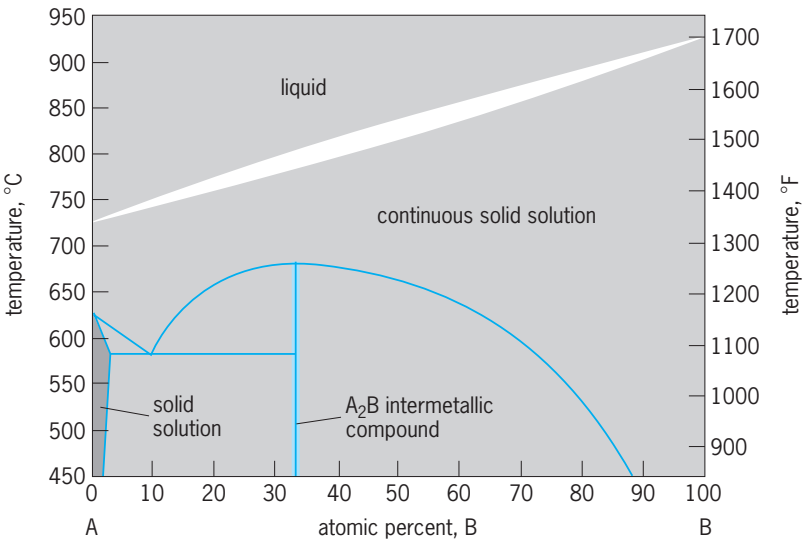


Fig. 1. Phase diagram at fixed pressure for a hypothetical alloy composed of elements A and B.

restrictions on the possible forms a phase diagram can take, and even a complicated diagram with many phases is composed of only a few geometrical features. *See* FREE ENERGY; PHASE EQUILIBRIUM; THERMODYNAMIC PRINCIPLES.

Figure 1 shows a phase diagram for a hypothetical alloy system composed of elements A and B. As with most phase diagrams, it is drawn at fixed pressure. Hence it is a two-dimensional figure with composition and temperature as the axes. There are four equilibrium phases in the A-B system: the liquid, two solid phases with crystal structures of the pure components, and an intermetallic compound. The intermetallic compound is an orderly arrangement of two A atoms for every B atom on a crystal lattice. The other solid phases are disordered solutions of A and B atoms on different crystal lattices, one with a complete range of solubility and one with only a small solubility of B in A. The phase diagram for this system maps the melting and solidification

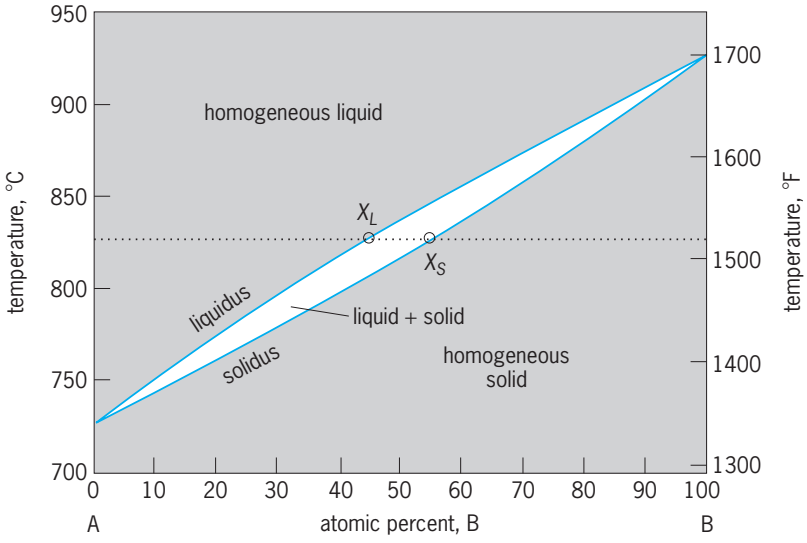


Fig. 2. Detail of the A-B system showing the liquidus and solidus.

temperatures, and the solubilities of the elements in each other and in the intermetallic compound. In the shaded regions the labeled phases are the only stable states of the alloys: in these regions the alloys are homogeneous liquid, solid solution, or A_2B compound. In the other regions of the diagram, two phases of different composition can coexist in two-phase, or heterogeneous, equilibrium. Two key features of the A-B diagram are the two-phase equilibrium and the three-phase reaction, and these are illustrated in more detail in **Figs. 2** and **3**. *See* SOLID SOLUTION; SOLID-STATE CHEMISTRY; THERMODYNAMIC PROCESSES.

Two-phase equilibrium. Figure 2 illustrates the concept of two-phase equilibrium using the specific example of the melting behavior of the A-B system. The pure elements A and B melt completely; each has a single melting-point temperature. The alloys, however, begin melting at one temperature but do not finish melting until they reach a higher temperature. In between, the alloy is slushy, and liquid and solid coexist in equilibrium with each other. Upon cooling the liquid alloy, the first small amount of solid to precipitate as the alloy solidifies is richer in element B than the original liquid. This phenomenon is called segregation of the two constituents. The lens region of the diagram is the region of the two-phase equilibrium between liquid and solid. The upper line where the liquid begins to solidify is the liquidus, and the lower line where the solid first begins to melt is the solidus.

For an alloy of overall composition x, the compositions of liquid and solid, x_l and x_s, in equilibrium at a particular temperature are found by reading off the compositions of the liquidus and solidus at that temperature. From the overall composition x, the amounts of liquid and solid, f_l and f_s, can also be calculated. For a two-phase alloy, the overall composition is obtained by averaging the compositions of liquid and solid, weighted by the amounts present. Since the overall composition is not changed during segregation, Eqs. (1) are valid. These equations are

$$\begin{aligned} x &= f_l x_l + f_s x_s \\ f_l &+ f_s = 1 \end{aligned} \tag{1}$$

a statement of the lever rule, one of the important rules for reading phase diagrams.

Liquid and solid equilibrium for fixed pressure over a range of temperatures does not occur in pure substances but only in alloys. The phase diagram can be used to predict how molten A-B alloys will solidify in different conditions. If the temperature of a liquid alloy is brought below the liquidus, a small amount of solid will form with a composition different from that of the liquid. How the solidification proceeds now depends on the rate of cooling. If the alloy is cooled very slowly through the two-phase region, equilibrium can be reached at each temperature and the compositions of the liquid and solid can continuously adjust to the values on the phase diagram. If the cooling rate is too high for equilibrium to be reached at each temperature, which is typically the

case, the phase diagram can still predict what can happen. The first solid to form is rich in element B compared to the liquid. The solid does not have time to readjust its composition by diffusion of element B back out of the solid, so the liquid becomes more and more enriched in element A. The solid ends up inhomogeneous in composition, and its microstructure depends on the phase diagram and the cooling rate.

The liquid-solid phase equilibria shown schematically in Fig. 2 make up the complete phase diagram for many systems, for example, silver-gold. Silver and gold are completely miscible in both the liquid and solid states and form no intermetallic compounds because of the close similarity of their chemical properties and atom sizes.

Three-phase reaction. The concept of the three-phase reaction is illustrated in Fig. 3 by the solidification behavior of the silver-copper alloy system. Silver and copper are not as similar to each other as are silver and gold. In the solid state, silver and copper do not mix in all proportions, and it is possible to have inhomogeneous alloys with two different solid phases in equilibrium, as well as solid phases in equilibrium with the liquid. Molten alloys can solidify to form copper-rich or silver-rich solids. The three two-phases fields intersect at a horizontal line, the eutectic horizontal. If a silver-rich liquid is slowly cooled, it separates into a silver-rich solid and a copper-rich liquid. However, below the eutectic temperature, the liquid is no longer a stable phase of the system. The silver-rich solid and the copper-rich liquid react to the eutectic temperature to form a new phase, solid copper. The reaction continues at the eutectic temperature until all the liquid is consumed, leaving a two-phase mixture of silver and copper. *See* EUTECTICS.

Minimization of Gibbs energy. The concepts of the two-phase equilibrium and the three-phase reaction describe all the ingredients of the phase diagrams of binary alloys. No more than three phases can ever coexist in equilibrium in a binary system at constant pressure. No matter how many complicated phases appear in the phase diagram, the regions are either single-phase or two-phase; horizontal lines represent the three-phase temperatures.

This property of binary phase diagrams stems from the thermodynamic principle of the minimization of the Gibbs energy. In alloy systems, each phase has a Gibbs energy function of composition and temperature at fixed pressure. For a pure substance, composition does not enter the picture, but there are several Gibbs energy functions of temperature for the different phases. Whichever function is lower at the temperature of interest is the stable phase. In a binary alloy system, the composition variable provides a new way for the system to lower the Gibbs free energy. **Figure 4** shows the Gibbs energy functions for liquid and solid for a hypothetical system. At the composition x, the solid phase has a lower Gibbs energy than the liquid and so is more stable. However, if the alloy segregates into a liquid of composition x_l and a solid of composition x_s, the Gibbs energy is lower yet than for either a uniform solid or a uniform liquid. The only constraint on the way the alloy can split up is that the slopes of the Gibbs energies of the two phases in equilibrium must be equal. The slope of the Gibbs energy regulates the composition exchange taking place between the two phases, and at equilibrium the compositions must be constant. The compositions can be determined by drawing a common tangent to the two Gibbs energy functions; the points of tangency determine x_l and x_s. This method of finding a two-phase equilibrium on the phase diagram is called the common tangent construction.

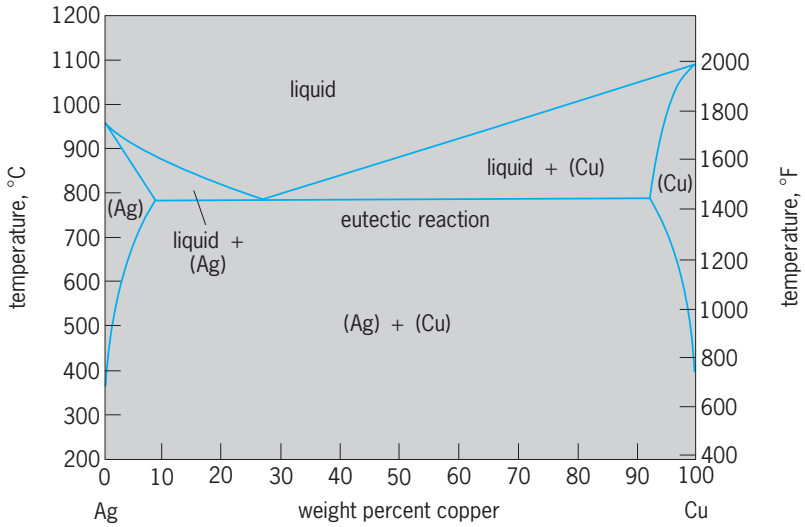


Fig. 3. Phase diagram of the silver-copper system, which exhibits the eutectic reaction. Silver-rich and copper-rich phases are denoted by (Ag) and (Cu), respectively.

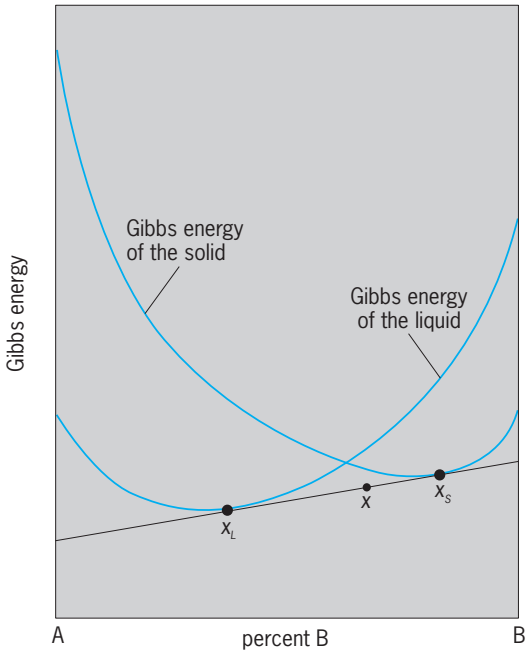


Fig. 4. Gibbs energies at fixed pressure and temperature as a function of concentration, and common tangent construction, for a hypothetical system.

The phase diagram is mapped out by determining the positions of common tangents to the Gibbs energy functions at various temperatures. If two Gibbs energy functions intersect, it is always possible to draw a common tangent. This is why the two-phase fields can extend over a range of temperature. To have three phases in equilibrium, however, three Gibbs energy functions must line up along a common tangent. This can only happen at one temperature, above and below which the Gibbs energies will move back out of line. Four phases cannot all be lined up. This is a statement of the Gibbs phase rule for binary systems. *See* PHASE RULE.

Ternary and multicomponent systems. In a ternary system such as silver-gold-copper, the composition is represented by two variables, for example, the percentage of silver and the percentage of gold. The temperature is also still a variable. The phase diagram must be drawn in three dimensions. The Gibbs energy functions are surfaces, and the common tangent is a tangent plane rather than a line. It is possible to find a plane tangent to three surfaces over a range of temperatures, so that for ternary systems the maximum number of phases in equilibrium is four, and increases by one for each element added. The same principles of construction of the three-dimensional phase diagram work for ternary systems, with one to four phases in equilibrium in each region, but the representation of the diagrams can become complicated.

Nonequilibrium processes. The state of a real alloy depends not only on the equilibrium shown on the phase diagram, but also on how far away from equilibrium the system has been brought by the temperature treatments it has been subjected to and on how close to equilibrium the system can approach in the time given it. The system always behaves in accordance with the thermodynamics and Gibbs energy functions that underlie the equilibrium phase diagram. If an equilibrium phase is not able to nucleate and grow, then a metastable equilibrium might be achieved between two other competing phases. A phase can sometimes be taken so far from its equilibrium range that it becomes very unstable and can rapidly transform to another structure by one of various nonequilibrium transformation processes. The materials scientist uses thermodynamic ideas to read from the equilibrium phase diagrams the possibilities of these processes.

In short, the phase diagram is a road map that guides the materials scientist through problems encountered in alloy design, design of processing techniques, failure analysis, and behavior of alloys in performance. Moreover, phase diagrams are not used only for metallurgical systems; the behavior of ceramics and minerals, and even mixtures of polymers, is also described by phase diagrams.

Joanne L. Murray; Kirit J. Bhansali

Size and Valence Factors

Comparison studies of a large number of alloys have led to the rules for alloy formation formulated by W. Hume-Rothery and others. These rules have recognized two general types of criteria regarding alloy formation. One relates to the relative size of the atomic constituents, the other to the valence, or the electronic behavior, of the two metals. These criteria have generally been treated independently of one another, although in reality some interaction can be expected. *See* VALENCE.

Size factor. The empirical rules for alloy formation state that, for substitutional alloys, the size of the constituents must be approximately the same, whereas one of the constituents of an interstitial alloy must be small compared to the other. Simply stated, the size factor recognizes the importance of choosing two metals which can fit together in a lattice structure. In order to define the concept of size, one must refer to the general ideas of the band theory of solids. According to this theory, the valence electrons of the metal atoms are detached from the immediate vicinity of the atom and contribute to the conduction band. In the conduction band the electrons have many of the characteristics of free particles similar to the atoms of a gas. The conduction electrons are spread over the metal, filling the interstices between the remaining ion cores of the metal atoms. The electrons of the inner shells remain tightly bound to the individual ions, however. *See* BAND THEORY OF SOLIDS; FREE-ELECTRON THEORY OF METALS.

It is assumed in any discussion of atomic size that the free electrons cause a rather weak attractive force drawing the metal together, while the ion cores strongly repel one another when one core overlaps a neighbor. The strength and form of the attractive force obviously depends on the crystal type; however, the ion cores are more nearly independent of their surroundings. To the extent that the size of the atom in the metal is given by the radius of the ion core, the concept of size is well defined. However, in those cases where some of the repulsive force is contributed by the structure-dependent free electrons, the size of an atom in an alloy will vary from one alloy to another.

To the extent that the size of an ion core can be rigorously defined, the alloy will contain large misfit stresses unless the ion-core sizes of the two metal atoms have a favorable ratio.

In the primary or terminal solid solution parts of the composition diagram, there are two simple possibilities. The atoms of the minority category (the solute atoms) are able to supplant the solvent atoms on the solvent atom lattice, or they may fit into the open spaces of the solvent lattice. The first case is known as a substitutional alloy, while the second is called an interstitial alloy.

Interstitial alloys. Examples of interstitial alloys are some of the alloys of the transition elements and, most familiarly, those of iron. The iron lattice is face-centered cubic at medium temperatures. If the face-centered lattice is considered to be made up of hard spheres of radius a_0, the largest atom which can be fitted into an interstitial position (the cube center) has a radius $0.59a_0$. There are only four neutral atoms which have smaller radii than this value for the transition-metal group. They are hydrogen, carbon,

nitrogen, and boron. Carbon is actually an exception to the rule for iron, as it has a radius of $0.63a_0$ for the iron lattice. Actually, these four elements are not metals in their normal state, even though they do form metallic alloys with the transition metals. *See* CRYSTAL DEFECTS; CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

Substitutional alloys. In the substitutional primary solutions the solute atom takes the place of one of the solvent atoms. In this case the size of the solvent and solute must be nearly the same. It is possible to calculate in a crude way the maximum permissible difference in size. The size difference is reflected in a lattice distortion, which is a contribution to the internal energy of the system.

In general, the formation of a homogeneous alloy is possible if the change in the free energy of the alloy relative to the two separate elements is negative, as in Eq. (2). Here F is the Helmholtz free energy, U is

$$\Delta F = \Delta U - T\Delta S \tag{2}$$

the internal energy, T is the absolute temperature, and S is the entropy. The distortion energy increases ΔU. On the other hand, there is an entropy of mixing when the solute atoms are distributed in a random fashion on the solvent lattice. Thus, if the distortion energy is not too large to overbalance the entropy change, the alloy can form. An estimate of the distortion energy as a function of size suggests that the size of the solute atom must be within about 15% of the size of the solvent atom. This number agrees with the empirically derived Hume-Rothery rule. *See* ENTROPY.

The substitutional alloys are the commonest type of alloy structure and have received the most study. Cu-Au is a good example of the substitutional alloy. The atomic sizes are very close, and the electronic structures of Cu and Au are very similar. As a result, this system forms a single primary solid solution system from one end of the composition diagram to the other. The only complexities are due to ordering phenomena, which will be discussed later.

Valence factor. In addition to the effects of the size of the ions on the formation of alloys, the electronic structure of the atoms involved also plays a role. It is the electronic configuration of the atoms which determines why mixtures of some elements form metallic alloys, whereas some form insulating compounds. Qualitatively, it is found that alloys are formed from the atoms of the middle of the electrochemical series of the elements. The reason is that there is always a tendency for the ions of the metallic state to polarize with respect to their neighbors and form an ionic solid instead of the metallic one. If the tendency is strong, no metallic alloy state can be formed; hence, only elements from the middle of the series, where there is little change in the ionization potential from one atom to another, can form successful alloys. *See* ELECTROCHEMICAL SERIES.

Stoichiometry. Closely related to the polarization effect is the reason why the alloys are nonstoichiometric (that is, why they form mixtures which do not consist of small-number ratios of one element to the other) and form homogeneous phases over wide ranges of composition. For example, copper and gold mix homogeneously for any composition. This result is in striking contrast to the behavior of the ionic compounds such as sodium chloride, NaCl. *See* STOICHIOMETRY.

In NaCl an excess Na ion is bound in the crystal with an energy less by about 1 eV than a normal ion, and this energy discrepancy is large enough so that, when the crystal becomes far from stoichiometric, the homogeneous compound is no longer formed. The excess constituent forms a separate phase, either as Na metal crystals embedded in the matrix or as bubbles of Cl_2. The energy discrepancy in the case of NaCl is primarily due to the excess of charge of one sign or the other when the wrong ion is on a lattice site. Since the lattice is completely ionized, the excess charge due to the excess ion amounts to the complete ionic charge. *See* IONIC CRYSTALS.

The situation for a metal is completely different. The charge of the ionic core of a metal is balanced by the charge of the free electrons which each atom contributes to the conduction band. Thus, in a perfect metal, the volume occupied in the lattice by each atom is neutral. If a single gold atom is placed substitutionally in a copper lattice, there will be a tendency for the charge in the gold lattice volume to become unbalanced because the copper and gold atoms have different ionization energies. Effectively, the copper and gold atoms in the metal have different affinities for the free electrons of the valence band; hence there will be a small charge imbalance induced around the gold lattice site. However, in the case of Au-Cu, the chemical valence is equal, and the amount of polarization inducted is a small fraction of a unit charge.

Screening length. A further effect is the screening of the polarization around the gold atom by the conduction electrons themselves. In this case, when the region around the gold is polarized, a voltage difference is also generated. Hence, the electrons of the metal tend to rush into the affected region to even out the discrepancy. However, it is not possible to redress the balance completely, and a region of the order of the size of the gold atomic volume itself remains polarized. The size of this region is indicated by the screening length of the electrons of the metal, which are here listed for several metals in units of lattice spacing a_0: Cu, 1.1; Al, 1.4; Tl, 1.0; Fe, 3.2; and Ni, 4.8.

The effect of replacing a copper atom with a gold atom in Cu-Au is thus seen to be very much less than the effect of replacing a Cl ion with an Na ion in NaCl. For this reason the copper lattice can accept an unlimited number of exchanges without breaking up its structure, and the alloy forms over the entire composition range. Cu-Au is a rather special example to pick, since the ionization potentials of Cu and Au are very close, and the size factor is very favorable. The discussion, however, does illustrate the reason why these metals do not form compounds in the chemical sense.

Much the same reasoning would be applicable for showing why covalent crystals should also form stoichiometric compounds, where dangling covalent bonds would have much the same effect on the energy of the crystal as does an excess charge in an ionic lattice. *See* CHEMICAL BONDING.

Electron Compounds

Even though no metal alloy forms a compound in the chemical sense, there is a characteristic behavior when the alloying constituents have differing valences, and these alloys go under the name of electron compounds. In the usual alloy the intermediate region of the phase diagram is covered by several intermediate phases, with more than the one such phase illustrated in Fig. 1. The term electron compound refers to the fact that there is considerable regularity in the types of intermediate phases which are observed.

Structure and electron density. It appears that in numerous cases there is a tendency for a particular type of crystal structure to correspond to a particular electron density in the conduction band. Thus, when zinc is alloyed to copper, near the 50% composition there is a phase change from the face-centered-cubic (fcc) lattice of copper to a body-centered-cubic (bcc) lattice. If it is assumed that all the valence electrons of all the atoms are contributed to the conduction band, there will be an average of 1.5 electrons per atom of the alloy in this band, since the valence of copper is 1 and the valence of zinc is 2. For the alloys Cu_3Al and Cu_5Sn, the crystal structure is also bcc for the same average free electron concentration. T. B. Massalsky has listed (see **table**) the intermediate phases which center about the characteristic electron concentrations shown. Note that there are three different compounds centered about the electron concentration 1.5 in the table. There are also two other well-developed electron compounds, one at about concentration 1.6 and one at about concentration 1.75. Also, there are two places where the hexagonal close-packed (hcp) structures occur, one at 1.5 and one at 1.75. For example, in the case of Cu-Si, the hcp structure is interrupted at 1.6 by a change to the phase called the γ-brass structure, which is a very complex crystal type with 52 atoms per unit cell.

Brillouin zones. The striking correlations of the various electron compounds in the empirical tabulation of the table has been interpreted on the basis of the free-electron theory of metals. H. Jones has used the free-electron theory of metals modified by perturbation theory to derive an expression for the difference in energy of the electrons of the fcc and bcc crystals. His result is that as the Brillouin zone is filled, at first the energy is very nearly equal for the two crystals. The energy of the fcc crystal then drops below that of the bcc crystal at an electron concentration of about 1 electron per atom; but at a concentration close to 1.5, the situation reverses and the bcc crystal becomes more stable. Similar discussions have been given for the hexagonal lattices; these show that certain distortions of the symmetry of the perfect hexagonal lattice which occur are due to electronic effects related to filling the Brillouin zone. *See* BRILLOUIN ZONE.

Binary systems with intermediate phases at concentrations of 1.5, 1.6, and 1.75 electrons per atom

1.5			1.6	1.75
Body centered cubic	Hexagonal close-packed	β-Manganese	$\gamma\gamma$ Brass	Hexagonal close-packed
Cu-Be	Cu-Ag	Cu-Si	Cu-Zn	Cu-Zn
Cu-Zn	Cu-Ge	Ag-Hg	Cu-Cd	Cu-Cd
Cu-Al	Cu-Si	Ag-Al	Cu-Hg	Cu-Sn
Cu-Ga	Ag-Al	Au-Al	Cu-Al	Cu-Si
Cu-In	Ag-In	Co-Zn	Cu-Ga	Ag-Zn
Cu-Si	Ag-Sn		Cu-In	Ag-Cd
Cu-Sn	Ag-Sb		Cu-Si	Ag-Al
Ag-Mg	Ag-Cd		Cu-Sn	Au-Zn
Ag-Zn	Au-Sn		Ag-Zn	Au-Cd
Ag-Cd	Au-In		Ag-Cd	Au-Sn
Ag-Al			Ag-Hg	Au-Al
Ag-In			Ag-In	
Au-Mg			Au-Zn	
Au-Zn			Au-Cd	
Au-Cd			Au-In	
Fe-Al			Mn-Zn	
Co-Al			Fe-Zn	
Ni-Al			Co-Zn	
Ni-In			Ni-Be	
Pd-In			Ni-Zn	
			Ni-Cd	
			Rh-Zn	
			Pd-Zn	
			Pt-Be	
			Pt-Zn	
			Na-Pb	

However, in spite of the seeming success of the theory in confirming the electron compound concept in terms of the interaction of the electrons with the boundaries of the Brillouin zone, the Jones treatment must be considered a very limited theory. First of all, Jones assumed, when he adopted the free-electron picture, that the role of the ions in the alloy is a minor one. On the other hand, when a zinc atom is placed in a copper lattice, it is necessary that the vicinity of the zinc atom be highly polarized, according to the discussion of the preceding section. In a 50% alloy of Cu-Zn, the atomic volume of a zinc ion will have only 1.5 electrons to cancel the charge of a double ionized core. Thus, the lattice cell of the zinc atom has a net charge of $+{}^1/_2$. Hence, the valence electrons from zinc atoms cannot be freely contributed to the free-electron cloud of the crystal. There must be a considerable clumping of this charge around the various different ions of the alloy in such a way that the excess charge of the ions is screened. In addition, there is some evidence from the experiments on nuclear magnetic resonance in dilute alloys that the electron density in the conduction band does not change as the divalent ion is added, in direct contradiction to the simple free-electron picture.

In view of these considerations, it is surprising that such simple results as those obtained by Jones are correct. In later work J. Friedel has been able to show that, although the description of the electrons in the conduction band is far from the free-electron picture

of Jones, the so-called rigid band model is valid for dilute alloys and leads to results similar to those of Jones. The theory of electron compounds is not yet firmly established.

Order-Disorder in Alloys

In the preceding discussion on the formation of alloys, it was tacitly assumed that the atoms of the two constituents were randomly distributed on the lattice of the alloy. In very dilute primary solutions the random distribution holds because the free energy of the system in the completely random state is lower than that of a more symmetric state. To be more specific, in a dilute solution of metal B in metal A, if there is no gain of internal energy by ordering B (placing the atoms of B in a regular arrangement on the lattice sides of A), the free energy, $U - TS$, is a minimum when B is completely randomized.

On the other hand, if an appreciable percentage of the lattice sites are occupied by B and if there is a difference in the interatomic forces between A-B nearest-neighbor atom pairs and those between A-A or B-B pairs, the situation is different. The internal energy of the alloy is then given by Eq. (3), where

$$U = N_{AB}E_{AB} + N_{AA}E_{AA} + N_{BB}E_{BB} \qquad (3)$$

N_{AB} is the number of A-B nearest-neighbor pairs in the alloy, E_{AB} is the energy of interaction between an A-B pair, and so on. It is assumed that the alloy atoms interact appreciably only when they are nearest neighbors. If the interaction between A-B pairs is greater than that between B-B or A-A pairs, then the state of lowest energy is that in which the number of like pairs is minimized. This state is also a completely ordered state of the crystal.

For any given temperature the equilibrium state of the crystal is that for which the free energy is minimized, and a balance is struck between the contradictory tendencies of the U and TS terms of the free energy. At low temperatures the U term can be expected to predominate, and ordering will usually occur, while at higher temperatures the entropy term will predominate, and a transition to a disordered state is usual.

Long- and short-range order. The ordering of a crystal is described in terms of two different order parameters. One is called local or short-range order, and the other is called long-range order. For the initial discussion of the long- and short-range order parameters, a 50% A-B alloy will be considered. It will also be supposed that the lattice structure is bcc and that the unlike pair interaction is the stronger one. The advantage of the bcc lattice for the discussion is that, in this lattice, all the neighbors of any given atom can be made unlike atoms.

In the completely ordered state every atom of the crystal is surrounded by unlike nearest neighbors, and the crystal becomes disordered when some of the A atoms exchange with B atoms. The long-range order parameter Σ is a measure of the number of such interchanges and expresses the number of sites occupied by the wrong atom, as in Eq. (4). Here n_a

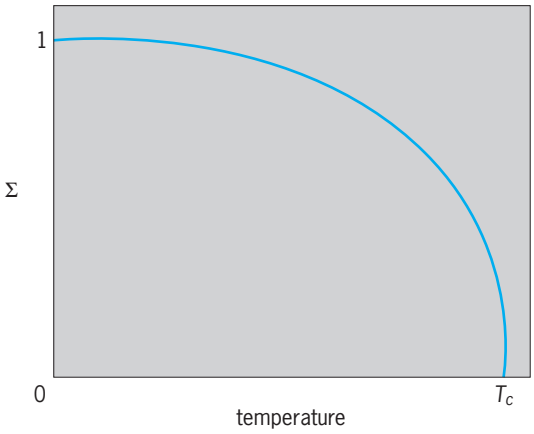


Fig. 5. Long-range order Σ of a crystal as a function of temperature. The order parameter has a rather precipitous drop at the transition temperature and is zero above that point.

$$\Sigma = \frac{2n_a}{N} - 1 \qquad (4)$$

is the number of A sites occupied by A atoms, and N is the total number of A lattice sites. The quantity Σ varies from -1 to $+1$, and the state of complete disorder corresponds to $\Sigma = 0$.

The short-range order expresses the fact that the neighbors surrounding any given A atom will have a tendency to be all B atoms. If this tendency is averaged over the lattice, the short-range order σ is defined in Eq. (5). Here σ is the local-order parameter, and q is the number of A-B pairs in the crystal divided by the total number of atom pairs in the crystal. The parameter σ has the range -1 to $+1$, with complete disorder at $\sigma = 0$.

$$\sigma = 2(q - {}^1/_2) \qquad (5)$$

Variation with temperature. A good qualitative notion of the behavior of the order of a crystal is obtained by observing the variation of the order parameter of a crystal as a function of temperature. Theory predicts the curves shown in **Figs. 5** and **6**. The long-range order decreases from a state of complete order at absolute zero temperature. At a critical transition

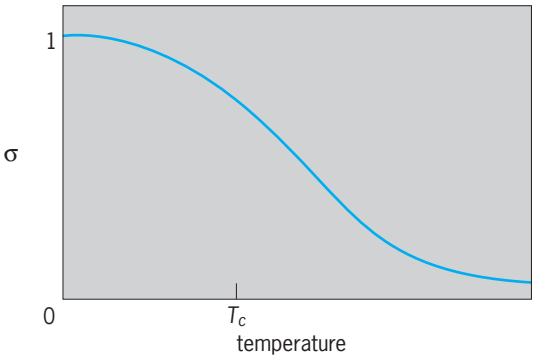


Fig. 6. Short-range order σ as a function of temperature. It varies more slowly than long-range order. Considerable order still exists above the transition temperature.

temperature which depends on the strength of the A-B bonds relative to the A-A bonds, the long-range order drops to zero. The local order also starts at absolute zero with a maximum value, but decreases more slowly than the long-range order does and remains finite for all temperatures.

The discussion up till now has been based on the model of a 50% alloy in a simple lattice. The theory for a material in which the composition is allowed to vary over the entire composition diagram is exceedingly complicated. The complication is due to the difficulty of specifying the types of order configurations which are possible, and computing their entropy. However, the general results are still comparable with the simple case. Long- and short-range order parameters can still be defined, with a transition temperature at the point where the long-range order disappears. The short-range order again persists to a considerable degree even above the transition temperature.

Transition temperature. The existence of the transition temperature amounts to a change of phase, because there is an attendant singularity in the specific-heat curve of the material. The phase change may be of the first or second order, depending on the type of singularity present in the long-range order at the transition temperature. The curve of Fig. 5 corresponds to a second-order transition; however, if there is a discontinuous drop of the order to zero at the transition point, the transition is first order. A general rule seems to be that for lattices in which the completely ordered state has only neighbors of A-B bonds, the transition is second order. The bcc lattice is such a case, as has already been explained. However, in close-packed lattices such as the fcc lattice, it is not possible for all the bonds of a 50% alloy to be of the sort A-B, and in such cases, it is found experimentally that the transition is first order. *See* PHASE TRANSITIONS.

Detection of order. The presence of order in a crystal and the transition from the ordered to the disordered state are detected by a variety of techniques. The resistance of the alloy at low temperatures varies with the order of the crystal because the electron waves of the crystal are sensitive to irregularities in the crystal, so that as disorder increases, the resistivity increases also. X-ray and neutron scattering of the lattice are also functions of the degree of order for the same reason. The specific heat of the crystal varies with the order, since as the crystal loses its order, energy must be supplied to the lattice to form the A-A and B-B bonds which have higher energy. At the transition temperature the specific heat rises to a sharp peak, which is easily detected. *See* NEUTRON DIFFRACTION; SPECIFIC HEAT OF SOLIDS; X-RAY DIFFRACTION.

Superlattices. The x-ray and neutron scattering of a completely ordered crystal have an interesting peculiarity. In the Bragg scattering processes new lines appear, called superlattice lines. The name is derived from the fact that the lines correspond to the appearance of a secondary crystalline structure in the lattice. Their explanation is clear in the light of the discussion of the ordered lattice. For a 50% A-B lattice in the bcc form, the A atoms form the corners of the cube, while the B atoms are at the center positions. The crystal is then composed of two interpenetrating simple cubic crystals, and is called a superlattice. The x-ray lines for two simple cubic structures appear instead of those for a bcc crystal. *See* INTERMETALLIC COMPOUNDS.

Robb M. Thomson

Bibliography. W. Hume-Rothery and R. E. Smallman, *The Structure of Metals and Alloys*, 1988; A. G. Khachaturyan, *Theory of Structural Transformations in Solids*, 1983; W. G. Moffatt, *Handbook of Binary Phase Diagrams*, 5 vols., 1981; W. F. Smith, *Structure and Properties of Engineering Alloys*, 2d ed., 1992.

Allspice

The dried, unripe fruits of a small, tropical, evergreen tree, *Pimenta officinalis*, of the myrtle family (Myrtaceae). This species (see **illus.**) is a native of the

Allspice. (*a*) Branch with fruit. (*b*) Flowers.

West Indies and parts of Central and South America. The spice, alone or in mixtures, is much used in sausages, pickles, sauces, and soups. The extracted oil is used for flavoring and in perfumery. Allspice is so named because its flavor resembles that of a combination of cloves, cinnamon, and nutmeg. *See* MYRTALES; SPICE AND FLAVORING.

Perry D. Strausbaugh; Earl L. Core

Almanac

A book that contains astronomical or meteorological data arranged according to days, weeks, and months of a given year and may also include diverse information of a nonastronomical character. This article is restricted to astronomical and navigational almanacs.

Development. The earliest known almanac material was computed by the Egyptians in A.D. 467, and astronomical ephemerides appeared irregularly thereafter. The first regular almanac appears to have been produced in Germany from 1475 to 1531. The first almanac to be published regularly by a national government was the *Connaissance des Temps* (1679) by France. In 1767 the British began publishing the *Nautical Almanac* to improve the art of navigation. The *Berliner Astronomisches Jahrbuch* was first published in Germany for 1776, and the *Efemerides Astronomicas* was published in Spain in 1791. In the United States *The American Ephemeris and Nautical Almanac* has been published since 1855. Beginning with the issue for 1981, the two series of publications titled *The Astronomical Ephemeris,* which had previously replaced *The Nautical Almanac and Astronomical Ephemeris,* and *The American Ephemeris and Nautical Almanac* were continued with the title of *The Astronomical Almanac.*

Over the years cooperation has developed between the organizations responsible for the preparation and publication of the almanacs. Under the auspices of the International Astronomical Union (IAU), agreements are reached concerning the bases and constants to be used in the publications and the exchange of data in different forms.

Astronomical Almanac. *The Astronomical Almanac* contains ephemerides, which are tabulations, at regular time intervals, of the orbital positions and rotational orientation of the Sun, Moon, planets, satellites, and some minor planets. It also contains mean places of stars, quasars, pulsars, galaxies, and radio sources, and the times for astronomical phenomena such as eclipses, conjunctions, occultations, sunrise, sunset, twilight, moonrise, and moonset. This volume contains the fundamental astronomical data needed by astronomers, geodesists, navigators, surveyors, and space scientists. The theory and methods on which *The Astronomical Almanac* is based are provided in the *Explanatory Supplement to the Astronomical Almanac. See* ASTRONOMICAL COORDINATE SYSTEMS; EPHEMERIS.

Navigational almanacs. While *The Astronomical Almanac* is basically designed for the determination of positions of astronomical objects as observed from the Earth, *The Nautical Almanac* and *The Air Almanac* are designed to determine the navigator's position from the tabulated position of the celestial object. The ground point of the celestial object is used with the North and South poles and the observer's assumed position to establish a spherical triangle known as the navigational triangle. By means of spherical trigonometry or navigational tables, the navigator determines a computed altitude and the direction of the ground point (azimuth) of the celestial object with respect to his or her position. The computed altitude and azimuth enable the navigator to plot a line of position that goes through or near this assumed position. The computed altitude is compared to the observed altitude of the celestial object. From the combination of two or more observations the navigator is able to determine position. *See* CELESTIAL NAVIGATION.

The Nautical Almanac contains hourly values of the Greenwich hour angle and declination of the Sun, Moon, Venus, Mars, Jupiter, and Saturn and the sidereal hour angle and declination of 57 stars for every third day. Monthly apparent positions are tabulated for an additional 173 navigational stars. The positions are tabulated to an angular accuracy of 0.1 minute of arc, which is equivalent to 0.1 nautical mile (0.2 km). Since tabular quantities must be combined to derive the navigational fix, this tabular accuracy is sufficient to produce a computed position with an error no greater than 0.3 to 0.4 nmi (0.6 to 0.7 km).*The Nautical Almanac* also contains the times of sunrise, sunset, moonrise, moonset, and twilight for various latitudes. A diary of astronomical phenomena, predictions of lunar and solar eclipses, visibility of planets, and other information of interest to the navigator are also included.

The Air Almanac is arranged with two pages of data back to back on a single sheet for each day; the positions of the Sun, first point of Aries, three planets, and the Moon are given at 10-min intervals. As necessary, information is adjusted so that the tabulated data at any given time can be used during the interval to the next entry, without interpolation, to an accuracy sufficient for practical air navigation. The times of sunrise, sunset, twilight, moonrise, and moonset are also given daily. Also provided are star recognition charts; a sky diagram; rising, setting, and depression graphs; standard times; and correction tables necessary for air navigation. While designed for air navigators, *The Air Almanac* is used by mariners who accept the reduced accuracy in exchange for its greater convenience compared with *The Nautical Almanac.* A number of countries publish in their languages air and nautical almanacs that are based on, or similar to, the English language versions.

Other almanacs. *The Astronomical Phenomena,* published annually, contains the times for phenomena such as eclipses, conjunctions, occultations, sunrise, sunset, moonrise, moonset, and moon phases, seasons, and visibility of planets. Also, the dates of civil and religious holidays are included.

For surveyors, the Royal Greenwich Observatory publishes *The Star Almanac,* designed to permit the surveyor to determine a geographical position from celestial observations.

Ephemerides of Minor Planets is prepared annually by the Institute of Applied Astronomy, St. Petersburg, and published by the Academy of Sciences of Russia. This volume contains the elements, opposition dates, and opposition ephemerides of all numbered minor planets. *See* ASTEROID.

Computer-based almanacs. The *Almanac for Computers* was published annually by the U.S. Naval Observatory from 1977 to 1991 to provide the coefficients for power series and Chebyshev polynomials for computation of any specific kind of astronomical data. The Royal Greenwich Observatory began publishing in 1981 *Compact Data for Navigation and Astronomy.* The Naval Observatory introduced the *Floppy Almanac* as an annual source of astronomical data by means of widely used personal-computer software. The latest source of high-precision astronomical data is the *Multi-Year Interactive Computer Almanac (MICA),* which provides data on compact disks. Thus, in addition to being able to compute the almanac data as published, the user is able to compute data for a particular location and time. The Naval Observatory also prepares a *Satellite Almanac,* which provides accurate positions for the planetary satellites.

The Naval Observatory provides software that can be used for computation of apparent places and transformation of coordinates. Astronomical data in the form of star catalogs are available from astronomical data centers in Strasbourg, France, and Goddard Space Flight Center in Maryland.

Publication and distribution. *The Astronomical Almanac, The Nautical Almanac, The Air Almanac*, and *Astronomical Phenomena* are cooperative publications of the U.S. Naval Observatory and Rutherford Appleton Laboratory, and are available from the U.S. Government Printing Office and the Stationery Office. These publications are used in many countries, and similar references, generally based on the same basic data, are published in various languages by the Spanish, Chinese, Russian, Japanese, German, French, Indian, Argentinian, Brazilian, Danish, Greek, Indonesian, Italian, Korean, Mexican, Norwegian, Peruvian, Philippine, and Swedish governments. Examples of publications comparable to *The Astronomical Almanac* are the *Astronomical Almanac* by Russia, the *Efemerides Astronomicas* by Spain, the *Chinese Astronomical Ephemeris,* the *Indian Ephemeris,* and the *Japanese Ephemeris.*

P. K. Seidelmann

Almond

A small deciduous tree, *Prunus amygdalus* (also known as *P. dulcis* or *Amygdalus communis*), closely related to the peach and other stone fruits and grown widely for its edible seeds. The fruit, classified botanically as a drupe (see **illus.**), is analogous to that of the peach except that the outer fleshy (mesocarp) layer does not enlarge in size during the latter part of the fruit growth-curve. Instead it splits (dehisces) as it nears maturity and separates from the shell (endocarp) and then dries. The nut with the kernel, or seed, within is then released. *See* PEACH; ROSALES.

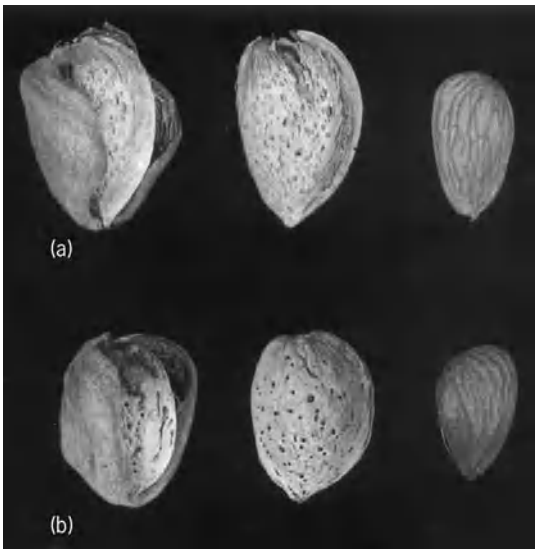
Origin. Numerous wild almond species are found in the mountainous and desert regions extending from southwestern Europe to Afghanistan, Turkistan, and western China, sometimes occurring as extensive forests and thickets. The cultivated almond species apparently originated in central Asia (Iran) from hybridization among native species followed by local seedling selection by native peoples. Its cultivation spread with civilization along the shores of the Mediterranean Sea into North Africa and to Italy, Spain, southern France, and Portugal. From there seeds and scions were used to introduce the crop to other parts of the world, including the United States, Australia, South Africa, and South America.

Production. In the United States, commercial production is limited to California; one-half or more of this production is exported. Spain is the second leading producer, but the amount produced is about one-half that of the United States. Italy has historically been a leading producer, primarily from the Bari and Sicily areas, but production has declined sharply.

Cultivation. Because of the early spring flowering habit, commercial culture of the almond is limited to areas with mild winters and few late spring frosts. The trees thrive under moderate to high summer temperatures.

Trees can withstand considerable drought due to natural adaptation to arid conditions. Thus traditional older culture in European and Asian areas has been largely restricted to hillside and otherwise marginal areas without irrigation. However, the almond responds so well to irrigation, improved nutrition, and intensification of culture (two- to threefold increases in yield) that modern culture systems in California, and now in other parts of the world, utilize the best soil areas, supplemental irrigation, fertilization, disease and insect control, herbicide application, and all other aspects of advanced agriculture.

Original establishment of almonds was by seeds, but modern cultivation includes selection of cultivars budded or grafted to rootstocks. Many cultivars exist, and in general these have arisen as local or chance seedlings from a particular production area, such as California, Spain, Sicily, Bari, and Australia.



Nut characteristics of the two major almond cultivars of California: (*a*) nonpareil, (*b*) mission. Left: Mature nut with dehisced hull; middle: nut with shell; right: kernels.

Most cultivars are self-incompatible, and orchards require provision for combination of cultivars to provide cross pollination. Pollinizer insects, primarily honeybees, are required to accomplish cross pollination. Some self-fertile cultivars have been discovered or produced by plant breeding. *See* BREEDING (PLANT).

Harvesting. In California, essentially all almond orchards are mechanical-harvested. Nuts are knocked to the ground when they (particularly in the center of the tree) are dry. They are then swept into windrows and taken by machines after careful soil preparation. Hulls are removed from the nuts by machines in a central huller and delivered to the processor and handler. There nuts are fumigated for worm control and stored. Most nuts are shelled, except for a few special varieties. Kernels are sorted electronically and graded into various shapes and sizes. *See* AGRICULTURAL MACHINERY.

In some parts of the world where culture is less intensive, almonds are knocked to the ground by mallets or poles onto canvases spread under the trees. Nuts are then collected by hand into boxes or bags, or may be lifted onto almond "boats" or "sleds."

Use. Almonds are used in a variety of products. Some are roasted whole and salted to be used as snacks. Others are blanched (the skin is removed) by steam and subjected to slicing, dicing, or halving. These may be roasted, and go into products such as candy bars, bakery products, ice cream, and almond paste, among many other uses.

Almonds are of two general types: the bitter type is a source of prussic acid and flavoring extracts, and the sweet type has various food uses as described above. The almond kernels contain approximately 50% fat or oil, 20% protein, 20% carbohydrate, and a variety of minerals and vitamins. *See* NUT CROP CULTURE.

D. E. Kester

Alpaca

A member of the camel family, Camelidae, which has been domesticated for more than 2000 years. The Camelidae belong to the mammalian order Artiodactyla, the even-toed ungulates. The alpaca (*Lama pacos*), more restricted in its range than other species of this family, is found at elevations above 12,000 ft (3600 m) along the shores of Lake Titicaca on the boundaries of Peru and Bolivia (see **illustration**).

As in other members of this family, the alpaca's neck and head are elongate and the upper lip has a deep cleft. The long, slender legs terminate in two toes; the feet are digitigrade, that is, the animals walk on the toes and not on the entire foot or the tip of the digits. Although the majority of the artiodactyl animals characteristically possess horns, the alpaca and other members of this family do not. The alpaca has 36 teeth, dental formula I 1/3 C 1/1 Pm 3/3 M 3/3.

The long, fine repellent hair, or wool, ranges from black to white and is highly prized for manufacturing cloth, particularly the white wool. Like many breeds of domesticated animals, the alpaca has been bred to produce pure strains for the wool. Although the alpaca is raised chiefly for its wool, its flesh is palatable.

Alpacas (*Lama pacos*). (*Courtesy of Brent Huffman/Ultimate Ungulate Images*)

The known extinct and extant species of the camel family provide an interesting example of discontinuous distribution. The family originated in North America and migrated centuries ago, so that the living species are now widely separated geographically. Among the living species, affinities and relationships are further complicated and obscured by their long history of domestication. *See* ARTIODACTYLA; BREEDING (ANIMAL); CAMEL; DENTITION; LLAMA; MAMMALIA; NATURAL FIBER.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Alpha Centauri

The third brightest star in the sky, apparent magnitude −0.3, and the brightest in the southern constellation Centaurus. It is the closest star to the Sun at a distance of 1.35 parsecs 2.59×10^{13} mi (or 4.16×10^{13} km), and its light takes more than 4 years to reach the Earth. It has an unusually large proper motion across the sky of 3.7 seconds of arc per year. *See* CENTAURUS.

Alpha Centauri is in reality a triple system, the two main components orbiting each other with period of nearly 80 years. They were discovered as a telescopic visual pair in 1689, and have now been followed for more than three complete revolutions. The orbit is eccentric, and their mean separation is approximately 23 astronomical units. α Cen A and B, as they are also known, are main-sequence stars of spectral types G2 and K1, respectively, and the brightest of the two is very similar to the Sun in mass (A = 1.16

and B = 0.97 solar masses) luminosity, and effective temperature. Chemical analyses indicate that the relative abundance of elements heavier than helium in α Cen A and B is almost twice as large as in the Sun. *See* ASTRONOMICAL UNIT; BINARY STAR; SPECTRAL TYPE.

The third component of this system, known as Proxima Centauri, is a faint reddish star of 11th magnitude, approximately 2.2° away in the sky. It was discovered in 1915, and has nearly the same proper motion and distance as the other two stars. Proxima is actually slightly closer to the Sun, which makes it the Earth's nearest neighbor in space. The linear separation from α Cen A and B is estimated to be about 13,000 astronomical units. Although the probability of such a close association in space and in projected motion occurring by chance among field stars is very small, measurements of the velocity of Proxima relative to the close binary cannot yet determine whether it is gravitationally bound. If its velocity exceeds the escape velocity of the system at Proxima's distance from the center of mass, it cannot be in orbit and is merely traveling together with the other stars in space. However, if it is indeed dynamically bound, the orbital period is probably on the order of 10^6 years. *See* STAR. David W. Latham

Alpha fetoprotein

A glycoprotein that is normally present in significant amounts only in the serum of the fetus. It is produced in the yolk sac, the liver, and other tissues of the gastrointestinal tract. Its role is unknown, but alpha fetoprotein may function as a carrier (or modulator of the concentration) of a small ligand, as an immunosuppressive, as a modulator of intracellular transport of unsaturated fatty acids, as a factor in estrogen transport, or as a means of binding retinoic acid.

Levels. Peak concentration of alpha fetoprotein in human fetal serum occurs at 13 weeks of gestation, and at that time it also reaches maximum levels in the amniotic fluid. Concentration of alpha fetoprotein in maternal serum peaks at approximately week 30–32 of gestation, reflecting the effect of both the production rate in the fetus, which is already declining at that age, and the mass of the growing fetal liver.

In pregnant women, there are large differences in the concentration of maternal serum alpha fetoprotein among individuals. The differences are attributable to the inbred differences in the production of alpha fetoprotein among fetuses, the sex of the fetus (males produce more alpha fetoprotein), the permeability of the intervening tissues, the weight of the mother (which reflects her blood volume), and maternal diabetes mellitus (which lowers the concentration). Ethnic differences in maternal serum alpha fetoprotein concentrations are most distinguishable between whites and blacks, the latter having a 15% higher concentration. There is a difference of only a few percent, if that, between whites and Asian or Indian populations.

Abnormal levels. Substantially increased levels of alpha fetoprotein were first observed in association with certain tumors, especially of the liver. Subsequently, increased concentrations were noted in the amniotic fluid and maternal serum of pregnant women carrying a fetus affected by an open-neural-tube defect. The two major types of neural tube defects are anencephaly, which results in a failure of the forebrain to form, and spina bifida, which is a failure of a variable part of the spinal cord to close. Both defects result in abnormalities in the overlying skin that allow excessive amounts of serum to leak into the amniotic fluid, and then into maternal blood. Thus, screening of maternal serum is now routine.

While the majority of pregnant women having a high concentration of serum alpha fetoprotein experience normal pregnancies, others manifest various problems, including multiple fetuses and placental abnormalities, or defects, such as gastroschisis, strictures of the fetal gastrointestinal or genitourinary tract, the Finnish type of nephrosis, occasional fetal tumors, and skin defects; these may result in fetal death. Further, about one out of three pregnancies in which alpha fetoprotein is increased has an adverse outcome, including fetal growth retardation, low birth weight, and increased perinatal mortality. In pregnancies where the fetus is affected by trisomy 21 (Down syndrome), there is lower concentration of alpha fetoprotein than usual. *See* DOWN SYNDROME.

Other genetic defects, including tyrosinemia, are also associated with increased alpha fetoprotein concentration. More frequently, acquired illness is responsible for the elevation. The most common cause of slight-to-moderate increases in serum alpha fetoprotein concentration is cirrhosis of the liver; the resumed production of alpha fetoprotein occurs in regenerating liver tissue. Higher concentrations are also associated with tumors, especially of the liver, gonads, and pancreas. A. Baumgarten

Bibliography. G. J. Mizejewski and H. I. Jacobson (eds.), *Biological Activities of Alpha-Fetoproteins*, vol. 1, 1989.

Alpha particles

Helium nuclei, which are abundant throughout the universe both as radioactive-decay products and as key participants in stellar fusion reactions. Alpha particles can also be generated in the laboratory, either by ionizing helium or from nuclear reactions. They expend their energy rapidly as they pass through matter, primarily by taking part in ionization processes, and consequently have short penetration ranges. Numerous technological applications of alpha particles can be found in fields as diverse as medicine, space exploration, and geology. Alpha particles are also major factors in the health concerns associated with nuclear waste and other radiation hazards.

The helium nucleus, or alpha particle (α), with mass 4.00150 atomic mass units (u) and charge +2,

is a strongly bound cluster of two protons (p) and two neutrons (n). Its stability is evident from mass-energy conservation in the hypothetical fusion reaction (1). The product mass (=4.00150 u) is less

$$2p + 2n \rightarrow \alpha \tag{1}$$

than the reactant mass (=2 × 1.00728 u + 2 × 1.00866 u) by 0.03038 u. By using Einstein's relation $E = mc^2$ (where c is the speed of light), this decrease in mass m (the alpha-particle binding energy) is equivalent to 28.3 MeV of energy E. The enormous magnitude of this energy is reflected in the fact that the fusion transformation of hydrogen into helium is the main process responsible for the Sun's energy. *See* CONSERVATION OF ENERGY; ENERGY; HELIUM; NUCLEAR BINDING ENERGY; NUCLEAR FUSION; PROTON-PROTON CHAIN; STELLAR EVOLUTION.

Discovery. Early absorption and deflection experiments with so-called rays from radioactive minerals revealed three distinct components, of which the least penetrating (named alpha rays by E. Rutherford in 1899) was shown to consist of positively charged particles, with a charge-to-mass ratio about half that of hydrogen. These particles were shown to be helium nuclei when helium gas was identified spectroscopically in a thin-walled glass vessel within which the rays had been trapped. *See* BETA PARTICLES; GAMMA RAYS.

Noting that a few alpha particles were strongly backscattered by thin foils while most passed directly through, Rutherford concluded that the occasional deflections were caused by coulombic repulsion between the alpha particles and compact positive entities which occupied only a small fraction of the total volume. These observations led Rutherford to propose his so-called planetary model of the atom in 1911. With subsequent modifications and refinements, Rutherford's visual concept of atomic structure has evolved to become the model of the atom accepted today. *See* ATOMIC STRUCTURE AND SPECTRA.

Alpha radioactivity. Coulombic repulsion between the protons within a nucleus leads to increasingly larger ratios of neutron number N to proton number Z for stable nuclei, as the mass numbers A ($= Z + N$) increase. Neutron-deficient nuclei can improve their N/Z ratios by means of alpha decay. The decay occurs because the parent nucleus has a total mass greater than the sum of the masses of the daughter nucleus and the alpha particle. The energy converted from mass energy to kinetic energy is called the Q value and is given by the Einstein mass-energy equivalency relation, Eq. (2), where c is the velocity of light in vacuum and the M is atomic mass. If the masses are given in atomic mass units, then the c^2 can be replaced with the conversion factor 931.501 MeV/u. The kinetic energy released is shared between the daughter nucleus and the alpha particle in accordance with the conservation of momentum, thus giving Eq. (3) for

$$Q = [M(\text{Parent}) - M(\text{Daughter}) - M(^4\text{He})]c^2 \tag{2}$$

$$\text{KE}_\alpha = Q\left[1 - \frac{M(^4\text{He})}{M(\text{Daughter})}\right] \tag{3}$$

the kinetic energy of the alpha particle, KE_α, in terms of the Q value and the masses. Thus, each radioactive alpha-emitting nuclide emits the alpha with a characteristic kinetic energy, which is one fingerprint in identification of the emitter. *See* NUCLEAR REACTION.

For example, in reaction (4), where the notation

$$^{235}_{92}\text{U}_{143} \rightarrow {}^{231}_{90}\text{Th}_{141} + {}^{4}_{2}\text{He}_2 \tag{4}$$

$^A{}_Z\text{U}_N$, and so forth, is used, an alpha particle (that is, a $^4{}_2\text{He}_2$ nucleus) is emitted from uranium-235, leaving thorium-231. For this decay, Q = (235.0439231 − 231.0362971 − 4.0026032)(931.501) MeV = 4.679 MeV, and thus the energy of the alpha is KE_α = 4.679 MeV (1 − 4/231) = 4.598 MeV.

An alpha particle interacts with the nucleus through the strong nuclear force, which has a very short range ($\sim 10^{-15}$ m) and is attractive, and the repulsive electrical (Coulomb) force. In alpha decay the Coulomb force is primarily responsible for accelerating the alpha to its final energy, but before that can happen the alpha must escape from the hold of the nuclear force. Relative to a potential energy of 0 at large distances, the alpha particle inside a heavy nucleus has a large positive component (∼50 MeV) due to the Coulomb force, but a much larger negative component (∼−100 MeV) due to the nuclear force. Just outside the nuclear surface, the part due to the nuclear force drops to zero, while that due to the Coulomb force is still large (∼25 MeV). The total energy (potential and kinetic) of the typical decay alpha particle (4–10 MeV) is smaller than the potential energy at distances slightly larger than the nuclear radius. Thus, the alpha particle must penetrate this potential energy barrier (Coulomb barrier) in order to escape from the nucleus. Such barrier penetration is forbidden in classical physics but is permitted in quantum-mechanical theory. In quantum mechanics there is a straightforward procedure to calculate the probability that the alpha can penetrate the Coulomb barrier; the penetration probability is sensitively dependent on the energy of the alpha and the height of the barrier. Other factors that influence the half-life are the orbital angular momentum of the emitted alpha particle, which in quantum theory adds to the height of the Coulomb barrier, and the similarity of the alpha-emitting nucleus and the resulting daughter nucleus. *See* NONRELATIVISTIC QUANTUM THEORY; RADIOACTIVITY.

There are three major natural series, or chains, through which isotopes of heavy elements decay by successions of alpha decays. Within these series, and with all reaction-produced alpha emitters as well, each isotope decays with a characteristic half-life and emits alpha particles of particular energies and intensities (see **table**).

The presence of these radioactive nuclides in nature depends upon either a continuous production mechanism, for example the interaction of cosmic

Energies and intensities of alpha-particle standard sources

Source	Half-life	E_α*	I_α†
^{147}Sm	1.06×10^{11} years	2.234 ± 0.003	
^{232}Th	1.41×10^{10} years	4.013 ± 0.003	77
		3.954 ± 0.008	23
^{238}U	4.468×10^{9} years	4.197 ± 0.005	77
		4.150 ± 0.005	23
^{235}U	7.04×10^{8} years	4.400 ± 0.002	55
^{230}Th	7.54×10^{4} years	4.6877 ± 0.0015	76.3
		4.6212 ± 0.0015	23.4
^{239}Pu	2.411×10^{4} years	5.1566 ± 0.0004	73.2
		5.1438 ± 0.0008	15.1
^{210}Po	138.38 days	5.30438 ± 0.00007	
^{241}Am	432.2 years	5.48560 ± 0.00012	85.2
		5.44290 ± 0.00013	12.8
^{238}Pu	87.7 years	5.49907 ± 0.00020	71.6
		5.4563 ± 0.0002	28.3
^{242}Cm	162.8 days	6.11277 ± 0.00008	74.0
		6.06942 ± 0.00012	25.0
^{253}Es	20.4 days	6.63257 ± 0.00005	89.8

*Alpha-particle energy in MeV.
†Alpha-particle intensity in number of alpha particles per 100 disintegrations of source.

rays with the atmosphere, or extremely long half-lives of heavy radioactive nuclides produced in past cataclysmic astrophysical events, which accounts for uranium and thorium ores in the Earth. The relative abundances of uranium-238, uranium-235, and their stable final decay products in ores of heavy elements can be used to calculate the age of the ore, and presumably the age of the Earth. *See* GEOCHRONOMETRY.

In addition to the study of alpha-particle emitters that appear in nature, alpha decay has provided a useful tool to study artificial nuclei, which do not exist in nature due to their short half-lives. Alpha decay is a very important decay mode for nuclei far from stability with a ratio of protons to neutrons that is too large to be stable, especially for nuclei with atomic mass greater than 150 u. Because of the ease of detecting and interpreting decay alpha particles, their observation has aided tremendously in studying these nuclei far from stability, extending the study of nuclei to the very edge of nuclear existence. The measured energy of the alpha particle can be used in Eqs. (2) and (3) to determine the mass difference of the parent and daughter nuclei, and if one of the masses is known by previous experiments, the absolute value for the other mass can be determined. The half-life, the measured time for the alpha-particle rate to be reduced by a factor of 2, is used to calculate a decay rate, which is a product of the barrier penetration probability and nuclear structure factors. By dividing out the barrier penetration probability, a number (the reduced width) is obtained which is relatively large if the parent and daughter nuclei have the same spins and parities and similar nuclear wave functions, and is smaller for decays requiring a change in spin or nuclear wave functions. Systematic measurement of reduced widths over a range of nuclei enables scientists to trace changes in nuclear structure as a function of proton and neutron number. Nuclear structure information for more than 400 nuclides has been obtained in this way (**Fig. 1**). In addition, fine structure peaks

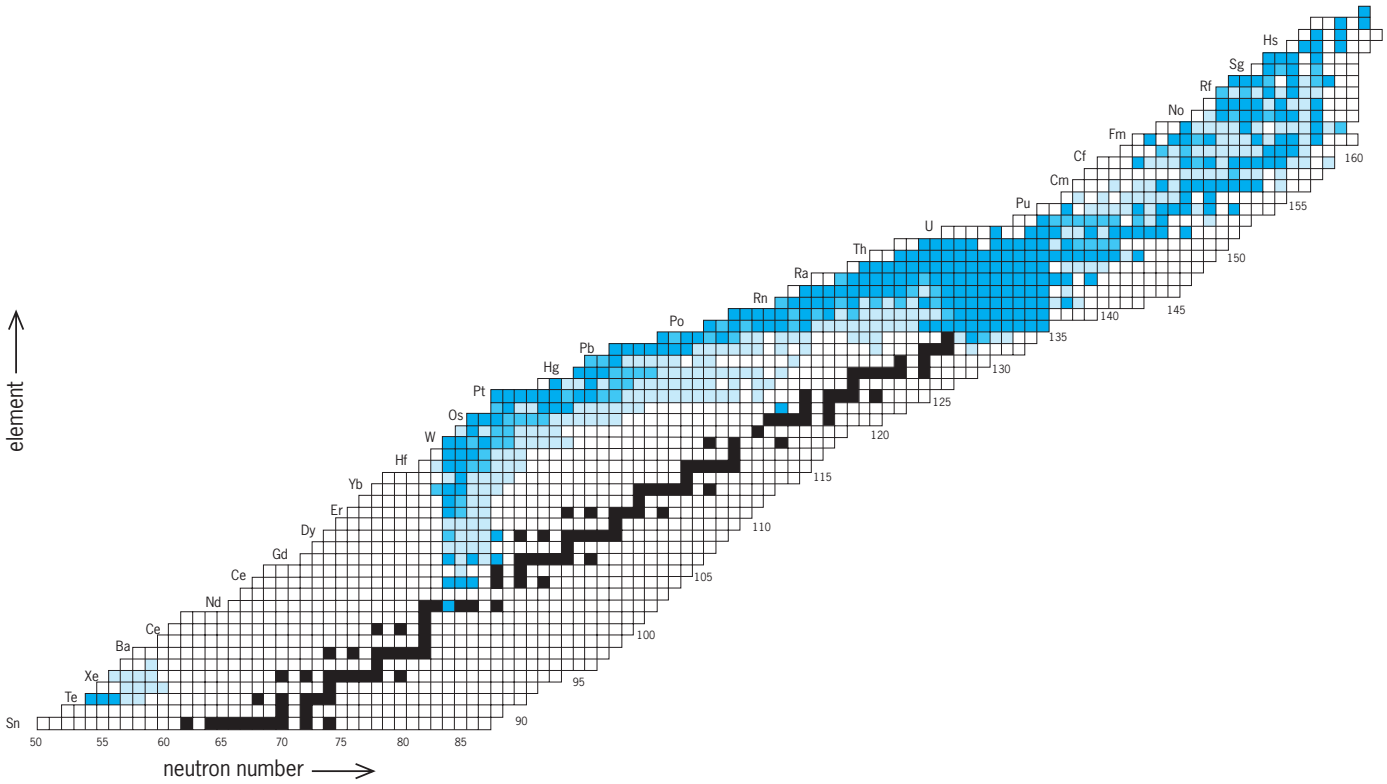


Fig. 1. Locations of known alpha emitters on the chart of nuclides are shown in color. Boxes are included for all nuclides with proton number $Z > 49$ which have been observed. The fraction of the time that a particular nuclide decays by alpha-particle emission is indicated by the shade of the colored square. Black boxes indicate the stable nuclides.

appear in the alpha-particle spectra for many of these nuclides; each such fine structure peak gives similar information about an excited state in the daughter nucleus.

Interactions with matter. By virtue of their kinetic energy, double positive charge, and large mass, alpha particles follow fairly straight paths in matter, interacting strongly with atomic electrons as they slow down and stop. These electrons may be excited to higher energy states in their host atoms, or they may be ejected, forming ion pairs in which the initial host atom becomes positively charged and the electron leaves. The more energetic ejected electrons, known as delta electrons, cause considerable secondary ionization, which accounts for 60–80% of the total ionization. A cascade of processes occurs along the alpha particle's track, leading to tens of thousands of disruptive events per alpha particle. *See* DELTA ELECTRONS; RADIATION DAMAGE TO MATERIALS.

The amount of energy expended by an alpha particle to form a single ion pair in passing through a medium is nearly independent of the alpha particle's energy, but it depends strongly on the absorbing medium. While it takes about 35 eV in air and 43 eV in helium to form an ion pair, an energy of only 2.9 eV is required in germanium and 3.6 eV in silicon. The energies expended in gases are roughly correlated to their ionization potentials. For germanium, silicon, and other semiconductors, the lower ion pair energy is, effectively, the amount required to raise an electron to the conduction band. *See* IONIZATION POTENTIAL; SEMICONDUCTOR.

The distance (or range) that an alpha particle travels before it stops depends both on the energy of the particle and on the absorbing medium. A 5.15-MeV alpha particle from plutonium-239 creates approximately 150,000 ion pairs, and travels about 1.5 in. (3.7 cm) in air before stopping. This same 5.15-MeV alpha particle has a range of about 0.0009 in. (0.0023 cm) in aluminum, 0.0014 in. (0.0035 cm) in water, and 0.0010 in. (0.0025 cm) in bone.

The passage of alpha particles through silicon is a particularly important example (**Fig. 2**). The semiconductor industry now produces chips so small that alpha particles from contaminants in the packaging materials can disrupt the memory-array areas of the chips, a serious problem which has been researched in considerable detail. About 1985, successful barriers were designed that could reduce the problem significantly, but there is concern that alpha particles might be a limiting factor in the increasing miniaturization of chips. The daily number of alpha particles to which these chips are exposed is typically less than 15/in.^{2} (2.3/cm^2). *See* INTEGRATED CIRCUITS; RADIATION HARDENING.

As alpha particles expend energy and slow down, the frequency of their interactions per unit distance increases. The rate at which an alpha particle loses energy is called its stopping power or specific ionization. A Bragg curve shows the stopping power of an alpha particle as a function of the distance to the end of its path (**Fig. 3**). The energy expended per

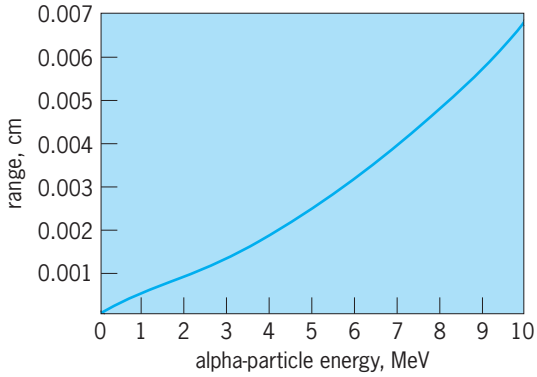


Fig. 2. Range-energy relationship for alpha particles in silicon. 1 cm = 0.4 in.

interval is seen to increase to the large, pronounced Bragg peak near the end of the path, and then to fall off to a small tail at the very end, due to range straggling.

An approximate relation for the range of alpha particles in air is given by Eq. (5). Here R is the

$$R = 0.309 \times E^{3/2} \tag{5}$$

range in centimeters (1 cm = 0.4 in.) and E is the alpha particle's initial energy in MeV. For an absorbing medium other than air, only the coefficient 0.309 would be different. According to another relationship, known as the Bragg-Kleeman rule, the relative stopping power per atom is approximately proportional to the square root of the mass number A.

In biological systems, the ionization and excitation produced by alpha particles can damage or kill cells. By rupturing chemical bonds and forming highly reactive free radicals, alpha particles can be far more destructive than other forms of radiation which interact less strongly with matter. *See* CHARGED PARTICLE BEAMS; RADIATION BIOLOGY; RADIATION CHEMISTRY; RADIATION INJURY (BIOLOGY).

Detectors. Alpha particles are detected by means of the ionization and excitation that they produce

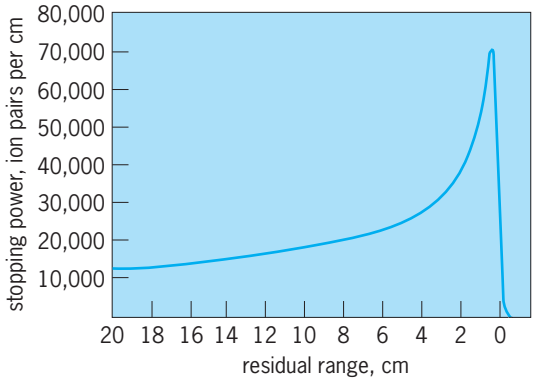


Fig. 3. Bragg curve for single alpha particle in air. The stopping power (specific ionization) in ion pairs per centimeter of air is plotted against residual range (distance from end of alpha particle's path) in centimeters. 1 cm = 0.4 in.

in matter. The first detectors were photographic plates, which turned dark when alpha particles sensitized and reduced grains of silver halide in the emulsions. In magnetic spectrometers, calibrated photographic emulsions have been used to determine alpha-particle energies and intensities.

In ionization chambers, the relationship between the number of ion pairs formed and the amount of charge collected is used in the detection process. In cloud chambers, tracks of droplets appear whenever alpha particles leave ion pairs to act as condensation centers. Finally, in present-day semiconductor detectors, the excitation of electrons from a valence band to the conduction band is detected electrically, and is then converted into precise energy and intensity values. *See* IONIZATION CHAMBER; JUNCTION DETECTOR; PARTICLE DETECTOR.

Applications. In the promising medical field of charged-particle radiotherapy, alpha particles are useful in the treatment of inaccessible tumors and vascular disorders. It is clear from the Bragg curve (Fig. 3) that the ionizing power of alpha particles is concentrated near the ends of their paths. Thus they can deliver destructive energy to a tumor while doing little damage to nearby healthy tissue. With proper acceleration, positioning, and dosage, the energy can be delivered so precisely that alpha-particle radiotherapy is uniquely suited for treating highly localized tumors near sensitive normal tissue (for example, the spinal cord). *See* RADIOLOGY.

The element-specific energies of backscattered (Rutherford-scattered) alpha particles are used in remote probes to analyze the mineral composition of geological formations. In particular, alpha particles scattered by light elements transfer more energy than those scattered by heavy elements. In another alpha-particle device, the energy from ^{238}Pu alpha decay is reliably harnessed in batteries based on the Brayton cycle, and used to power scientific equipment left on the Moon. Large power systems of this type are contemplated for use in space stations. These examples are just a few of the widespread applications of alpha particles. *See* BRAYTON CYCLE; ION-SOLID INTERACTIONS; NUCLEAR BATTERY; SPACE POWER SYSTEMS.

Hazards. Since an energetic alpha particle usually loses its energy by ionizing the atoms it encounters, it causes severe damage to living cells in its path. However, because of its short range, an alpha-particle source external to the body does not generally pose a problem. The major hazard for humans results if an alpha-particle emitter is ingested and becomes a long-term component of the body. Great care must be taken when working with alpha emitters to prevent such ingestion.

Research on indoor air quality in homes and buildings has brought attention to the long-standing health hazard associated with radon. Small amounts of this alpha-emitting gas, which is generated in the decay of heavy elements in granite and building materials, can diffuse within buildings, to be ingested by occupants. This problem is exacerbated by the fact that increasing numbers of modern buildings are tightly sealed, concentrating the radon. Environmental sources of alpha-particle exposure also include radon in mines, polonium in cigarette smoke, and thorium in coal. *See* ENVIRONMENTAL RADIOACTIVITY; RADON.

Along with the many other dangerous components, alpha-emitting isotopes in nuclear waste present serious health hazards. Efforts to achieve safe disposal of the growing accumulations of nuclear waste are receiving increased attention. *See* RADIOACTIVE WASTE MANAGEMENT.

Carrol Bingham

Bibliography. Y. A. Akovali, Review of alpha-decay data from doubly-even nuclei, *Nucl. Data Sheets*, 84:1–113, 1998; M. Eisenbud and T. F. Gesell, *Environmental Radioactivity: From Natural, Industrial, and Military Sources*, 4th ed., Academic Press, 1997; K. S. Krane, *Introductory Nuclear Physics*, Wiley, 1987; S. S. M. Wong, *Introductory Nuclear Physics*, 2d ed., Wiley, 1998; F. Yang and J. H. Hamilton, *Modern Atomic and Nuclear Physics*, McGraw-Hill, 1996.

Alpine vegetation

Plant growth forms characteristic of upper reaches of forests on mountain slopes. In such an environment, trees undergo gradual changes that, though subtle at first, may become dramatic beyond the dense forest as the zone of transition leads into the nonforested zone of the alpine tundra. In varying degrees, depending on the particular mountain setting, the forest is transformed from a closed-canopy forest to one of deformed and dwarfed trees interspersed with alpine tundra species (**Fig. 1**). This zone of transition is referred to as the forest-alpine tundra ecotone. The trees within the ecotone are stunted, often shrublike, and do not have the symmetrical shape of most trees within the forest interior. The classic image is one of twisted, stunted, and struggling individual trees clinging to a windswept ridge (**Fig. 2**). The ecotone in which these trees exist is visually one of the most striking vegetational transition areas known.

Fig. 1. Forest-alpine tundra near Hyndman Cirque, Idaho. The trees change in density and form within the ecotone beyond the boundary of the timberline upslope toward the alpine tundra. The distribution of timberline, treeline, and the ecotone is irregular because of climatic, geomorphic, and topographic controls. *(Courtesy of T. Crawford)*

Fig. 2. Whitebark pine (*Pinus albicaulis*) showing the classic dwarfed and contorted image of trees within the windswept forest-alpine tundra ecotone of the Beartooth Plateau, Wyoming. *(Courtesy of T. Crawford)*

These trees are often referred to as krummholz, a German term meaning crooked wood. In the correct scientific usage, krummholz refers only to those ecotonal trees of the European Alps that have inherited their deformed shape genetically. Trees that have not been proven to have the genetic deformation should be referred to as environmental dwarfs, cripples, elfin wood, or (in reference to their precise form) flag, flag-mat, and mat tree species. Whether or not the deformation is inherited, the ultimate cause is the presence of severe environmental conditions.

The forest-alpine tundra ecotone is a mosaic of both tree and alpine tundra species; and it extends from timberline (the upper limit of the closed-canopy forest of symmetrically shaped, usually evergreen trees) to treeline (the uppermost limit of tree species) and the exposed alpine tundra. With elevational increases, tree deformation is magnified, tree height is reduced, and the total area occupied by trees becomes smaller as the alpine shrub, grass, and herbaceous perennials become more dominant. *See* PLANTS, LIFE FORMS OF.

Distribution. The forest-alpine tundra ecotone is generally at its highest elevation in the tropics and lowest in the polar regions. It is also higher in continental areas than in marine locations. The precise distribution is controlled by a combination of climatic, topographic, and geomorphic processes. The effect of these active processes can be seen by an irregular distribution of timberlines and treelines (Fig. 1). Coniferous trees on many steep slopes are eliminated by avalanches; and they often are replaced by flexible deciduous species or herbaceous meadow cover, leaving trees on the ridges where, undisturbed, they may attain great ages. *See* ALTITUDINAL VEGETATION ZONES; PLANT GEOGRAPHY.

Environmental controls. The environment in which these tenacious individuals survive is harsh and involves a complex interaction of many factors, with the major controlling factor often being climate. The climate is characterized by a short growing season, low air temperatures, frozen soils, drought, high levels of ultraviolet radiation, irregular accumulation of snow, and strong winds. The interaction of all these factors produces varying levels of stress within the trees.

The ultimate cause of the tree deformations and of the eventual complete cessation of tree growth lies in the inability of the tissues of the shoots and the needles to mature and prepare for the harsh environmental conditions. As the length of the growing season decreases with elevation, new needles often do not mature; they have thinner cuticles (the waxlike covering on the needles that protects against desiccation and wind abrasion), and they are less acclimated against low air temperatures. Factors that particularly affect the length of the growing season include air and soil temperatures, and the depth and distribution of snow.

Low air temperatures restrict growth by limiting net uptake and assimilation of carbon dioxide, affecting the ratio of photosynthesis to respiration. With a certain duration of heat, enzymes accumulate and initiate cell division and growth. With an adequately long growing season, the tissues of the trees ripen enough to tolerate the seasonally adverse conditions. If the season is inadequate, the tissues are not well prepared.

Snow serves as the primary source of moisture for the trees; and it insulates the soil, keeping temperatures somewhat moderate. If an adequate amount of snow accumulates around the tree prior to the arrival of the low air temperatures of winter, the soil temperatures may remain above 32°F (0°C), providing liquid water. Availability of water for root uptake and an adequately developed cuticle are critical for preventing winter desiccation, considered a prime cause of tissue loss, deformation, and eventual death. Late-lying snow restricts the warming of soil in early summer, slowing the initiation of growth. Late-lying snow also provides a habitat for a parasitic snow fungus that kills needles covered by snow.

Wind is a major factor in sculpturing the trees in the ecotone. Mechanical abrasion by wind-transported snow, rock, and ice particles pits needle surfaces and breaks stems. The exposure of pitted tissue to a dry, winter atmosphere guarantees desiccation, which is usually lethal. Wind can blow snow away from the trees, exposing them to desiccation,

low air temperatures, and high amounts of radiation. *See* WIND.

Winter air temperatures become critical, depending on the timing and the status of the tree tissues. When adequately prepared, the tree tissues can tolerate temperatures as low as −40°F (−40°C). This survival mechanism, known as cold-hardy acclimation, is the seasonal transition from the tender, frost-susceptible condition to the hardy, non-frost-susceptible one. If the tissues have not adequately acclimated because of a poor growing season, the cells are lethally sensitive to temperatures above the minimums tolerated by well-prepared tissues. *See* ADAPTATION (BIOLOGY); AIR TEMPERATURE.

Radiation is critical, because the greater amounts of ultraviolet at higher elevations are magnified by the reflective snow cover. Deactivation of the chlorophyll may occur, resulting in the weakening or death of needle tissue. Large amounts of radiation can be detrimental to photosynthesis, again restricting the growth of viable tissue tolerant of the seasonal climates.

Tree forms. Changes in tree form reflect increases in severity of climate with elevational rise and topographic exposure. A common progression of form changes, upslope from timberline to treeline, is from a flag, to a flag-mat, to a mat form. The trees with a flag form have branching mainly on the leeward, protected side of the upper part of the trunk. The flag branching is exposed above the winter snow cover, and growth on the windward side is eroded away by mechanical abrasion. Snow often accumulates to great depths around the base of these trees, inducing the parasitic snow fungus there. The combination results in a distorted flag form with barren branches on the lower trunk (**Fig. 3**). The trees of flag-mat form support scrawny trunks that are flagged. Close to the ground, a mat, or shrublike growth, survives protected under the winter snowpack. At the highest elevations within the ecotone, the mat (cushion-tree) form exists. The tree is dwarfed to a mat. Rarely are vertical branches or upright leaders found. Each mat has a size and shape that is adapted to its specific topographic environment. The wind side of the mat is a contorted mass of broken needles and distorted branches. Survival of the mat is dependent on winter snow cover, as shoots that project above the snow are abraded mechanically.

Fig. 3. Flagged tree, mainly supporting branching on the leeside of the upper trunk; needles have been lost from the lower trunk to the parasitic snow fungus. ***(Courtesy of T. Crawford)***

Importance of trees. The environmentally dwarfed and contorted trees of the forest-alpine tundra ecotone are important for many reasons. Snow is maintained in the ecotone by low air temperatures and tree shading, providing a water source late into the summer season. The trees afford shade, wind protection, and the moisture of snowbanks, critical for the establishment and survival of seedlings. Many of the oldest trees in the world have been found in the ecotone. These offer yearly historical records of events (such as climate changes, volcanic eruptions, pollution, and fire), as recorded in tree rings. Finally, the ecotone offers a unique habitat for other plants and animals to thrive as the two communities combine resources.

Ecosystems of the forest-alpine tundra ecotone are increasingly subjected to disturbances from human activities. Large numbers of visitors enter mountain areas and forest-tundra ecotones in search of recreation and wilderness experience. The impact can be large. Vegetation is trampled, water often becomes polluted, and soil is eroded along trails, contributing to a progressive degradation of ecotones. Trees that grow only a few millimeters each year are sometimes stripped of their wood for campfires or even for decorative items. Then much time is needed to replenish the growth; and in many cases regrowth may not be possible under the present climate. Strategies for management must respond to the human use within these ecosystems, as well as to the impacts of acid rain and additional climatic variability (due to increasing amounts of carbon dioxide and other chemicals in the atmosphere and to depletion of ozone). A combination of good management and better understanding of the processes and dynamics of the forest-alpine tundra ecotone ecosystem would provide a safeguard against potential environmental degradation. *See* ACID RAIN; FOREST ECOSYSTEM; TREE; TUNDRA; VEGETATION AND ECOSYSTEM MAPPING.

Katherine J. Hansen

Bibliography. J. D. Ives (ed.), *Mountains*, 1994; J. D. Ives and R. G. Barry (eds.), *Arctic and Alpine Environments*, 1974; L. W. Price, *Mountains and Man: A Study of Process and Environment*, 1981; W. Tranquillini, *Physiological Ecology of the Alpine Timberline*, 1979.

Alternating current

Electric current that reverses direction periodically, usually many times per second. Electrical energy is ordinarily generated by a public or a private utility organization and provided to a customer, whether industrial or domestic, as alternating current. *See* ELECTRIC POWER GENERATION.

One complete period, with current flow first in one direction and then in the other, is called a cycle, and 60 cycles per second (60 Hz) is the

customary frequency of alternation in the United States and in the rest of North America. In Europe and in many other parts of the world, 50 Hz is the standard frequency of alternation. On aircraft a higher frequency, often 400 Hz, is used to make possible lighter electrical machines.

When the term alternating current is used as an adjective, it is commonly abbreviated to ac, as in ac motor. Similarly, direct current as an adjective is abbreviated dc.

Advantages. The voltage of an alternating current can be changed by a transformer. This simple, inexpensive, static device permits generation of electric power at moderate voltage, efficient transmission for many miles at high voltage, and distribution and consumption at a conveniently low voltage. With direct (constant) current it is not possible to use a transformer to change voltage. On a few power lines, electrical energy is transmitted for great distances as direct current, but the electrical energy is generated as alternating current, transformed to a high voltage, rectified to direct current and transmitted, and then changed back to alternating current by an inverter, to be transformed down to a lower voltage for distribution and use. *See* DIRECT-CURRENT TRANSMISSION.

In addition to permitting efficient transmission of energy, alternating current provides advantages in the design of generators and motors, and for some purposes gives better operating characteristics. Certain devices involving chokes and transformers could not be operated on direct current. Also, the operation of large switches (called circuit breakers) is facilitated because the instantaneous value of alternating current automatically becomes zero twice in each cycle, and an opening circuit breaker need not interrupt the current but only prevent current from starting again after its instant of zero value. *See* ALTERNATING-CURRENT GENERATOR; ALTERNATING-CURRENT MOTOR; CIRCUIT BREAKER.

Sinusoidal form. An alternating current waveform is shown diagrammatically in **Fig. 1**. Time is measured horizontally (beginning at any arbitrary moment) and the current at each instant is measured vertically. In this diagram it is assumed that the current is alternating sinusoidally; that is, the current i is described by Eq. (1), where I_m is the maximum

$$i = I_m \sin 2\pi f t \tag{1}$$

instantaneous current, f is the frequency in cycles per second (hertz), and t is the time in seconds. *See* SINE WAVE.

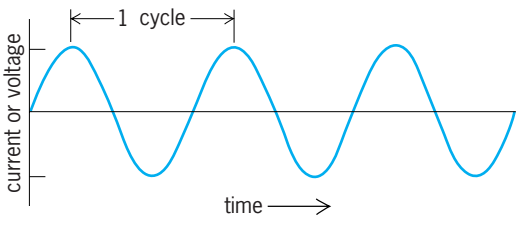


Fig. 1. Diagram of sinusoidal alternating current.

Measurement. Quantities commonly measured by ac meters and instruments are energy, power, voltage, and current. Other quantities less commonly measured are reactive volt-amperes, power factor, frequency, and demand (of energy during a given interval such as 15 min).

Energy is measured on a watt-hour meter. There is usually such a meter where an electric line enters a customer's premises. The meter may be single-phase (usual in residences) or three-phase (customary in industrial installations), and it displays on a register of dials the energy that has passed, to date, to the system beyond the meter. The customer frequently pays for energy consumed per the reading of such a meter. *See* ELECTRICAL ENERGY MEASUREMENT; WATT-HOUR METER.

Power is measured on a wattmeter. Since power is the rate of consumption of energy, the reading of the wattmeter is proportional to the rate of increase of the reading of a watt-hour meter. The same relation is expressed by saying that the reading of the watt-hour meter, which measures energy, is the integral (through time) of the reading of the wattmeter, which measures power. A wattmeter usually measures power in a single-phase circuit, although three-phase wattmeters are sometimes used. *See* ELECTRIC POWER MEASUREMENT; WATTMETER.

Current is measured by an ammeter. Power absorbed by an element is the product of current through the element, voltage across it, and the power factor between the current and the voltage, as shown in Eq. (5). With unidirectional (direct) current, the amount of current is the rate of flow of electricity; it is proportional to the number of electrons passing a specified cross section of a wire per second. This is likewise the definition of current at each instant of an alternating-current cycle, as current varies from a maximum in one direction to zero and then to a maximum in the other direction (Fig. 1). An oscilloscope will indicate instantaneous current, but the value of instantaneous current is not often useful for analysis purposes. A dc (d'Arsonval-type) ammeter will measure average current, but this is useless in an ac circuit, for the average of sinusoidal current is zero. A useful measure of alternating current is found in the ability of the current to do work, and the amount of current is correspondingly defined as the square root of the average of the square of instantaneous current, the average being taken over an integer number of cycles. This value is known as the root-mean-square (rms) or effective current. It is measured in amperes. It is a useful measure for current of any frequency. The rms value of direct current is identical to its dc value. The rms value of sinusoidally alternating current is $I_m/2$, where I_m is the maximum instantaneous current. [See Fig. 1 and Eq. (1).] *See* AMMETER; CURRENT MEASUREMENT; OSCILLOSCOPE.

Voltage is measured by a voltmeter. Voltage is the electrical pressure. It is measured between one point and another in an electric circuit, often between the

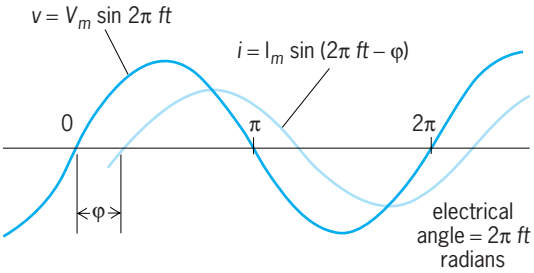


Fig. 2. Phase angle φ.

two wires of the circuit. As with current, instantaneous voltage in an ac circuit reverses each half cycle and the average of sinusoidal voltage is zero. Therefore the rms or effective value of voltage is used in ac systems. The rms value of sinusoidally alternating voltage is $V_m/2$, where V_m is the maximum instantaneous voltage. This rms voltage, together with rms current and the circuit power factor, is used to compute electric power, as in Eqs. (4) and (5). *See* VOLTAGE MEASUREMENT; VOLTMETER.

The ordinary voltmeter is connected by wires to the two points between which voltage is to be measured, and voltage is proportional to the current that results through a very high electrical resistance within the voltmeter itself. The voltmeter, actuated by this current, is calibrated in volts.

Phase difference. Phase difference is a measure of the fraction of a cycle by which one sinusoidally alternating quantity leads or lags another. **Figure 2** shows a voltage waveform v which is described in Eq. (2) and a current i which is described in Eq. (3).

$$v = V_m \sin 2\pi f t \tag{2}$$

$$i = I_m \sin (2\pi f t - \varphi) \tag{3}$$

The angle φ is called the phase difference between the voltage and the current; this current is said to lag (behind this voltage) by the angle φ. It would be equally correct to say that the voltage leads the current by the phase angle φ. Phase difference can be expressed as a fraction of a cycle, as an angle in degrees, or as an angle in radians, as shown in Eq. (3). If there is no phase difference, and $\varphi = 0$, voltage and current are in phase. If the phase difference is a quarter cycle, and $\varphi = \pm 90^\circ$, the quantities are said to be in quadrature.

Power factor. Power factor is defined in terms of the phase angle. If the rms value of sinusoidal current from a power source to a load is I and the rms value of sinusoidal voltage between the two wires connecting the power source to the load is V, the average power P passing from the source to the load is shown as Eq. (4). The cosine of the phase angle,

$$P = VI \cos \varphi \tag{4}$$

cos φ, is called the power factor. Thus the rms voltage, the rms current, and the power factor are the components of power.

The foregoing definition of power factor has meaning only if voltage and current are sinusoidal. Whether they are sinusoidal or not, average power, rms voltage, and rms current can be measured, and a value for power factor is implicit in Eq. (5). This

$$P = VI \text{ (power factor)} \tag{5}$$

gives a definition of power factor when V and I are not sinusoidal. If the voltage across the element and the current through it are in phase (and of the same waveform), power factor equals 1. If voltage and current are out of phase, power factor is less than 1. If voltage and current are sinusoidal and in quadrature, then the power factor is zero.

The phase angle and power factor of voltage and current in a circuit that supplies a load are determined by the load. Thus a load of pure resistance, such as an electric heater, has unity power factor. An inductive load, such as an induction motor, has a power factor less than 1 and the current lags behind the applied voltage. A capacitive load, such as a bank of capacitors, also has a power factor less than 1, but the current leads the voltage, and the phase angle φ is negative.

If a load that draws lagging current (such as an induction motor) and a load that draws leading current (such as a bank of capacitors) are both connected to a source of electric power, the power factor of the two loads together can be higher than that of either one alone, and the current to the combined loads may have a smaller phase angle from the applied voltage than would currents to either of the two loads individually. Although power to the combined loads is equal to the arithmetic sum of power to the two individual loads, the total current will be less than the arithmetic sum of the two individual currents (and may, indeed, actually be less than either of the two individual currents alone). It is often practical to reduce the total incoming current by installing a bank of capacitors near an inductive load, and thus to reduce power lost in the incoming distribution lines and transformers, thereby improving efficiency. This process is called power-factor correction.

Three-phase system. Three-phase systems are commonly used for generation, transmission, and distribution of electric power. A customer may be supplied with three-phase power, particularly if a large amount of power is needed or if the customer wishes to use three-phase loads. Small domestic customers are usually supplied with single-phase power.

A three-phase system is essentially the same as three ordinary single-phase systems with the three voltages of the three single-phase systems out of phase with each other by one-third of a cycle (120°), as shown in **Fig. 3**. The three voltages may be written

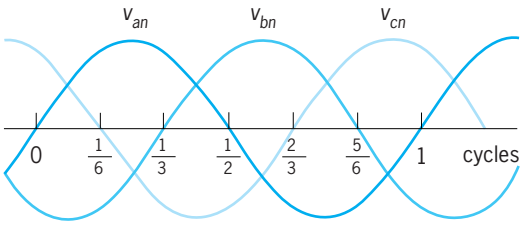


Fig. 3. Voltages of a balanced three-phase system.

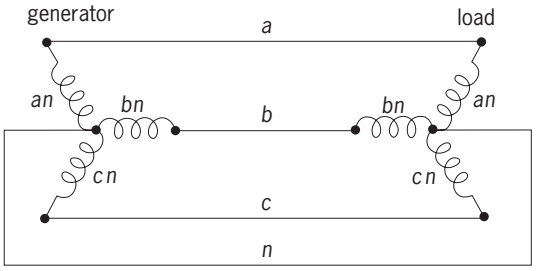


Fig. 4. Connections of a simple three-phase system.

as Eqs. (6), (7), and (8), where $V_{an(\max)}$ is the maxi-

$$v_{an} = V_{an(\max)} \sin 2\pi ft \tag{6}$$

$$v_{bn} = V_{bn(\max)} \sin 2\pi (ft - {}^{1}/_{3}) \tag{7}$$

$$v_{cn} = V_{cn(\max)} \sin 2\pi (ft - {}^{2}/_{3}) \tag{8}$$

mum value of voltage in phase *an*, and so on. The three-phase system is balanced if relation (9) holds,

$$V_{an(\max)} = V_{bn(\max)} = V_{cn(\max)} \tag{9}$$

and if the three phase angles are equal, one-third cycle each as shown.

If a three-phase system were actually three separate single-phase systems, there would be two wires between the generator and the load of each system, requiring a total of six wires. In fact, however, a single wire can be common to all three systems, so that it is only necessary to have three wires for a three-phase system (**Fig. 4***a–c*) plus a fourth wire *n* serve as a common return or neutral conductor. On some systems the Earth is used as the common or neutral conductor.

Each phase of a three-phase system carries current and conveys power and energy. If the three loads on the three phases of the three-phase system are identical and the voltages are balanced, then the currents are balanced also. Figure 2 can then apply to any one of the three phases. It will be recognized that the three currents in a balanced system are equal in rms (or maximum) value and that they are separated one from the other by phase angles of one-third cycle and two-thirds cycle. Thus the currents (in a balanced system) are themselves symmetrical. Note, however, that the three currents will not necessarily be in phase with their respective voltages; the corresponding voltages and currents will be in phase with each other only if the load is pure resistance and the phase angle between voltage and current is zero. Otherwise some such relation as that of Fig. 2 will apply to each phase.

It is significant that, if the three currents of a three-phase system are balanced, their sum is zero at every instant. In practice, the three currents are not usually exactly balanced, and one of two situations arises. Either the common neutral wire *n* is used, in which case it carries little current (and may be of high resistance compared to the other three line wires), or else the common neutral wire *n* is not used, only three line wires being installed, and the three phase currents are thereby forced to add to zero even though this requirement results in some imbalance of phase voltages at the load.

It is also significant that the total instantaneous power from generator to load is constant (does not vary with time) in a balanced, sinusoidal, three-phase system. Power in a single-phase system that has current in phase with voltage is maximum when voltage and current are maximum, and is instantaneously zero when voltage and current are zero; if the current of the single-phase system is not in phase with the voltage, the power will reverse its direction of flow during part of each half cycle. However, in a balanced three-phase system, regardless of the phase angle, the flow of power is unvarying from instant to instant. This results in smoother operation and less vibration of motors and other ac devices.

Three-phase systems are almost universally used for large amounts of power. In addition to providing smooth flow of power, three-phase motors and generators are more economical than single-phase machines. Polyphase systems with two, four, or other numbers of phases are possible, but they are little used except when a large number of phases, such as 12, is desired for economical operation of a rectifier.

H. H. Skilling

Symmetrical (0, 1, 2) components. When three coils are equally spaced around the periphery of the stator of a generator, a properly shaped magnetic field rotating in a forward direction at uniform velocity will induce voltages in these coils. The three voltages may be written as Eqs. (6), (7), and (8). For analytical purposes it is more convenient to express the voltages as phasors, as in Eqs. (10). For a discus-

$$\begin{aligned} \mathbf{V}_{a1} &= \mathbf{V}_{a1} \\ \mathbf{V}_{b1} &= \mathbf{V}_{a1}\, e^{-j2\pi/3} \\ \mathbf{V}_{c1} &= \mathbf{V}_{a1}\, e^{+j2\pi/3} \end{aligned} \tag{10}$$

sion of phasor notation and complex representation. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

In order to simplify notation, it has become accepted practice to introduce a standard set of unit three-phase phasors as in Eqs. (11), so that Eqs. (10)

$$1 = e^{j0} \quad \mathbf{a} = e^{j2\pi/3} \quad \mathbf{a}^2 = e^{-j2\pi/3} \tag{11}$$

may be written more succinctly in matrix format as in Eq. (12).

$$\begin{bmatrix} \mathbf{V}_{a1} \\ \mathbf{V}_{b1} \\ \mathbf{V}_{c1} \end{bmatrix} = (\mathbf{V}_{a1}) \begin{bmatrix} 1 \\ \mathbf{a}^2 \\ \mathbf{a} \end{bmatrix} \tag{12}$$

See MATRIX THEORY.

When the voltages proceed in time-phase in the order $\mathbf{V}_a$, $\mathbf{V}_b$, $\mathbf{V}_c$, as reflected in Eqs. (6), (7), and (8), the set of three-phase voltage is called a positive-sequence set of voltages and is identified by the subscript 1, as in Eqs. (10) and (12). When three coils are equally spaced around the periphery of the stator of a generator, a properly shaped magnetic field rotating in a backward direction at uniform velocity will induce voltages in these coils which proceed in time-phase in the order $\mathbf{V}_a$, $\mathbf{V}_c$, $\mathbf{V}_b$, and the set of

three-phase voltages is called a negative-sequence set of voltages and identified by the subscript 2. In terms of phasor notation, the voltages are written as in Eqs. (13).

$$\begin{bmatrix} \mathbf{V}_{a2} \\ \mathbf{V}_{b2} \\ \mathbf{V}_{c2} \end{bmatrix} = (\mathbf{V}_{a2}) \begin{bmatrix} 1 \\ \mathbf{a} \\ \mathbf{a}^2 \end{bmatrix} \tag{13}$$

If coils b and c are placed in the same slot with coil a, then $\mathbf{V}_a = \mathbf{V}_b = \mathbf{V}_c$, and the set of three-phase voltages is called a zero-sequence set of voltages and identified by the subscript 0. In terms of phasor notation the voltages are written as in Eqs. (14).

$$\begin{bmatrix} \mathbf{V}_{a0} \\ \mathbf{V}_{b0} \\ \mathbf{V}_{c0} \end{bmatrix} = (\mathbf{V}_{a0}) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{14}$$

The set of positive-, negative-, and zero-sequence components are collectively referred to as symmetrical components.

If the voltages $\mathbf{V}_a$, $\mathbf{V}_b$, and $\mathbf{V}_c$ are written as in Eqs. (15), and, in matrix format, as in Eq. (16), then

$$\begin{aligned} \mathbf{V}_a &= \mathbf{V}_{a0} + \mathbf{V}_{a1} + \mathbf{V}_{a2} = \mathbf{V}_{a0} + \mathbf{V}_{a1} + \mathbf{V}_{a2} \\ \mathbf{V}_b &= \mathbf{V}_{b0} + \mathbf{V}_{b1} + \mathbf{V}_{b2} = \mathbf{V}_{a0} + \mathbf{a}^2\mathbf{V}_{a1} + \mathbf{a}\mathbf{V}_{a2} \\ \mathbf{V}_c &= \mathbf{V}_{c0} + \mathbf{V}_{c1} + \mathbf{V}_{c2} = \mathbf{V}_{a0} + \mathbf{a}\mathbf{V}_{a1} + \mathbf{a}^2\mathbf{V}_{a2} \end{aligned} \tag{15}$$

$$\begin{bmatrix} \mathbf{V}_a \\ \mathbf{V}_b \\ \mathbf{V}_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \mathbf{a}^2 & \mathbf{a} \\ 1 & \mathbf{a} & \mathbf{a}^2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{a0} \\ \mathbf{V}_{a1} \\ \mathbf{V}_{a2} \end{bmatrix} \tag{16}$$

by selecting complex values for $\mathbf{V}_{a0}$, $\mathbf{V}_{a1}$, and $\mathbf{V}_{a2}$, respectively, in an arbitrary manner, it is possible to establish an infinite number of unbalanced phase voltages $\mathbf{V}_a$, $\mathbf{V}_b$, and $\mathbf{V}_c$. Conversely, and more important, since the inverse relationship given by Eq. (17)

$$\begin{bmatrix} \mathbf{V}_{a0} \\ \mathbf{V}_{a1} \\ \mathbf{V}_{a2} \end{bmatrix} = (1/3) \begin{bmatrix} 1 & 1 & 1 \\ 1 & \mathbf{a} & \mathbf{a}^2 \\ 1 & \mathbf{a}^2 & \mathbf{a} \end{bmatrix} \begin{bmatrix} \mathbf{V}_a \\ \mathbf{V}_b \\ \mathbf{V}_c \end{bmatrix} \tag{17}$$

exists, for any given unbalanced set of three-phase voltages, it is always possible to determine a corresponding unique set of zero-, positive-, and negative-sequence voltages. Equally well, for any given unbalanced set of three-phase currents, it is always possible to determine a corresponding unique set of zero-, positive-, and negative-sequence currents.

Three-phase power system components, such as transmission lines, transformers, loads, motors, and generators, may, in many practical instances, be considered geometrically symmetrical among the three phases. In this event, it can be shown that when a zero-, positive-, or negative-sequence voltage is applied to such a component, the resultant current will be a zero-, positive-, or negative-sequence component, respectively. The associated zero-, positive-, and negative-sequence impedances can be readily identified.

Consequently, because of the symmetrical nature of each of the symmetrical-component voltages and currents, it is necessary to consider only one of the three phases of the three-phase power system in an analysis. By accepted convention, phase "a" is the selected phase.

Perhaps the most important virtue of symmetrical components lies in the fact that if a common type of unsymmetrical fault (single-line-to-ground fault, line-to-line fault, one open conductor, two open conductors, and so forth) occurs at one point in an otherwise symmetrical three-phase network, a relatively simple interconnection occurs between the symmetrical component networks at the fault location.

M. Harry Hesse

Power and information. Although this article has emphasized electric power, ac circuits are also used to convey information. An information circuit, such as telephone, radio, or control, employs varying voltage, current, waveform, frequency, and phase. Efficiency is often low, the chief requirement being to convey accurate information even though little of the transmitted power reaches the receiving end. For further consideration of the transmission of information. *See* ELECTRICAL COMMUNICATIONS; RADIO; TELEPHONE; WAVEFORM.

An ideal power circuit should provide the customer with electrical energy always available at unchanging voltage of constant waveform and frequency, the amount of current being determined by the customer's load. High efficiency is greatly desired. *See* CAPACITANCE; CIRCUIT (ELECTRICITY); ELECTRIC CURRENT; ELECTRIC FILTER; ELECTRICAL IMPEDANCE; ELECTRICAL RESISTANCE; INDUCTANCE; JOULE'S LAW; NETWORK THEORY; OHM'S LAW; RESONANCE (ALTERNATING-CURRENT CIRCUITS).

H. H. Skilling

Bibliography. J. Arrillaga and C. P. Arnold, *Computer Analysis of Power Systems*, 1990; O. I. Elgerd, *Electric Energy Systems Theory*, 2d ed., 1982; J. J. Grainger and W. D. Stevenson, Jr., *Power System Analysis*, 1994; C. P. Paul, S. A. Nasar, and L. E. Unnewehr, *Introduction to Electrical Engineering*, 2d ed., 1992.

Alternating-current circuit theory

The mathematical theory for the analysis of electric circuits when the currents and voltages are alternating functions of time.

In an electric circuit carrying direct current only (**Fig. 1**), the current I (amperes) flowing through a circuit of total resistance R (ohms) when a steady voltage V (volts) is applied is given by Ohm's law, Eq. (1).

$$I = \frac{V}{R} \tag{1}$$

See CIRCUIT (ELECTRICITY); DIRECT-CURRENT CIRCUIT THEORY; OHM'S LAW.

In many practical situations, electrical energy is generated, transmitted, and distributed in a form in which the currents and voltages are not constant but are sinusoidally varying or "alternating" functions of time (**Fig. 2**). Public electricity supply systems are

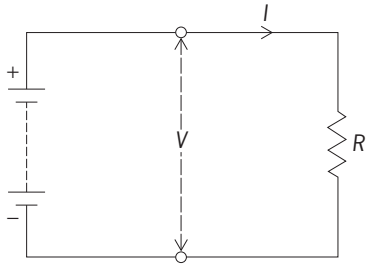


Fig. 1. Direct-current circuit.

of this type, and many telecommunications systems use electrical signals of sinusoidal form as carriers of information. It is necessary to have a comprehensive theory to analyze such systems and to provide the basis for their design.

This is the object of alternating-current circuit theory. It differs from direct-current theory in that there are circuit elements other than resistors (particularly inductors and capacitors) which respond to (impede) the flow of current: these are known generally as impedances. Also, it is possible to couple energy from one circuit to another by means of mutual inductances or transformers without the need for a linking conductor. Whereas in the dc case the sizes of voltages, currents, and resistors can all be described by single numbers, in the ac case pairs of numbers are needed to describe amplitude and phase or, for an impedance, amplitude ratio and phase shift. Mathematically, the properties and interrelations of pairs of numbers can be dealt with succinctly by using complex algebra. Alternating-current circuit theory is treated using simple time-domain analysis of waveforms, illustrated by phasor diagrams. It will be shown how the use of complex algebra provides a more powerful and elegant theory.

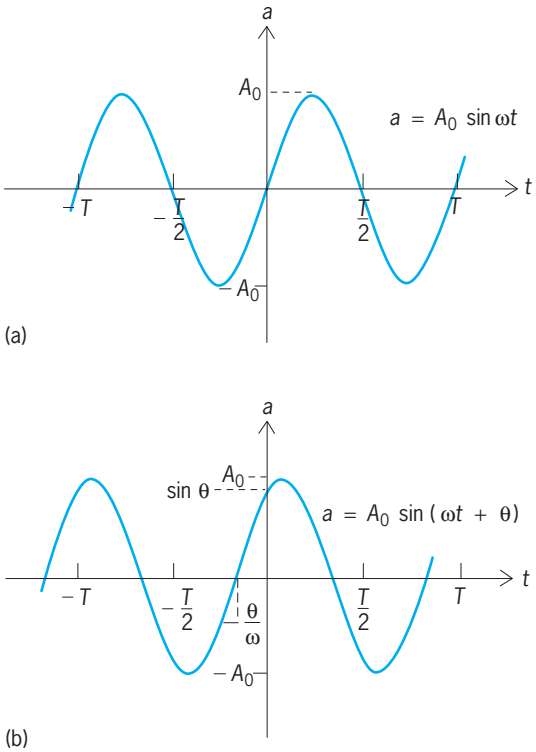


Fig. 2. Sinusoidal waveforms. (*a*) Simple sinusoid without phase shift (*b*) Sinusoid with phase shift.

Waveforms. A sinusoidal (or alternating) function a of time t is shown in Fig. 2*a*. It is described by Eq. (2), where A_0 is the maximum value (or amplitude) of a and ω is known as the angular or radian frequency. The period T of the function is the time after which the function repeats itself. During one period the function takes on all its possible values and is said to go through one cycle. The frequency f of the sinusoid as usually defined (for example, 60 Hz) is the number of cycles that occur in 1 s, so $f = 1/T$. Since $\sin(x + 2\pi) = \sin x$, $\omega T = 2\pi$, and hence $\omega = 2\pi f$.

$$a(t) = A_0 \sin \omega t \tag{2}$$

It is necessary to consider waveforms which do not cross the zero level at $t = 0$ but which include a phase shift. Figure 2*b* shows the waveform of Eq. (3), which waveform is identical in shape with

$$a(t) = A_0 \sin(\omega t + \theta) \tag{3}$$

the waveform described by Eq. (2) but shifted in time so that it crosses the zero level at a time θ/ω before $t = 0$. This is referred to as a phase advance; the corresponding waveform with θ negative would illustrate a phase delay.

Since $\sin(x + \pi/2) = \cos x$, Eq. (3) could equally well be written as Eq. (4), where the phase angle is

$$\begin{aligned} a &= A_0 \sin\left(\omega t + \frac{\pi}{2} + \theta - \frac{\pi}{2}\right) \\ &= A_0 \cos(\omega t + \psi) \end{aligned} \tag{4}$$

now $\psi = \theta - \pi/2$ radians. Henceforth, either of the forms in Eqs. (3) and (4) will be used, as convenient, to represent an alternating quantity. An alternating voltage is written as Eq. (5), and an alternating current as Eq. (6), where v and i represent the instantaneous values, and V_0 and I_0 the maximum values or amplitudes, of voltage and current respectively. *See* ALTERNATING CURRENT.

$$v = V_0 \sin(\omega t + \theta) \text{ or } v = V_0 \cos(\omega t + \psi) \tag{5}$$

$$i = I_0 \sin(\omega t + \theta) \text{ or } i = I_0 \cos(\omega t + \psi) \tag{6}$$

Phasor diagrams. It is helpful to visualize an alternating waveform as the projection of a rotating vector. This may be considered as corresponding in some sense to the generation of an alternating current by the rotation of a coil in a magnetic field.

In **Fig. 3***a*, OR is a vector of length A_0 rotating in the anticlockwise direction about the origin of the x-y plane with angular velocity ω radians per second. Then the projection of the vector OR on the x axis is $OX = A_0 \cos \omega t$, and the projection on the y axis is $OY = A_0 \sin \omega t$. The rotating vector can therefore be thought of as generating a sinusoidal waveform.

If, instead of the vector OR, the vector OR' in Fig. 3*b* is considered, displaced from OR by an angle ϕ and rotating with it, then $OX' = \cos(\omega t + \phi)$ and

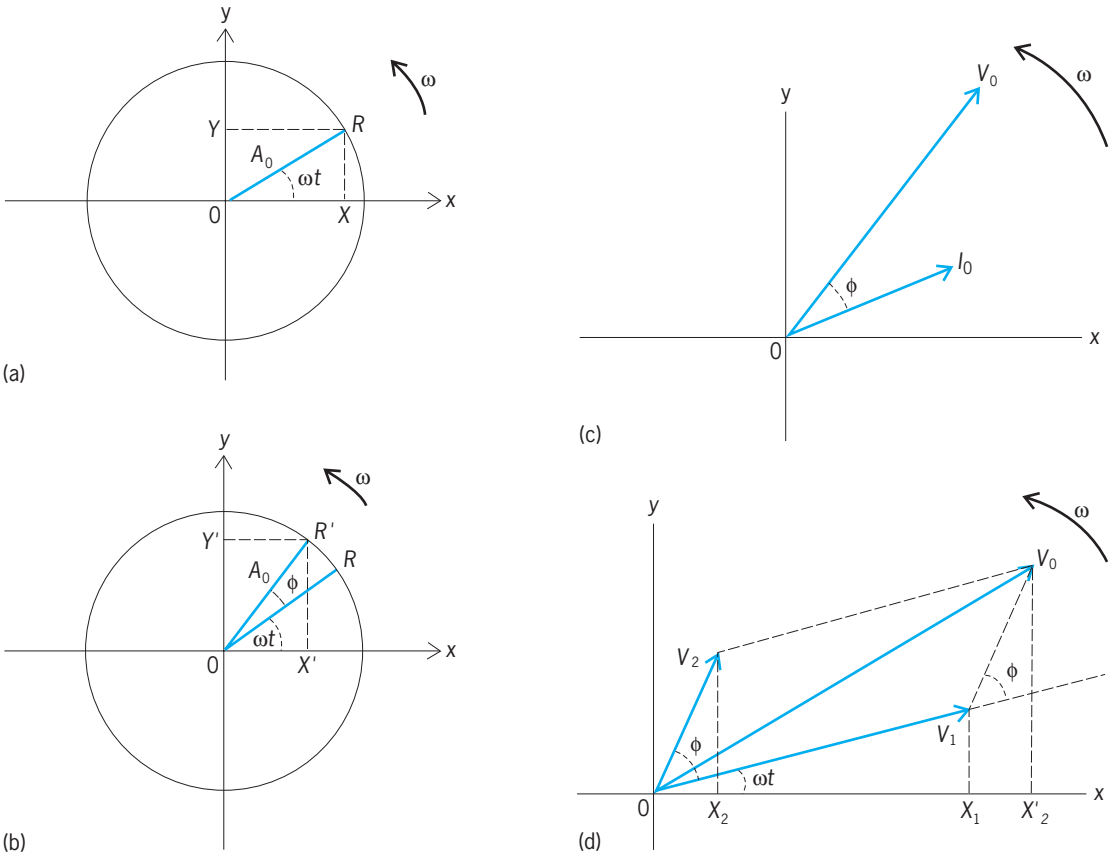


Fig. 3. Phasors. (*a*) Simple phasor. (*b*) Phasor with phase shift. (*c*) Voltage and current phasors. (*d*) Phasor for the sum of two voltages.

$OY' = \sin(\omega t + \phi)$, so that rotation of the phasor (the name given to the vector in this type of diagram) corresponds to a phase shift in the waveform.

In a linear circuit, the voltage and current waveforms have related amplitudes and often a phase shift between them, but both are of the same frequency. They can thus be represented, as in Fig. 3*c*, as the projections of phasors of amplitudes V_0 and I_0 with an angle ϕ between them, rotating together in the $x-y$ plane. Circuit theory enables the ratio V_0/I_0 phase shift ϕ to be calculated from the circuit parameters.

If there are two voltages (or two currents) of different amplitudes and with a phase shift between them, say $v_1 = V_1 \cos \omega t$ and $v_2 = V_2 \cos(\omega t + \phi)$, represented as the projection on the x-axis of two phasors of amplitudes V_1 and V_2, it is easy to see from Fig. 3*d* that their sum can be represented as the projection of the resultant phasor V_0, the vector sum of V_1 and V_2. For $OX_1 = v_1$, $X_1 X'_2 = OX_2 = v_2$, and therefore $OX'_2 = v_1 + v_2$.

Complex representation. While a phasor diagram helps to visualize the relationship between current and voltage in an ac circuit, it does little to help with the mathematical calculations. These are greatly simplified if complex algebra is used to describe the current, voltage, and circuit parameters. Only the formulas of complex algebra essential for circuit theory are reviewed here. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

A complex number z is defined by a pair of ordinary numbers. It can be written as in Eq. (7), where

$$z = x + jy \tag{7}$$

x and y are real numbers and $j = -1$. (Mathematicians often use i for this quantity, but in electrical work the symbol i is reserved for electric current.) Here x is called the real part of z, written Re (z), and y the imaginary part, written Im (z). If x and y are regarded as coordinates in a (complex) plane, z defines a point in the plane, as in **Fig. 4***a*. The distance r from the origin O to z is known as the modulus of z, written $|z|$; and the angle θ between the line from O to z and the real (x) axis is known as the argument of z. These quantities are given by Eqs. (8) and (9), and x and y are expressed in terms of them by Eqs. (10).

$$|z| = r = \sqrt{x^2 + y^2} \tag{8}$$

$$\theta = \tan^{-1} y/x \tag{9}$$

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned} \tag{10}$$

De Moivre's theorem states that Eq. (11) is valid,

$$e^{j\theta} = \cos\theta + j\sin\theta \tag{11}$$

so an alternative way of writing z is Eq. (12),

and in giving numerical values it is often convenient to write Eq. (13).

$$z = re^{j\theta} \tag{12}$$

$$z = r\angle\theta \tag{13}$$

The complex conjugate of z, z^* is defined by Eq. (14). This is illustrated in Fig. 4*b*.

$$z^* = x - jy = re^{-j\theta} \tag{14}$$

Multiplication by j is equivalent to rotation through $\pi/2$ radians, or 90°, as Fig. 4*c* illustrates, for

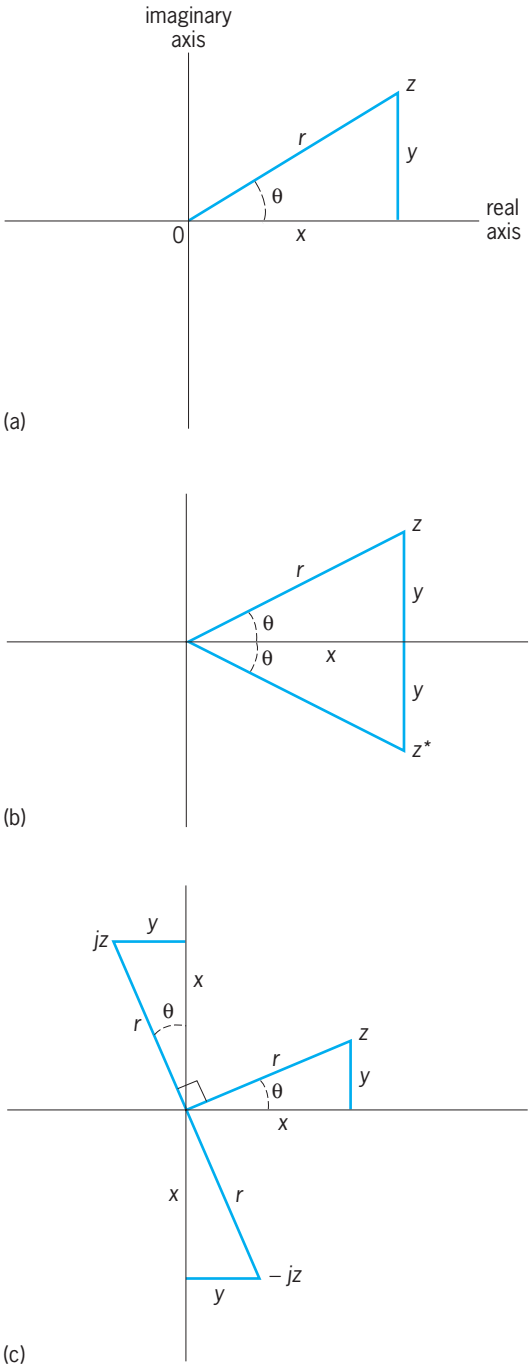


Fig. 4. Complex numbers. (*a*) The number $z = x + jy = re^{j\theta}$. (*b*) Its complex conjugate $z^* = x - jy = re^{j\theta}$. (*c*) Multiplication of z by $\pm j$, equivalent to rotation through 90°.

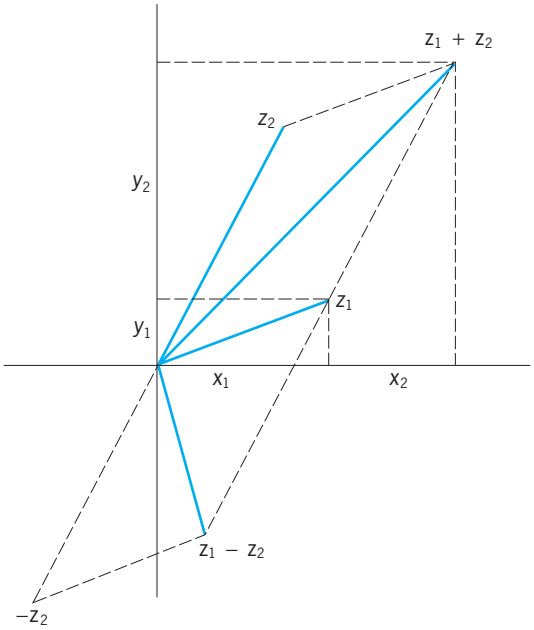


Fig. 5. Addition and subtraction of complex numbers.

Eq. (15) is valid since $j^2 = -1$, and the real and imaginary parts of the number interchange as shown. Similarly, multiplication by $-j$ corresponds to rotation through 90° in the opposite (clockwise) direction. From this it is apparent that $j = e^{j\pi/2}$.

$$jz = j(x + jy) = -y + jx \tag{15}$$

In circuit analysis it is necessary to combine complex numbers in various ways. The rules of complex algebra for combining the numbers $z_1 = x_1 + jy_1$ and $z_2 = z_2 + jy_2$ are as follows.

The rule for addition is given by Eq. (16). This can

$$z_1 + z_2 = (x_1 + x_2) + j(y_1 + y_2) \tag{16}$$

be interpreted geometrically by saying that $z_1 + z_2$ is represented by the vector sum of the two vectors Oz_1 and Oz_2, as shown in **Fig. 5**.

The rule for subtraction is given by Eq. (17).

$$z_1 - z_2 = (x_1 - x_2) + j(y_1 - y_2) \tag{17}$$

Similarly, $z_1 - z_2$ can be represented as the vector sum of the two vectors Oz_1 and $O(-z_2)$, as in Fig. 5.

The rule for multiplication is given by Eq. (18), and that for division by Eq. (19). The process

$$\begin{aligned} z_1 z_2 &= (x_1 + jy_1)(x_2 + jy_2) \\ &= (x_1x_2 - y_1y_2) + j(x_1y_2 + x_2y_1) \\ &= r_1r_2e^{j(\theta_1+\theta_2)} \end{aligned} \tag{18}$$

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{x_1 + jy_1}{x_2 + jy_2} \\ &= \frac{(x_1 + jy_1)(x_2 - jy_2)}{(x_2 + jy_2)(x_2 - jy_2)} \\ &= \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + j\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2} \\ &= (r_1/r_2)e^{j(\theta_1-\theta_2)} \end{aligned} \tag{19}$$

illustrated in Eq. (19) for converting the quotient of two complex numbers into a number of form $a + jb$ by multiplying both numerator and denominator by the complex conjugate of the denominator is known as rationalization.

A comparison of Figs. 3*a* and 4*a* shows that alternating waveforms can be represented by the real or imaginary parts of complex numbers of argument ωt, as in Eqs. (20).

$$A_0 \cos \omega t = A_0 \text{ Re } e^{j\omega t}$$
$$A_0 \sin \omega t = A_0 \text{ Im } e^{j\omega t} \tag{20}$$

A sinusoid of arbitrary phase may be written as in Eq. (21), where B is given by Eq. (22) [by De Moivre's theorem]. B contains all the information needed in circuit analysis, since A_0 and θ can be calculated from B, and the frequency is given by ω. It is therefore possible to formulate a circuit analysis in which the driving waveforms are of the form given by Eqs. (23), where V and I are complex quantities.

$$A_0 \cos(\omega t + \theta) = {}^1/_2 A_0 e^{j(\omega t+\theta)} + {}^1/_2 A_0 e^{-j(\omega t+\theta)}$$
$$= Be^{j\omega t} + B^* e^{-j\omega t} \tag{21}$$

$$B = {}^1/_2 A_0 e^{j\theta} \tag{22}$$

$$v = Ve^{j\omega t} = V_0 e^{j\theta} e^{j\omega t}$$
$$i = Ie^{j\omega t} = I_0 e^{j\theta} e^{j\omega t} \tag{23}$$

The circuit parameters are also in complex form, and from these a response in similar form can be calculated which can be directly interpreted as a sinusoidal waveform.

Circuit elements. With constant currents, the only circuit element of significance is resistance. Voltage, current, and resistance are related by Ohm's law: circuit theory is needed to calculate the effective resistance of various series and parallel combinations of individual resistors. With alternating currents, inductance and capacitance have also to be taken into account. In these elements the rate of change of current or voltage is important. The relation between alternating current and voltage in the various types of circuit elements will be discussed. For purposes of theory, these are regarded as ideal elements of one kind only. In practice it is impossible to make an inductor, for example, without some residual resistance, but this will be ignored in the discussion below.

Resistance. As in the dc case, the instantaneous voltage v across a resistance R and the current i through it are related by Ohm's law, Eq. (24). If v is given by Eq. (25), then i is given by Eq. (26). Current and voltage are in phase and can be represented in a phasor diagram as in **Fig. 6**. *See* ELECTRICAL RESISTANCE; RESISTOR.

$$v = iR \tag{24}$$

$$v = V_0 \cos \omega t \tag{25}$$

$$i = \frac{V_0}{R} \cos \omega t \tag{26}$$

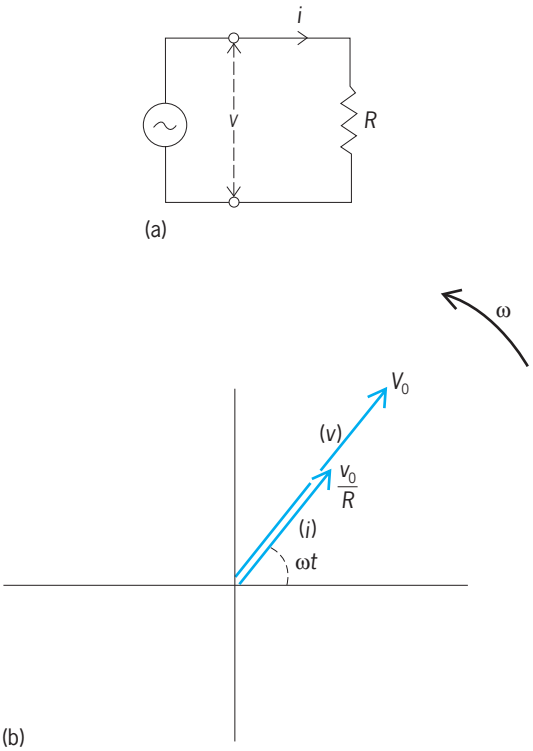


Fig. 6. Resistance. (*a*) Circuit diagram of resistor and source of alternating voltage. (*b*) Phasor diagram of voltage and current through a resistance.

Self-inductance. An inductor typically takes the form of a coil of wire, with or without a core of high-permeability magnetic material depending on the value of inductance required and the frequency of operation. The essential property of a single, or self-, inductance (so called to distinguish it from mutual inductance involving two coils, discussed below) is that the voltage across its terminals is related to the rate of change of current through it by Eq. (27). The

$$v = L\frac{di}{dt} \tag{27}$$

inductance L is measured in henrys when the voltage and current are in volts and amperes.

If the current is alternating and given by Eq. (28), the voltage is then given by Eq. (29). Thus the ratio

$$i = I_0 \cos \omega t \tag{28}$$

$$v = LI_0(-\omega \sin \omega t) = \omega L I_0 \cos\left(\omega t + \frac{\pi}{2}\right) \tag{29}$$

of the amplitudes of voltage and current is ωL and the voltage waveform is in advance of the current by a phase shift of $\pi/2$, or 90° (**Fig. 7**).

If the current is instead represented by a complex quantity as in Eq. (30), then Eq. (31) is obtained.

$$i = Ie^{j\omega t} \tag{30}$$

$$v = j\omega L I e^{j\omega t} \tag{31}$$

This represents the same result, since multiplying by j is equivalent to a positive (anticlockwise) rotation through a right angle. *See* INDUCTANCE; INDUCTOR.

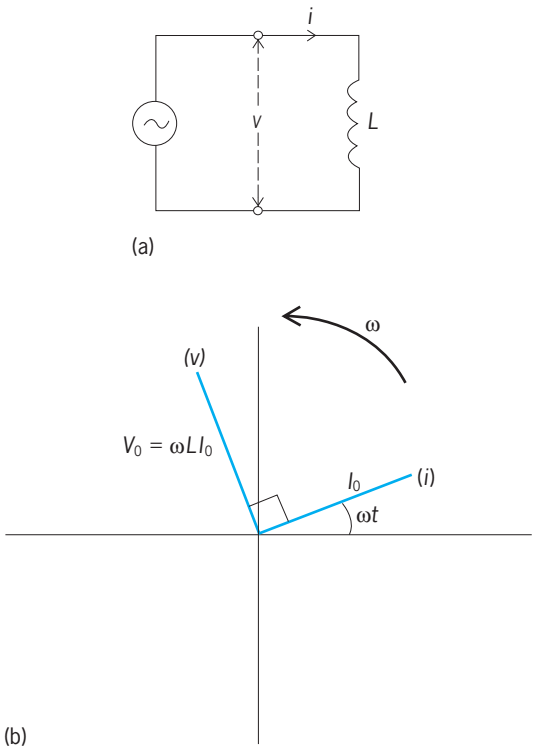


Fig. 7. Inductance. (*a*) Circuit diagram of inductor and source of alternating voltage. (*b*) Phasor diagram of voltage and current through an inductance.

Capacitance. An inductor stores energy in the magnetic field associated with an electric current. A capacitor stores energy in the electric field produced by the separation of charge in a dielectric medium. Typically a capacitor consists of two conducting plates separated by a dielectric. The current flowing in is related to the rate of change of voltage across the capacitor by Eq. (32), where C is the capacitance in farads.

$$i = C\frac{dv}{dt} \tag{32}$$

If the voltage is given by Eq. (33), then the current is given by Eq. (34). So in this case the ratio of voltage to current amplitudes is $1/\omega C$, and the voltage waveform lags behind the current by $\pi/2$ (**Fig. 8**).

$$v = V_0 \cos \omega t \tag{33}$$

$$i = CV_0(-\omega \sin \omega t) = \omega C V_0 \cos\left(\omega t + \frac{\pi}{2}\right) \tag{34}$$

In complex form, if the voltage is given by Eq. (35), the current is given by Eq. (36), representing the same result. *See* CAPACITANCE; CAPACITOR.

$$v = Ve^{j\omega t} \tag{35}$$

$$i = j\omega CVe^{j\omega t} \tag{36}$$

Mutual inductance. If two coils are linked by a common magnetic flux, the voltage v_2 across the second coil is related to the rate of change of current i_1 in the first coil by Eq. (37), where M is the mutual inductance, measured in henrys.

$$v_2 = M\frac{di_1}{dt} \tag{37}$$

If i_1 is given by Eq. (38), then v_2 is given by Eq. (39), and the voltage is 90° out of phase with the primary current.

$$i_1 = I_1 e^{j\omega t} \tag{38}$$

$$v_2 = j\omega L I_1 e^{j\omega t} \tag{39}$$

For further discussion of mutual inductance. *See* COUPLED CIRCUITS.

Circuit analysis. The main problem of conventional ac circuit theory is the calculation of the relationships between currents and voltages in circuits containing various combinations of the circuit elements described above.

In order to analyze any physical electric circuit, it is necessary to construct a mathematical model. Such a model is an idealization of the real circuit, but the behavior of the model may be made to approximate that of the real circuit sufficiently closely so that analysis of the model yields results which may be applied with confidence to the real circuit. In the vast majority of cases the circuits to be analyzed are linear. In a linear circuit the most important property is that if the amplitude of the source voltages or currents is changed, then the amplitudes of all other currents and voltages are changed in the same proportion. In the case of sinusoidal currents and voltages every current and voltage in the circuit has

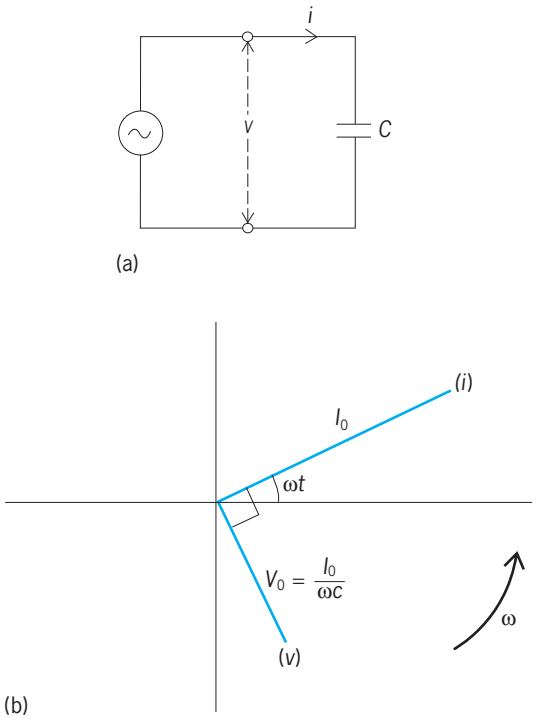


Fig. 8. Capacitance. (*a*) Circuit diagram of capacitor and source of alternating voltage. (*b*) Phasor diagram of voltage and current through a capacitance.

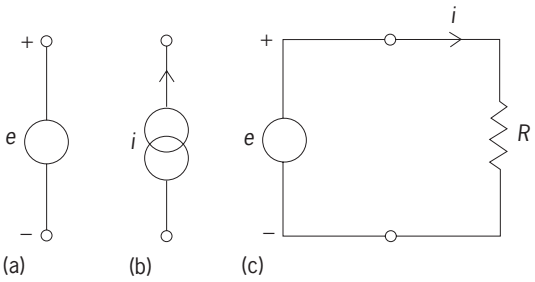


Fig. 9. Sources and signs. (*a*) Voltage source. (*b*) Current source. (*c*) Sign convention.

the same frequency, ω. In the rest of this article only linear circuits are discussed.

In an electric circuit the elements are connected together in some given fashion. Sources (or generators) of external current or voltage are applied to the circuit, and the problem is to calculate the resulting current and voltage associated with each circuit element. **Figure 9***a* and *b* illustrates the symbols used for voltage and current sources respectively. Although voltage and current change direction every half-cycle, it is necessary to have a sign convention to relate voltage and direction of current flow, compatible of course with that used in dc circuits. This is illustrated in Fig. 9*c* which is to be understood in the sense that, during the half-cycle in which the upper terminal of the source in the diagram carries the positive voltage, current flows through the resistor in the direction shown, from top to bottom of the diagram. In the next half-cycle both voltage and current reverse. The power dissipated in the resistor is vi, so the convention ensures that the power is positive in each half-cycle. *See* GENERATOR.

The analysis of any circuit is based on Kirchhoff's laws, which state that:

1. The sum of the instantaneous currents entering the junction of two or more circuit elements is zero.
2. The sum of the instantaneous voltages around any closed loop formed by two or more circuit elements is zero.

In both cases, proper account must be taken of the convention regarding positive and negative quantities. *See* KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS.

Analysis in the time domain. Consider the simple circuit shown in **Fig. 10**, which consists of a voltage source, a resistor, and an inductor connected in series. (Components are in series when the same current flows through each.) Kirchhoff's law indicates that the sum of the voltages around the loop is zero; that is, Eq. (40) holds. Now the voltage source e

$$v_R + v_L - e = 0 \tag{40}$$

is $E_0 \sin(\omega t + \theta)$, and the (unknown) current is assumed to be $I_0 \sin(\omega t + \beta)$, so that Eq. (41) is valid, and must be solved to find I_0 and β.

$$RI_0 \sin(\omega t + \beta) + L\omega I_0 \cos(\omega t + \beta) - E_0 \sin(\omega t + \theta) = 0 \tag{41}$$

When $\omega t + \theta = 0$, Eq. (41) reduces to Eq. (42) or Eq. (43).

$$R \sin(\beta - \theta) + L\omega \cos(\beta - \theta) = 0 \tag{42}$$

$$\tan(\beta - \theta) = -\frac{L\omega}{R} \tag{43}$$

When $\omega t + \theta = \pi/2$, Eq. (41) reduces to Eq. (44) or (45). Substituting from Eq. (43) for $\sin(\beta - \theta)$ and $\cos(\beta - \theta)$ yields Eqs. (46) or (47).

$$RI_0 \sin\left(\frac{\pi}{2} + \beta - \theta\right) + L\omega I_0 \cos\left(\frac{\pi}{2} + \beta - \theta\right) = E_0 \tag{44}$$

$$I_0[R \cos(\beta - \theta) - L\omega \sin(\beta - \theta)] = E_0 \tag{45}$$

$$I_0\left(\frac{R^2}{\sqrt{R^2 + L^2\omega^2}} + \frac{L^2\omega^2}{\sqrt{R^2 + L^2\omega^2}}\right) = E_0 \tag{46}$$

$$I_0 = \frac{E_0}{\sqrt{R^2 + L^2\omega^2}} \tag{47}$$

Hence the voltages across the resistor and the inductor are given by Eqs. (48) and (49). The phasor diagram for this circuit is given in Fig. 3*b*. The voltages across the resistor and the inductor are $\pi/2$

$$v_R = \frac{E_0 R}{\sqrt{R^2 + \omega^2 L^2}} \cdot \sin\left(\omega t + \theta - \tan^{-1}\frac{\omega L}{R}\right) \tag{48}$$

$$v_L = \frac{E_0 \omega L}{\sqrt{R^2 + \omega^2 L^2}} \cdot \sin\left(\omega t + \theta - \tan^{-1}\frac{\omega L}{R} + \frac{\pi}{2}\right) \tag{49}$$

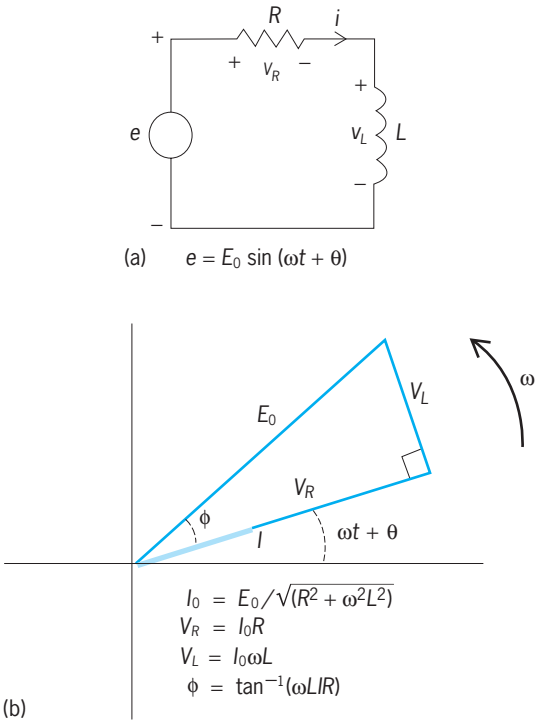


Fig. 10. Circuit with resistance and inductance. (*a*) Circuit diagram. (*b*) Phasor diagram.

radians out of phase, and the voltage across the circuit leads the current through it by the phase angle $\tan^{-1}(\omega L/R)$.

The analysis of any other circuit proceeds in the same way. The necessary equations are written by using Kirchhoff's laws and the relationships of Eqs. (24), (29), and (34) as appropriate. The solutions for the unknown currents or voltages are sinusoids with the same frequency as the applied voltages and currents but unknown amplitudes and phase angles. By evaluating the equations at particular values of time, as in the above examples, the unknown amplitudes and phase angles are then determined. The phase angles are found as differences relative to the phase angle of the applied generator. In the above example, $\beta - \theta$ is found, but not β as such. This is because one phase angle (here θ) must be taken as a reference angle, and the only important quantity is the value of other phase angles relative to this. Another way of looking at this is to note that the choice of origin on the time axis is quite arbitrary since all voltages and currents extend from $t = -\infty$ to $t = \infty$, which is an idealization of the real situation where the generator must start at some definite time. The choice of the reference phase angle corresponds to a particular choice of time origin.

While this method can, in principle, be used to analyze any circuit, it becomes extremely cumbersome for a circuit of even moderate complexity. Simplification of the analysis by the use of complex algebra will be discussed, but first the dissipation of power in an ac circuit must be considered.

Power. Consider the current and voltage associated with an arbitrary circuit (**Fig. 11**). The instantaneous power p is equal to vi. Let the voltage and current be given by Eqs. (50) and (51). Then the instantaneous

$$v = V_0 \sin(\omega t + \theta) \tag{50}$$

$$i = I_0 \sin(\omega t + \beta) \tag{51}$$

power is given by Eq. (52), which is the sum of a

$$p = V_0 I_0 \sin(\omega t + \theta) \sin(\omega t + \theta)$$

$$= {}^1/_2 V_0 I_0 [\cos(2\omega t + \theta + \beta) + \cos(\beta - \theta)] \tag{52}$$

sinusoid at radian frequency 2ω and a constant independent of t. The instantaneous power is generally of much less interest than the average power P. The average value of a sinusoid is zero so that the average power is given by Eq. (53), where $\hat{V} = V/\sqrt{2}$ and

$$P = {}^1/_2 V_0 I_0 \cos(\beta - \theta) = \hat{V}\hat{I} \cos(\beta - \theta) \tag{53}$$

$\hat{I} = I/\sqrt{2}$ are called the root-mean-square (rms) values of v and i respectively, and $\cos(\beta - \theta)$ is called the power factor of the circuit. Thus the power depends not only on the amplitudes of the current and voltage but also on the phase angle between them. This is illustrated by the cases of the resistor, given by Eqs. (54); the inductor, given by Eqs. (55); and the capacitor, given by Eqs. (56). For both the inductor and the capacitor the average power is zero.

$$\left.\begin{aligned} V_0 &= RI_0,\ \beta - \theta = 0 \\ P &= \hat{V}\hat{I} = R\hat{I}^2 = \hat{V}^2/R \end{aligned}\right\} \text{resistor} \tag{54}$$

$$\left.\begin{aligned} V_0 &= L\omega I_0,\ \beta - \theta = -\tfrac{\pi}{2} \\ P &= 0 \end{aligned}\right\} \text{inductor} \tag{55}$$

$$\left.\begin{aligned} I_0 &= C\omega V_0,\ \beta - \theta = \tfrac{\pi}{2} \\ P &= 0 \end{aligned}\right\} \text{capacitor} \tag{56}$$

It follows therefore that in an ac circuit, power is dissipated only in resistive elements, not in inductors or capacitors. The voltage across these latter elements is 90° out of phase with the current through them. The total voltage across a series circuit is made up of the vector sum of the phasors representing the individual element voltages. It is common to speak of in-phase components of the voltage (having the same phase as the current) and out-of-phase or quadrature components with a 90° phase difference. Because no power is dissipated in these components, the current through an inductor or capacitor is sometimes referred to as a wattless current.

In electric power systems the values quoted for current and voltage are usually the rms values since the power available is related directly to these quantities. However, it must be borne in mind that they must be multipled by 2 to obtain the actual amplitudes of current and voltage.

Analysis using complex impedances. The method of time domain analysis described above is quite general, and any circuit may be solved with those techniques. However, the method is cumbersome and does not give much insight into circuit behavior. Further, and perhaps more important, there are a wide variety of circuit theorems developed for use in dc circuit analysis which can both simplify analysis and provide considerable insight, but which cannot be used in ac circuits if one is restricted to the methods of time domain analysis. These considerations led to the development of a method of analysis based on complex notation which makes these techniques available in ac circuit analysis.

In an ac circuit every current and voltage is a sinusoid of the same frequency ω, and each is characterized by two quantities, namely the amplitude and the phase angle relative to some reference angle. Thus two independent qualities are needed to describe each current and voltage. As discussed above, an alternating current or voltage can be represented by

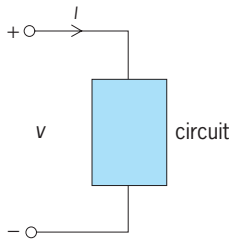


Fig. 11. Generalized circuit.

an expression of the form $Ie^{j\omega t}$ or $Ve^{j\omega t}$, and this representation can be applied to circuit analysis.

In the previous example of Fig. 3, Eq. (40) leads to Eq. (57) or (58), from which the modulus and argument of I are given by Eqs. (59) and (60). Thus if the voltage source is given by Eq. (61), then the current is given by Eq. (62), which is the result previously

$$RIe^{j\omega t} + jL\omega Ie^{j\omega t} - Ee^{j\omega t} = 0 \tag{57}$$

$$I = \frac{E}{R + jL\omega} \tag{58}$$

$$|I| = \frac{E}{\sqrt{R^2 + L^2\omega^2}} \tag{59}$$

$$\text{Argument } I = \text{ Argument } E - \tan^{-1}\frac{L\omega}{R} = \beta \tag{60}$$

$$e = E_0 \sin(\omega t + \theta) \tag{61}$$

$$i = |I| \sin\left(\omega t + \theta - \tan^{-1}\frac{L\omega}{R}\right) \tag{62}$$

found. However, now the analysis is very simple by using the complex quantities.

In the case of the resistor, inductor, and capacitor and in the example, equations are obtained where every term is multiplied by $e^{j\omega t}$, and this is true in general. Thus, this multiplier could just as well be omitted in every case and equations could be written involving only the complex quantities V and I. Once V and I are found, the answer for sinusoidal generators can be found immediately. Thus, for a resistor, $V = RI$; for an inductor, $V = j\omega LI$; and for a capacitor, $V = I/jC\omega$. For the complex voltages and currents corresponding to a frequency ω, an inductor behaves as if it were a resistor of value $j\omega L$, and a capacitor behaves as if it were a resistor of value $1/j\omega C$. The problem of analysis is thereby reduced to the analysis of dc circuits, except that all currents and voltages are complex numbers and some "resistors" are real while others are imaginary. *See* NETWORK THEORY.

When these complex currents and voltages are used, the ratio V/I in a circuit is called the (complex) impedance of the circuit (usually written Z; **Fig. 12**), and the ratio I/V is called the (complex) admittance of the circuit (usually written Y). The units of impedance are ohms, and of admittance are siemens (mhos). The impedance and admittance are expressed in terms of their real and imaginary parts by Eqs. (63) and (64) and are related to each other by Eq. (65) or (66). Similarly, Eq. (67) is valid.

$$\frac{V}{I} = Z = R + jX \tag{63}$$

$$\frac{I}{V} = Y = G + jB \tag{64}$$

$$Y = \frac{1}{Z} \tag{65}$$

$$G + jB = \frac{1}{R + jX} = \frac{R}{R^2 + X^2} - j\frac{X}{R^2 + X^2} \tag{66}$$

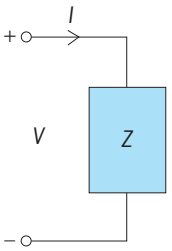


Fig. 12. General impedance.

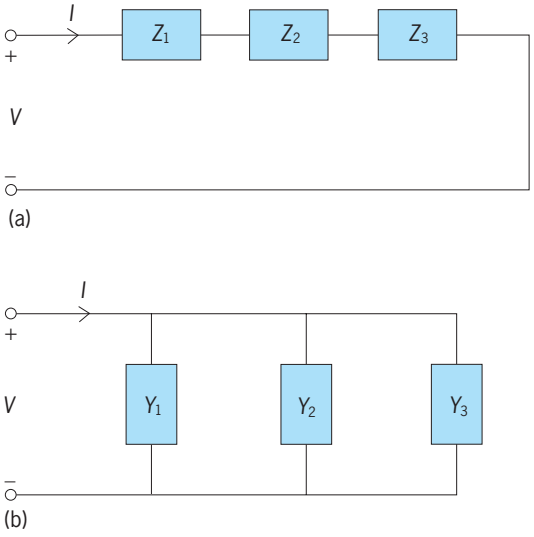


Fig. 13. Combination of impedances and admittances. (*a*) Impedances in series. (*b*) Admittances in parallel.

$$R + jX = \frac{1}{G + jB} = \frac{G}{G^2 + B^2} - j\frac{B}{G^2 + B^2} \tag{67}$$

Impedances, like resistances, can be combined in series and parallel. **Figure 13***a* shows three impedances in series. By definition the voltage across each impedance is ZI, and application of Kirchhoff's law to the loop yields Eq. (68) or (69), so the equivalent impedance Z_{eq} is the sum of the impedances in series.

$$Z_1I + Z_2I + Z_3I - V = 0 \tag{68}$$

$$\frac{V}{I} = Z_1 + Z_2 + Z_3 = Z_{eq} \tag{69}$$

By definition the current in each admittance is YV, and application of Kirchhoff's law yields Eq. (70) or Eq. (71). Thus the reciprocal of the equivalent impedance Z_{eq} is the sum of the admittances in parallel.

$$Y_1V + Y_2V + Y_3V - I = 0 \tag{70}$$

$$\frac{V}{I} = \frac{1}{Y_1 + Y_2 + Y_3} = Z_{eq} \tag{71}$$

The real and imaginary components of impedance and admittance are given special names. For impedance, $Z = R + jX$, the real part R is resistance

in the usual way; the imaginary part X (ωL for an inductor or $-1/\omega C$ for a capacitor) is reactance.

For admittance, $Y = G + jB$, the real part G is called conductance, and the imaginary part B, susceptance.

In a series circuit having resistance R and reactance X, the conductance is not simply the reciprocal of the resistance but, from Eq. (66), a quantity involving both the resistance and the reactance, namely $R/(R^2 + X^2)$.

Occasionally, it is useful to be able to refer to either impedance or admittance without specifying which in particular is meant. The general term immittance is used to cover both. For obvious reasons it is not possible to specify units for it. *See* ADMITTANCE; CONDUCTANCE; ELECTRICAL IMPEDANCE; IMMITTANCE; REACTANCE; SUSCEPTANCE.

Power dissipated. As discussed above, the power entering a circuit, such as Fig. 11, is given by Eq. (72),

$$P = {}^1/_2 V_0 I_0 \cos(\theta - \beta) \tag{72}$$

where V_0 and I_0 are the amplitudes of the voltage and current sinusoids, and $\theta - \beta$ is the angular phase difference between them. This power can also be expressed in terms of the complex quantities V and I.

If V and I are written as in Eqs. (73) and (74), then Eqs. (75) and (76) are valid. Thus, the power is given by Eq. (77).

$$V = V_0 e^{j\theta} \tag{73}$$

$$I = I_0 e^{j\beta} \tag{74}$$

$$\begin{aligned} VI^* &= V_0 I_0 e^{j(\theta-\beta)} \\ &= V_0 I_0 [\cos(\theta - \beta) + j \sin(\theta - \beta)] \end{aligned} \tag{75}$$

$$V^* I = V_0 I_0 [\cos(\theta - \beta) - j \sin(\theta - \beta)] \tag{76}$$

$$P = {}^1/_4 (VI^* + V^* I) = {}^1/_2 V_0 I_0 \cos(\theta - \beta) \tag{77}$$

If $V = ZI$, then the power is also given by Eq. (78), and if $I = YV$, then it is also given by Eq. (79).

$$P = {}^1/_4 (ZII^* + Z^* II^*) = {}^1/_4 I_0^2 (Z + Z^*) \tag{78}$$

$$P = {}^1/_4 (YVV^* + Y^* VV^*) = {}^1/_4 V_0^2 (Y + Y^*) \tag{79}$$

These equations give alternative ways of calculating the power from the voltage or current and circuit parameters.

Resonant circuit. Consider the circuit of **Fig. 14***a* with a voltage generator driving a capacitor in parallel with an inductor and resistance in series. The generator produces a voltage $e = E_0 \sin(\omega t + \theta)$. Suppose it is necessary to calculate the currents in the capacitor, resistor, and inductor, and the power dissipated in the circuit.

The circuit can be redrawn as in Fig. 14*b* Y_1 as the admittance of the capacitor and $Y_2 = 1/Z_2$, where Z_2 is the impedance of the resistor and inductor in series. Now Z_2 is given by Eq. (80), so Y_2 is given by Eq. (81) and Y_1 is given by Eq. (82). The equivalent impedance of the circuit, Z_{eq}, is given by Eq. (83), and

$$Z_2 = R + jL\omega \tag{80}$$

$$Y_2 = \frac{1}{R + jL\omega} = \frac{R}{R^2 + L^2\omega^2} - j\frac{L\omega}{R^2 + L^2\omega^2} \tag{81}$$

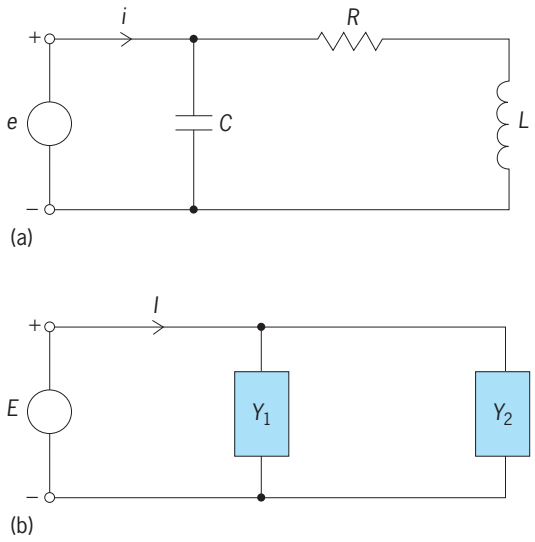


Fig. 14. Resonant circuit. (*a*) Diagram showing components. (*b*) Diagram showing admittances.

$$Y_1 = jC\omega \tag{82}$$

$$Z_{eq} = \frac{1}{Y_1 + Y_2} \tag{83}$$

the generator voltage is related to the total current by Eq. (84) or Eq. (85) so that the total current is given by Eq. (86).

$$E = Z_{eq} I \tag{84}$$

$$I = E/Z_{eq} \tag{85}$$

$$\begin{aligned} I &= E(Y_1 + Y_2) \\ &= E\left[\frac{R}{R^2 + L^2\omega^2} + j\left(C\omega - \frac{L\omega}{R^2 + L^2\omega^2}\right)\right] \end{aligned} \tag{86}$$

The current through the capacitor is $Y_1 E = j\omega CE$, while the current through the series combination of the resistor and inductor is $Y_2 E$, given by Eq. (87).

$$Y_2 E = E\left(\frac{R}{R^2 + L^2\omega^2} - j\frac{L\omega}{R^2 + L^2\omega^2}\right) \tag{87}$$

The power dissipated in the circuit is found, by using Eq. (79), to be given by Eq. (88).

$$\begin{aligned} P &= {}^1/_4 E_0^2 (Y_{eq} + Y_{eq}^*) \\ &= {}^1/_4 E_0^2 (Y_1 + Y_2 + Y_1^* + Y_2^*) \\ &= {}^1/_2 E_0^2 \frac{R}{R^2 + L^2\omega^2} \end{aligned} \tag{88}$$

This circuit exhibits the phenomenon of resonance. Suppose the frequency ω is varied. Then in Eq. (86) the imaginary part becomes zero when Eq. (89) is satisfied, which gives Eq. (90) for ω^2. This

$$C\omega = \frac{L\omega}{R^2 + L^2\omega^2} \tag{89}$$

$$\omega^2 = \frac{1}{LC} - \left(\frac{R}{L}\right)^2 = \omega_0^2 \tag{90}$$

is called the resonant frequency, and at this value of ω the power dissipated is given by Eq. (91). The quantity $L\omega_0/R$ is called the quality factor or Q-factor of the circuit and is given by Eq. (92). Since Q is often very large, it is usually approximated by Eq. (93).

$$P_0 = {}^1/_2 E_0^2 \frac{RC}{L} \tag{91}$$

$$Q = \sqrt{\frac{L}{R^2 C} - 1} \tag{92}$$

$$Q = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{93}$$

Having found these various complex voltages and currents, the corresponding sinusoidal quantities could be written immediately, but this is rarely done in practice since all the information is already contained in the complex quantities. *See* RESONANCE (ALTERNATING-CURRENT CIRCUITS).

Practical circuit elements. Alternating-current circuit theory is, of course, applicable at higher, radio frequencies so long as the circuit elements can be considered as "lumped," that is, so long as their dimensions are small compared with a wavelength of the current and its phase is the same at all points in the component. But particularly at the higher frequencies, the idealized representations of circuit elements which were given earlier need to be replaced by representations which more accurately reflect the properties of actual components.

An inductor, for example, in the form of a wire-wound coil, will have resistance as well as inductance, so it should be shown as a resistance and inductance in series, with impedance $R + j\omega L$. For a well-designed coil, R is much smaller than ωL at the frequency of use. There will also be capacitance between the turns of the coil which becomes important at the higher frequencies. As discussed above, this is a resonant circuit, and there is for any coil a frequency at which it is self-resonant; above this frequency its effective impedance is capacitative rather than inductive. *See* CORE LOSS; SKIN EFFECT (ELECTRICITY).

A capacitor will generally suffer some power loss in the dielectric medium, which can be represented by a parallel high resistance, and the connecting leads will have low resistance and inductance which may be significant at very high frequencies. *See* PERMITTIVITY.

Other stray impedances may also have to be taken into account. At high frequencies it is particularly important to remember that even lengths of connecting wire may have significant inductance and capacitance to ground.

Microwave circuits. At microwave frequencies the assumption of lumped components whose size is small compared with the wavelength breaks down. The basic principles of ac circuit theory can still be applied, but it is necessary to assume that the properties of resistance, inductance, and capacitance are no longer localized but are distributed throughout the circuits, and to take account of the finite time of travel of a wave from one part of the circuit to another. The concept of impedance then has to be generalized. *See* MICROWAVE; TRANSMISSION LINES.

Active circuits. In addition to the passive circuits discussed in this article, circuit theory can be extended to cover the cases of amplification and feedback. Again the concept of impedance may have to be extended: it is possible, for example, to produce circuits whose effective impedance is that of a negative capacitor, or which do not correspond to any possible combination of simple circuit elements. *See* AMPLIFIER; CONTROL SYSTEMS; FEEDBACK CIRCUIT.

Nonsinusoidal waveforms. The above analysis of circuits has been entirely in terms of the response to single-frequency sinusoidal waveforms. There are many cases where it is necessary to deal with other waveforms. Two methods are available:

1. When it is necessary to find the response to a repetitive waveform or perhaps a single pulse whose Fourier transform can easily be calculated as $F(\omega)$, the impedance or transfer function (circuit property relating input and output voltage or current) can be calculated in the usual way as $Z(\omega)$, and then the inverse transform of $F(\omega) \cdot Z(\omega)$ can be extracted to give the response waveform as a function of time. *See* FOURIER SERIES AND TRANSFORMS.

2. An alternative method, particularly useful in dealing with transient waveforms, is (for example) to take the Laplace transform of the input waveform, $L(s)$, and to multiply it by $Z'(s)$, given by the impedance or transfer function calculated in the usual way but with $j\omega$ replaced by s. The output is then the inverse Laplace transform of $L(s) \cdot Z'(s)$. *See* LAPLACE TRANSFORM.

J. O. Scanlan; A. E. Bailey

Bibliography. C. K. Alexander and M. N. O. Sadiku, *Fundamentals of Electric Circuits*, 3d ed., McGraw-Hill, 2007; J. W. Nilsson and S. A. Riedel, *Electric Circuits w/PSpice*, 7th ed., Prentice Hall, 2005.

Alternating-current generator

A machine that converts mechanical power into alternating-current electric power. Almost all electric power is produced by alternating-current (ac) generators that are driven by rotating prime movers. Most of the prime movers are steam turbines whose thermal energy comes from either fossil or nuclear fuel. Combustion turbines are often used for the smaller units and in cases where gas or oil is the available fuel. Where water power is available from dams, hydroelectric ac generators are powered by hydraulic turbines. Small sites may also use diesel or gasoline engines to drive the generator, but these units are usually used only for standby generation or to provide electric power in remote areas. *See* DIESEL ENGINE; GAS TURBINE; HYDRAULIC TURBINE; HYDROELECTRIC GENERATOR; INTERNAL COMBUSTION ENGINE; STEAM ELECTRIC GENERATOR; STEAM TURBINE.

Alternating-current generators are used because ac power can easily be stepped up in voltage, by using transformers, for more efficient transmission of power over long distances and in larger amounts.

Similarly, transformers step the voltage down again at the utilization site to safer and more convenient levels. *See* DIRECT-CURRENT GENERATOR; ELECTRIC POWER SYSTEMS; TRANSFORMER.

Principles of operation. Most ac generators are synchronous machines, that is, the rotor is driven at a speed that is exactly related to the rated frequency of the ac network. Generators of this type have a stationary armature with three windings that are displaced at regular intervals around the machine to produce three-phase voltages. These machines also have a field winding that is attached to the rotor. This winding provides magnetic flux that crosses the air gap and links the stator coils to produce a voltage according to Faraday's law. The field winding is supplied with direct current, usually through slip rings. *See* ARMATURE; SLIP RINGS; WINDINGS IN ELECTRIC MACHINERY.

To understand the action of an ac generator, it is helpful to visualize a rotating magnetic flux-density wave in the air gap of the machine (**Fig. 1**). This wave links the stator winding, causing each coil to experience an alternating flux. This is the mechanism for inducing an alternating voltage.

It is usually assumed that the flux density in the air gap has a sinusoidal distribution, which may be written as Eq. (1), where θ is the angular position mea-

$$B = B_{\max} \cos \frac{p\theta}{2} = B_{\max} \cos \theta_e \tag{1}$$

sured around the air gap of the machine in a certain direction. The angle θ_e is defined by Eq. (2), where p

$$\theta_e = \frac{p}{2}\theta \quad \text{electrical radians} \tag{2}$$

is the number of poles. This angle is defined so that 360 electrical degrees corresponds to an angle that passes through flux defining a north pole and then a south pole. *See* ELECTRICAL DEGREE.

The total flux linking the coil is given by Eq. (3), where the differential area is given by Eq. (4), and

$$\phi_c = \int \int B dA \tag{3}$$

$$dA = LrD\theta = \frac{2Lr}{p} d\theta_e \tag{4}$$

L is the coil length, r is the radius of the air gap in the cylindrical geometry of the machine, and θ is the angle defined above. The generator shaft rotates at synchronous speed with velocity given by Eq. (5), where ω_e, defined by Eq. (6), is the speed in electri-

$$\omega_s = \frac{2\pi f}{p/2} = \frac{2}{p}\omega_e \text{ radians per second} \tag{5}$$

$$\omega_e = 2\pi f \tag{6}$$

cal radians per second and f is the system frequency.

The flux density may be regarded as a traveling wave, given by Eq. (7). Substituting the expressions

$$B(\theta, t) = B_{\max} \cos \frac{p}{2}(\theta - \omega_s t)$$

$$= B_{\max} \cos (\theta_e - \omega_e t) \tag{7}$$

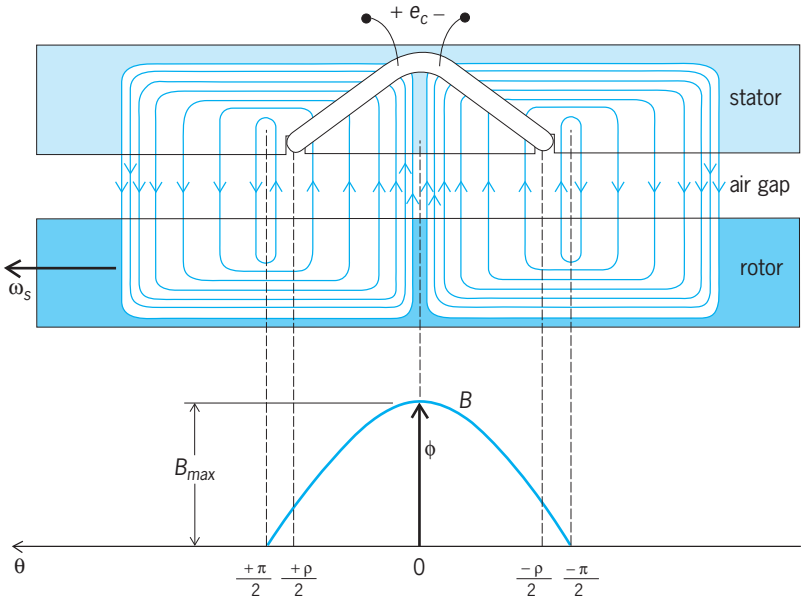


Fig. 1. End view of one coil of an alternating-current generator, linked by the moving airgap flux.

in Eqs. (4) and (7) for dA and B into the integrand of Eq. (3) and evaluating between the limits $\pm\rho/2$ yields Eq. (8). Then by using the identity of Eq. (9),

$$\phi_c = \int_{-\rho/2}^{\rho/2} \frac{2Lr}{p} [B_{\max} \cos (\theta_e - \omega_e t)] \, d\theta_e \tag{8}$$

$$\cos (x - y) = \cos x \cos y - \sin x \sin y \tag{9}$$

the integration is readily performed. The flux for one coil is found to be given by Eq. (10), where k_p and ϕ_p are defined by Eqs. (11) and (12).

$$\phi_c = k_p \phi_p \cos \omega_e t \tag{10}$$

$$k_p = \text{ pitch factor } = \sin \frac{\rho}{2} \tag{11}$$

$$\phi_p = \text{flux per pole} = \frac{4B_{\max} Lr}{p} \tag{12}$$

The induced electromotive force (emf) for the coil is computed from Faraday's law, which states that the emf is equal to the rate of change of flux linkages; that is, Eq. (13) is valid where N_c is the number of turns in

$$e_c = -\frac{d\lambda_c}{dt} = -\frac{d(N_c \phi_c)}{dt} = \omega_e N_c k_p \phi_p \sin \omega_e t \tag{13}$$

the coil. It is convenient to write the coil voltage as in Eq. (14), where E_c is the root-mean-square (rms)

$$e_c = \sqrt{2} E_c \sin \omega_e t \tag{14}$$

value of the coil voltage. The total pitch of the coil ($\pi - \rho$) is less than one pole pitch (π). This difference has the effect of reducing harmonics more than it reduces the fundamental component of voltage. This reduction is expressed in terms of the pitch factor. *See* ELECTROMAGNETIC INDUCTION.

The voltage e_c is the induced voltage in only one coil. (Fig. 1). Usually, the voltages of a group of coils are added together by connecting the coil ends in

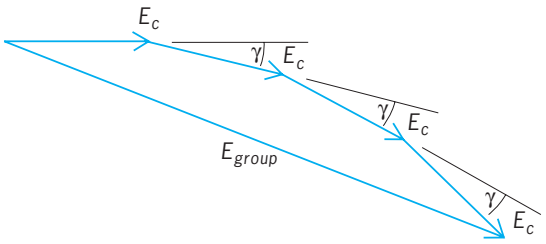


Fig. 2. Phasor diagram for the total root-mean-square emf of a group of coils, E_{group}.

series. The coils in the group are not all in the same slots, however, but are displaced by the slot pitch γ. Therefore, the voltage induced in the individual coils will be out of phase by this angle. This means that the addition of the voltages is not a simple arithmetic addition, but is usually performed as a phasor addition to compute the total rms emf of the group of coils, E_{group}. This addition is shown in **Fig. 2**, where the number of coils, n, in the group is assumed to be four. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

From the geometry of Fig. 2, Eq. (15) may be computed, where a new constant k_d is defined for convenience. Finally, the total phase voltage is composed of p groups in series and is thus given by Eq. (16),

$$E_{group} = nE_c \frac{\sin \frac{n\gamma}{2}}{n \sin \frac{\gamma}{2}} = nE_c k_d \tag{15}$$

$$E_{phase} = pE_{group} \tag{16}$$

where p is the number of poles. For steam-turbine-driven generators p is usually 2 or 4. Hydroelectric generators may have a much larger number of poles, depending on the shaft speed.

Three-phase generators produce three-phase voltages that are equal in magnitude but displaced in phase by 120 electrical degrees. The phasor diagram for the voltages of a three-phase ac generator is shown in **Fig. 3**.

Armature reaction. Armature reaction is a demagnetizing effect that limits the output of an ac generator. The magnetomotive forces (mmf's) of the coils combine to produce a rotating mmf that opposes the field flux. Thus, more excitation is required to maintain the flux magnitude as the armature current is increased. The mmf of one coil per pole is a square wave that has a sine-wave fundamental component with a maximum value given by Eq. (17), where I_c is the rms armature current in the coil. The mmf factor for all coils in one phase is $F_{max}k_d$.

$$F_{max} = \sqrt{2}\frac{4}{\pi} N_c I_c k_p \tag{17}$$

To compute the resultant rotating mmf for a polyphase winding, the single alternating wave is divided into two equal components of one-half value each, rotating in opposite directions. For balanced phase currents, the forward components combine and the reverse components cancel. Thus, for three-phase windings, the resultant rotating wave has a maximum of $^3/_2$ the single phase value and is usually referred to as F_1, where the subscript 1 indicates the fundamental component of the sine wave. *See* ARMATURE REACTION.

Characteristics. Certain constants define the operation of the ac generator on the system to which it is connected. One important set of machine constants is the reactances of the windings. There are several different reactances that are derived to describe the machine behavior under different conditions. *See* REACTANCE.

The synchronous reactance x_d is the unsaturated steady-state reactance, including leakage and armature reaction, and is used to describe steady-state generator performance. The transient reactance x'_d is useful in determining how the machine responds to sudden changes, after the first few cycles. This reactance is often used to compute the fault current for circuit-breaker opening. The subtransient reactance x''_d determines the initial fault current, which is a current that the machine windings and connecting bus-work must be able to withstand mechanically.

The generator operates in synchronism with the system to which it is connected. As load is increased, the shaft angle moves ahead by the angle δ. The power delivered is given by Eq. (18), where the subscript s stands for system quantities and the subscript g stands for generator quantities. The voltage E_g for steady-state conditions is the rated excitation voltage, and x_g is the synchronous reactance.

$$P = \frac{E_s E_g}{x_s + x_d} \sin \delta \tag{18}$$

To calculate system swings due to sudden changes, x_d can be taken as the transient reactance and E_d as the effective voltage behind that reactance. This

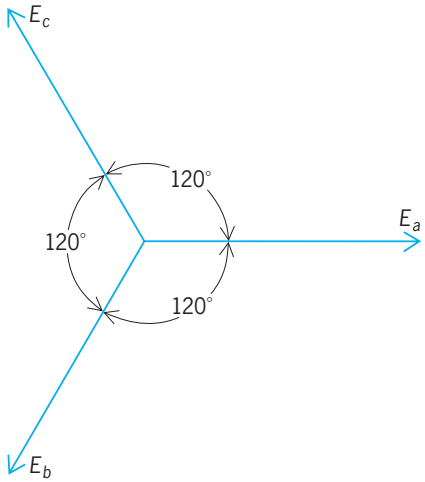


Fig. 3. Phasor diagram of the balanced three-phase generated voltages, E_a, E_b, and E_c, for a synchronous alternating-current generator.

voltage changes only slowly but can be corrected for the action of the excitation system, which affects the control of the voltage and plays an important role in the dynamic performance of the machine.

The calculation of dynamic swings during a short circuit can be performed step by step or by computer simulations, the latter method being the usual choice for systems containing more than one machine. The computation gives the change in power output of the machine and the resulting acceleration and angle, taking into account the moment of inertia of the turbine-generator combination.

Growth in size. The size of ac generators increased rapidly up to the mid-1970s. This increase resulted in considerable economy in terms of the cost per megawatt of machine rating and a concurrent improvement in turbine efficiency. The maximum ratings doubled about every 10 years for many years, which is about the same rate at which the total electrical load in North America was increasing. The largest units now manufactured are rated at about 1800 megavolt-amperes. This rapid growth in unit size was a major challenge to machine designers. Better rotor materials were required to withstand higher centrifugal forces associated with larger rotor diameters. Major changes in the method of cooling were required, resulting first in hydrogen cooling and later in water-cooled stator windings. These refinements were necessary to keep the efficiency of the generator at about 99%. *See* VOLT-AMPERE.

Other generator types. The foregoing discussion has dealt entirely with the class of generators called synchronous machines. These are by far the most common, but are not the only type of ac generators.

Induction generators have been used in remote applications where maintenance of the excitation system is a problem. These units are essentially like induction motors, but are driven by a prime mover at speeds slightly above synchronous speed, forcing the unit to generate power due to the reverse slip. Induction generator units draw reactive power from the system and are not as efficient as synchronous generators. *See* INDUCTION MOTOR.

High-frequency single-phase generators have been built as induction alternators, usually with twice as many stator poles (teeth) as rotor poles, and with a constant air-gap flux supplied from a homopolar field coil in the center of the machine, pushing flux into the stator at one end and out at the other. Their effectiveness is lower than that of the synchronous machine because the flux is a pulsating unidirectional field, rather than an alternating field. *See* ALTERNATING CURRENT; ELECTRIC ROTATING MACHINERY; GENERATOR.

Lee A. Kilgore; Paul M. Anderson

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 6th ed., 2003; G. McPherson and R. D. Laramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., 1990.

Alternating-current motor

An electrical machine that converts alternating-current (ac) electric energy to mechanical energy. Alternating-current motors are widely used because of the general availability of ac electric power and because they can be readily built with a variety of characteristics and in a large range of sizes, from a few watts to many thousands of kilowatts. They can be broadly classified into three groups—induction motors, synchronous motors, and ac series motors:

Induction motor
- Single-phase
  - Split-phase
  - Capacitor-start
  - Capacitor-run
- Polyphase

Synchronous motor
- Single-phase
  - Permanent-magnet (PM)
  - Reluctance
  - Hysteresis
- Polyphase
  - Wound-field
  - Permanent-magnet (PM)
  - Reluctance

AC series or universal motor (single-phase)

See ALTERNATING CURRENT; DIRECT-CURRENT MOTOR.

Induction motor. The most common type of ac motor, both in total number and in total power, is the induction motor. In larger sizes these machines employ a polyphase stator winding, which creates a rotating magnetic field when supplied with polyphase ac power. The speed of rotation depends upon the frequency of the supply and the number of magnetic poles created by the winding; thus, only a discrete number of speeds are possible with a fixed frequency supply.

Currents are induced in the closed coils of the rotor for any rotor speed different from the speed of the rotating field. The difference in speed is called the slip speed, and efficient energy conversion occurs only when the slip speed is small. These machines are, therefore, nearly constant-speed machines when operated from a constant-frequency supply. They are, however, routinely started from zero speed and accelerated through the inefficient high-slip-speed region to reach operating speed. *See* SLIP (ELECTRICITY).

Single-phase induction machines are generally inferior in conversion efficiency and have zero torque at stand-still (zero speed). Auxiliary starting means usually involving a second winding and some technique for creating a time phase difference in the second winding current must be employed. Capacitors are commonly used to create a large phase shift, although extra resistance in the starting winding (in the split-phase motor) can be used if reduced starting

torque is acceptable. Typically, the second winding is switched off by a rotation-speed-actuated switch, but in some low-starting-torque applications the second winding with a series capacitor is left connected to improve the running condition (in the capacitor-run motor). Single-phase machines are widely used in situations where only single-phase ac power is available, especially for appliances in homes and small commercial installations. *See* INDUCTION MOTOR.

Synchronous motor. In contrast to an induction motor, the rotor of a synchronous motor runs exactly at the rotating field speed and there are no induced rotor currents. Torque is produced by the interaction of the rotating field with a direct-current (dc) field created by injected dc rotor current or permanent magnets, or with a rotor magnetic structure that has an easy direction for magnetization (in the reluctance motor). Since for any frequency of excitation there is only one speed for synchronous torque, synchronous machines have no starting torque unless the frequency is variable. When the motor is used in fixed-frequency applications, an induction-machine winding is also placed on the rotor to allow starting as an induction motor and running as a synchronous motor. *See* SYNCHRONOUS MOTOR.

Large synchronous motors for industrial applications have dc windings on the rotor and always have polyphase windings on the stator. These machines characteristically have high efficiency and a power factor that is adjustable via the rotor dc field current. Both leading and lagging power factors are possible. Smaller polyphase synchronous motors using permanent magnets on the rotor are widely used in self-synchronous, variable-frequency drive systems (brushless dc motors and electronically commutated motors). In these drives the rotor position is detected and used to determine the frequency and phase of the excitation. The advantages include high efficiency, fast response, and very good control characteristics.

In smaller sizes, synchronous reluctance motors are used where precise speed control is required, such as in timing motors or where many small motors must be operated at exactly the same speed. Reluctance machines generally have lower efficiency and power factor but are inexpensive. Improvements in design may make the reluctance machine competitive with induction machines in variable-speed applications using variable-frequency excitation. *See* RELUCTANCE MOTOR.

For special, low-power applications requiring precise speed control and very smooth torque, the hysteresis motor is used. These machines use a special permanent-magnet rotor material which is remagnetized each time the motor is started. Starting torque results from the lag between magnetizing force and flux density and is essentially independent of rotor speed. At synchronism the hysteresis motor runs as a permanent-magnet synchronous motor. *See* HYSTERESIS MOTOR.

AC series motor. A dc motor with the armature and field windings in series will run on ac since both magnetic fields reverse when the current reverses. Since these machines run on ac or dc, they are commonly called universal motors. The speed can be controlled by varying the voltage, and these machines are therefore widely used in small sizes for domestic appliances that require speed control or higher speeds than can be attained with 60-Hz induction motors. *See* ELECTRIC ROTATING MACHINERY; MOTOR; UNIVERSAL MOTOR; WINDINGS IN ELECTRIC MACHINERY.

Donald W. Novotny

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 6th ed., 2003; G. R. Slemon, *Electric Machines and Drives*, 1992.

Alternative fuels for vehicles

Conventional fuels such as gasoline and diesel are gradually being replaced by alternative fuels such as gaseous fuels (natural gas and propane), alcohol (methanol and ethanol), and hydrogen. Conventional fuels can also be modified to a reformulated gasoline to help reduce toxic emissions. Technological advances in the automotive industry (such as in fuel cells and hybrid-powered vehicles) are helping to increase the demand for alternative fuels.

Two key issues associated with the use of conventional fuels for vehicles prompt interest in alternative fuels: (1) gasoline- and diesel-powered vehicles are considered to be a significant source of air pollution; and (2) conventional fuels are produced from crude oil, a nonrenewable energy resource. The success of alternative fuels in the marketplace will depend on numerous factors, including public understanding and consumer awareness; economics; automobile performance; availability of fuels, vehicles, and distribution and marketing systems; and changes in technology. *See* AIR POLLUTION.

Gaseous fuels. Vehicle emissions from natural gas and propane are expected to be lower and less harmful to the environment than those of conventional gasoline. Because natural gas and propane are less complex hydrocarbons, the levels of volatile organic compounds and ozone emissions should be reduced. Both of these fuels are introduced to the engine as a gas under most operating conditions and require minimal fuel enrichment during warm-up. Leaner burning fuels, they also achieve lower carbon dioxide and carbon monoxide levels than gasoline. However, because they burn at higher temperatures, emissions of nitrogen oxide are higher. An important property of gaseous fuels is their degree of resistance to engine knock. Because of their higher-octane value relative to gasoline, there is less of a tendency for these fuels to knock in spark-ignition engines. To achieve the optimal performance and maximum environmental benefits of natural gas and propane, technological advancements must continue to reduce the costs of dedicated vehicles to be competitive with conventional vehicles, and the necessary fueling infrastructure must be ensured.

Natural gas. Natural gas is a colorless, odorless hydrocarbon that is neither carcinogenic nor corrosive.

Once processed at a gas plant, natural gas is composed of about 97% methane (CH_4). Being lighter than air, it does not pool on the ground when leaked but dissipates into the atmosphere, reducing the hazard of fire. One drawback as a transportation fuel is its very low energy density relative to gasoline or diesel fuel. To eliminate bulkiness, the gas is compressed and stored in cylindrical tanks. At a pressure of 3000 pounds per square inch (20 megapascals), the volumetric energy density of a typical natural gas cylinder is only 20% that of a similar volume of gasoline. Thus, the driving range of a natural gas vehicle carrying the same volume of fuel as a gasoline-fueled vehicle will be less. Larger vehicles, carrying two or three natural gas tanks, can achieve an adequate driving range for most personal or fleet purposes, but at the cost of space and the additional weight of the cylinders. *See* METHANE.

Natural gas engine technology is fairly well developed, including heavy-duty applications. Vehicle refueling appliances have reduced the constraints associated with the lack of distribution and refueling infrastructure; however, the associated costs in addition to those related to conversion tend to limit natural gas usage to high-mileage vehicles. Significant advances have been made with regard to lighter and stronger fuel cylinders, and there is more variety of natural gas vehicles produced by original equipment manufacturers than for any other alternative fuel. Still required for market competitiveness of natural gas is a refueling infrastructure with widespread access and low-cost fast-fill capabilities. *See* LIQUEFIED NATURAL GAS (LNG); NATURAL GAS.

Propane. Propane (C_3H_8), a hydrocarbon produced from crude oil and natural gas, has been one of the more successful alternative transportation fuels. A cleaner-burning fuel than gasoline or diesel fuel, it can result in lower maintenance costs, requiring fewer spark plug and oil changes, and less wear on such components as piston rings and bearings. Since propane has a higher octane rating than gasoline, propane-fueled vehicles can use higher engine compression ratios, resulting in more power and better fuel efficiency.

Propane has a lower energy density than gasoline by volume but a higher energy density by weight; therefore, the weight of a tank full of propane is similar to that of one filled with gasoline. Compared to ethanol, methanol, or natural gas, propane has a higher energy content by weight and volume, so it can take a vehicle farther on an average-capacity tank than a similar tank of any of the other alternative fuels. Refueling time for propane is similar to conventional fuels.

Several factors related to customer convenience and conversion costs have been restrictive to automotive propane market development. The physical characteristics of propane require modifications to conventional vehicles by either the automotive manufacturer at the factory or after-market conversions; either way, the cost of the modifications adds to the price paid by the consumer for a gasoline or diesel vehicle. In order to maintain a liquid state, propane must be stored under pressures of up to 200 psi (1.3 MPa). Because of this fact and the lower energy density of propane relative to gasoline, propane-powered vehicles must be fitted with storage tanks that impose weight and space penalties in order for the vehicles to travel the same distance as gasoline counterparts. Also, the vapor pressure of propane varies in relation to the atmospheric temperature. At temperatures of $-43^\circ C$ ($-45^\circ F$) and colder, the pressure in a fuel tank is essentially zero, the product becomes liquid, and propane fuel systems simply do not operate. Another inconvenience arises due to the potential hazard of vapor accumulation at ground level; for this reason, some city and municipality bylaws forbid propane vehicles from parking in underground facilities.

On the basis of a grams per unit of distance traveled, propane full fuel cycle emissions compare favorably with gasoline, except for nitrogen oxides. With further development of propane vehicles by original equipment manufacturers, the establishment of certification standards for propane conversions, and continued technological improvements in emission control systems, the environmental impact of propane could prove even greater. The potential economic benefits from conversion to propane are greatest in high-fuel-consumption fleets such as taxis, where payback occurs in 1–2 years.

Alcohol fuels. The most significant advantage of alcohol fuels over gasoline is their potential to reduce ozone concentrations and to lower levels of carbon monoxide. Another important advantage is their very low emissions of particulates in diesel engine applications. In comparison with hydrocarbon-based fuels, the exhaust emissions from vehicles burning low-level alcohol blends (such as gasohol containing 10% alcohol by volume) contain negligible amounts of aromatics and reduced levels of hydrocarbons and carbon monoxide but higher nitrogen oxide content.

Exposure to aldehydes, in particular formaldehyde which is considered carcinogenic, is an important air-pollution concern. The aldehyde fraction of unburned fuel, particularly for methanol, is appreciably greater than for hydrocarbon-based fuels; therefore, catalytic converters are required on methanol vehicles to reduce the level of formaldehyde to those associated with gasoline. *See* ALCOHOL FUEL.

Methanol. Methanol (CH_3OH) is a clear, colorless, high-performance liquid fuel that can be used in both spark-ignition and diesel engines. It can be burned as a neat (100% methanol) or near-neat fuel in internal combustion engines. A blend of 85% methanol and 15% gasoline, M85, is the most common form of methanol fuel used in light-duty vehicles. M100, or neat methanol, is used in some heavy-duty trucks and buses.

The appeal of methanol as an alternative to gasoline stems from its being a relatively inexpensive clean-burning liquid fuel. A constraint is the fact that it has only half the energy density of gasoline, compensated to some degree by its better thermal efficiency. Safety concerns relate to the fact that

methanol is toxic and can be absorbed through the skin. Also, neat methanol burns with an invisible flame, making methanol fires without a colorant difficult to detect.

Besides being used as a neat or near-neat transportation fuel, methanol has been used to produce methyl tertiary butyl ether (MTBE), an oxygenate which, if blended with gasoline up to 10%, adds one octane number to the fuel. Concerns related to MTBE leaking and contaminating ground water in some areas in the United States resulted in a reexamination of the use of this oxygenate, and the phaseout of its use in California by December 31, 2002. Methanol is also being tested in the transportation industry as a source of the hydrogen in hybrid fuel cell vehicles. *See* ETHER; METHANOL.

Ethanol. Ethanol (C_2H_5OH) produced from biomass is a renewable fuel source currently being marketed in neat, near-neat, and low-level premium fuel blends. In addition to providing environmental benefits, use of ethanol fuel produced from fermentable starch or sugar crops provides economic advantages to the agriculture sector. However, ethanol production costs are very high relative to gasoline and must be reduced substantially if ethanol is to compete on an economic basis.

To date, in the fuel industry ethanol has been used mainly as an additive in low-level gasohol blends. These blends have been successfully used in unmodified gasoline vehicles with warranty coverage provided for automobiles sold in North America. Neat ethanol, generally used in heavy-duty applications, is less common and requires costly engine modifications. Flexible fuel vehicles can operate on a mixture of gasoline and up to 85% ethanol (E85). Ethanol has also been used in the production of ethyl tertiary butyl ether (ETBE), but to date it has not been economically competitive with MTBE.

The oxygen content of ethanol fuels used in properly modified vehicles results in increased energy efficiency and engine performance. Extended spark-plug life and lower carbon deposits are also expected because ethanol burns cleaner than gasoline.

A significant drawback of ethanol is its much lower energy density than gasoline (approximately 34% less). To travel distances similar to a gasoline-powered vehicle without refueling would require a much bigger fuel tank; this requirement is slightly offset by the greater energy efficiency of ethanol. *See* ETHYL ALCOHOL.

Hydrogen. Hydrogen is the lightest and most abundant element in the universe. It can be produced from a number of feedstocks in a variety of ways. The production method thought to be most environmentally benign is the electrolysis of water, but probably the most common source of hydrogen is the steam reforming of natural gas. Once produced, hydrogen can be stored as a gas, liquid, or solid and distributed as required. Liquid storage is currently the preferred method, but it is very costly. Metal hydride and compressed storage are also being investigated. *See* METAL HYDRIDES.

Hydrogen-powered vehicles can use internal combustion engines or fuel cells. They can also be hybrid vehicles of various combinations. When hydrogen is used as a gaseous fuel in an internal combustion engine, its very low energy density compared to liquid fuels is a major drawback requiring greater storage space for the vehicle to travel a similar distance to gasoline. Although hybrid vehicles can be more efficient than conventional vehicles and result in lower emissions, the greatest potential to alleviate air-pollution problems is thought to be in the use of hydrogen-powered fuel cell vehicles. Though currently very expensive, fuel cells are more efficient than conventional internal combustion engines. They can operate with a variety of fuels, but the fuel of choice is gaseous hydrogen since it optimizes fuel cell performance and does not require on-board modification.

Numerous hydrogen fuel cell vehicle models are being developed and tested with the objective of having such vehicles available for sale to the public in the early years of the twenty-first century. However, for gaseous hydrogen to become a fuel of choice in the transportation industry, numerous technical and economic constraints related to hydrogen production, on-board and site storage, and marketing and distribution systems must be overcome. Safety issues and perception must also be addressed. *See* FUEL CELL; HYDROGEN.

Reformulated fuels. Conventional gasoline and diesel fuels are complex mixtures of many different chemical compounds. Over time these fuels have undergone reformulation to improve their value to the transportation industry; this reformulation ranged from mild, like the backing-out of butane to reduce the volatility, to severe adjustments that substantially change the composition.

The U.S. Clean Air Act Amendments (CAAA) have served to increase interest in using regulated changes to motor fuel characteristics as a means of achieving environmental goals. The reformulated gasoline (RFG) program was designed to resolve ground-level ozone problems in urban areas. Under this program, compared to conventional gasoline, the amount of heavy hydrocarbons is limited in reformulated gasoline, and the fuel must include oxygenates and contain fewer olefins, aromatics, and volatile organic compounds.

Diesel-powered heavy-duty trucks and buses are said to account for about a quarter of the nitrogen oxides (NO_x) that result in smog formation and well over half of the particulates produced by mobile sources. Particulates, unburned fuel particles, can stick to the lungs if inhaled, causing increased susceptibility to respiratory illness. Research is under way to develop diesel fuel that will reduce both the particulate and NO_x levels. These fuels will have higher cetane ratings than conventional diesel fuel and lower sulfur and aromatics. Michelle Heath

Bibliography. R. L. Bechtold, *Alternative Fuels Guidebook: Properties, Storage, Dispensing, and Vehicle Facility Modifications*, 1997; R. L. Busby,

Hydrogen and Fuel Cells: A Comprehensive Guide, 2005; Society of Automotive Engineers, *State of Alternative Fuel Technologies 2000*, 2000.

Altimeter

Any device which measures the height of an aircraft. The two chief types are the pressure altimeter, which measures the aircraft's distance above sea level, and the radar altimeter, which measures distance above the ground.

Pressure Altimeter

A pressure altimeter precisely measures the pressure of the air at the level an aircraft is flying and converts the pressure measurement to an indication of height above sea level according to a standard pressure-altitude relationship. In essence, a pressure altimeter is a highly refined aneroid barometer since it utilizes an evacuated capsule whose movement or force is directly related to the pressure on the outside of the capsule (**Fig. 1**). Various methods are used to sense the capsule function and cause a display to respond such that the pilot sees the altitude level much as one looks at a watch.

Because altitude measured in this manner is also subject to changes in local barometric pressure, altimeters are provided with a barosetting that allows the pilot to compensate for these weather changes, the sea-level air pressure to which the altimeter is adjusted appearing in a window of the dial. Flights below 18,000 ft (5486 m) must constantly contact the nearest traffic center to keep the altimeters so updated. Flights above 18,000 ft and over international waters utilize a constant altimeter setting of 29.92 in. Hg, or 1013.2 millibars (101.32 kilopascals), so that all high-flying aircraft have the same reference and will be interrelated, providing an extra margin of safety. *See* AIR NAVIGATION.

Electromechanical computers are provided that correct the sensed static pressure for the static systems defect error, which is an inherent aircraft error due to the effect of high-speed flying on the air mass immediately around the aircraft fuselage in the vicinity of the static ports. The corrected altitude information is usually transmitted to electromechanical indicators for the pilot's display. Much work has been done to improve the readability of the display, so that pilots may make rapid readings with a low probability of reading error. The present accuracy of altimeters is on the order of 0.5% of the indicated altitude for all mechanical indicators and 0.2% of the indicated altitude for computer systems. *See* BAROMETER.

James W. Angus

Radar Altimeter

A radar altimeter is a low-power radar that measures the distance of an aircraft (or other aerospace vehicle) above the ground. It differs from a pressure altimeter, which measures distance above sea level when properly adjusted for atmospheric pressure variations.

Radar altimeters are often used in aircraft during bad-weather landings. They are an essential part of many blind-landing and automatic navigation systems and are used over mountains to indicate terrain clearance. Special types are used in surveying for quick determination of profiles. Radar altimeters are used in bombs, missiles, and shells as proximity fuses to cause detonation or to initiate other functions at set altitudes. Radar altimeters have been used on

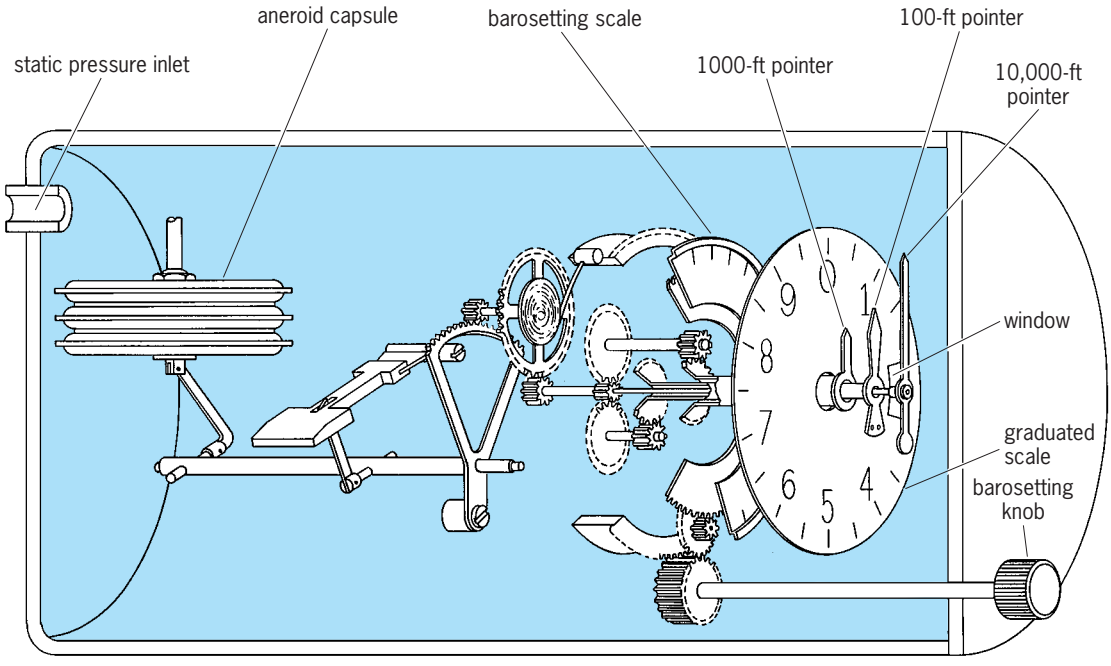


Fig. 1. Schematic arrangement of sensitive pressure altimeter. 1 ft = 0.3 m.

various spacecraft, starting with *Skylab* in 1973, to measure the shape of the geoid and heights of waves and tides over the oceans. Notable spacecraft altimeters over the Earth were those aboard the *Seasat* and *Geosat* American satellites, the American-French *TOPEX/POSEIDON,* and the European *ERS 1.* Other spacecraft altimeters provide topographic information on other planets, particularly Venus. *See* AUTOMATIC LANDING SYSTEM; GROUND PROXIMITY WARNING SYSTEM.

Principle of operation. Like other radar devices, the altimeter measures distance by determining the time required for a radio wave to travel to and from a target, in this case the Earth's surface. If the Earth were a perfectly flat horizontal plane or smooth sphere, the signal would come only from the closest point (**Fig. 2***a*), and would be a true measure of altitude. Actually, the Earth is not smooth, and energy is scattered back to the radar from all parts of the surface illuminated by the transmitter (Fig. 2*b*). For the radar to measure distance to the ground accurately, it must distinguish between the energy from points near the vertical and that from more distant points.

Radar altimeters can seldom have highly directive antennas, for they must be able to function regardless of the aircraft attitude. Thus the antenna cannot discriminate against off-vertical signals. The wide beam width means the antenna gain is small, so the power required is much greater than for a comparable directive radar working against a large, close target. Spaceborne altimeters used to measure distances to the ocean and ice sheets, however, are mounted on stable platforms that allow them to use very narrow antenna beams. *See* ANTENNA (ELECTROMAGNETISM).

Most radio altimeters use either pulse or frequency modulation, the former being more popular for high altitudes, and the latter for low altitudes. *See* FREQUENCY MODULATION; PULSE MODULATOR.

Pulse altimeters. In a typical pulse altimeter (**Fig. 3***a*) the radio-frequency carrier is modulated with short pulses (under 0.25 microsecond). The short pulse permits measurements, even at low altitudes, of the time delay between the leading edge of the transmitted pulse and that of the pulse returned from the ground. Early pulse altimeters displayed the received signal on a cathode-ray tube with circular sweep, allowing the pilot to determine the leading-edge position of the echo signal. Modern pulse altimeters use a tracking gate system. One gate is kept close to the leading edge by a servo system that adjusts the position of the gate to the optimum delay point. A simple single-gate system can be used (Fig. 3*a*), but most pulse altimeters use two or three gates to achieve better distance measurement in the presence of noise and fading. The time delay is displayed for the pilot as a range-dial reading or digital readout. The delay is usually converted to a digital signal that is delivered to the autopilot and flight-control computer.

The return power for a pulse altimeter (the mean of many pulses; Fig. 2*c*) is a combination of scattered returns from the entire illuminated area and specular return from the nearest point to which the

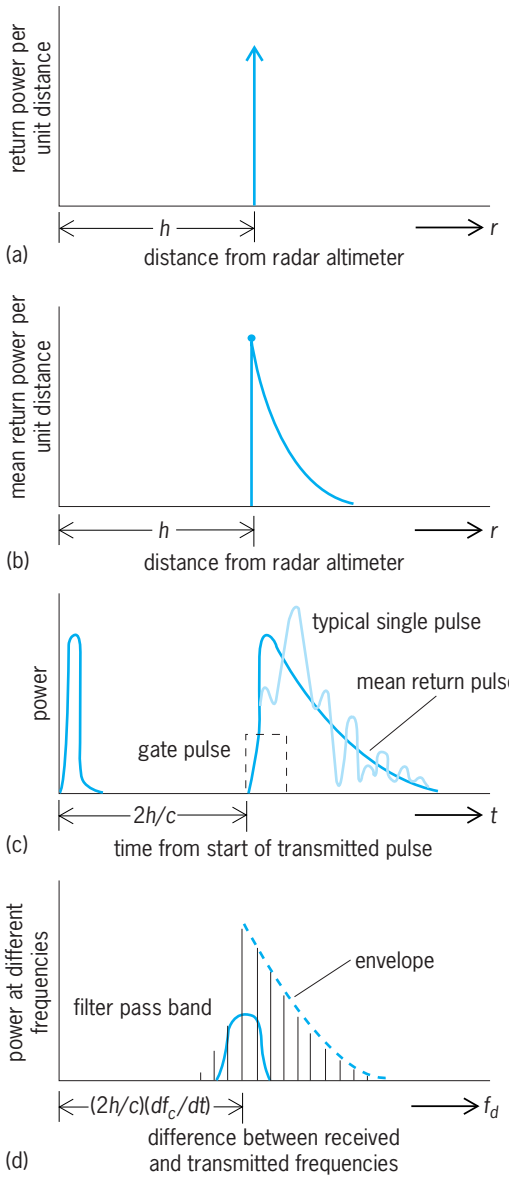


Fig. 2. Radar altimeter principles. (*a*) Ideal return power distribution, which would be observed if the signal came only from the closest point to the radar. (*b*) Typical distribution of average returned power as a function of distance from the radar. (*c*) Pulse shapes in a wide-beam, narrow-pulse altimeter with a single tracking gate. (*d*) Typical FM altimeter difference-frequency spectrum. Spacing between lines is the sweep repetition frequency.

beam is perpendicular. The specular contribution is important only for very smooth water, pavements, or desert playas. Usually the maximum value of the return power varies with height h as $1/h^2$ to $1/h^3$. Interference between different scattering centers causes fading, so the pulses must be averaged over a considerable time to get a precise indication. *See* INTERFERENCE OF WAVES.

FM altimeters. In a frequency-modulated (FM) altimeter (Fig. 3*b*), the frequency of a continuous carrier is swept in some manner, usually to give a triangular frequency-time curve. The difference in frequency between that received from the ground (but transmitted earlier) and that being transmitted is

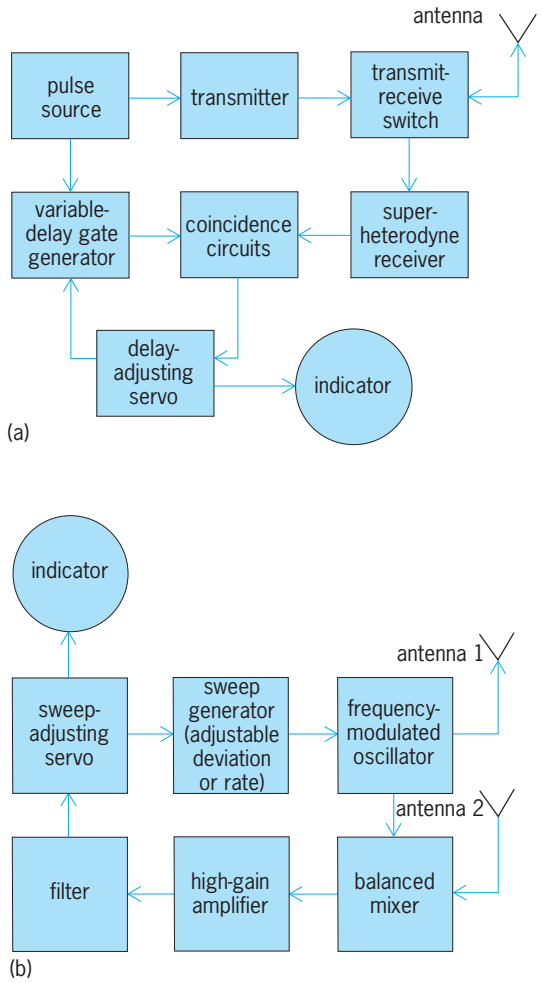


Fig. 3. Block diagrams of typical radar altimeters. (*a*) Automatic pulse altimeter. Control, tracking, and output circuits may all be digital. (*b*) Servo-adjusted-sweep FM altimeter.

a measure of the time delay (Fig. 2*d*). This difference frequency is given by the equation below, where r

$$f_d = \frac{df_c}{dt}(2r/c)$$

is the distance from the radar, c is the speed of light, and f_c is the carrier frequency. Some radars with fixed df_c/dt (lacking a servo and fixed filter) use electronic frequency meters calibrated in range as indicators. Because the return comes from many ranges besides the altitude, an unsophisticated radar of this type may read an effective range frequency more than 10% high in some cases. Other types (Fig. 3*b*) use a relatively narrow filter and keep the minimum difference frequency (altitude frequency) centered in this filter by a servo which adjusts the sweep rate df_c/dt. The sweep rate is indicated, as range, on a dial and also digitized for supply to autopilot and flight computer.

The frequency spectrum (Fig. 2*d*) is composed of individual lines, spaced by the sweep frequency. Suitable signal processing permits recovery of the envelope of the spectrum, which is analogous to the mean return pulse of a pulse altimeter (Fig. 2*c*). Because of problems with short pulses at short ranges, FM altimeters are normally used where altitude measurements must be below about 100 ft (30 m).

Other radio devices may also serve as altimeters. The proximity fuse is a crude device for measuring altitude by field strength alone. Various correlation devices have been proposed, some of which employ noise modulation of either amplitude or frequency.

Spaceborne altimeters. Pulse altimeters are used on spacecraft to observe properties of the ocean and continental ice sheets (Greenland and Antarctica). For these applications, the exact location of the spacecraft must be determined by accurate tracking from land. The distance from altimeter to surface then becomes a measure of the variation of the surface itself. The altimeters must achieve great accuracy. To do so, they transmit pulses with very large bandwidths, thereby achieving resolutions of only a few feet. The actual distance measurement accuracy can be better than 5 cm (2 in.), provided suitable corrections are made for slight variations in the refractive index of (and therefore velocity of propagation through) the atmosphere. *TOPEX/POSEIDON* uses two frequencies to aid in correcting for the minute effects of the ionosphere.

To achieve this great precision without the use of extremely short pulses that would need excessive peak power, pulse-compression techniques are employed. Normally these methods use a binary phase-coded pulse.

These spaceborne systems have allowed major improvements in knowledge of the shape of the geoid over the oceans, where surveying methods that work with great precision over land perform poorly or cannot be used at all. Moreover, with the precision now available, oceanographers can locate and track changes in the major current systems in the ocean by the slight bulge associated with the current flow. Glaciologists use the altimeter data taken over the continental ice sheets to detect growth and decay of the ice, which can be an indication of global warming or cooling. Since the ice sheets contain most of the fresh water on Earth, significant melting resulting from global warming could result in major catastrophes along coastlines, making monitoring of glaciers an important concern. *See* GEODESY; GLACIOLOGY; OCEAN CIRCULATION.

The same altimeters also can measure wave height at points along the ocean surface beneath the satellite (the subsatellite track). When ocean waves are high, the leading edge of the average pulse (Fig. 2*c*) is stretched out, since the signal returns first from a relatively small area near the crests of the waves and then builds up as the entire surface is illuminated. The altimeter samples the pulse at close intervals (typically about 10 nanoseconds), and the averaged samples are used to determine the shape of the received pulse. This can be compared with theoretical shapes to achieve good estimates of the root-mean-square wave height beneath the spacecraft.

Altimeters can also be used as scatterometers to estimate the wind speed beneath the spacecraft. This is possible because the received signal decreases as the wind increases. Since meteorologists require wind

measurements over wide areas, this subsatellite wind measurement is less useful than that from spaceborne radar scatterometers looking for hundreds of miles to the side of the spacecraft. Moreover, the wind sensitivity of radar scatterometers beyond an angle of 20° from the vertical is much greater than that at vertical incidence. Nevertheless, the altimeters have proved useful during intervals when no spaceborne scatterometer was active. *See* REMOTE SENSING.

Richard K. Moore

Bibliography. D. E. Barrick, Analysis and interpretation of altimeter sea echo, *Adv. Geophys.,* 37:60–98, 1985; B. C. Douglas and R. E. Cheney, Geosat: Beginning of a new era in satellite oceanography, *J. Geophys. Res.,* 95:2833–2836, 1990; Seasat Special Issue II: Scientific Results, *J. Geophys. Res.,* 88(C3):1531–1745, 1983; M. I. Skolnik, *Introduction to Radar Systems,* 2d ed., 1980; G. I. Sonnenberg, *Radar and Electronics Navigation,* 6th ed., 1988; F. T. Ulaby, R. K. Moore, and A. K. Fung, *Microwave Remote Sensing,* vol. 2, 1982.

Altitudinal vegetation zones

Intergrading regions on mountain slopes characterized by specific plant life forms or species composition, and determined by complex environmental gradients.

Along an altitudinal transect of a mountain, there are sequential changes in the physiognomy (growth form) of the plants and in the species composition of the communities. This sequential zonation of mountain vegetation has been recognized for centuries. Vertical zonation was fully developed as an ecological concept by the work of C. H. Merriam with the U.S. Biological Survey of 1889. He described a series of life zones on the slopes of the San Francisco Peaks in Arizona, based on characteristic species of the flora and fauna. Other patterns of plant physiognomic and community zonation have now been cataloged in mountain ranges throughout the world. *See* LIFE ZONES.

Merriam associated his life zones with temperature gradients present along mountain slopes. Later research on patterns of altitudinal zonation has centered on the response of species and groups of species to a complex of environmental gradients. Measurements of a species along a gradient, for example, the number of individuals, biomass, or ground coverage, generally form a bell-shaped curve. Peak response of a species occurs under optimum conditions and falls off at both ends of the gradient. The unique response of each species is determined by its physiological, reproductive, growth, and genetic characteristics.

Zones of vegetation along mountain slopes are formed by intergrading combinations of species that differ in their tolerance to environmental conditions. Zones are usually indistinct entities rather than discrete groupings of species. However, under some conditions of localized disjunctions, very steep sections of gradients, or competitive exclusion, discontinuities in the vegetation can create discrete communities. *See* ECOLOGICAL COMMUNITIES.

Vegetation zones are often defined by the distributions of species having the dominant growth form, most frequently trees. For example, in the Cascade Mountains of Washington and Oregon, tundra above treeline has been called the sedge–grass zone, but successively lower zones are spruce–fir, arborvitae–hemlock, Douglas-fir, ponderosa pine, and finally several fescue–wheatgrass–sagebrush zones in which no trees grow. Another set of zones could be designated if other criteria were used. The boundaries of the middle elevation zones might be different if they had been defined instead by distributions of the dominant shrubs.

Environmental gradients in mountains. Altitudinal vegetation zonation, therefore, is an expression of the response of individual species to environmental conditions. Plants along an altitudinal transect are exposed, not to a single environmental gradient, but to a complex of gradients, the most important of which are solar radiation, temperature, and precipitation. Although these major environmental gradients exist in most mountain ranges of the world, the gradients along a single altitudinal transect are not always smooth because of topographic and climatic variability. This environmental variation can result in irregular vegetation zones.

The solar energy received by mountain surfaces increases with altitude, associated with decreases in air density and the amount of dust and water vapor. Global radiation levels in the European Alps have been shown to be 21% greater at 10,000 ft (3000 m) than at 650 ft (200 m) under a cloudless sky. An overcast sky is more efficient at reducing short-wave energy reaching low elevations and can increase the difference in energy input to 160%. However, more frequent clouds over high elevations relative to sunnier lower slopes commonly reduces this difference. *See* SOLAR RADIATION.

Vegetation patterns are also strongly influenced by the decline in air temperature with increasing altitude, called the adiabatic lapse rate. Lapse rates are generally between 1.8°F to 3.6°F per 1000 ft (1°C to 2°C per 300 m), but vary with the amount of moisture present; wet air has a lower lapse rate. Thus, plants occurring at higher elevations generally experience cooler temperatures and shorter growing periods than low-elevation plants. Variation in the temperature gradient can be caused by differences in slope, aspect, radiation input, clouds, and air drainage patterns. *See* AIR TEMPERATURE.

The precipitation gradient in most mountains is the reverse of the temperature gradient: precipitation increases with altitude. Moist air from low elevations is forced upward by the blocking action of the mountains into regions of cooler temperatures. Air holds less moisture at low temperatures than at warmer ones. When the atmosphere reaches the point of saturation (the dew point), condensation occurs, forming clouds or precipitation. This moisture gradient is most pronounced on windward slopes and in ranges close to warm and moist oceanic

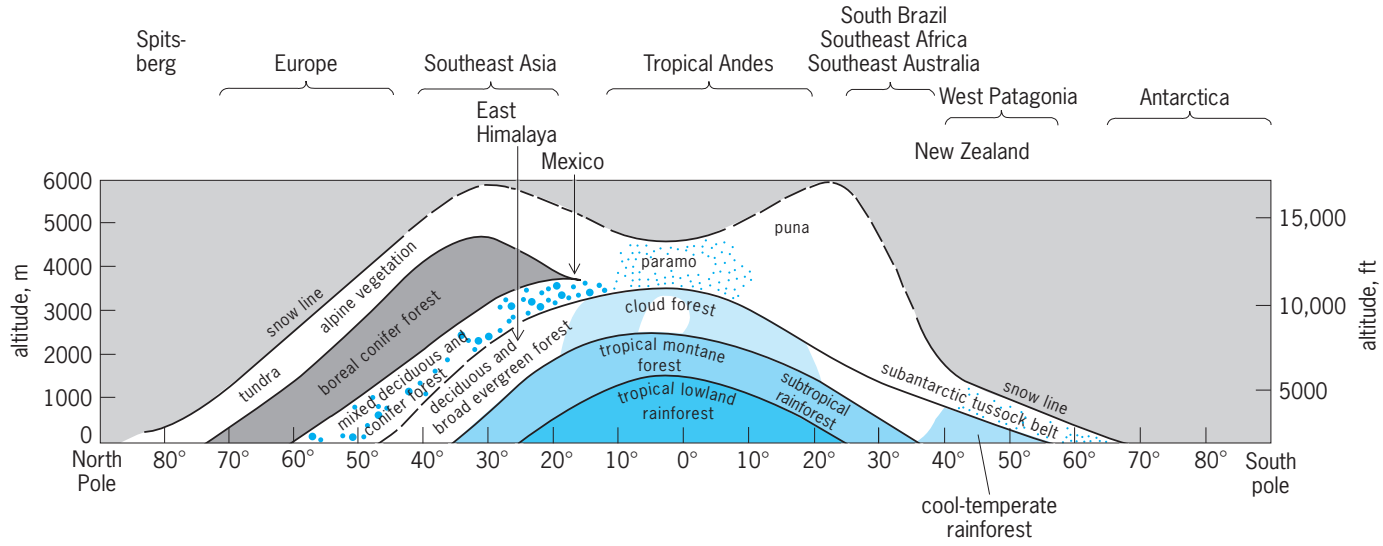


Fig. 1. Generalized patterns of altitudinal vegetation zonation in the Northern and Southern hemispheres. (*After L. W. Price, Mountains and Man, University of California Press, 1981*)

winds. As the moving air passes over the peaks of the mountain range and moisture in the air is depleted, precipitation is reduced, creating a rain "shadow" on the lee side of the range. *See* PRECIPITATION (METEOROLOGY).

Temperate, tropical, and high latitudes. General changes in vegetation with increases in altitude include reduction in plant size, slower growth rates, lower production, communities composed of fewer species, and less interspecific competition. However, many regional exceptions to these trends exist. In western North America, for example, the lowest zone is often a treeless shrubland or prairie, so plant size initially increases with altitude as trees become important. Above this zone, the trend toward smaller size prevails.

Characteristics of vegetation zones also vary with latitude (**Fig. 1**). Mountains at higher latitudes have predominantly seasonal climates, with major temperature and radiation extremes between summer and winter. Equatorial and tropical mountains have a strong diurnal pattern of temperature and radiation input with little seasonal variation. The upper altitudinal limit of trees, and the maximum elevation of plant growth generally, decreases with distance from the Equator, with the exception of a depression near the Equator.

The following examples of altitudinal zonation in temperate, tropical, and high-latitude regions illustrate patterns that exist in each region. These patterns are not comprehensive, and many others exist locally and worldwide.

Temperate. The vegetation of the Santa Catalina Mountains in Arizona illustrates the variety of zones found in a temperate mountain range (**Fig. 2**). Typically, precipitation increases with altitude, while air and soil temperatures decrease. Community diversity, the number of species present, decreases with altitude in all zones except the scrub desert, where arid conditions support fewer species. Biomass production is less at low elevations than on higher slopes, the reverse of the general trend, because of limitations to growth under arid conditions.

Hot and dry conditions at lower elevations support a treeless desert shrub and grassland, dominated by such shrubs as palo verde (*Cercidium microphyllum*) and mesquite (*Prosopis juliflora*). With increasing elevation, more moisture allows growth of broadleaf deciduous and evergreen oak trees (*Quercus* spp.). Needle-leaf evergreen trees, including ponderosa pine (*Pinus ponderosa*) and border limber pine (*P. strobiformis*), become increasingly important in the mid-elevation vegetation zones where temperatures are cooler, moisture more readily available, and the growing season longer. Broadleaf trees decrease in importance in these zones. Finally, in the upper zones, where the coolest temperatures and highest moisture levels prevail, conifers that commonly extend farther north including Douglas-fir (*Pseudostuga menziesii*) and Engelmann spruce (*Picea engelmanni*) replace species better adapted to lower-elevation environments. At higher latitudes and where mountains are high enough, an alpine zone of grasses, forbs, cushion plants, and dwarf shrubs occurs above treeline. A similar sequential replacement of species in the altitudinal zones is repeated in both the herb and shrub layer.

The Southern Hemisphere has fewer mountains in the temperate regions than areas north of the Equator. The oceanic climate of these mountains is cooler and moister because of the small land mass relative to the nearby oceans. A prolonged cold season is absent. In New Zealand, broadleaf evergreen trees such as evergreen beech (*Nothofagus* spp.) dominate moist mountain forests, with needle-leaf evergreens such as *Podocarpus* spp. being of lesser importance. Zones at all elevations have a richer, more diverse flora than their Northern Hemisphere counterparts. Within each zone, shrubs are more luxurious, and epiphytes and hanging vines are common.

Tropical. Vegetation zones at all elevations in tropical mountains are dominated by a diurnal climate

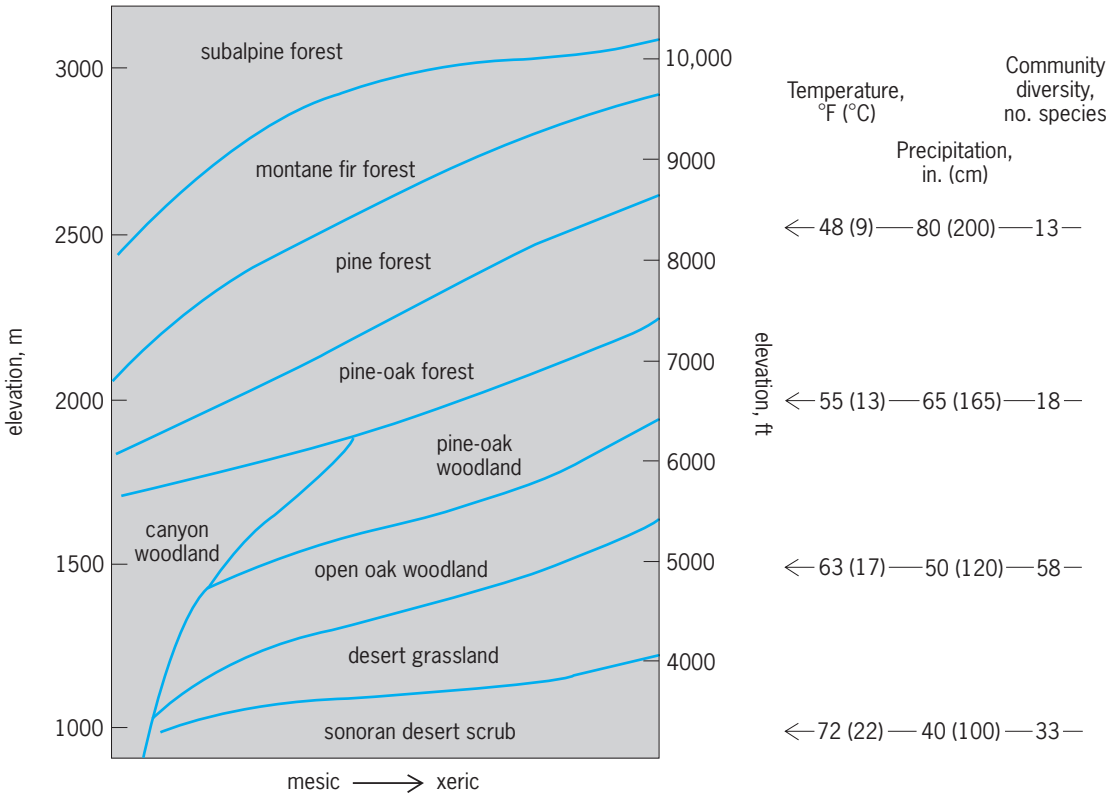


Fig. 2. Vegetation zones and gradients of mean annual soil temperature, annual precipitation, and community diversity on the south slopes of the Santa Catalina Mountains, Arizona, a temperate mountain range. (*After R. H. Whittaker, and W. A. Niering, Vegetation of the Santa Catalina Mountains: A gradient analysis of the south slope, Ecology, 46: 429–451, 1965*)

with little seasonality, because sun angles are high throughout the year. Growing seasons are determined by moisture availability rather than by temperature. At lower elevations in the Andes, tropical rainforests with a species-rich multilayered and luxurious canopy of broadleaf evergreen trees cover mountain slopes. These trees are accompanied by a lush mixture of hanging vines, epiphytes, shrubs, and flowering herbs. In the submontane and montane zones at higher elevations, the complexity of the forest is reduced. Trees are shorter and are adapted to drier conditions, and the canopy has fewer layers. Vines, epiphytes, and shrubs are present in fewer numbers, but mosses and lichens become more common.

The subalpine zone is an elfin woodland or cloud forest of short, stunted trees covered with an abundance of mosses and lichens. Vegetation is adapted to the higher moisture and lower light levels created by clouds, which almost continuously envelop the zone. Plant communities in this zone are less complex than those lower on the mountain; fewer species are present and canopy structure is simpler. In the northern Andes of Colombia, the alpine zone consists of a low vegetation called paramo, which is dominated by arborescent members of the sunflower family, Compositae. Growth forms consist of tall tussock grasses, dwarf shrubs, herbs, and two forms unique to the tropics of the Southern Hemisphere: dendroid or tufted-leaf stemmed plants, and woolly candlelike plants. *See* PARAMO.

Farther from the Equator in the central and southern Andes, the climate is drier and lower zones are dominated by forests of deciduous trees or grassland savannas. The alpine zone consists of puna, a vegetation of smaller tussock grasses and many cushion and rosette plants. *See* PUNA; SAVANNA.

High latitude. At high latitudes in both Northern and Southern hemispheres, fewer vegetation zones exist. Short growing seasons permit no tree growth, and shrubs and herbs are limited at higher elevations by a permanent cover of ice or snow, called the nival zone. Plant communities consist of fewer species than in temperate or tropical regions. In the Olgivie Mountains, Yukon Territory, Canada, valley bottoms are covered by willow (*Salix* spp.) and birch (*Betula* spp.) shrubs with a ground layer of grasses and forbs in drier sites, and tussock tundra dominated by cottongrass (*Eriophorum* spp.) in wet sites. At higher elevations, tall woody shrubs are absent, and this zone consists of a tundra of dwarf heath shrubs including mountain heather (*Cassiope* spp.) and mountain avens (*Dryas* spp.), grasses, forbs, and plants with cushion or rosette growth forms. *See* TERRESTRIAL ECOSYSTEM.

John S. Campbell

Bibliography. E. W. Beals, Vegetational change along altitudinal gradients, *Science*, 165:981-985, 1969; W. Lauer, The altitudinal belts of the vegetation in the central Mexican highlands and their climatic condition, *Arctic Alpine Res.*, 5(pt. 2):A99-114, 1973; L. W. Price, *Mountains and Man*, 1981; C. Troll (ed.), *Geoecology of the Mountainous Regions of the Tropical Americas*, 1968; H. Walter, *Vegetation of the Earth and Ecological Systems of the Geo-Biosphere*, 1979; R. H. Whittaker and W. A. Niering,

Vegetation of the Santa Catalina Mountains: A gradient analysis of the south slope, *Ecology*, 46:429-451, 1965.

Alum

A colorless to white crystalline substance which occurs naturally as the mineral kalunite and is a constituent of the mineral alunite. Alum is produced as aluminum sulfate by treating bauxite with sulfuric acid to yield alum cake or by treating the bauxite with caustic soda to yield papermaker's alum. Other industrial alums are potash alum, ammonium alum, sodium alum, and chrom alum (potassium chromium sulfate). Major uses of alum are as an astringent, styptic, and emetic. For water purification alum is dissolved; it then crystallizes out into positively charged crystals that attract negatively charged organic impurities to form an aggregate sufficiently heavy to settle out. Alum is also used in sizing paper, dyeing fabrics, and tanning leather. With sodium bicarbonate it is used in baking powder and in some fire extinguishers. *See* ALUMINUM; COLLOID.

Frank H. Rockett

Aluminum

A metallic chemical element, symbol Al, atomic number 13, atomic weight 26.98154, in group 13 of the periodic system. Pure aluminum is soft and lacks strength, but it can be alloyed with other elements to increase strength and impart a number of useful properties. Alloys of aluminum are light, strong, and readily formable by many metalworking processes; they can be easily joined, cast, or machined, and accept a wide variety of finishes. Because of its many desirable physical, chemical, and metallurgical properties, aluminum has become the most widely used nonferrous metal. *See* PERIODIC TABLE.

1												13	14	15	16	17	18
1 H	2											5 B	6 C	7 N	8 O	9 F	2 He
3 Li	4 Be											13 Al	14 Si	15 P	16 S	17 Cl	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	31 Ga	32 Ge	33 As	34 Se	35 Br	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	49 In	50 Sn	51 Sb	52 Te	53 I	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	81 Tl	82 Pb	83 Bi	84 Po	85 At	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	113					86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112						

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Aluminum is the most abundant metallic element on the Earth and Moon but is never found free in nature. The element is widely distributed in plants, and nearly all rocks, particularly igneous rocks, contain aluminum in the form of aluminum silicate minerals. When these minerals go intosolution, depending upon the chemical conditions, aluminum can be precipitated out of the solution as clay minerals or aluminum hydroxides, or both. Under such conditions bauxites are formed. Bauxites serve as principal raw materials for aluminum production. *See* BAUXITE; CLAY MINERALS; IGNEOUS ROCKS; WEATHERING PROCESSES.

Physiochemical properties of pure aluminum

Property	Value
Atomic number	13
Atomic weight	26.98154
Crystal structure	Face-centered cubic
Density, g/cm^3 at 25°C	2.698
Melting point	660.37°C (1220.7°F)
Boiling point	2447°C (4436.6°F)
Latent heat of fusion, cal/g	94.9
Latent heat of vaporization, cal/g	2576
Heat of combustion, cal/g at 25°C	7420
Specific heat, cal/g/°C at 25°C	0.215
Thermal conductivity, cal/cm/°C/s at 25°C	0.566
Coefficient of thermal expansion, 10^{-6}/°C	23
Thermal diffusivity, cm^2/s at 25°C	0.969
Electrical resistivity, microhm-cm at 25°C	2.7
Hardness, Mohs	2–2.9
Reflectivity	85–90%

Aluminum is a silvery metal. Naturally occurring aluminum consists of a single isotope, $^{27}_{13}$Al. Aluminum crystallizes in the face-centered cubic structure with edge of the unit lattice cube of 4.0495 angstroms. Aluminum is known for its high electrical and thermal conductivities and its high reflectivity. A summary of some of the important properties of pure aluminum is given the **table**.

The electronic configuration of the element is $1s^2 2s^2 2p^6 3s^2 3p^1$. Aluminum exhibits a valence of +3 in all compounds, with the exception of a few high-temperature monovalent and divalent gaseous species.

Aluminum is stable in air and resistant to corrosion by seawater and many aqueous solutions and other chemical agents. This is due to protection of the metal by a tough, impervious film of oxide. At a purity greater than 99.95%, aluminum resists attack by most acids but dissolves in aqua regia. Its oxide film dissolves in alkaline solutions, and corrosion is rapid. *See* CORROSION.

Aluminum is amphoteric and can react with mineral acids to form soluble salts and to evolve hydrogen.

Molten aluminum can react explosively with water. The molten metal should not be allowed to contact damp tools or containers.

At high temperatures aluminum reduces many compounds containing oxygen, particularly metal oxides. These reactions are used in the manufacture of certain metals and alloys.

Applications in building and construction represent the largest single market of the aluminum industry. Millions of homes use aluminum doors, siding,

windows, screening, and down-spouts and gutters. Aluminum is also a major industrial building product. Transportation is the second largest market. In automobiles, aluminum is apparent in interior and exterior trim, grilles, wheels, air conditioners, automatic transmissions, and some radiators, engine blocks, and body panels. Aluminum is also found in rapid-transit car bodies, rail cars, forged truck wheels, cargo containers, and in highway signs, divider rails, and lighting standards. In aerospace, aluminum is found in aircraft engines, frames, skins, landing gear, and interiors. The food packaging industry is a fast-growing market.

In electrical applications, aluminum wire and cable are major products. Aluminum appears in the home as cooking utensils, cooking foil, hardware, tools, portable appliances, air conditioners, freezers, and refrigerators.

There are hundreds of chemical uses of aluminum and aluminum compounds. Aluminum powder is used in paints, rocket fuels, and explosives, and as a chemical reductant. *See* ALUMINUM ALLOYS.

Allen S. Russell

Bibliography. F. A. Cotton, et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004.

Aluminum (metallurgy)

The separation, extraction, and purification of alumina from ores, followed by the production of aluminum. Aluminum is the most abundant metallic element on the Earth and Moon but is never found free in nature. It makes up more than 8% of the solid portion of the Earth's surface. Seawater contains an average of only 0.5 ppm aluminum. The element is widely distributed in plants, where it may be present in significant concentrations, particularly in vegetation in marshy places and acid soils.

Nearly all rocks, particularly igneous rocks, contain aluminum in the form of aluminum silicate minerals. When subjected to weathering under the right conditions, these minerals go into solution; then, depending upon the chemical conditions, aluminum can be precipitated out of the solution as clay minerals or aluminum hydroxides, or both. Any quartz in the parent rock is relatively unattacked by the weathering processes and remains as sand in the deposit. Under conditions which precipitate clay minerals and leach out iron and the alkali and alkaline-earth metals, deposits of high-grade kaolin clay can result. Sometimes natural elutriation or mechanical separation by washing occurs, which separates the clay from the sand and enhances the purity of the deposit. Under conditions which precipitate aluminum principally as hydroxide minerals and leach out alkali and alkaline-earth metals completely and iron to varying degrees, bauxites are formed. It is also possible for clay minerals to be desilicified to aluminum hydroxides and for aluminum hydroxides to be resilicified to clay minerals. *See* BAUXITE; CLAY MINERALS; IGNEOUS ROCKS; WEATHERING PROCESSES.

Bauxites usually consist of mixtures of the following minerals in varying proportions: gibbsite, also known as hydrargillite [$Al(OH)_3$]; boehmite and diaspore [AlO(OH)]; clay minerals such as kaolinite [$Al_2Si_2O_5(OH)_4$]; quartz [SiO_2], and anatase, rutile, and brookite [TiO_2]. Small amounts of magnetite [Fe_3O_4], ilmenite [$FeTiO_3$], and corundum [Al_2O_3] are sometimes present. In addition to the above minerals, bauxites usually contain traces of other insoluble oxides such as zirconium, vanadium, gallium, chromium, and manganese. The term alumina trihydrate is often applied to the mineral gibbsite, and the term alumina monohydrate to the minerals boehmite and diaspore. However, the minerals are not hydrates in the true sense of the word. However, for processing purposes, bauxites are often classified as trihydrate bauxites and monohydrate bauxites, depending upon their content of the principal mineral or minerals containing the extractable alumina. In general, European bauxites are of the monohydrate type and are geologically older than bauxites in tropical countries, which occur generally as the trihydrate type. This, however, is not an absolute rule. Some European bauxites contain gibbsite along with boehmite and diaspore, and Caribbean bauxites contain small-to-medium amounts of boehmite along with the gibbsite. Typical bauxites which are used for aluminum production contain 40–60% total alumina (Al_2O_3), 1–15% total silica (SiO_2), 7–30% total Fe_2O_3, 1–5% titania (TiO_2), and 12–30% combined H_2O.

The suitability of a bauxite as a raw material for aluminum production depends not only on its alumina content but also on its content of combined silica, which is usually in the form of the mineral kaolinite. Kaolinite not only contains aluminum that cannot be extracted in the Bayer process, but also reacts with the sodium–aluminate solution to cause a loss of sodium hydroxide. In some bauxites, the clay minerals are concentrated in the fine-particle size range. In this case, if the economics are favorable, the bauxite can be beneficiated by washing it on screens to remove substantial amounts of the clay and thus make a product containing higher alumina and lower combined silica contents than the original bauxite (**Fig. 1**).

History. H. C. Oersted probably prepared the first metallic aluminum (impure) in 1824 by reducing aluminum chloride with potassium amalgam. F. Wöhler is generally credited with the discovery of the metal by virtue of the first isolation of less impure aluminum in 1827 and the first description of a number of its properties. The metal remained a laboratory curiosity until 1854, when Henri Sainte-Claire Deville improved the earlier method of preparation by employing sodium as reductant and was successful in producing larger quantities of relatively pure metal. The first electrolytic preparation of aluminum, which also occurred in 1854, was accomplished by electrolysis of fused sodium aluminum chloride independently in the laboratories of Deville in France and of R. Bunsen in Germany. *See* ELECTROLYSIS.

The modern industrial electrolytic method of production was discovered simultaneously and independently by Charles Martin Hall in the United States

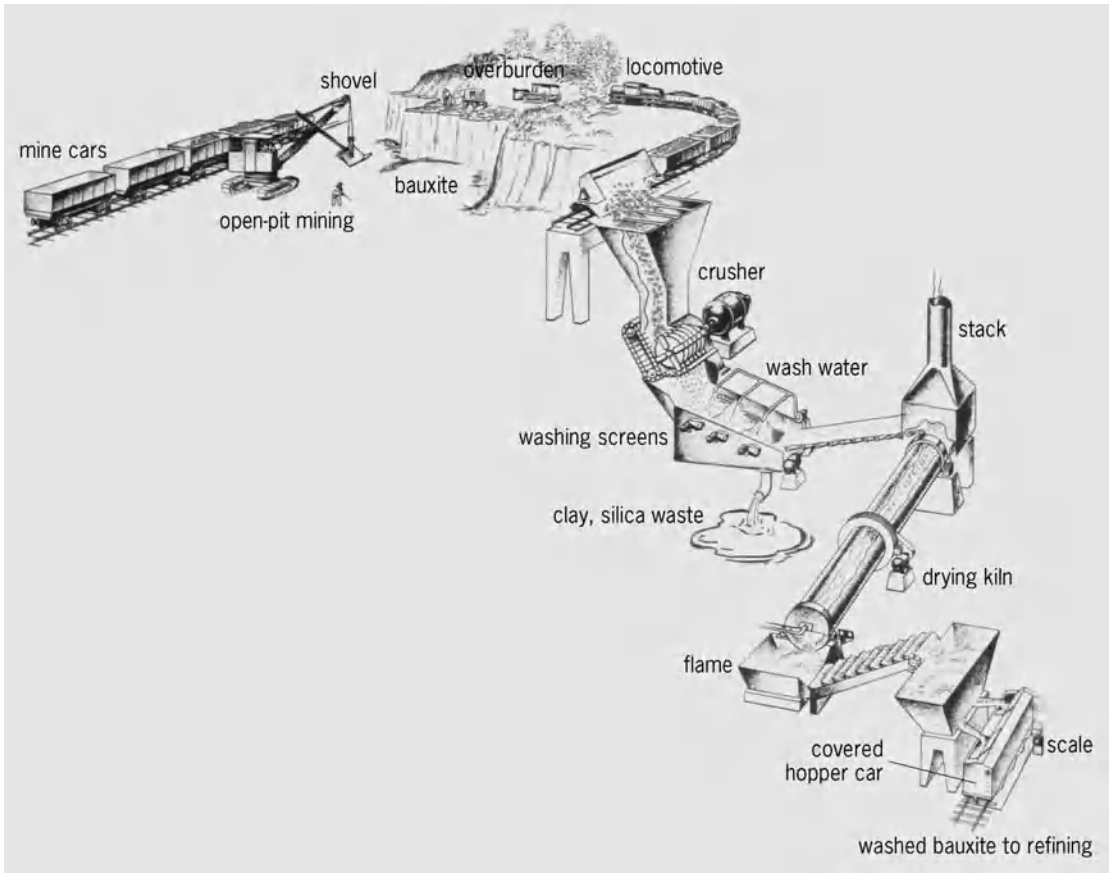


Fig. 1. Bauxite washing. (*Aluminum Company of America*)

and Paul-Louis Héroult of France in 1886. The essentials of their discovery remain the basis for today's aluminum industry. Alumina dissolved in a fluoride fusion (largely cryolite, Na_3AlF_6) is electrolyzed with direct current. Carbon dioxide is discharged at the anode, while the metal is deposited on molten aluminum lying on the carbon lining at the bottom of the cell. The technology for extracting aluminum from its ores was further improved in 1888, when Karl Josef Bayer patented in Germany a method for making pure alumina from bauxite.

Production. The most widely used technology for producing aluminum involves two steps: extraction and purification of alumina from ores, and electrolysis of the oxide after it has been dissolved in fused cryolite. Bauxite is by far the most used raw material. Technically feasible processes exist for extracting alumina from other raw materials such as clays, anorthosite, nepheline syenite, and alunite; however, these processes have not been competitive with the Bayer process in the United States. With the present rapidly changing economics of bauxite production, some of these processes may become competitive in the future. A few of these alternate raw materials are used as a source for alumina in Europe and Asia.

Alumina extraction. A general flow sheet for the Bayer process is shown in **Fig. 2.** In the process, bauxite is crushed and ground, then digested at elevated temperature (280–450°F or 140–230°C) and pressure in a strong solution of caustic soda (80–110 g Na_2O/liter or 0.7–0.9 lb Na_2O/gal). For monohydrate-type bauxites, in which the alumina occurs in forms which are more difficult to dissolve than in trihydrate-type bauxites, stronger solutions (up to 220 g Na_2O/liter or 1.8 lb Na_2O/gal), higher temperatures (up to 570°F or 300°C) and pressures (as high as 150 atm or 1.52×10^7 pascals), and sometimes longer digestion times are required. The gibbsite, boehmite, or diaspore in the bauxites reacts with the sodium hydroxide to form soluble sodium aluminate. The residue, known as red mud, contains the insoluble impurities and the sodium aluminum silicate compound, referred to as desilication product, formed by the reaction of clay minerals with the sodium aluminate–sodium hydroxide solution. The red mud is separated from the solution by countercurrent decantation and filtration. After cooling, the solution is supersaturated with respect to alumina. It is seeded with recycled synthetic gibbsite (alumina trihydrate) and agitated. A large part of the alumina in solution thus crystallizes out as gibbsite. This gibbsite is classified into product and seed, the seed being recycled and the product washed. Wash water and spent liquor, after concentration by evaporation, are recycled to the digestion system. For metal production, the product is calcined at temperatures up to 2400°F (1300°C) to produce alumina containing about 0.3–0.8% soda, less than 0.1% iron oxide plus silica, and trace amounts of other oxides. Some of the precipitated gibbsite is sold for use in the chemicals industry. Specially precipitated pigment-grade gibbsite is used for rubber,

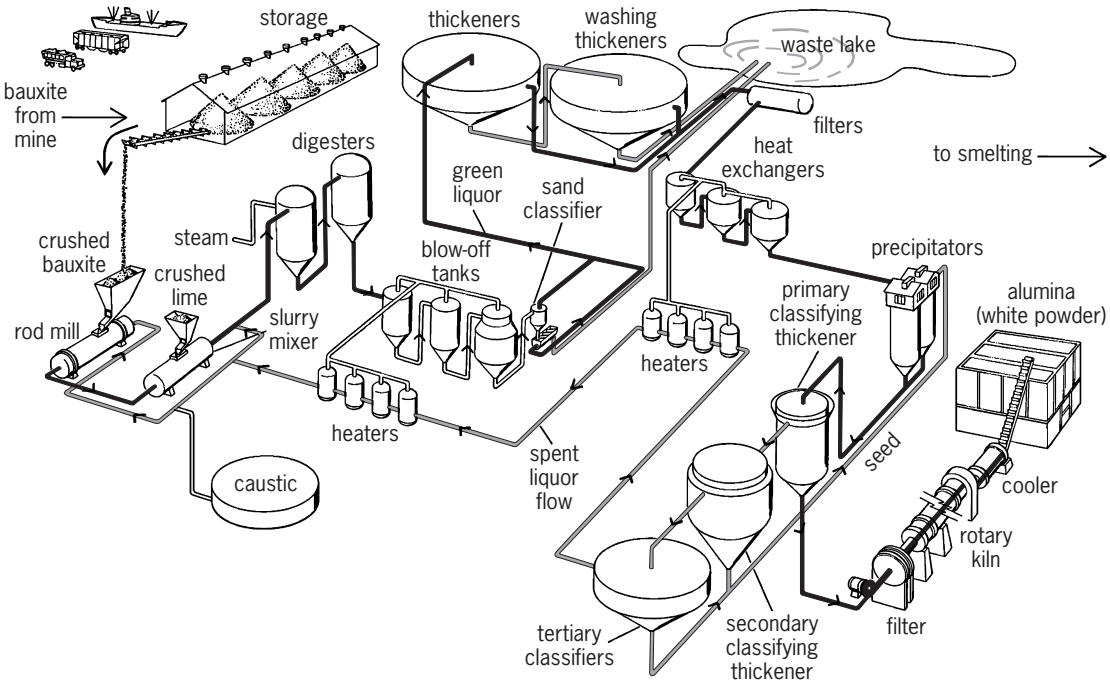


Fig. 2. Processing of alumina from bauxite. (*Aluminum Company of America*)

plastic and paper fillers, and paper coating. Activated aluminas having high internal surface area are made from the gibbsite by calcination at low-to-moderate temperatures. They are used as desiccants and catalysts. Fully calcined aluminas are used in the production of ceramics, abrasives, and refractories.

Alumina can be recovered from low-grade bauxites containing high concentrations of combined silica by the so-called combination or lime–soda sinter process. The process consists of treatment by the Bayer process, following which the red mud is mixed with calcium oxide and sodium carbonate and calcined. This treatment forms sodium aluminate from the aluminous phases in the red mud. The sodium aluminate is then leached out with water, and alumina is recovered from the resulting solution.

Processes for treating other ores have been developed which consist of calcination or fusion of the ore with limestone to make calcium aluminate, from which alumina is leached out with sodium carbonate solution and recovered. Such a process is used in Russia for treating nepheline syenite to yield alumina, and cement as a by-product.

High-iron bauxites have been smelted with limestone in the Pedersen process to yield pig iron and a calcium aluminate slag that can be leached with sodium carbonate solution to recover alumina.

Electrolytic reduction (smelting). Although unchanged in principle, the smelting process of today differs in detail and in scale from the original process discovered by Hall and Héroult. Technology has effected substantial improvements in equipment, materials, and control of the process, and has lowered the energy and labor requirements and the final cost of primary metal.

In a modern smelter (**Fig. 3**), alumina is dissolved in cells (pots)—rectangular steel shells lined with carbon—containing a molten electrolyte (bath) consisting mostly of cryolite. The bath usually contains 2–8% alumina. Excess aluminum fluoride and calcium fluoride are added to lower the melting point and to improve operation. Carbon anodes are hung from above the cells with their lower ends extending to within about 1.5 in. (3.8 cm) of the molten metal, which forms a layer under the molten bath. The heat required to keep the bath molten is supplied by the electrical resistance of the bath as current passes through it. The amount of heat developed with a given current depends on the length of the current path through the electrolyte, that is, anode–cathode distance, which is adjusted to maintain the desired operating temperature, usually 1760–1780°F (960–970°C). A crust of frozen bath, 1–3 in. (2.5–7.5 cm) thick, forms on the top surface of the bath and on the sidewalls of the cell. Alumina is added to the bath or on the crust, where its sorbed moisture is driven off by heat from the cell. While preheating on the crust, the alumina charge serves as thermal insulation. Periodically, the crust is broken and the alumina is stirred into the bath to maintain proper concentration. *See* CRYOLITE.

The passage of direct current through the electrolyte decomposes the dissolved alumina. Metal is deposited on the cathode, and oxygen on the gradually consumed anode. About 0.5 lb (0.23 kg) of carbon is consumed for every pound of aluminum produced. The smelting process is continuous. Alumina is added, anodes are replaced, and molten aluminum is periodically siphoned off without interrupting current to the cells.

Current efficiencies in the industrial electrolytic process are about 85–92%, and the energy efficiency is about 40%. The voltage at the cell terminals is 4–6 V, depending on the size and condition of the

cell. Voltage is required to force the current through the entire cell, and the corresponding power (voltage × amperage) is largely converted into heat in the bath. The amount of power required to maintain the temperature is a smaller proportion of the total power input in large cells than in small ones because of the lower ratio of surface to volume. Thus, power consumed per pound of metal is somewhat less in large than in small cells. Consumption of 6–8 kWh/lb (13–18 kWh/kg) of aluminum produced includes bus-bar, transformer, and rectifier losses.

A potline may consist of 50–200 cells with a total line voltage of up to 1000 V at current loads of 50,000–225,000 A. Electric power is one of the most costly raw materials in aluminum production. Aluminum producers have continually searched for sources of cheap hydroelectric power, but have also had to construct facilities that produce power from fossil fuels. In the past half century, technological advances have significantly reduced the amount of electrical energy necessary to produce a pound of aluminum.

Current is led out of the cell to the anode bus-bar by a number of carbon block anodes suspended in parallel rows on vertical conducting rods of copper or aluminum. Because impurities in the anodes dissolve in the bath as they are consumed, pure carbon (calcined petroleum coke or pitch coke) is used as raw material. The ground coke is mixed hot with enough coal tar or petroleum pitch to bond it into a block when pressed in a mold to form the "green anode." This is then baked slowly at temperatures up to 2000–2200°F (1100–1200°C). In a cavity molded in the top of each block, a steel stub is embedded by casting molten iron around it or by using a carbonaceous paste; the conducting bar is bolted to this stub. Such an electrode is termed a prebaked anode to distinguish it from the Soderberg anode, in which the electrode (single large anode to a cell) is formed in place from a carbonaceous paste that is baked by heat from the pot as it gradually descends into the electrolyte.

The steel shell of the pot is thermally insulated and lined with carbon. The carbon bottom, covered with molten aluminum, serves as cathode of the cell. Electrical connection is made to the carbon cathode by steel bars running through the base of the cell and embedded in the carbon lining.

Molten aluminum is siphoned from the smelting cells into large crucibles. From there the metal may be poured directly into molds to produce foundry ingot, or transferred to holding furnaces for further refining or alloying with other metals to form fabricating ingot. As it comes from the cell, primary aluminum averages about 99.8% purity.

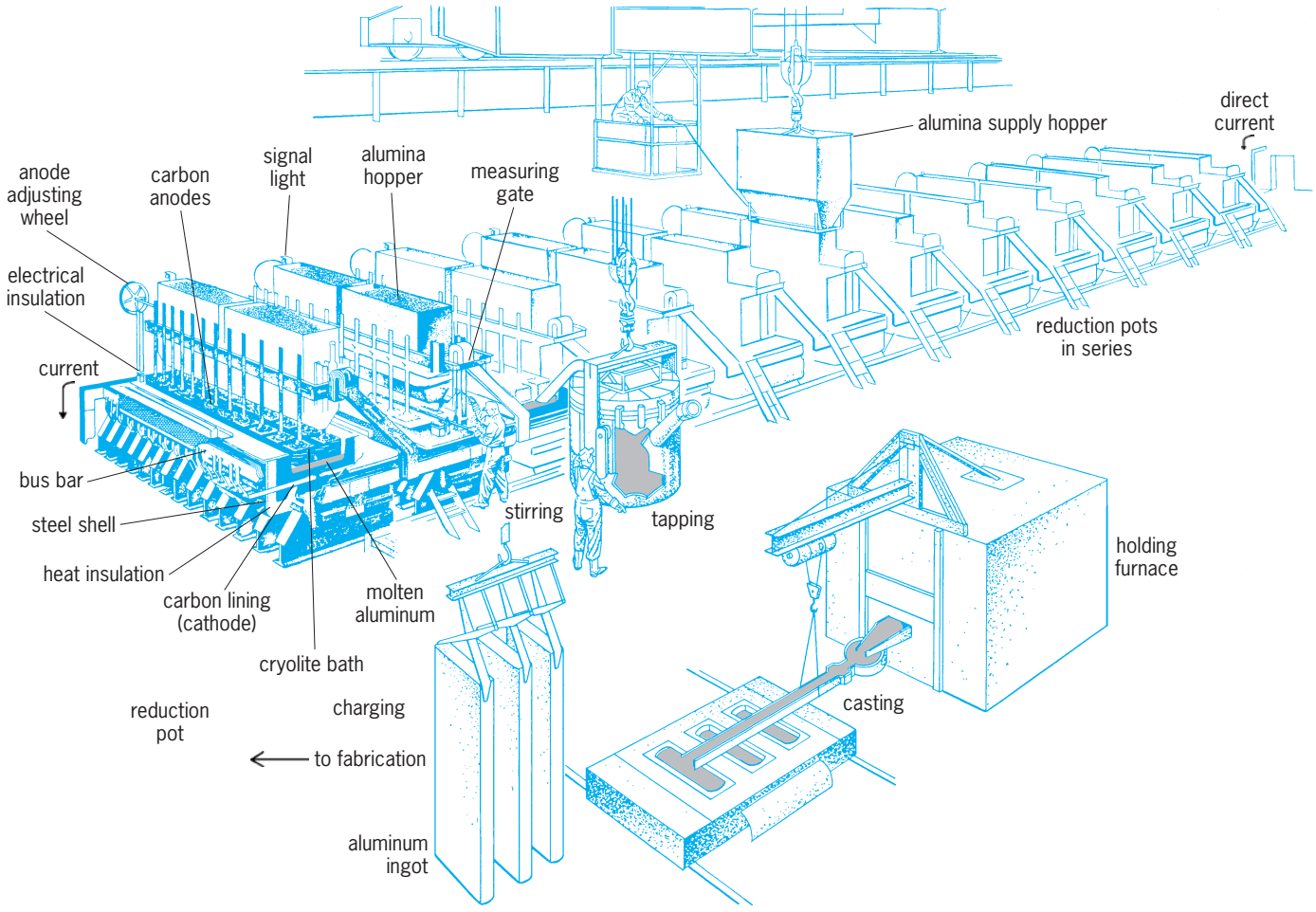


Fig. 3. Production of aluminum metal. (*Aluminum Company of America*)

See ELECTROMETALLURGY; PYROMETALLURGY, NONFERROUS.

Melting. In plants not adjacent to a smelter, it is necessary to remelt charges, usually consisting of mill scrap or returned scrap with enough primary metal to provide composition control. For casting ingot for fabrication, gas- or oil-fired reverberatory furnaces are commonly used. Molten metal is generally transferred from the melting furnace to a reverberatory holding furnace or to a ladle for subsequent casting. Metal for foundry use in producing sand, permanent mold, and die castings is usually remelted in gas- or oil-fired crucible furnaces, although in large-scale operations reverberatory furnaces may be used.

Scrap recycling has grown through the years to considerable economic importance, and provides a valuable source of aluminum at a much lower energy expenditure than for primary metal. For many years, secondary producers have purchased scrap and reclaimed it, generally into foundry ingot for remelting. Primary producers have also purchased large amounts of mill scrap, as well as discarded cans, and recycled it into ingot used to fabricate sheet for more cans. This closed-circuit type of recycling saves energy, helps to preserve the environment, and conserves metal to the greatest degree.

After remelting, the molten aluminum alloy must be treated to remove dissolved hydrogen, inclusions such as oxides formed during remelting or from the surface of the charge material, and undesirable trace elements, such as sodium, in some alloys. Classically, the practice has been to bubble chlorine or a mixture of nitrogen and chlorine through the melt by means of graphite tubes. Modern practice makes use of a wide variety of in-line systems in which the metal is treated by filtering or fluxing, or both, during transfer between the holding furnace and the casting unit. These processes are more efficient than furnace fluxing, and their use minimizes air pollution, improves metal quality, and saves production time.

Although many ingot casting methods have been used historically, today most ingots for fabricating are cast by the direct chill process or some modification thereof. This process is accomplished on a semicontinuous basis in vertical casting and on a continuous basis in horizontal casting. The direct chill process consists of pouring the metal into a short mold. The base of the mold is a separate platform that is lowered gradually into a pit while the metal is solidifying. The frozen outer shell of metal retains the still-liquid portion of the metal when the shell is past the mold wall. Water is used to cool the mold, and water impingement on the ingot shell also cools the ingot as it is lowered. Ingot length is limited only by the depth to which the platform can be lowered. In horizontal casting, the same principles are applied, but the mold and base are turned so that the base moves horizontally, removing the constraint of pit depth. Much longer ingot can be practically cast by the horizontal method. By sawing the ingot while casting is progressing, the process can be made continuous.

Fabrication methods. Aluminum alloys are fabricated by all the methods known to metalworking and are commercially available in the form of castings; in wrought forms produced by rolling, extruding, forging, or drawing; in compacted and sintered shapes made from powder; and in other monolithic forms made by a wide variety of processes. Some of the common fabrication processes are represented in **Fig. 4**. In addition, pieces produced separately can be assembled by common joining procedures to build up complex shapes and structures. *See* ALUMINUM ALLOYS; METAL FORMING.

Casting. In casting aluminum alloys, molten metal is poured, or forced by pressure, into a mold where it solidifies. The temperature is usually 1150–1400°F (620–760°C) depending on the alloy composition and on the type of casting. For die casting and permanent mold casting, the molds are metal and are used repeatedly. Other processes use expendable molds made of sand or plaster. *See* METAL CASTING.

Rolling. Rolling is the process of reducing material by passing it between pairs of rolls. Aluminum alloy rolled products include plate (0.250 in. or 0.6 cm or more thick), sheet (0.006–0.249 in. or 0.02–0.632 cm thick), and foil (less than 0.006 in. thick). The starting product is ingot, which may be up to 360 in. (914 cm) long, 72 in. (182 cm) wide, and 26 in. (66 cm) thick. Initial reduction is hot (800–1100°F, or 426–593°C, depending on alloy) and is done in a reversing mill. Subsequent rolling operations may be in multistand continuous mills or single-stand mills, and may be hot or cold depending on the thickness and properties required in the final product. Other aluminum alloy rolled products include rod and bar, which are passed through grooved rolls. *See* METAL ROLLING.

Extruding. In extruding aluminum alloys, an ingot or billet is forced by high pressure to flow from a container through a die opening to form an elongated shape or tube, usually at metal temperatures between 550 and 1050°F (287 and 565°C). Hydraulic presses with capacities of 500 to 14,000 tons (4.5×10^6 to 1.2×10^8 newtons) are used. The extrusion process is capable of producing sections with weights of a few ounces to more than 200 lb/ft (300 kg/m), with thicknesses from a few tenths of an inch to about 10 in. (25 cm), with circumscribing circle diameters of 0.25 in. (0.64 cm) to about 3 ft (0.9 m), and lengths in excess of 100 ft (30 m). Tubing in diameters of 0.25–33 in. (0.64–0.84 cm) and pipe up to 20 in. (50 cm) are also produced. Stock that is produced by extrusion is also fabricated to tubing by drawing and to forgings by subsequent metalworking operations. *See* EXTRUSION.

Forging. Forgings are produced by pressing or hammering aluminum alloy ingots or billets into simple rectangular or round shapes on flat dies and into complex forms in cavity dies. Hydraulic presses capable of forces up to 75,000 tons (6.60×10^8 N) and mechanical presses with capacities up to 16,000 tons are used. Forging hammers have ram weights of 500–110,000 lb (227–50,000 kg). Metal temperatures of 600–880°F (315–471°C) are generally used,

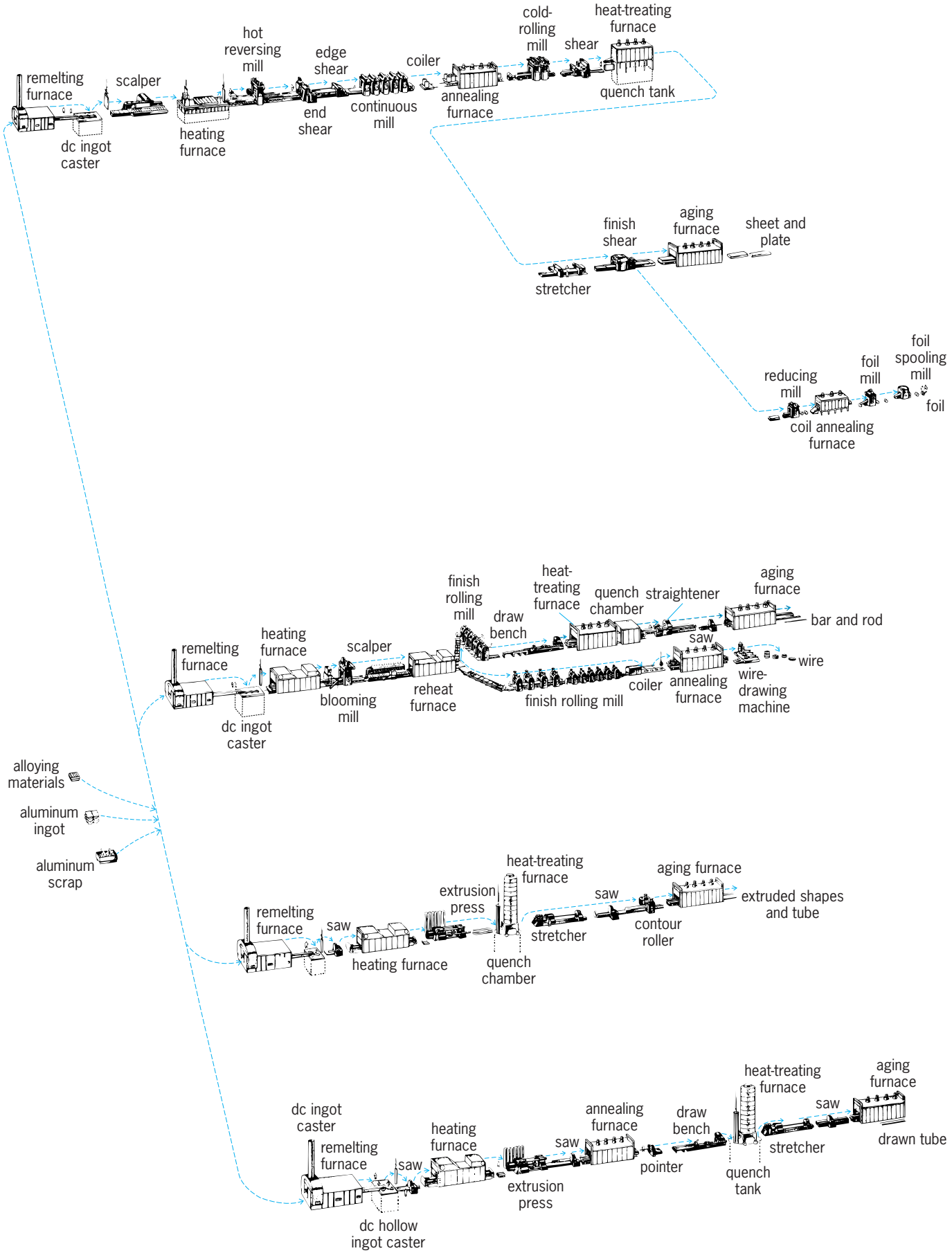


Fig. 4. Fabrication operations. (*Aluminum Company of America*)

depending on alloy and type of forging. *See* FORGING.

Drawing. Aluminum alloys can be made into deep-drawn shapes, of which pots and pans and beverage cans are common examples, by forcing sheet or foil into deep holes or cavities of the desired shape. Other drawn products, such as wire and tubing, are produced by pulling rolled or extruded stock through dies. *See* DRAWING OF METAL.

Compacting and sintering. Aluminum powders produced by atomizing can be compacted, either alone or mixed with powders of other elements, in dies of complex shape. The compact is then heated to a high temperature to promote additional bonding between the particles to result in products with desired shapes. *See* POWDER METALLURGY.

Other processes. Aluminum products are machined to final shape and dimension on lathes, joiners, routers, drill presses, shears, grinders, and many other high-speed metal-cutting and metalworking machines. Other fabricating techniques include impact extrusion, forming, stamping, embossing, coining, bending, rotary swaging, and cold heading.

Joining. Aluminum products formed into shapes by any of the above processes can be joined by welding, brazing, soldering, and adhesive bonding, and by mechanical means such as crimping, seaming, screwing, bolting, and riveting.

Products and uses. Applications in building and construction represent the largest single market of the aluminum industry. Millions of homes use aluminum doors, siding, windows, screening, and downspouts and gutters, which require little maintenance and provide a long life. Aluminum is also a major industrial building product. Excellent weather resistance makes it ideal for all climates and locations. Selected aluminum alloys are suitable near the seacoast, where salt spray may be deleterious to other metals. Colored and given additional protection by an electrolytic process called anodizing, aluminum appears in the curtain wall construction of many of the world's tallest buildings.

Transportation is the second largest market. In automobiles, aluminum is apparent in interior and exterior trim, grilles, wheels, and air conditioners. Another major use is in automatic transmissions. Some auto radiators, engine blocks, and body panels are made of aluminum to hold down weight and improve fuel economy. Aluminum is also found in rapid-transit car bodies, engine parts for diesel locomotives, rail freight and tank cars, bus and truck engines, forged truck wheels, cargo containers, and in highway signs, divider rails, and lighting standards. Aluminum's light weight and corrosion resistance are responsible for applications in pleasure-boat hulls, fishing boats, tanks for ships transporting liquefied natural gas, and deck houses for naval vessels.

In aerospace, aluminum is found in aircraft engines, frames, skins, landing gear, and interiors. Toughness, lightness, and heat-reflective characteristics have made it the preferred material for satellites and other spacecraft. *See* AIRFRAME; SPACECRAFT STRUCTURE.

In the packaging industry, aluminum helps in the preparation of foods and beverages and keeps them pure during distribution and storage. The fast-cooling, easy-opening, lightweight, and recyclable features of the all-aluminum can have resulted in the use of billions of aluminum beverage containers. Foil pouches and bags, twist-off closures, and easy-open ends revolutionized the food and beverage packaging industries. *See* FOOD ENGINEERING.

In electrical applications, aluminum wire and cable are major products. Underground electrical cables require large amounts of aluminum. Aluminum wiring is also used in residential, commercial, and industrial buildings. *See* TRANSMISSION LINES; WIRING.

Consumer use began before the 1900s with aluminum cooking utensils. In modern homes, aluminum appears as cooking foil, hardware, tools, portable appliances, air conditioners, freezers, and refrigerators.

There are hundreds of chemical uses of aluminum and aluminum compounds. Aluminum powder is used in paints, rocket fuels, and explosives, and as a chemical reductant.

Allen S. Russell

Bibliography. Aluminum Corporation of America, *Melting and Casting Aluminum*, 1995; G. S. Brady, H. R. Clauser, and J. A. Vaccari, *Materials Handbook*, 15th ed., 2002; J. E. Hatch (ed.), *Aluminum: Properties and Physical Metallurgy*, 1984; G. E. Totten and D. S. MacKenzie (eds.), *Handbook of Aluminum*, vol. 1: *Physical Metallurgy and Processes*, 2003.

Aluminum alloys

Substances formed by the addition of one or more elements, usually metals, to aluminum. The principal alloying elements in aluminum-base alloys are magnesium (Mg), silicon (Si), copper (Cu), zinc (Zn), and manganese (Mn). In wrought products, which constitute the greatest use of aluminum, the alloys are identified by four-digit numbers of the form NXXX, where the value of N denotes the alloy type and the principal alloying element(s) as follows: 1 (Al; at least 99% aluminum by weight), 2 (Cu), 3 (Mn), 4 (Si), 5 (Mg), 6 (Mg + Si), 7 (Zn), 8 (other). *See* COPPER; MAGNESIUM; MANGANESE; SILICON; ZINC.

Iron and silicon are commonly present as impurities in aluminum alloys, although the amounts may be controlled to achieve specific mechanical or physical properties. Minor amounts of other elements, such as chromium (Cr), zirconium (Zr), vanadium (V), lead (Pb), and bismuth (Bi), are added to specific alloys for special purposes. Titanium additions are frequently employed to produce a refined cast structure. *See* IRON; TITANIUM.

Aluminum-base alloys are generally prepared by making the alloying additions to molten aluminum, forming a liquid solution. As the alloy freezes, phase separation occurs to satisfy phase equilibria requirements and the decrease in solubility as the temperature is lowered. The resultant solidified structure consists of grains of aluminum-rich solid solution

TABLE 1. Nominal composition and forming processes for common wrought aluminum alloys

Alloy	Form*	Si	Cu	Mg	Mn	Zn	Other
1100	b–d	—	—	0.1	—	—	99.00 Al (min.)
1350	b–d	—	—	—	—	—	99.5 Al (min.)
2011	b–d	—	5.5	—	—	—	0.5 Pb, 0.5 Bi
2014	b–e	0.8	4.4	0.4	0.8	—	—
2024	b–e	—	4.4	1.5	0.6	—	—
2036	b	—	2.6	0.45	0.25	—	—
2219	b, e	—	6.3	—	0.3	—	0.1 V, 0.1 Zr
3003	b–d	—	—	—	1.2	—	—
3004	b–d	—	—	1.0	1.2	—	—
5052	b–d	—	—	2.5	—	—	0.25 Cr
5083	b, e	—	—	4.5	0.7	—	0.15 Cr
5086	b–d	—	—	4.0	0.5	—	0.15 Cr
5182	b	—	—	4.5	0.35	—	—
5657	b, c	—	—	0.8	—	—	—
6061	b–e	0.6	0.25	1.0	—	—	0.25 Cr
6063	b–e	0.4	—	0.7	—	—	—
7005	b–e	—	—	1.5	0.5	4.5	0.15 Cr, 0.14 Zr
7050	b, e	—	2.4	2.3	—	6.2	0.12 Zr
7075	b–e	—	1.6	2.5	—	5.6	0.25 Cr

*b = rolled; c = drawn; d = extruded; and e = forged.

and crystals of intermetallic compounds. Elements which lower the freezing point of aluminum, such as copper, magnesium, and silicon, tend to segregate to the portions of the grains which freeze last, such as the cell boundaries. Elements which raise the freezing point, such as chromium or titanium (Ti), segregate in the opposite manner. *See* EUTECTICS; SOLID SOLUTION.

A decrease in solubility with falling temperature also provides the basis for heat treatment of solid aluminum alloys. In this operation, the alloy is held for some time at a high temperature to promote dissolution of soluble phases and homogenization of the alloy by diffusion processes. The limiting temperature is the melting point of the lowest melting phase present. The time required depends both on temperature and on the distances over which diffusion must occur to achieve the desired degree of homogenization. Times of several hours can be necessary with coarse structures such as sand castings. Only a few minutes may be adequate, however, for rapidly heated thin sheet. *See* HEAT TREATMENT (METALLURGY).

The solution heat treatment is followed by a quenching operation in which the article is rapidly cooled, for example, by plunging it into cold or hot water or by the use of an air blast. This produces a supersaturated metallic solid solution that is thermodynamically unstable at room temperature. In several important alloy classes, such as 2XXX, 6XXX, and 7XXX, the supersaturated solution decomposes at room temperature to form fine, submicroscopic segregates or precipitates that are precursors to the equilibrium phases predicted by phase diagrams. The precipitation phenomenon, occurring over periods of days to years, produces substantial increases in strength. Additional precipitation strengthening can be obtained by heating the alloy at temperatures in the range 250–450°F (121–232°C), the time and temperature varying with alloy composition and the objectives with respect to mechanical properties and other characteristics such as corrosion resistance.

Wrought alloys. **Table 1** lists the nominal compositions of a number of commercially important wrought alloys and the type of products for which they are used. The alloys are generally classified in two broad categories depending upon their response to heat treatment as described in the preceding paragraph. Those having no or minor response are identified as non-heat-treatable alloys and include the 1XXX, 3XXX, and 5XXX compositions. Those that do respond are known as heat-treatable alloys and include the 2XXX, 6XXX, and 7XXX compositions.

Included in the non-heat-treatable group is 1350, a special grade used for electrical conductor products. Alloy 1100 is a grade of 99.0% minimum aluminum content with particular controls on iron (Fe), silicon, and copper contents, available in a variety of product forms such as sheet, foil, wire, rod, and tube, used for packaging, fin stock, and a variety of sheet metal applications. The manganese-containing alloy 3003 is a moderate-strength, very workable alloy for cooking utensils, tube, packaging, and lithographic sheet applications. The stronger aluminum-manganese-magnesium alloy 3004 is used for architectural applications, for storage tanks, and especially for drawn and ironed beer and beverage containers.

Alloy 5052 is a workable, corrosion-resistant aluminum-magnesium alloy for many metalworking and metal-forming purposes and marine applications. Where higher strength is required, the higher-Mg-content, weldable 5086 or 5083 alloys may be used. The latter is employed in construction of welded tanks for liquefied gas (cryogenic) transport and storage and for armor plate in military vehicles. Alloy 5182 is also a high-strength aluminum-magnesium alloy that is employed primarily in a highly strain-hardened condition for beverage can ends. Alloy 5657 is a lower-strength material

TABLE 2. Nominal composition and casting procedure for common aluminum casting alloys

Alloy	Form*	Si	Cu	Mg
413.0	D	12		
B†443.0	B, C	5.3	.15 max	
F†332.0	C	9.5	3.0	1.0
355.0	B, C	5.0	1.3	0.5
356.0	B, C	7.0		0.3
380.0	D	8.5	3.5	
390.0	C, D	17	4.5	0.55

* B = sand casting; C = permanent mold casting; and D = die casting.
† The letter indicates modifications of alloys of the same general composition or differences in impurity limits, from alloys having the same four-digit numerical designations.

produced with a bright anodized finish for automobile trim and other decorative applications.

The heat-treatable alloys 2014, 2024, and 7075 have high strengths and are employed in aircraft and other transportation applications. Modifications of these basic alloys, such as 2124 and 7475, were developed to provide increased fracture toughness. High toughness at high strength levels is achieved in thick-section products with 7050. Where elevated temperatures are involved, the 2XXX alloys are preferred. One such alloy, 2219, also has good toughness at cryogenic temperatures and is weldable. This alloy was prominently employed in the fuel and oxidizer tanks serving as the primary structure of the Saturn space vehicle boosters.

Alloy 6061 is used for structural applications where somewhat lower strengths are acceptable. For example, 6061 may be used for trailers, trucks, and other transportation applications. Alloy 6063 is a still-lower-strength, extrudable, heat-treatable alloy for furniture and architectural applications. The use of lead and bismuth in 2011 produces a heat-treatable, free-machining alloy for screw-machine products. Alloy 2036 is a moderate-strength alloy with good workability and formability, employed as body sheet in automotive applications.

Casting alloys. **Table 2** shows the nominal compositions of several important casting alloys. The major alloying addition is silicon, which improves the castability of aluminum and provides moderate strength. Other elements are added primarily to increase the tensile strength. Most die castings are made of alloy 413.0 or 380.0. Alloy 443.0 has been very popular in architectural work, while 355.0 and 356.0 are the principal alloys for sand castings. Number 390.0 is employed for die-cast automotive engine cylinder blocks, while alloy F332.0 is used for pistons for internal combustion engines.

Casting alloys are significant users of secondary metal (recovered from scrap for reuse). Thus, casting alloys usually contain minor amounts of a variety of elements; these do no harm as long as they are kept within certain limits. The use of secondary metal is also of increasing importance in wrought alloy manufacturing as producers take steps to reduce the energy that is required in producing fabricated aluminum products. *See* ALLOY; ALLOY STRUCTURES; METAL CASTING.

Allen S. Russell

High-strength alloys. Since aluminum comprises 70–80% of the weight of an airframe, metallurgists have been pursuing aluminum alloy development programs directed toward producing materials which would be characterized by stronger, stiffer, and lighter-weight properties. In addition, the titanium alloys of aircraft gas turbines are prime targets for replacement by lighter-weight alloys. Researchers have improved aluminum alloys by using techniques that add lithium and by using three processing technologies comprising powder metallurgy, mechanical alloying, and a process which involves blending aluminum alloy powders and silicon carbide fibers to form a composite.

Aluminum-lithium alloys. Lithium is the lightest metal in existence. For each weight percent of lithium added to aluminum, there is a corresponding decrease of 3% in the alloy's weight. Therefore, aluminum-lithium alloys offer the attractive property of low density. A second beneficial effect of lithium additions is the increase in elastic modulus (stiffness). Also, as the amount of lithium is increased, there is a corresponding increase in strength due to the presence of very small precipitates which act as strengthening agents to the aluminum. As the precipitates grow during heat treatment, the strength increases to a limit, then begins to decrease. Aluminum-lithium alloys therefore come under the classification of precipitation-strengthening alloys. Since the size and distribution of the precipitates can be controlled by heat treating, these alloys are also referred to as heat-treatable.

Since the addition of lithium to aluminum has also been shown to result in an alloy with unacceptable levels of ductility, a necessary engineering property for many aerospace applications, other elements such as copper, magnesium, and zirconium have been added to offset the loss in ductility. However, these alloy additions, particularly copper, increase the density. Consequently, schemes of alloy development have focused on balancing the various positive and negative attributes of the different elements, to arrive at a composition with suitable properties. **Figure 1** is a schematic relating yield strength and

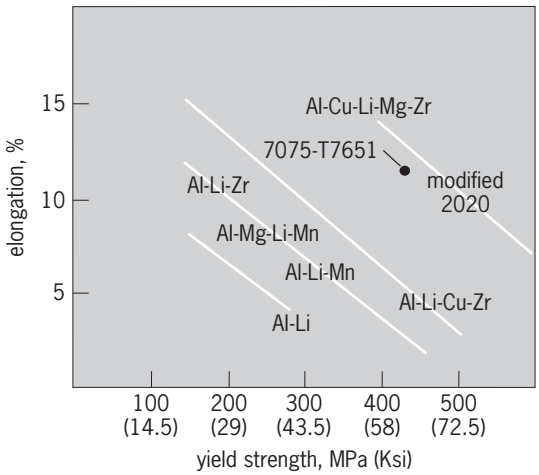


Fig. 1. Strength–ductility relationships for various Al-Li-X alloys. The alloy 7075-T7651 is included as a reference point.

elongation for a variety of aluminum-lithium alloys. The alloy 7075–T7651, commonly used for such applications, is included as a reference point. In order to be competitive, the new alloys should have strength-ductility relations equal to or better than 7075. *See* LITHIUM.

Powder metallurgy. As early as 1947, investigators showed interest in the development of high-modulus, high-temperature aluminum alloys. Elements such as iron, cobalt, and nickel have been added to aluminum in large quantities to produce alloys which have relatively stable structures at elevated temperatures and higher elastic moduli than conventional aluminum alloys.

The titanium in the fan sections of gas turbines has to withstand operating temperatures up to 450°F (235°C). The substitution of a suitable aluminum alloy for titanium could result in a 31% reduction in weight for that part. In order to develop aluminum alloys with sufficient strength to withstand the operating temperatures of a gas turbine and remain thermally stable, researchers have considered alloy additions to aluminum which move slowly (have low diffusivities) in it, so that the characteristics of that structure do not change with time at the service temperature. Since the elements which have low diffusivities in aluminum also have very limited solubilities, processes other than conventional ingot casting must be utilized.

An alternative for producing a candidate alloy is the powder metallurgy process, which is a rapid solidification process. Rapid solidification processing involves the transformation of finely divided liquid either in the form of droplets or thin sheets which become solid at high solidification rates (on the order of 10^4 K/s). Solidification occurs as heat is removed from the molten metal. The primary mechanisms of heat transfer can be divided into two broad categories—conductive and convective. Certain rapid-solidification-process forms such as ribbons and fibers are cooled primarily by conduction. The heat is transferred away by a substrate—the surface on which the liquid is cast. The substrate is generally made out of copper and can be either a rotating drum or a rotating wheel. On the other hand, another rapid-solidification-process form, powder, is cooled primarily by convection. The heat is transferred away from the fine liquid droplets in a gas stream. Regardless of the mechanisms of heat transfer, the resulting structures are more homogeneous and are on a finer scale than those formed by conventional ingot-casting techniques. The rapidly solidified particulates are then consolidated by forging or extrusion into a final shape. Even after consolidation, the various rapid-solidification-process forms are more homogeneous and have finer microstructures than a corresponding ingot-cast and worked product.

Research centered on utilizing the natural benefit of rapid-solidification processing has produced metastable microstructures which decompose and precipitate during consolidation and working. Thus elements which have limited solubility and low diffusivity in aluminum (such as iron, cobalt, and nickel—the transition metals), in conjunction with elements such as molybdenum or cerium, can be used to maximum advantage. However, the use of these alloying elements necessitates control of thermal history of the particulate both in its formation and upon subsequent processing. For example, in aluminum–transition metal alloys, the precipitates can form directly from the melt rather than from the solid. Consequently, the precipitates in these systems tend to be coarser than those in aluminum-lithium systems.

As in the case of aluminum-lithium alloys, the yield strength is affected primarily by the interparticle spacing and the fineness of the precipitates which form. The ductility and fracture toughness are affected by macroparticles that concentrate stress, thereby reducing the toughness. Since ductility and fracture toughness are important properties, rapid-solidification-process alloys will be limited until a control of the volume fraction of coarse particles can be achieved.

Rapid-solidification-process aluminum alloys containing iron and cerium appear to result in alloys which have refined particle-size distribution and improved high-temperature stability. These powder metallurgy alloys appear competitive with titanium up to 375°F (191°C). *See* METALLIC GLASSES; POWDER METALLURGY.

Mechanical alloying. This process circumvents the limitations of conventional ingot casting. Blends of powders are mixed in a ball mill. A drum is mounted horizontally and half-filled with steel balls and blends of elemental metal powders. As the drum rotates, the balls drop on the metal powder. The degree of homogeneity of the powder mixture is determined by the size of the particles in the mixture. If the powder's particles are too coarse, the different constituent elements in the blend will not interdiffuse during consolidation. Consequently, high rates of grinding must be used.

In a high-energy mill the particles of metal are flattened, fractured, and rewelded. Every time two balls collide, they trap powder particles between them. The impact creates atomically clean fresh surfaces. When these clean surfaces meet, they reweld. Since such surfaces readily oxidize, the milling operation is generally conducted in an inert atmosphere. During the initial stages of the mechanical alloying process, the particles are layered composites of the constituent powders and the starting constituents are easily identifiable within the composite particle. The composition and shape of the individual particles are different. As alloying progresses, the particles are further fractured and rewelded. The intimate mixture of the starting constituents decreases the diffusion distance, and the particle tends to become more homogeneous. Metastable and stable phases are beginning to form in the individual particles. Diffusion of the elements is accelerated by the strain introduced during the milling. During the final stage of processing, the individual particles became similar in composition and similar to one another. Completion of the process occurs when the tendency to fracture is balanced by the tendency to

TABLE 3. Data for several loading fractions of silicon carbide (SiC)

Tensile properties of SiC/2024 Al-T4 composites[*]				
Vol. % SiC:	0	15	20	25
Ultimate tensile strength, MPa (Ksi)[†]	400–434 (58–64)	NA[‡]	455–524 (66–76)	552–641 (80–93)
Elongation, GPa (Msi)[§]	73 (10.6)	89–97 (13–14)	97–117 (14–17)	117–151 (17–22)
Elongation to failure, %	19–20	1–2	1–2	1–2
Tensile properties of SiC/2024 Al composites, as extruded[¶]				
Vol. % SiC:	0	20		25
Ultimate tensile strength, MPa (Ksi)[†]	186 (27)	331–372 (48–54)		400–448 (58–65)
Elongation, GPa (Msi)[§]	73 (10.6)	97–117 (14–17)		117–151 (17–22)
Elongation to failure, %	20–22	1–2		1–2

[*]18 specimens. [†]Ksi = thousand pounds per square inch. [‡]Not available. [§]Msi = million pounds per square inch. [¶]20 specimens.

weld and the particle size distribution is constant. The relationship between the thickness and layers within the composite is schematically illustrated in **Fig. 2.**

Aluminum–silicon carbide composites. A composite is a material in which two or more constituents are combined to result in a material which has properties different from those of either constituent. Typical composites are from materials in which one of the components has very high strength and modulus and the other has high ductility. Their properties generally follow a rule of mixtures. For example, if elastic modulus is the property of interest, the elastic modulus of the composite is approximately the weighted sum of the elastic moduli of the constituents.

In one example, silicon carbide whiskers are mixed with aluminum alloy powder, compacted under pressure at elevated temperatures, and extruded or forged into a final product. The very high-modulus silicon carbide is incorporated in the ductile aluminum alloy matrix. The resulting properties depend on the volume fraction of silicon carbide, and to a large degree on the fabricating methods. The primary advantage of an aluminum–silicon carbide composite is the high elastic modulus and strength (**Table 3**). *See* ALUMINUM; COMPOSITE MATERIAL.

Tom H. Sanders, Jr.

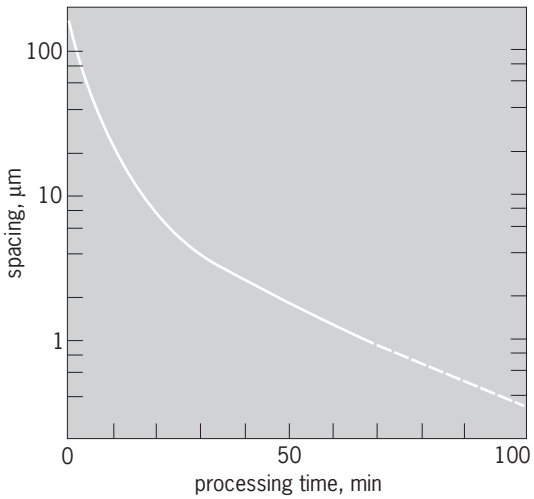


Fig. 2. Effect of processing time on the spacing between the layers of mechanically alloyed powder. The smaller the spacing, the more homogeneous the powder particle.

Bibliography. American Society for Metals, *Source Book on Selection and Fabrication of Aluminum Alloys*, 1978; A. P Divecha, S. G. Fishman, and S. D. Karmarkar, Silicon carbide reinforced aluminum: A formable composite, *J. Metals*, 33:12–17, September 1981; P. S. Gilman and J. S. Benjamin, Mechanical alloying, *Annu. Rev. Mater. Sci.*, 13: 279–300, 1983; F. King, *Aluminum and Its Alloys*, 1987; S. L. Langenbeck, Elevated temperature aluminum alloy development, *Interim Technical Report for Period April–September 1981*, U.S. Airforce Rep. LR 29977, October 1981; E. A. Starke, Jr., T. H. Sanders, Jr., and I. G. Palmer, New approaches to alloy development in the Al-Li system, *J. Metals*, 33:24–33, August 1981.

Alunite

A mineral of composition $KAl_3(SO_4)_2(OH)_6$. Alunite occurs in white to gray rhombohedral crystals or in fine-grained, compact masses. It has a hardness of 4 and a specific gravity of 2.6–2.8.

Alunite is produced by sulfurous vapors on acid volcanic rocks and also by sulfated meteoric waters affecting aluminous rocks. The mineral is generally found associated with quartz and kaolinite.

Alunite is used as a source of potash or for making alum. Alum has been manufactured from the well-known alunite deposits at Tolfa, near Civita Vecchia, Italy, since the mid-fifteenth century. In the United States alunite is widespread in the West. Large deposits occur in Mineral and Hindsdale counties, Colorado, and in Piute County, Utah. *See* ALUM; FERTILIZER; POTASSIUM.

Edward C. T. Chao

Alzheimer's disease

A disease of the nervous system characterized by a progressive dementia that leads to profound impairment in cognition and behavior. Dementia occurs in a number of brain diseases where the impairment in cognitive abilities represents a decline from prior levels of function and interferes with the ability to perform routine daily activities (for example, balancing a checkbook or remembering appointments). Alzheimer's disease is the most common form of dementia, affecting 5% of individuals over age 65. The onset of the dementia typically occurs in middle to late life, and the prevalence of the illness increases with advancing age to include 25–35% of individuals over age 85. *See* AGING.

Clinical characteristics. Memory loss, including difficulty in remembering recent events and learning new information, is typically the earliest clinical feature of Alzheimer's disease. As the illness progresses, memory of remote events and overlearned information (for example, date and place of birth) declines together with other cognitive abilities. Orientation to time (for example, date, day of week, or season) is often impaired, followed by a decline in orientation to personal information and immediate environment. Language deficits usually begin with difficulty in naming objects and in word-finding ability and gradually progress to a loss of all language skills. The ability to use visuospatial skills to draw simple figures and shapes is increasingly impaired. Declines in abstract reasoning, judgment, and problem solving interfere with the ability to perform mathematical calculations and to make appropriate decisions in social situations. In the later stages of Alzheimer's disease, there is increasing loss of cognitive function to the point where the individual is bedridden and requires full-time assistance with basic living skills (for example, eating and bathing). *See* MEMORY.

Although Alzheimer's disease typically has a slowly progressing course, the rate of cognitive decline is variable among individuals. The average duration from the onset of symptoms to death is 8–10 years. In individuals reaching the end stage of dementia, the cause of death is often related to secondary medical conditions (for example, respiratory infections). Behavioral disturbances that can accompany Alzheimer's disease include agitation, aggression, depressive mood, sleep disorder, and anxiety. Changes in personality, increased social withdrawal, and irritability are observed in some individuals. Delusions and hallucinations can occur. The onset of clinical symptoms can vary from middle to advanced age. Although a distinction between individuals with an early onset (65 years of age and younger) and with a late onset (older than 65) of dementia has been made, the same neuropathological disease process occurs in both groups.

Neuropathological features. The major neuropathological features of Alzheimer's disease include the presence of senile plaques, neurofibrillary tangles, and neuronal cell loss. Although the regional distribution of brain pathology varies among individuals, the areas commonly affected include the association cortical and limbic regions. Senile plaques are extracellular and contain a core of an amyloid beta peptide that is derived from the cleavage of a much larger beta amyloid precursor protein. Although the amyloid precursor protein occurs normally in neural cell membranes, the reason why individuals with Alzheimer's disease produce excessive amounts of the amyloid beta peptide is unknown. Neurofibrillary tangles represent intracellular pathology in which paired helical filaments accumulate within neurons. The basic component of neurofibrillary tangles is a hyperphosphorylated tau protein producing a highly insoluble cytoskeletal structure that is associated with cell death. Prominent loss of pyramidal neurons from cerebral cortical layers III and V has been observed in Alzheimer's disease. Other microscopic features include neuropil threads, Hirano bodies, granulovascular degeneration, and inflammation around plaque formations. The brain regions of the medial temporal lobes are commonly affected early in the course of Alzheimer's disease. In vivo neuroimaging studies that measure functional brain activity by assessing cerebral blood flow and metabolism have shown that the temporal, parietal, and frontal association cortices appear most consistently and severely affected in the early to middle stages of the illness.

Alzheimer's disease has been associated with dysfunction of several neurotransmitter systems. These neurochemical systems are responsible for the transfer of information between neurons via the synapse and are important for cognitive and behavioral functions. In Alzheimer's disease, disruption of information transfer between neurons can occur with loss of presynaptic elements and reduction of the recognition sites at which the neurotransmitter acts. It is also thought that loss of or abnormalities in the receptor-operated ion channels and second messenger systems within cells may be related to neuronal dysfunction in Alzheimer's disease. Deficits in cholinergic, serotonergic, noradrenergic, and peptidergic (for example, somatostatin) neurotransmitters have been demonstrated. Dysfunction of the cholinergic neurotransmitter system has been specifically implicated in the early occurrence of memory impairment in Alzheimer's disease, and it has been a target in the development of potential therapeutic agents. The deficiency in choline acetyltransferase, an enzyme important for the synthesis of acetylcholine, has been associated with dementia severity. The cortical projection of the cholinergic system is especially concentrated in key structures such as the hippocampus in the medial temporal lobes, brain structures that are thought to be important in aspects of learning and memory. *See* ACETYLCHOLINE; NEUROBIOLOGY; NORADRENERGIC SYSTEM.

Diagnosis. A definite diagnosis of Alzheimer's disease is made only by direct examination of brain tissue obtained at autopsy or by biopsy to determine the presence of senile plaques and neurofibrillary tangles. A clinical evaluation, however, can provide

a correct diagnosis in more than 80% of cases. The clinical diagnosis of Alzheimer's disease requires a thorough evaluation to exclude all other medical, neurological, and psychiatric causes of the observed decline in memory and other cognitive abilities. This evaluation begins with a clinical history obtained from the patient and caregivers concerning the onset and course of symptoms and whether there are other medical conditions, head injury, excessive use of alcohol, or use of medications that could cause or exacerbate the cognitive decline. Neurological and psychiatric examinations are performed to characterize the cognitive deficits and to evaluate the presence of other disorders, such as Parkinson's disease, Huntington's disease, and cerebrovascular disease, that can cause dementia. The presence of behavioral disturbances such as depression and psychosis that can compromise cognitive function must also be evaluated. Laboratory tests are used to screen for medical conditions, including diabetes, renal dysfunction, hypothyroidism, infection, and vitamin deficiencies. A detailed neuropsychological evaluation is made to quantify and describe the cognitive deficits and to assist in staging the severity of dementia. Neuroimaging studies of the brain are performed to help rule out the presence of stroke, brain tumor, or other central nervous system abnormalities.

Risk factors. Although the cause of Alzheimer's disease is unknown, a number of factors that increase the risk of developing this form of dementia have been identified. Age is the most prominent risk factor, with the prevalence of the illness increasing twofold for each decade of life after age 60. Research in molecular genetics has shown that Alzheimer's disease is etiologically heterogeneous. Gene mutations on several different chromosomes are associated with familial inherited forms of Alzheimer's disease. Mutations of the amyloid precursor protein gene located on chromosome 21 have been linked to early-onset Alzheimer's disease in a small number of families. Additionally, virtually all individuals with Down syndrome (where there is an extra copy of chromosome 21) develop the brain pathology of Alzheimer's disease by the fourth decade of life. Gene defects on chromosomes 1 and 14 have also been implicated in familial early-onset forms of the illness, and they are thought to account for the vast majority of familially inherited cases of Alzheimer's disease. In the more common sporadic, late-onset form of Alzheimer's, specific allele types of the apolipoprotein gene on chromosome 19 act as a susceptibility factor that influences the age of onset of the disease.

Treatment. A major strategy for the treatment of Alzheimer's disease has focused on the relation between memory impairment and dysfunction of the acetylcholine neurotransmitter system. Tetrahydroaminoacridine (Tacrine) and donepezil hydrochloride (Aricept) were approved by the U.S. Food and Drug Administration for the treatment of Alzheimer's disease. These medications serve to increase the amount of acetylcholine available in the synapse by preventing the breakdown of the neurotransmitter that naturally occurs with presence of the enzyme acetylcholinesterase. Other treatment strategies to delay or diminish the progression of Alzheimer's disease are being explored, including efforts to modulate multiple neurotransmitter systems, to reduce the accumulation of amyloid in the brain, and to diminish the inflammatory response. Behavioral and pharmacological interventions are also available to treat the specific behavioral disturbances that can occur in Alzheimer's disease, such as depression, agitation, and sleep disorder. Gene Alexander

Bibliography. M. M. Esiri and J. H. Morris (eds.), *The Neuropathology of Dementia*, Cambridge University Press, 1997; B. A. Lawlor (ed.), *Behavioral Complications in Alzheimer's Disease*, American Psychiatric Press, 1995; R. D. Terry, R. Katzman, and K. C. Bick (eds.), *Alzheimer's Disease*, Raven Press, 1994; W. Wasco and R. E. Tanzi (eds.), *Molecular Mechanisms of Dementia*, Humana Press, 1997.

Amalgam

An alloy of mercury. Practically all metals will form alloys or amalgams with mercury, with the notable exception of iron. Amalgams are used as dental materials, in the concentration of gold and silver from their ores, and as electrodes in various industrial and laboratory electrolytic processes.

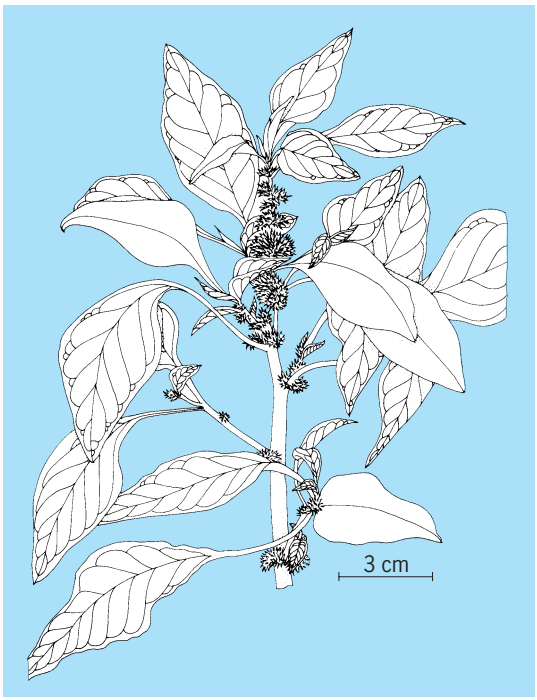
Amalgams used in dental work require the following composition: silver, 65% minimum; copper, 6% maximum; zinc, 2% maximum; and tin, 25% minimum. These amalgams are prepared by the dentist as needed, and harden within 3–5 min, but may be shaped by carving for 15 min or so. *See* ALLOY; GOLD; MERCURY (ELEMENT); SILVER. E. Eugene Weaver

Amaranth

An annual plant (seldom perennial) of the genus *Amaranthus* (family Amaranthaceae), distributed worldwide in warm and humid regions. Amaranths are botanically distinguished by their small chaffy flowers, arranged in dense, green or red, monoecious or dioecious inflorescences (spikes; see **illus**.), with zero to five perianth segments and two or three styles and stigmata, and by their dry membranous, indehiscent, one-seeded fruit. *See* FLOWER; FRUIT.

Physiological, genetic, and nutritional studies have revealed their potential economic value. Of particular interest are high rate of productivity as a rapidly growing summer crop, the large amounts of protein in both seed and leaf with high lysine, the overall high nutritional value, and the water use efficiency for the C_4 photosynthetic pathway. Amaranths are important in the culture, diet, and agricultural economy of the people of Mexico, Central and South America, Africa, and northern India. Genetic, ethnobotanical, and agronomic research has been undertaken to develop amaranths as an important food plant in modern agriculture.

Classification and species relationships. Up to 200 or more species of *Amaranthus* have been



Flowers on branch of *Amaranthus tricolor*.

described, of which nearly 50 are distributed in the New World. Some of the more widespread weedy species are red amaranth or pigweed (*A. hybridus*), thorny amaranth (*A. spinosus*), tumbleweed (*A. albus*), and wild beet (*A. hybridus* ssp. *hypochondriacus*); Joseph's coat (*A. tricolor*) and love-lies-bleeding (*A. caudatus*) are recognized for their ornamental spikes and foliage colors.

Two sections of the genus are *Amaranthotypus*, with large terminal inflorescence, five tepals and stamens, and fruit opening circularly; and *Blitopsis*, with axillary cymes, often two to four tepals and stamens, and fruits opening irregularly. The cultivated grain species (*Amaranthus cruentus, A. caudatus*, and *A. hypochondriacus* = *A. leucocarpus*) with ornamental and crop forms belong to the section *Amaranthotypus*, whereas two important cultivated vegetable species (*A. tricolor* and *A. lividus*) belong to the section *Blitopsis*. Several Asian natives, such as *A. tricolor, A. gangeticus*, and *A. melancholicus*, have been grown as pot herbs or ornamentals but never as a grain crop.

The cultivated grain species are largely pale-seeded, whereas the ornamentals and weedy forms are black-seeded, presumably an important trait used by early peoples in distinguishing them. Interspecific hybridization, reported in various regions, is considered to be a primary source of evolutionary diversity. The grain crop forms and their weedy relatives include both 32- and 34-chromosome species; much of the preliminary biosystematic and cytogenetic research has not yet provided conclusive species relationships. However, the weedy and cultivated species of *Amaranthotypus* are separable into two distinct groups on the basis of taxonomic and protein-variation studies.

Domestication and cultivation. Amaranths are of ancient New World origin and were important as grains in the Aztec, Incan, and Mayan civilizations where wheat and rice were not grown. Archeological and ethnobotanical evidence suggests that the earliest domestication took place in South America at least 4000 years ago. About A.D. 500, grain amaranths were a widely cultivated food crop in Mexico. Some of the ornamental forms moved from the New World to Africa and Asia in the early 1800s, and may have spread from India to the East Indies, China, Manchuria, and Polynesia as a leaf vegetable. Ceremonial sacrifice and use in rituals drew the attention of the Spaniards, who suppressed the Aztec culture and amaranth cultivation.

Amaranths are a widely grown vegetable in tropical regions, including southern India, West and Equatorial Africa, and New Guinea. Amaranths have been found to be the most popular vegetable in southern Benin, in West Africa. Varieties such as Chinese spinach, Ceylon spinach, and African spinach suggest use as a salad ingredient; however, somewhat high oxalic acid content make the leaves unpleasant for this purpose.

Food value. Leaf amaranths are rich in provitamin A, vitamin C, iron, calcium, and protein, with lysine constituting as much as 5.9% of protein (equal to soymeal, and more than some of the best maize strains); glutamic acid constitutes as much as 10.8% of protein in *A. edulis*. Amaranth grains, with starch of very high quality and quantity, high protein (up to 17%), and high digestibility, rate better in nutritive value than all other cereals. They have been used in Mexico in modern times to make alegrias (cakes of popped seed), flour (of popped seed) used in the preparation of pinole, and paste used in the preparation of chuale; the black-seeded and variegated forms have been used to make amaranth milk (atole). In India, amaranth seed is used in alegrias and in confectioneries (flour) and as a cereal (popped seed). *See* CARYOPHYLLALES.

Subodh K. Jain

Bibliography. R. Bressani, *Amaranth Newsletters*, Arch. Latino Americanos de Nutricion, Guatemala, 1984–1997; G. J. H. Grubben, *The Cultivation of Amaranth as a Tropical Leaf Vegetable*, Roy. Trop. Inst. Commun. 67, 1976; S. K. Jain, P. A. Kulakow, and I. Peters, Genetics and breeding of grain amaranth: Some research issues and findings, *Proceedings of the 3d Amaranth Conference, 1984*, pp. 174–191, 1986; National Research Council, *Amaranth*: *Modern Prospects for an Ancient Crop*, 1984; R. M. Saunders and R. Becker, *Amaranthus*: A potential food and feed resource, *Adv. Cereal Sci. Technol.*, 6:358–396, 1984.

Amateur radio

Two-way radio communications by individuals as a leisure-time activity. Amateur, or "ham," radio is defined by international treaty as a "service of self-training, intercommunications, and technical investigation carried on by amateurs; that is, by duly

authorized persons interested in radio technique solely with a personal aim and without pecuniary interest."

Activities. The government allows amateur operators many privileges because the hobby is partially based on service to the general public, and hams can be relied on to assist during emergencies. Groups of amateur operators meet annually to practice handling emergency communications in the field and to compete against other groups nationwide in performing certain emergency-related tasks. Amateur operators may set up warning and relief networks during the hurricane and tornado seasons, and handle communication when phone lines or cellular links are damaged by disasters.

In addition to public service activities, amateurs enjoy many recreational activities, including DXing (where the objective is to contact amateurs in as many foreign countries as possible), contesting (where the amateurs compete for the maximum number of contacts in a given time span), and foxhunting (where the objective is to use radio skills to locate a hidden transmitter).

Since high-frequency (HF) signals (below 30 MHz) are reflected from the Kennelly-Heaviside ionosphere layers, amateurs are commonly able to carry out international communication. Very high frequencies (VHF; 30–300 MHz) and ultrahigh frequencies (UHF; 300–3000 MHz) have been subject to exploration by amateurs at the leading edge of communications technology. Amateurs bounce signals off the Moon or ionized meteor trails, and communicate through amateur operator-built earth satellites called OSCAR (orbiting satellites carrying amateur radio).

Equipment. For reliable, versatile, long-distance communications, some operators use large, directional beam antennas, but many have much simpler wire antennas strung between trees. Many operators have inexpensive equipment consisting of a wire antenna and just one transceiver (a radio that both receives and transmits) that functions on five or more frequency bands. *See* ANTENNA (ELECTROMAGNETISM).

Unlike most radio services, amateurs may design and build their equipment as long as it meets FCC requirements for spectral purity and bandwidth. They may assemble it from kits or may purchase it ready to go. As a result of the proliferation of commercial equipment, operators generally build only accessory items or experiment with modifications to equipment they have purchased. A few operators design equipment for frequency bands where commercial choices are limited, or experiment with new devices originally designed for the commercial frequencies.

Privileges. In the United States, the Federal Communications Commission (FCC) issues five classes of license, progressively allowing more frequencies and greater privileges. The first license level is the Novice Class, followed, in order of increasing privileges, by the Technician Class, General Class, Advanced Class, and Amateur Extra Class.

In 1991 the FCC eliminated the Morse code requirement from the Technician Class license, which shortly thereafter became the license of choice for most entry-level amateurs. Technician Class licensees have operating privileges on all amateur radio frequencies above 30 MHz.

Novice Class licensees may use voice, Morse code, and digital communications on a limited number of frequencies that allow worldwide HF voice communications and short-distance VHF voice communications on frequencies where readily available handheld radios allow mobility and portability.

Higher-class licensees have more frequency privileges and are permitted 1500 W peak envelope power output on most bands. Higher-class licensees may use any of the following modes: code (CW); amplitude-modulated (AM) and single-sideband (SSB) suppressed-carrier voice; frequency modulation (FM); radioteletype (RTTY), using either the Baudot, ASCII, or AMTOR (amateur teletype over radio) digital codes; facsimile; television; or pulse modulation. *See* AMPLITUDE-MODULATION RADIO; FREQUENCY-MODULATION RADIO; PULSE MODULATION; SINGLE SIDEBAND; TELETYPEWRITER; TELEVISION.

Amateur operators with General Class and higher licenses can communicate on portions of or all of the following frequency bands, depending on their specific license level:

1.8–2.0 MHz	1240–1300
3.5–4.0	2300–2310
7.0–7.3	2390–2450
10.10–10.15	3300–3500
14.00–14.35	5650–5925
18.068–18.168	10,000–10,500
21.000–21.450	24,000–25,250
24.890–24.990	48,000–50,000
28.0–29.7	71,000–76,000
50.0–54.0	165,000–170,000
144.0–148.0	240,000–250,000
222–225	Others above
420–450	300,000 MHz
902–928	

See RADIO SPECTRUM ALLOCATIONS.

Requirements. Questions found on each license class examination involve basic radio theory, rules that the FCC requires amateur radio operators to abide by, and a knowledge of international Morse code (except the Technician Class license). In each license class exam the questions get progressively more difficult to cover topics that befit the accompanying privileges of that license. For example, because the General Class license allows RTTY, there are questions covering RTTY on that exam but not on the Novice Class exam.

The Novice Class license requires the ability to send and receive Morse code at the rate of five words per minute (WPM); the General Class license requires 13 WPM; and the Amateur Extra, 20 WPM.

Technical developments. The open policies toward amateur radio by the government have made pioneering possible in several areas. In 1923, for instance, amateurs led the way in the use of short

waves after an amateur in Connecticut talked with a ham in France on the wavelength of 110 m (360 ft). The accepted theory had been that only wavelengths above 200 m (660 ft) were really useful for reliable communications.

By the early 1930s communications between points on opposite sides of the globe had become commonplace on wavelengths of 80–20 m (264–66 ft; frequencies of 3.5–14.4 MHz). Many hams then turned their attention to higher frequencies. Again amateurs shattered previous conceptions, especially those concerning so-called line-of-sight limitations on communications at 56 MHz and above. These pioneers discovered means of radio propagation which reached far beyond the usual horizon: reflections from the aurora borealis, from meteor trails, from sporadic E layers (patches of ionized particles about 70 mi or 110 km above the Earth), and from the Moon; bending of radio waves through layers of stable air; and a phenomenon called transequatorial scatter, by which stations on one side of the Equator may communicate with stations on the other side over distances of more than 1000 mi (1600 km) at times when such communication would otherwise be considered impossible. This mode and several of the others were subjects of an International Geophysical Year study undertaken by amateurs through the American Radio Relay League. *See* RADIO-WAVE PROPAGATION; TROPOSPHERIC SCATTER.

Other technical accomplishments and experimentation by amateurs include computer communications multiplexed on common radio channels, or packet radio; computer station control and interfacing as a result of the popularity and availability of personal computers; and spread spectrum techniques, enabling hundreds of simultaneous conversations to be carried on one frequency band without any mutual interference. Amateurs have installed relay repeater stations on mountaintops in many areas and use them to extend the range of communication on the VHF and UHF bands from tens to hundreds of miles. Under the direction of the Radio Amateur Satellite Corporation (AMSAT), amateurs have built and launched as secondary payloads on scheduled space shots a series of highly advanced satellites, some into low earth orbit and others into high elliptical orbits, containing refirable hydrazine engines, master on-board computers, and high-power VHF-UHF transponders. These satellites act as relay stations, again increasing the range of possible communications on the frequencies of the transponders. *See* SPREAD SPECTRUM COMMUNICATION.

Associations. The developments in the hobby over the years required a clearinghouse, an information exchange, which is embodied in the American Radio Relay League (ARRL), an association of radio amateurs in North America founded in 1914. It publishes a monthly magazine, a handbook, and other publications and operator aids covering various aspects of the hobby. ARRL also serves to represent amateurs in government regulatory matters; presents new technical developments and sponsors operating contests and other activities; and organizes networks to handle messages and coordinates emergency communications training. The League is also Secretariat for the International Amateur Radio Union, composed of over 125 similar national radio societies around the world. *See* RADIO.

Steve Mansfield

Bibliography. American Radio Relay League: *Advanced Class License Manual, Extra Class License Manual, The FCC Rule Book, Radio Amateur's Handbook, General Class License Manual,Now You're Talking, Radio Amateur's Handbook*, all annually.

Amber

Most commonly, a generic name for all fossil resins, although it has been restricted by some to refer only to succinite, the mineralogical species of fossil resin making up most of the Baltic Coast deposits. Resins generally are complex mixtures of mono-, sesqui-, di-, and triterpenoids; however, some resins contain aromatic phenols or are even predominantly composed of these compounds. Resins are synthesized in appreciable quantity by about 10% of present-day plant families. Among the plants, primarily trees, that produce copious amounts of resin that may fossilize to become amber, about two-thirds are tropical or subtropical. Members of the families Pinaceae, Araucariaceae, and Taxodiaceae are the most prominent copious producers among the conifers, and Leguminosae, Dipterocarpaceae, Burseraceae, and Hamamelidaceae among the angiosperms. *See* RESIN.

Fossilization. Under natural forest conditions resins harden with varying degrees of rapidity, and can become fossilized if they have the requisite chemical composition for polymerization, are sufficiently stable to withstand both oxidative and microbial degradation, and appropriate depositional conditions are available. The fossilization process is still not completely understood but appears primarily to involve polymerization, with some modification of the original constituents, especially due to the temperature during the maturation of the resin in the sediments.

Occurrence. Although ambers occur throughout the world in deposits from Carboniferous to Pleistocene in age, they have been reported most commonly from Cretaceous and Tertiary strata and often are associated with coal or lignites. Amber may contain beautifully preserved insects, spiders, flowers, leaves, and even small animals (**Fig. 1**). The most extensively studied deposits are those from the Baltic Coast, Alaska, Canada, North American Atlantic coast, southeast Asia, Dominican Republic, and Mexico (**Fig. 2**).

Uses. Since the earliest stages of human social development, ambers have had esthetic appeal, have been used to ward off evil powers, and have been thought to cure certain illnesses. The attribution of these special powers to amber may result partially from the negative electrical properties exhibited when ambers are rubbed. Thales recognized

Fig. 1. ***Pseudosphegina carpenteri*** **(Diptera), syrphid fly in Baltic amber, Oligocene. (*Courtesy of F. M. Carpenter*)**

these properties and, in fact, the term electricity evolved from *electron,* the Greek name for amber. *See* ELECTRIC CHARGE; ELECTRICITY.

Carvings and beads of Baltic amber found in caves and tombs indicate its significance in human activities since the Stone Age. Amber provided a distinctive and imperishable barter item for trade routes that crossed Europe from the Baltic to the Adriatic and Black seas to the Bronze and Iron ages as well as during Roman and Greek periods. When used for jewelry, it usually is transparent yellow, reddish-brown, or "amber" color. Translucent or semitranslucent amber is used for pipe stems, decorating small boxes, and other ornamental purposes. Specific gravity varies from 1.05 to 1.10, and hardness from 1 to 3 on Mohs scale.

Large amounts of amber occur in some coal deposits, which are being modified for some industrial uses, such as jet fuel. Furthermore, fossil resins, probably from members of the Dipterocarpaceae, have made marked contributions to petroleum in numerous southeast Asian sites.

Chemistry and origin. At one time, chemical studies of amber were mineralogically oriented because the purpose was to describe and classify amber as a semiprecious gem. However, phytochemical studies comparing fossil and present-day resins, employing such techniques as infrared spectroscopy, x-ray diffraction, nuclear magnetic resonance, and pyrolysis gas chromatography–mass spectrometry are providing information regarding the botanical origins of ambers. For example, amber from Chiapas, Mexico, and from the Dominican Republic has been shown to be derived from the leguminous genus *Hymenaea*; that from Sumatra from *Shorea* (Dipterocarpaceae); that from Ecuador from members of the Burseraceae; that from Australia and New Zealand

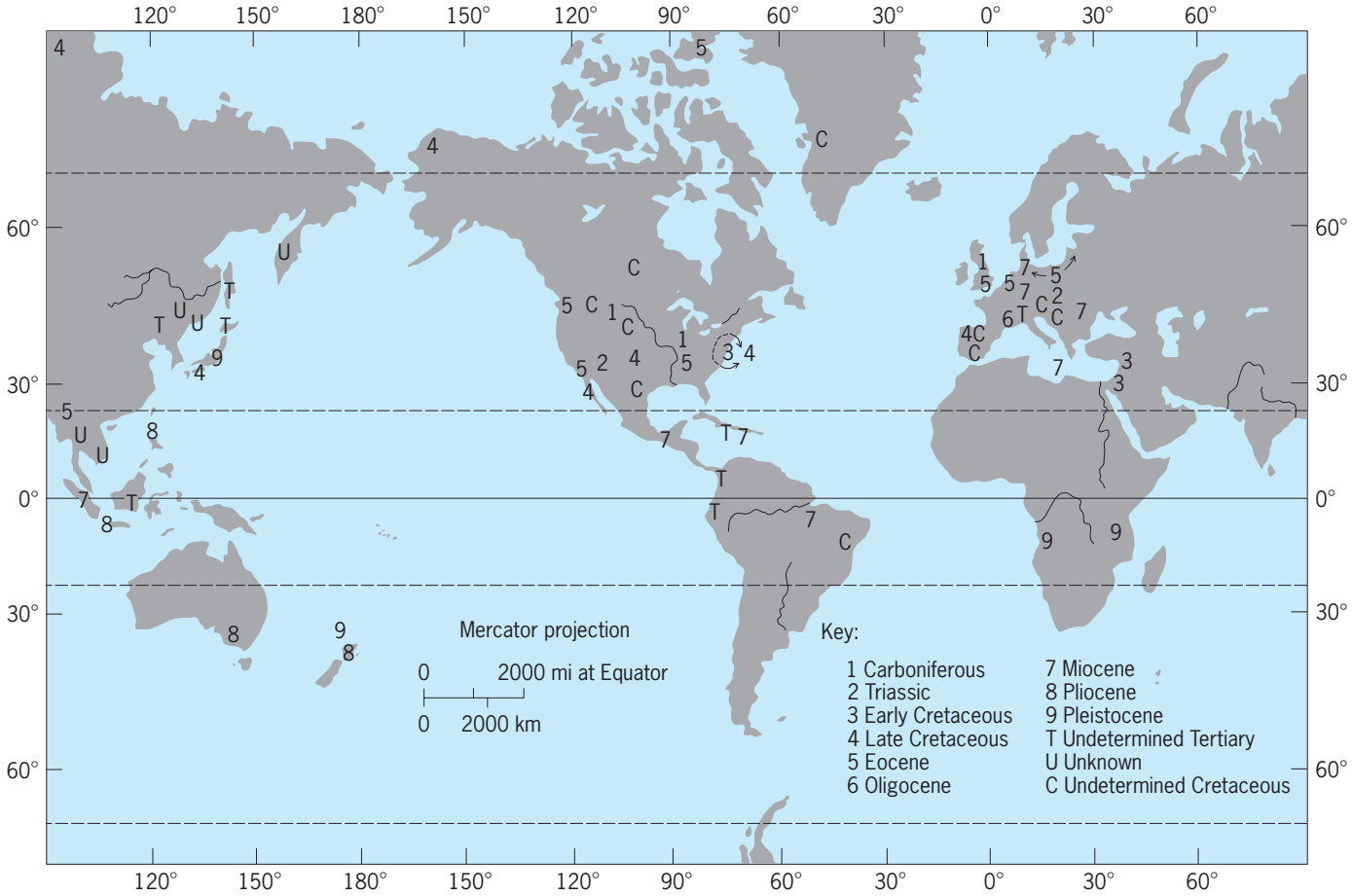


Fig. 2. Important amber deposits of the world. (*After J. H. Langenheim, Amber: A botanical inquiry, Science, 163:1157–1169, 1969*)

from *Agathis* (Araucariaceae); and that from Alaska probably from genera, such as *Metasequoia*, in the Taxodiaceae. The origin of the amber from the extensive Baltic deposits remains a mystery as the chemistry of the resin is more closely related to *Agathis* than to *Pinus* (Pinaceae), which early was considered to be the source from parts of the plant found in the amber. Then *Agathis* was considered because it has many chemical similarities to Baltic amber; however, it lacks succinic acid, a characteristic component, as well as plant remains of it in the amber. Recently amber from a Canadian Arctic site that contains succinic acid has been found within cone scales of *Pseudostrobus* (Pinaceae). As with *Agathis*, no parts of *Pseudostrobus* have been found in the amber. Although some chemical aspects differ from Baltic amber, the presence of succinic acid in the *Pseudostrobus* amber may provide a clue for future searches regarding the enigmatic source of Baltic amber. The predominantly tropical or subtropical occurrence of amber-producing plants through geologic time has led to evolutionary studies of the natural purpose of resins and their possible defensive role for trees against injury and disease inflicted by the high diversity of insects and fungi in tropical environments. *See* ARCHEOLOGICAL CHEMISTRY.

Jean H. Langenheim

Bibliography. D. A. Grimaldi, *Amber: Window to the Past*, Harry Abrams, New York, 1996; J. H. Langenheim, Biology of amber-producing trees: Focus on case studies of *Hymenaea* and *Agathis*, in K. B. Anderson and J. C. Crelling (eds.), *Amber, Retinite and Fossil Resins*, ACS Ser., no. 617, 1995; G. O. Poinar, Jr., and R. Poinar, *The Amber Forest*, Princeton University Press, 1999; P. C. Rice, *Amber: The Golden Gem of the Ages*, Kosciusko Foundation, New York, 1987.

Ambergris

A fatty substance formed in the intestinal tract of the sperm whale (*Physeter catodon*). There is a question as to whether the origin of ambergris is normal or pathological, but it does serve as protection from the horny indigestible portions of the squid and cuttlefish that constitute much of the whale's diet.

Ambergris contains acids, alkaloids, and a fatty substance, its main constituent, called ambrein. Although fresh ambergris is soft and black and has an offensive odor, it hardens into pleasantly fragrant gray or yellow masses when exposed to the air, sun, and sea. Being lighter than water, it is found in lumps, weighing from $^1/_2$ oz (15 g) to over 100 lb (45 kg) floating on tropical seas or cast up on the shores. It is also gathered directly from the abdomens of dead or captured whales when they are slaughtered. Collecting grounds for ambergris are principally on the shores of China, Japan, Africa, the Americas, tropical islands, and the Bahamas.

Known and highly prized since the earliest times, ambergris was used chiefly for medicinal purposes. In modern times this rare commodity is valued in the manufacture of perfumes. The ambergris is ground and used in the form of a tincture, dissolved in a dilute solution of alcohol, which when added to perfume acts as a fixative, increasing the duration of the fragrance while adding its own sweet, earthy scent.

Sybil P. Parker

Amblygonite

A lithium aluminum phosphate mineral of basic formula $LiAl(PO_4)(F)$. The structure of amblygonite consists of phosphate (PO_4) groups of tetrahedra and AlO_6 groups of octahedra. Each PO_4 tetrahedron is connected to an AlO_6 octahedron. Corner-sharing octahedra form zig-zag chains along the b axis. Lithium (Li) is in fivefold coordination, and lies between the PO_4 tetrahedra and nearest AlO_6 octahedra. The dominant substitution in this mineral structure is hydroxyl (OH) for fluorine (F). This substitution gives rise to the amblygonite-montebrasite [$LiAlPO_4(OH)$] solid solution series. When OH is greater than F, the mineral is known as montebrasite. Appreciable amounts of sodium substitute for lithium in the five-coordinated polyhedra. The sodium-rich varieties such as natromontebrasite and hedronite are rare.

Amblygonite crystallizes in the triclinic system. Its color is commonly white or gray with tints of blue, green, and yellow. Amblygonite is transparent to translucent and has a vitreous to pearly luster. It often exhibits polysynthetic twinning on both a hand specimen and a microscopic scale. Two cleavage directions are prominent. Cleavage fragments of amblygonite-montebrasite may be confused with potassium feldspar, but they are distinguished by their higher density (specific gravity = 3.0–3.1) and cleavage angles that are not at 90°.

Members of the amblygonite-montebrasite series typically occur as coarse, cleavable nodules in the inner zones and quartz cores of zoned lithium-rich, granitic pegmatites. Quartz, spodumene, petalite, and feldspar are the most common minerals associated with amblygonite. Minerals such as lepidolite, beryl, cassiterite, tourmaline, and niobium-tantalum oxides are commonly associated with amblygonite-bearing mineral assemblages in smaller abundances. These large masses of amblygonite-montebrasite are magmatic in origin and generally contain 4–7 wt % fluorine. Clear, equant crystals are rare and occur in vugs in pegmatites. Crystals associated with vugs and altered (secondary) magmatic amblygonite contain 0.3–4 wt % fluorine. A single mass of amblygonite with dimensions of 7.62 m × 2.44 m × 1.83 m (25 ft × 8 ft × 6 ft) has been documented in the Hugo granitic pegmatite in the Black Hills of South Dakota. Larger masses have been reported. The best-known occurrences of amblygonite are in Montebras, France; the Black Hills of South Dakota; the White Picacho District in Arizona; pegmatite districts in Maine; the Tanco pegmatite in Manitoba, Canada; and Portland, Connecticut. While

amblygonite has been mined as an ore of lithium, it is not a major ore. *See* PHOSPHATE MINERALS.

Charles K. Shearer

Bibliography. C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy,* 21st ed., 1999; P. B. Moore, *Pegmatite Minerals of P (V) and B (III): Granitic Pegmatities in Science and Industry,* 1982.

Amblypygi

The tailless whip scorpions or whip spiders, an order of the class Arachnida. There are about 80 species in the tropics and subtropics. They are flattened, red to brown, and range from 5 to 45 mm (0.2 to 1.8 in.) in body length. All are nocturnal, hiding under bark or stones, among leaves, or in caves during the day.

The cephalothorax (prosoma) and segmented abdomen (opisthosoma) are narrowly attached. There is no tail. The cephalothoracic shield (carapace) has three simple eyes on each side and a pair of eyes anterior. The chelicerae are like the jaws of a spider. The raptorial pedipalps are strong and have long, hard spines used to surround and crush insect prey. The long, whiplike first legs are sense organs used to locate prey. There are booklungs on the underside of the abdomen. There are no venom or repellent glands.

During courtship the male uses his long first legs to stroke the female; he may tremble while stroking. After some time he deposits a spermatophore (a package of sperm) on the ground. He pulls or guides the female over the spermatophore by using his chelicerae or first legs. The female picks up the spermatophore with her gonopore. Up to 60 eggs are laid in a flexible sac secreted by the female reproductive organs; the sac remains attached to the underside of the female's abdomen. On hatching, the young climb up on their mother's abdomen and are carried until their next molt. *See* ARACHNIDA.

H. W. Levi

Ameba

Any protozoon moving by means of protoplasmic flow. In their entirety, the ameboid protozoa include naked amebas, those enclosed within a shell or test, as well as more highly developed representatives such as the heliozoians, radiolarians, and foraminiferans. Ameboid movement is accomplished by pseudopods—cellular extensions which channel the flow of protoplasm. Pseudopods take varied forms and help distinguish among the different groups. A lobe-shaped extension or lobopod is perhaps the simplest type of pseudopod. The shapelessness and plasticity of these locomotory organelles impart an asymmetric, continually changing aspect to the organism. Other, more developed, representatives have pseudopodial extensions containing fibrous supporting elements (axopods) or forming an extensive network of anastomosing channels (reticulopods). Though involved in locomotion, these organelles are also functional in phagocytosis—the trapping and ingesting of food organisms (usually bacteria, algae, or other protozoa) or detritus. *See* FORAMINIFERIDA; HELIOZOIA; PHAGOCYTOSIS; RADIOLARIA.

Amebas are found in a variety of habitats, including fresh-water and marine environments, soil, and as symbionts and parasites in body cavities and tissues of vertebrates and invertebrates. Because of their manner of locomotion, amebas typically occur on surfaces, such as the bottom of a pond, on submerged vegetation, or floating debris. In soil, they are a significant component of the microfauna, feeding extensively on bacteria and small fungi. Amebas in marine habitats may be found as planktonic forms adapted for floating at the surface (having oil droplets to increase bouyancy and projections to increase surface area), where they feed upon bacteria, algae, and other protozoa. Several species of amebas may be found in the human intestinal tract as harmless commensals (for example, *Entamoeba coli*) or as important parasites responsible for amebic dysentery (*E. histolytica*).

Amebas range from small soil organisms, such as *Acanthamoeba* (20 micrometers), to the large fresh-water forms *Amoeba proteus* (600 μm; see **illus.**) and *Pelomyxa* (1 mm, or more). Some types, such as *Amoeba,* are uninucleate; others are multinucleate. Reproduction is by mitosis with nuclear division preceding cytoplasmic division to produce two daughters. Multinucleate forms have more unusual patterns of division, since nuclear division is not immediately nor necessarily followed by cytoplasmic division. Transformation of the actively feeding ameba into a dormant cyst occurs in many species, particularly those found in soil or as symbionts. The resting stages allow survival over periods of desiccation, food scarcity, or transmission between hosts. *See* REPRODUCTION (ANIMAL).

Much attention has focused on the mechanism of ameboid movement in *Amoeba proteus,* a favorite research species because of its large size and ease of maintenance in the laboratory. Under the light microscope, the organism is seen to contain a granular endoplasmic core in a sol (fluid) state, enclosed by a gel-like ectoplasm. Endoplasm flows forward as the organism progresses; as it reaches the front end of the advancing pseudopod, the endoplasm is deflected sideways and rearward, where it converts to ectoplasmic gel. This pattern has been termed fountain streaming. At the posterior end of the organism, the reverse occurs, with ectoplasm reverting to a forward-flowing endoplasm. The ameba resembles an ectoplasmic cylinder enclosing a fluid endoplasmic core. It is generally accepted that movement is brought about by the contraction of protoplasm at the anterior end of the pseudopod, pulling the ameba along the substrate. The contraction involves cytoplasmic microfilaments, adenosine triphosphate (ATP) as the energy source, and divalent

Phase-contrast photomicrograph of *Amoeba proteus*, a large fresh-water ameba. The organism is seen moving by means of a single lobose pseudopod.

cations, for example, Ca^{2+} and Mg^{2+}. *See* CELL MOTILITY.

Cytoplasm of *Amoeba* contains, besides food vacuoles, a contractile vacuole functioning for osmoregulation—a microkidney, which expells water. Energy for this, as well as other functions, is derived from numerous mitochondria present in the cytoplasm, which may also contain crystals and endosymbiotic bacteria. *See* MITOCHONDRIA; OSMOREGULATORY MECHANISMS; VACUOLE.

Ameboid protozoa are believed to have evolved from flagellated protozoan ancestors through loss of the flagellar apparatus. Some ameboid organisms, the ameboflagellates, can change from creeping amebas to swimming flagellates during their life cycle, providing support for this evolutionary pathway. Among certain specialized ameboid protozoa, production of flagellated gametes prior to sexual reproduction may occur. *See* AMOEBIDA; PROTOZOA.

Frederick L. Schuster

Americium

A chemical element, symbol Am, atomic number 95. The isotope ^{241}Am is an alpha emitter with a half-life of 433 years. Other isotopes of americium range in mass from 232 to 247, but only the isotopes of mass 241 and 243 are important. The isotope ^{241}Am is routinely separated from "old" plutonium and sold for a variety of industrial uses, such as 59-keV gamma sources and as a component in neutron sources. The longer-lived ^{243}Am (half-life 7400 years) is a precursor in ^{244}Cm production.

1												13	14	15	16	17	18
1 H	2																2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

In its most prominent aqueous oxidation state, 3+, americium closely resembles the tripositive rare earths. The formal analogy to the rare earths is also marked in anhydrous compounds of both tripositive and tetrapositive americium. Americium is different in that it is possible to oxidize Am^{3+} to both the 5+ and 6+ states.

Americium metal has a vapor pressure markedly higher than that of its neighboring elements and can be purified by distillation. The metal is nonmagnetic and superconducting at 0.79 K. Under high pressure the metal has been compressed to 80% of its room-temperature volume and displays the α-uranium structure. Americium-241 is used as an alpha-particle source in ionization-type smoke detectors. *See* ACTINIDE ELEMENTS; ALPHA PARTICLES; BERKELIUM; CURIUM; FIRE DETECTOR; NUCLEAR REACTION; PERIODIC TABLE; TRANSURANIUM ELEMENTS.

R. A. Penneman

Bibliography. N. M. Edelstein, J. O. Navratil, and W. W. Schulz, *Americium and Curium Chemistry and Technology*, 1985; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; J. J. Katz, G. T. Seaborg, and L. R. Morss (eds.), *The Chemistry of the Actinide Elements*, 2 vols., 2d ed., 1986; W. W. Schulz, *The Chemistry of Americium*, ERDA Crit. Rev. Ser. Rep. TID-26971, 1976.

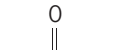
Amethyst

The transparent purple to violet variety of the mineral quartz. Although quartz is perhaps the commonest gem mineral known, amethyst is rare in the deep colors that characterize fine quality. Amethyst is usually colored unevenly and is often heated slightly in an effort to distribute the color more evenly. Heating at higher temperatures usually changes it to yellow or brown (rarely green), and further heating removes all color. The principal sources are Brazil, Arizona, Uruguay, and Russia. Amethyst is often cut in step or brilliant shapes, and drilled or carved for beads. Carvings are made both from transparent and nontransparent material. *See* GEM; QUARTZ.

Richard T. Liddicoat, Jr.

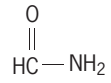
Amide

A derivative of a carboxylic acid with general formula (**1**), where R is hydrogen or an alkyl or

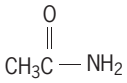


(**1**)

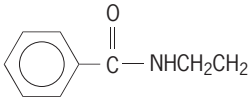
aryl radical. Amides are divided into subclasses, depending on the number of substituents on nitrogen. The simple, or primary, amides are considered to be derivatives formed by replacement of the carboxylic hydroxyl group by the amino group, NH_2. They are named by dropping the "-ic acid" or "-oic acid" from the name of the parent carboxylic acid and replacing it with the suffix "amide," as shown in examples (**2**–**5**). In the secondary and tertiary amides, one or



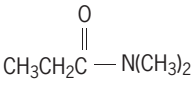
Formamide (**2**)



Acetamide (**3**)



N-Ethylbenzamide (**4**)

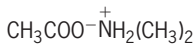


N,N-Dimethylpropanamide (**5**)

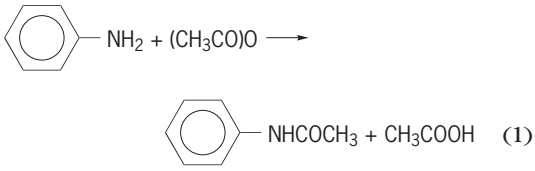
both hydrogens are replaced by other groups. The presence of such groups is designated by the prefix capital *N* (for nitrogen), as shown in the examples.

Except for formamide, all simple amides are relatively low-melting solids, stable, and weakly acidic. They are strongly associated through hydrogen bonding, and hence soluble in hydroxylic solvents, such as alcohol. Because of ease of formation and sharp melting points, amides are frequently used for the identification of organic acids and, conversely, for the identification of amines.

Formation and properties. Commercial preparation of amides involves thermal dehydration of ammonium salts of carboxylic acids. Thus, slow pyrolysis of ammonium acetate, $CH_3COO^-NH_4^+$, forms water and acetamide. *N,N*-dimethylacetamide may be similarly prepared from dimethylammonium acetate:

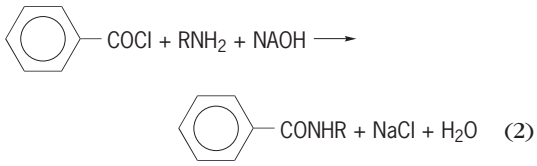


Acid anhydrides react with ammonia or with primary or secondary amines to form amides. The preparation of acetanilide [reaction (1)] is an example.



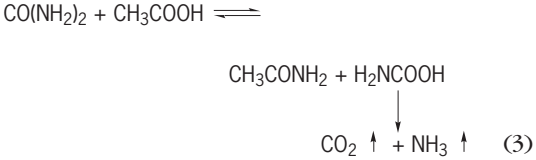
(**1**)

The reaction of amines with acid chlorides in the presence of aqueous sodium hydroxide (Schotten–Baumann reaction) is often used [reaction (2)].



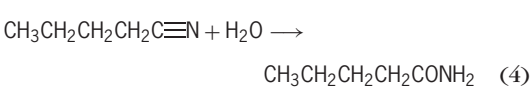
(**2**)

When urea, the diamide of carbonic acid, is heated with a carboxylic acid, a new amide is formed by an exchange reaction. The other product of the reaction, carbamic acid, decomposes to permit the reaction to go to completion [reaction (3)].



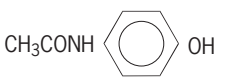
(**3**)

The partial hydrolysis of nitriles also affords a convenient synthesis of a variety of amides [reaction (4)]. The reaction is catalyzed by both acid and base.

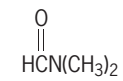


(**4**)

Uses. Amides are important chemical intermediates since they can be hydrolyzed to acids, dehydrated to nitriles, and degraded to amines containing one less carbon atom by the Hofmann reaction. In pharmacology, acetaminophen (**6**) is a popular analgesic. *N,N*-Dimethylformamide (DMF; **7**) is a useful aprotic, highly polar solvent.



(**6**)

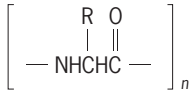


(**7**)

The amides of the straight-chain fatty acids having 12, 14, 16, or 18 carbon atoms are especially useful industrially because of the variety of properties

that can be obtained by substitution on the nitrogen atoms. Such products find extensive use as waterproofing agents, lubricant additives, detergents, emulsifiers, and wetting agents.

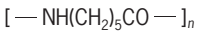
Amides are very prevalent in nature, since all peptides and proteins are polymers of the natural α-amino acids, as represented by (**8**).



(**8**)

See PEPTIDE; PROTEIN.

Nylon is the generic term for any synthetic, long-chain polymer with an amide linkage as a recurring, integral part of the chain. Nylon-6 (**9**) is one of the most useful.



(**9**)

See ACID ANHYDRIDE; ACID HALIDE; AMINE; CARBOXYLIC ACID; NITRILE; POLYAMIDE RESINS.

Paul E. Fanta

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998.

Amiiformes

An order of actinopterygian fishes in the subclass Neopterygii. Amiiformes plus several related fossil orders comprise the Halecomorphi, a well-developed group known from the middle Mesozoic. The Amiiformes comprise several families that persisted into the Cenozoic era; but only one, the Amiidae, survived to the present and it consists of a single species, *Amia calva*, the bowfin. The fossil amiids are from freshwater deposits, the oldest of which are of Jurassic age. The other halecomorphs were marine species.

Amia calva (see **illustration**) is characterized by an abbreviated heterocercal tail; fusiform body; large mouth equipped with sharp teeth; an elongate dorsal fin, with each ray supported by a single pterygiophore; scales with a ganoin (enamel-like) surface but thin and overlapping; no spiracles; a vascular swim bladder that can act as a lung; and a large gular plate. The maximum length of the bowfin is about 90 cm (35 in.). Its usual habitat is still waters with an abundance of rooted vegetation in lowlands of eastern North America from the St. Lawrence River to the Gulf Slope and from the Mississippi Basin to the Atlantic Slope, avoiding Appalachian streams. *See* ACTINOPTERYGII; OSTEICHTHYES; SWIM BLADDER; TELEOSTEI.

Herbert Boschung

***Amia calva*, the bowfin. (*Courtesy of Frank Teigler, Hippocampus Bildarchiv*)**

Bibliography. L. Grande and W. E. Bemis, A comprehensive phylogenetic study of amiid fishes (Amiidae) based on comparative skeletal anatomy, *J. Vert. Paleontol.*, Spec. Mem. 4 (suppl. to vol. 18), 1998; J. G. Maisey, *Santana Fossils: An Illustrated Atlas*, T. F. H. Publications, Neptune City, NJ, 1991; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

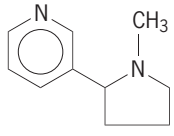
Amine

A member of a group of organic compounds which can be considered as derived from ammonia by replacement of one or more hydrogens by organic radicals. Generally amines are bases of widely varying strengths, but a few which are acidic are known.

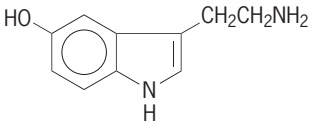
Amines constitute one of the most important classes of organic compounds. The lone pair of electrons on the amine nitrogen enables amines to participate in a large variety of reactions as a base or a nucleophile. Amines play prominent roles in biochemical systems; they are widely distributed in nature in the form of amino acids, alkaloids, and vitamins. Many complex amines have pronounced physiological activity, for example, epinephrine (adrenaline), thiamin or vitamin B_1, and Novocaine. The odor of decaying fish is due to simple amines produced by bacterial action. Amines are used to manufacture many medicinal chemicals, such as sulfa drugs and anesthetics. The important fiber nylon is an amine derivative.

Amines are classified according to the number of hydrogens of ammonia which are replaced by radicals. Replacement of one hydrogen results in a primary amine (RNH_2), replacement of two hydrogens results in a secondary amine (R_2NH), and replacement of all three hydrogens results in a tertiary amine (R_3N). The substituent groups (R) may be alkyl, aryl, or aralkyl. In another group of amines the nitrogen forms part of a ring (heterocyclic amines).

Examples of such compounds are nicotine (**1**), which is obtained commercially from tobacco for use as an insecticide, and serotonin (**2**), which plays



(**1**)



(**2**)

a key role as a chemical mediator in the central nervous system.

Many aromatic and heterocyclic amines are known by trivial names, and derivatives are named as substitution products of the parent amine. Thus, $C_6H_5NH_2$ is aniline and $C_6H_5NHC_2H_5$ is *N*-ethylaniline. For a discussion of definitive rules for naming amines *see* ORGANIC CHEMISTRY

According to the Brönsted-Lowry theory of acids and bases, amines are basic because they accept protons from acids. In water the equilibrium shown in reaction (1) lies predominantly to the left. The ex-

$$RNH_2 + H_2O \rightleftharpoons RNH_3^+ + OH^- \quad (1)$$

tent to which the amine is successful in picking up a proton from water is given by Eq. (2), where the

$$K_b = \frac{[RNH_3]^+[OH]^-}{RNH_2} \quad (2)$$

quantities in brackets signify concentrations of the species given. For short-chain aliphatic amines the basic ionization constant K_b lies near 10^{-4}; for aromatic amines $K_b < 10^{-9}$; for ammonia $K_b = 1.8 \times 10^{-5}$. Stable salts suitable for the identification of amines are in general formed only with strong acids, such as hydrochloric, sulfuric, oxalic, chloroplatinic, or picric.

Reactions and identification. Several test-tube reactions for recognition and characterization of amines are known: the Schotten-Baumann reaction, the Hinsberg test, the carbylamine reaction, and the action of nitrous acid.

The Schotten-Baumann reaction involves treatment of an amine with benzoyl chloride in basic solution. It serves to distinguish tertiary amines from primary and secondary amines by the formation of substituted benzamides from the primary and secondary amines [reactions (3) and (4)]. Tertiary

$$RNH_2 + C_6H_5COCl + OH^- \longrightarrow C_6H_5CONHR + Cl^- + H_2O \quad (3)$$

$$R_2NH + C_6H_5COCl + OH^- \longrightarrow C_6H_5CONR_2 + Cl^- + H_2O \quad (4)$$

amines do not react with benzoyl chloride. The substituted benzamides are generally insoluble in water, solid, and easily purified, and they have characteristic melting points. Thus they serve to identify the amines.

The more reactive acylating agents, acetic anhydride and acetyl chloride, give substituted acetamides without added base. This reaction gives the same type of information as the Schotten-Baumann reaction.

Closely related to the Schotten-Baumann reaction is the Hinsberg test. This has the added advantage of distinguishing between primary and secondary amines. It involves reaction of an amine with benzenesulfonyl chloride in alkaline solution. Both it and the Schotten-Baumann test are applicable to both aliphatic and aromatic amines with the exception of those amines which are substantially nonbasic in character. Primary amines give sulfonamides that are soluble in basic solutions; secondary amines give insoluble derivatives; and tertiary amines with no replaceable hydrogen do not react with the reagent. In general, the sulfonamides are solids and are useful for identification of the amines.

Carbylamines (isocyanides) possess a very unpleasant, nauseating odor and are formed by the reaction of any primary amine with chloroform in basic solution [reaction (5)].

$$RNH_2 + CHCl_3 + 3OH^- \longrightarrow R{-}N{\equiv}C + 3Cl^- + 3H_2O \quad (5)$$

Reaction with nitrous acid serves as a further method for distinguishing between various classes of amines. Primary aliphatic amines evolve nitrogen, whereas primary aromatic amines give diazonium compounds, which may be recognized by dye formation on coupling with a suitable second component. Secondary amines of both series give nitrosamines, generally as yellow oils. Tertiary aliphatic amines do not react with nitrous acid, and mixed aliphatic aromatic tertiary amines undergo nuclear nitrosation. Nitrosamines have been identified as potent carcinogenic substances. *See* DIAZOTIZATION.

In the infrared absorption spectrum, amines exhibit characteristic bands due to N-H stretching and bending, as well as C-N stretching vibrations. In the proton magnetic resonance spectrum, the appearance of the proton on nitrogen is complicated by the rate of exchange and the electrical quadrupole of the ^{14}N nucleus.

In alkaline solutions tertiary amines are oxidized by hydrogen peroxide to amine oxides which, although still basic, are not strong bases as are the quaternary ammonium types [reaction (6)].

$$R_3N + H_2O_2 \longrightarrow R_3NO + H_2O \quad (6)$$

Aromatic amines undergo halogenation and sulfonation on the ring. However, because of their susceptibility to oxidation, nitration cannot be accomplished without prior protection of the labilizing amino group. This is commonly done by acetylation.

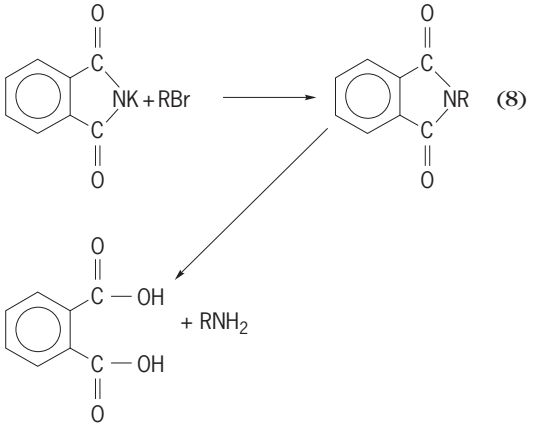
Exhaustive methylation is one of a number of reactions of amines associated with the name of A. W. Hofmann. It involves a sequence of reactions terminating in the thermal decomposition of a quaternary ammonium hydroxide to yield a tertiary amine, usually trimethylamine, water, and an olefin. It has been widely used as a tool in the determination of the structures of complex compounds and, to a lesser extent, for the synthesis of olefins. The sequence of reactions is shown in (7).

$$RCH_2CH_2NH_2 \xrightarrow{3CH_3I} RCH_2CH_2\overset{+}{N}(CH_3)_3L^- \xrightarrow{AgOH} RCH_2CH_2\overset{+}{N}(CH_3)_3OH^- \xrightarrow{heat} RCH{=}CH + (CH_3)_3N + H_2O \quad (7)$$

See ALKALOID.

Preparation. Commercial preparation of aliphatic amines can be accomplished by direct alkylation of ammonia (Hofmann method, 1849) or by catalytic alkylation of amines with alcohols at elevated temperatures. Reduction of various nitrogen functions carrying the nitrogen in a higher state of oxidation also leads to amines. Such functions are nitro, oximino, nitroso, and cyano. For the preparation of pure primary amines, Gabriel's synthesis and Hofmann's hypohalite reaction are preferred methods. The Bucherer reaction is satisfactory for the preparation of polynuclear primary aromatic amines.

Gabriel's synthesis. This is a method for the synthesis of pure primary aliphatic amines by the hydrolysis of an *N*-alkyl phthalimide. The *N*-alkyl phthalimides are prepared by reaction of potassium phthalimide with an alkyl halide (preferably a bromide) [reaction (8)].



Hofmann hypohalite reaction. This is another reaction associated with Hofmann, which furnishes a convenient method for the preparation of pure primary amines of either the aliphatic or aromatic series. In the overall sense, it involves conversion of an acid amide to an amine with loss of one carbon atom. The reaction proceeds through the stages shown in reactions (9)–(11). The amine arises by hydrolysis of

$$RCONH_2 + NaOH \longrightarrow RCONHX + NaOH \quad (9)$$

$$RCONHX + NaOH \longrightarrow R{-}N{=}C{=}O + NaX + H_2O \quad (10)$$

$$R{-}N{=}C{=}O + 2NaOH \longrightarrow RNH_2 + Na_2CO_3 \quad (11)$$

the isocyanate formed by migration of R from carbon to nitrogen in reaction (10) . Sodium hypobromite is the common laboratory reagent, but the cheaper calcium hypochlorite is used in commercial applications, such as the manufacture of anthranilic acid from phthalimide. *See* AMINO ACIDS; QUATERNARY AMMONIUM SALTS.

Paul E. Fanta

Bibliography. N. L. Allinger et al., *Organic Chemistry*, 2d ed., 1976; W. H. Brown and E. P. Rogers, *General, Organic, and Biochemistry*, 3d ed., 1987.

Amino acid dating

Determination of the relative or absolute age of materials or objects by measurement of the degree of racemization of the amino acids present. With the exception of glycine, the amino acids found in proteins can exist in two isomeric forms called D- and L-enantiomers. Although the enantiomers of an amino acid rotate plane-polarized light in equal but opposite directions, their other chemical and physical properties are identical. Amino acid handedness or homochirality is one of the most distinctive features of terrestrial life. It was discovered by L. Pasteur around 1850 that only L-amino acids are generally found in living organisms, but scientists still have not formulated a convincing reason to explain why life on Earth is based on only L-amino acids. *See* AMINO ACIDS; ENANTIOMER.

Racemization. Under conditions of chemical equilibrium, equal amounts of both enantiomers are present ($D/L = 1.0$); this is called a racemic mixture. Living organisms maintain a state of disequilibrium through a system of enzymes that selectively utilize only the L-enantiomers. Once a protein has been synthesized and isolated from active metabolic processes, the L-amino acids are subject to a racemization reaction that converts them into a racemic mixture. The racemization reaction can be written as (1), where k_i is the first-order rate constant for the

$$\text{L-amino acid} \underset{k_i}{\overset{k_i}{\rightleftharpoons}} \text{D-amino acid} \quad (1)$$

interversion of the enantiomers. The rate of racemization is equal to $2k_i$. The kinetic equation of the racemization reaction is given by (2), where D/L is

$$\ln\left(\frac{1+(D/L)}{1-(D/L)}\right) - \ln\left(\frac{1+(D/L)}{1-(D/L)}\right)_{t=0} = 2k_i t \quad (2)$$

the ratio of the enantiomers of a certain amino acid at a particular time t. The $t = 0$ term is necessary to account for some racemization that occurs during sample preparation.

Since racemization is a chemical process, the extent of racemization is dependent not only on the time that has elapsed since the L-amino acids were synthesized but also on the exposure temperature: the higher the temperature, the faster the rate of racemization. The rate of racemization is also different for most of the various amino acids. The half-life (for example, the time required to reach a D/L ratio of 0.33) at neutral pH for aspartic acid, an amino acid with one of the fastest racemization rates, is approximately 3500 years at 25°C (77°F) but is only 35 days at 100°C (212°F). Amino acids with the slowest racemization rates have half-lives roughly 10 times greater than those of aspartic acid. *See* HALF-LIFE.

A variety of analytical procedures can be used to separate amino acid enantiomers; gas chromatography and high-performance liquid chromatography are the most widely used. Since these techniques have sensitivities in the parts-per-billion range, only a few hundred milligrams of sample material are

normally required. Samples are first hydrolyzed in hydrochloric acid to break down the proteins into free amino acids, which are then isolated by cation-exchange chromatography. *See* CHROMATOGRAPHY.

Since the late 1960s, the geochemical and biological significance of amino acid racemization has been extensively investigated. Geochemical uses of amino acid racemization include the dating of fossils or, in the case of known age specimens, the determination of their temperature history. Fossil types such as bones, teeth, and shells have been studied, and racemization has been found to be particularly useful for dating specimens that were difficult to date by other methods. Racemization has also been observed in the metabolically inert tissues of living mammals. Racemization can be studied in certain organisms and used to assess the biological age of a variety of mammalian species; in addition, it may be important in determining the biological lifetime of certain proteins. *See* RACEMIZATION.

Dating fossils. Fossils have been found to contain both D- and L-amino acids, and the extent of racemization generally increases with geologic age. The amount of racemization in a fossil is determined by the value of k_i and its age. The magnitude of k_i in a fossil from a particular locality is primarily a function of a value known as the integrated exposure temperature, although protein diagenetic processes (for example, geochemical degradation and alteration) are also important. The rate constant k_i can be estimated by using a calibration procedure, wherein the *D/L* ratio is measured in a fossil of known age from the study area. This *D/L* ratio and the calibration age of the sample are substituted into Eq. (2); an in-place k_i value is thus calculated. This calibrated rate constant integrates variations in exposure temperature and other environmental parameters of the locality over the time period represented by the age of the calibration sample. Following calibration, the k_i value can be used, with certain limitations, to date other samples from the surrounding region. The principal limitation is that the calibration sample should have a similar amino acid composition and content, and should be from roughly the same geologic period, as the sample being dated.

The Olduvai Gorge region in the north-central Tanzanian Rift Valley offers an excellent opportunity to study the racemization reaction of amino acids in fossil bones and teeth over a time period extending back several million years. Since the geology of the region has been extensively studied as a result of the discovery of numerous hominid fossils and artifacts, the relative and absolute ages of the various stratigraphic units in the Olduvai region are well established. Thus, amino acid racemization can be investigated over several geologic periods.

Aspartic acid racemization can be used to date only the Holocene and perhaps upper Pleistocene deposits in the Olduvai region. The k_i values determined for aspartic acid at Olduvai suggest that bones and teeth older than about 60,000 years should contain only racemic aspartic acid, for example, *D/L* aspartic acid = 1.0. However, fossil bones from the older stratigraphic units at Olduvai have been found to contain nonracemic aspartic acid. This indicates that these bones contain secondary aspartic acid, which probably is incorporated into fossil bone by percolating ground waters. Aspartic acid racemization of bones is of limited use in dating the important anthropological deposits in the Olduvai Gorge region because it racemizes rapidly and is prone to contamination problems.

In contrast, the epimerization of L-isoleucine, which yields the nonprotein amino acid D-alloisoleucine, provides a much better chronological tool for dating the deposits in the Olduvai region. Isoleucine has two centers of asymmetry, and thus the reaction is termed epimerization rather than racemization; the equilibrium ratio of D-alloisoleucine to L-isoleucine is 1.3, and so Eq. (2) must be slightly modified to account for this. In fossil bones and teeth ranging in age from the upper Pleistocene to the Pliocene, the aile/ile ratio in tooth enamel steadily increases until an approximate equilibrium ratio of 1.3 is attained in the oldest deposits. This indicates that isoleucine epimerization in teeth is effectively a closed system with respect to the introduction of secondary isoleucine for a period of 3–4 million years. Isoleucine may be less susceptible to contamination than aspartic acid because the peptide bonds involving this amino acid are more stable with respect to hydrolysis than are those containing aspartic acid.

By using the enamel sample from the base of Olduvai Bed I for calibration, isoleucine epimerization ages for the middle and lower Pleistocene deposits are obtained that are in good agreement with stratigraphic and radiometric age determinations. Some bones also yield reasonable ages. For example, the isoleucine epimerization age for a bone from the lower unit of the Masek beds is about 500,000 years, which is consistent with the age estimated from other evidence. Bones from older stratigraphic units, however, have been found to have aile/ile ratios less than those measured in tooth enamel. This implies that isoleucine in bones is more susceptible to contamination than is isoleucine in enamel. Isoleucine epimerization ages of bones from the older deposits should probably be viewed with caution and should be considered minimum age estimates. Contamination should consist primarily of L-amino acids; thus, incorporation of secondary amino acids into a fossil would lower the *D/L* ratio and yield an age estimate that is too young.

The k_i values for isoleucine epimerization in the upper Pleistocene deposits are about four times greater than those determined by using the lower Pleistocene Bed I sample. The faster rate of isoleucine epimerization in the younger deposits may be due to several factors, such as different exposure temperatures and the complex process of protein diagenesis. The upper Pleistocene k_i value should be used only to date upper Pleistocene samples; it is not applicable to dating the older stratigraphic units.

In general, the ages based on racemization and epimerization at Olduvai Gorge are consistent with those estimated from other evidence. The best results are obtained by using tooth enamel. On the basis of these results, the racemization and epimerization reactions in bones and teeth have been investigated at many localities throughout the world. Several hominid remains have been directly dated, and the technique has been used to evaluate dates derived from geologic, radiometric, and faunal evidence.

Other racemization studies have utilized both terrestrial and marine fossil shells to estimate the age of various archeological sites as well as uplifted marine terrace deposits. In these cases, the limit for racemization dating is in the 100,000-year range, although at high latitudes where temperatures are cold the age range is extended nearly 1 million years. *See* FOSSIL.

Racemization in living mammals. Enamel, dentine, and the proteins present in the nucleus of the eye lens are metabolically inert and are thus incubated at approximately 37°C (98.6°F) throughout a mammal's life span. In mammals that live for decades, detectable racemization of aspartic acid has been found to take place in these tissues. Although the extent of racemization is quite small (for example, *D/L* ratios are in the range 0.02–0.10), it can be easily and precisely measured with available analytical techniques. Only aspartic acid racemization has been detected, which is expected since this amino acid has the fastest racemization rate.

The racemization of amino acids thus can be used as a biochronological tool. This aging method is particularly useful for determining the ages of humans and certain other mammals, whose ages are not well documented. An interesting application of the racemization technique has been the determination of the ages of various marine mammals. For example, the ages of narwhals (*Mondon monoceros*), Arctic cetaceans that are characterized by the long (about 6.5 ft or 2 m) counterclockwise-spiraling tusk of the male, have been deduced by using the extent of aspartic acid racemization in their teeth. The method is particularly useful for determining the ages of female narwhals whose ages previously were largely unknown. The aspartic acid measurements indicate that both female and male narwhals can live more than 50 years.

The method has also been used to estimate the ages of bowhead whales that live in the Arctic. During the recent hunting of these animals by Arctic peoples, some whales were found to have embedded in their skin stone harpoon tips, which were no longer used in the Arctic after around 1880. This finding suggested that bowhead whales might have unusually long life spans and, in order to help verify this surmise, a systematic study of the racemization in the eye lens nucleus was carried out. Extensive racemization was observed in the eye lens nucleus of some animals, and based on these results their ages were estimated to be in excess of 100 years. In the case of one whale, the calculated age was close to 200 years. These findings suggest that bowhead whales may be some of the most aged mammals in the world.

The extent of aspartic acid racemization in enamel and dentine can also be used to estimate the death age of humans from ancient burials. The main requirement in this application is that either the burial should be fairly recent or the burial environmental temperature should be relatively low, so that the extent of postmortem racemization is small compared with that which occurred during the lifetime of the individual. The racemization method has been used to calculate the age at death of mummified Alaskan Eskimos, of individuals in a medieval cemetery in Czechoslovakia, and of corpses in forensic cases. The racemization ages have greater accuracy (about ±5%) than those estimated by other means, and the method is especially useful in determining the ages of older individuals, whose age at death is often difficult to estimate from anatomical evidence. *See* ARCHEOLOGICAL CHEMISTRY; GEOCHRONOMETRY; RADIOCARBON DATING; ROCK AGE DETERMINATION.

Other applications of amino acid racemization. Amino acid racemization provides a general indicator of the level of preservation of biomolecules in geological specimens. The extent of racemization of amino acids correlates with the level of overall degradation of labile molecules. One example deals with the preservation of ancient DNA. It has been found that only in specimens in which the extent of racemization is minimal ($D/L < 0.1$) are fragments of original DNA preserved. This has proved to be a valuable index for judging if DNA in fossils is original or derived from contamination. *See* DEOXYRIBONUCLEIC ACID (DNA).

The reaction has implications with respect to protein turnover in mammalian tissues. It has been shown that there is an increasing level of water-insoluble protein present in the eye lens nucleus, of long-lived mammals with increasing age. This is correlated with increasing racemization in the eye lens nucleus, and it has been suggested that structural changes induced by racemization result in protein insolubility. Amino acid racemization has also been found to take place during the processing of various food materials. This is caused by both heat exposure during cooking and the treatment of food components with alkali in order to make them more palatable. The racemization induced by these treatments may have deleterious nutritional effects because it has been found that proteins containing racemized amino acids are less digestible.

The racemization of amino acids has implications in the search for life beyond Earth. It is assumed that life elsewhere would require homochiral amino acids, although whether it would be based on L- or D-amino acids is considered to be mainly a matter of a chance selection event that occurred either at the time of the origin of life or during early evolution. Detection of this amino acid homochirality signature would depend on whether life still existed on another world or, if it became extinct, whether racemization has erased the signature. In the case of

Mars, using racemization rates determined in terrestrial environments, the survival time of an amino acid homochiral signature has been estimated for the conditions relevant to the surface of Mars throughout the planet's history. Homochiral amino acids associated with a Martian biota that became extinct during the early history of the planet would still be preserved today in dry, cold (<250 K) Martian environments. In contrast, if liquid water was present for extended periods (several millions of years) after life on Mars became extinct, racemization would have converted any originally homochiral amino acids into a racemic mixture. These considerations are important in the search for evidence of life on Mars and other solar system bodies during future space missions designed to search for evidence of extraterrestrial life. *See* MARS.

Jeffrey L. Bada

Bibliography. J. L. Bada, Amino acid racemization dating of fossil bones, *Annu. Rev. Earth Planet. Sci.*, 13:241–268, 1985; J. L. Bada, *In vivo* racemization in mammalian proteins, *Meth. Enzymol.*, 106:98–115, 1984; J. L. Bada and G. D. McDonald, Amino acid racemization on Mars: Implications for the preservation of biomolecules from an extinct Martian biota, *Icarus*, 114:139–143, 1995; J. C. George et al., Age and growth estimates of bowhead whales (*Balaena mysticetus*) via aspartic acid racemization, *Can. J. Zool.*, 77:571–580, 1999; H. Y. Goksu, M. Oberhofer, and D. Regulla (eds.), *Scientific Dating Methods*, 1991; W. C. Mahaney (ed.), *Quaternary Dating Methods*, 1984; H. N. Poinar et al., Amino acid racemization and the preservation of ancient DNA, *Science*, 272:864–866, 1996; M. R. Zimmerman and J. L. Angel (eds.), *Dating and Age Determination of Biological Materials*, 1986.

Amino acids

Organic compounds possessing one or more basic amino groups and one or more acidic carboxyl groups. Of the more than 80 amino acids which have been found in living organisms, about 20 serve as the building blocks for the proteins.

All the amino acids of proteins, and most of the others which occur naturally, are α-amino acids, meaning that an amino group (—NH_2) and a carboxyl group (—COOH) are attached to the same carbon atom. This carbon (the α carbon, being adjacent to the carboxyl group) also carries a hydrogen atom; its fourth valence is satisfied by any of a wide variety of substituent groups, represented by the letter R in **Fig. 1**.

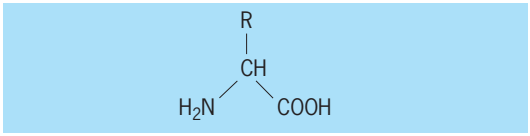


Fig. 1. Structural formula for an amino acid.

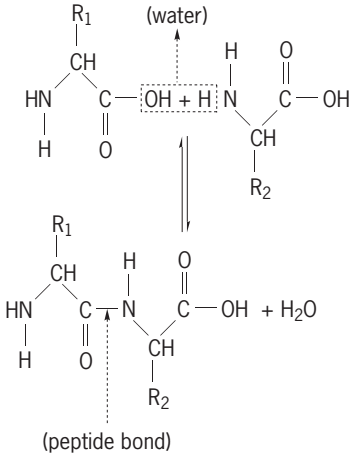
Amino acids of proteins, grouped according to the nature of R

Amino acids*	R
Glycine	Hydrogen
Alanine, valine, leucine, isoleucine	Unsubstituted aliphatic chain
Serine, threonine	Aliphatic chain bearing a hydroxyl group
Aspartic acid, glutamic acid	Aliphatic chain terminating in an acidic carboxyl group
Asparagine, glutamine	Aliphatic chain terminating in an amide group
Arginine, lysine	Aliphatic chain terminating in a basic amino group
Cysteine, cystine, methionine	Sulfur-containing aliphatic chain
Phenylalanine, tyrosine	Terminates in an aromatic ring
Tryptophan, proline, histidine	Terminates in a heterocyclic ring

* See articles on the individual amino acids listed in the table.

In the simplest amino acid, glycine, R is a hydrogen atom. In all other amino acids, R is an organic radical; for example, in alanine it is a methyl group (—CH_3), while in glutamic acid it is an aliphatic chain terminating in a second carboxyl group (—CH_2—CH—COOH). Chemically, the amino acids can be considered as falling roughly into nine categories based on the nature of R (see **table**).

Occurrence of conjugated amino acids. Amino acids occur in living tissues principally in the conjugated form. Most conjugated amino acids are peptides, in which the amino group of one amino acid is linked to the carboxyl group of another. This type of linkage is known as a peptide bond; a molecule of water is split out when a peptide bond is formed, and a molecule of water must be added when a peptide bond is broken, as shown in the reaction below.



Since each amino acid possesses both an amino group and a carboxyl group, the acids are capable of linking together to form chains of various lengths, called polypeptides. Proteins are polypeptides ranging in size from about 50 to many thousand amino acid residues. The process by which peptides are

formed from free amino acids actually cannot be as simple as pictured in the equation, for a considerable amount of energy is required. This process is discussed later in this article.

Although most of the conjugated amino acids in nature are proteins, numerous smaller conjugates occur naturally, many with important biological activity. The line between large peptides and small proteins is difficult to draw, with insulin (molecular weight = 7000; 50 amino acids) usually being considered a small protein and adrenocorticotropic hormone (molecular weight = 5000; 39 amino acids) being considered a large peptide. In addition to their role as hormones, peptides often occur in coenzymes (such as folic acid and glutathione), bacterial capsules (the polyglutamic acid capsule which contributes to the pathogenicity of *Bacillus anthracis*), fungal toxins (the tomato wilt toxin of *Fusarium* and the toxins of the poisonous mushroom *Amanita phalloides*), and antibiotics (chloramphenicol, penicillin, bacitracin, and polymixins). Elucidation of the structure of bacterial cell walls has shown that they are composed in part of a series of cross-linked peptides, the cross-linking providing the wall with a large part of its rigidity. The action of penicillin in inhibiting this cross-linking reaction accounts for its antibiotic activity. Finally, a considerable part of the phospholipid fraction of any organism contains serine linked by phosphoester bond to glycerol phosphate. *See* ANTIBIOTIC; COENZYME; TOXIN.

Occurrence of free amino acids. Free amino acids are found in living cells, as well as the body fluids of higher animals, in amounts which vary according to the tissue and to the amino acid. The amino acids which play key roles in the incorporation and transfer of ammonia, such as glutamic acid, aspartic acid, and their amides, are often present in relatively high amounts, but the concentrations of the other amino acids of proteins are extremely low, ranging from a fraction of a milligram to several milligrams per 100 g wet weight of tissue. In view of the fact that amino acid and protein sythesis go on constantly in most of these tissues, the presence of free amino acids in only trace amounts points to the existence of extraordinarily efficient regulation mechanisms. Each amino acid is ordinarily synthesized at precisely the rate needed for protein synthesis. The regulation mechanism has been found most often to be one of feedback control; each amino acid acts as an inhibitor of its own biosynthesis. If any amino acid is formed in excess of that required for protein synthesis, the biosynthesis of that amino acid is slowed down until the excess has been used.

In addition to the amino acids of protein, a variety of other free amino acids occurs naturally. Some of these are metabolic products of the amino acids of proteins; for example, γ-aminobutyric acid occurs as the decarboxylation product of glutamic acid. Others, such as homoserine and ornithine, are biosynthetic precursors of the amino acids of protein. However, the origin and role of many unusual free amino acids is not yet known.

General properties. The amino acids are characterized physically by the following: (1) the pK_1, or the dissociation constant of the various titratable groups; (2) the isoelectric point, or pH at which a dipolar ion does not migrate in an electric field; (3) the optical rotation, or the rotation imparted to a beam of plane-polarized light (frequently the D line of the sodium spectrum) passing through 1 decimeter of a solution of 100 grams in 100 milliliters; and (4) solubility. *See* IONIC EQUILIBRIUM; ISOELECTRIC POINT; OPTICAL ACTIVITY.

At ordinary temperatures, the amino acids are white crystalline solids; when heated to high temperatures, they decompose rather than melt. They are stable in aqueous solution, and with few exceptions can be heated as high as 120°C (248°F) for short periods without decomposition, even in acid or alkaline solution. Thus, the hydrolysis of proteins can be carried out under such conditions with the complete recovery of most of the constituent free amino acids. The exceptions are as follows: Acid hydrolysis of protein destroys most of the tryptophan and some of the serine and threonine, oxidizes cysteine to cystine, and deamidates glutamine and asparagine; alkaline hydrolysis destroys serine, threonine, cystine, cysteine, and arginine, and also causes deamidations.

Enantiomorphs. Since all of the amino acids except glycine possess a center of asymmetry at the α carbon atom, they can exist in either of two optically active, mirror-image forms or enantiomorphs. All of the common amino acids of proteins appear to have the same configuration about the α carbon; this configuration is symbolized by the prefix L-. The opposite, generally unnatural, form is given the prefix D-. Some amino acids, such as isoleucine, threonine, and hydroxyproline, have a second center of asymmetry and can exist in four stereoisomeric forms. The prefix allo- is used to indicate one of the two alternative configurations at the second asymmetric center; thus, isoleucine, for example, can exist in the L, L-allo, D, and D-allo forms. *See* STEREOCHEMISTRY.

Unlike chemical syntheses, which lead to mixtures of D and L forms, biosynthetic processes invariably produce optically active amino acids. For most amino acids, only the L isomer occurs naturally; but in a few cases, the D isomer is found also. For example, the cell walls of certain bacteria contain D-alanine and D-glutamic acids, and the D isomers of phenylalanine, leucine, serine, and valine occur in some antibiotic peptides.

Ionic state. Another important general property of all amino acids is their ionic state (**Fig. 2**). The basic amino group can bind a proton from solution and become a cation; the acidic carboxyl group can release a proton into solution and become an anion. At the isoelectric point (the pH at which the molecule has no net charge), amino acids exist as dipolar ions or zwitterions, while in strong acid solution, the carboxyl group exists in the undissociated form, and the molecule becomes a cation. If such an acidic solution is titrated with strong alkali, two

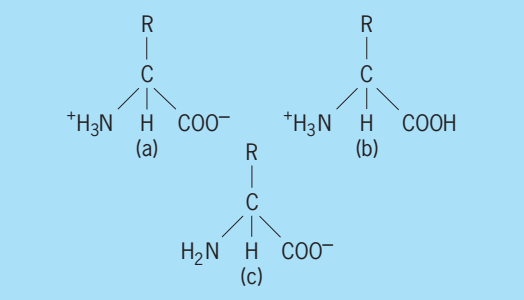


Fig. 2. Ionic states of an amino acid. (*a*) Zwitterion. (*b*) Cation. (*c*) Anion.

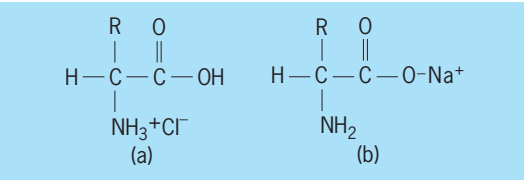


Fig. 3. Salts of amino acids. (*a*) Amino acid hydrochloride. (*b*) Sodium amino acid.

dissociations of protons are observed. The carboxyl group, having the weakest affinity for its proton, dissociates at a fairly low pH; its p*K* (the pH at which half of the molecules are dissociated) in most cases is close to 2.0. As more alkali is added, the proton on the amino group begins to dissociate; p*K* values for this dissociation are generally found close to 9.5. When sufficient alkali has been added to pull off all the dissociable protons, the amino acid exists as an anion.

Since the amino acids are ions, they can be prepared as their salts. For example, the titration of an amino acid solution with hydrochloric acid (HCl) leads to formation of the amino acid hydrochloride, while titration with sodium hydroxide (NaOH) forms the sodium salt (**Fig. 3**).

When the R radical contains an ionizable group, the amino acid will have correspondingly more ionic forms. Those amino acids whose radicals contain carboxyl groups (aspartic and glutamic acids) are known as the acidic amino acids, since a solution of the zwitterion will be strongly acidic. Similarly, histidine, lysine, and arginine are known as the basic amino acids, and the rest as the neutral amino acids. (It is important to note that a solution of the zwitterion of a neutral amino acid will in fact be slightly acidic.)

The salts are, in general, more soluble in water or alcohol than the corresponding zwitterions.

Isolation and determination. Since most amino acids occur in conjugated form, their isolation usually requires their prior release in free form by acid or alkaline hydrolysis. Hydrolysates of proteins or other polypeptides, or crude extracts of plants, animal, or microbial materials, serve as the starting point for the isolation in pure form of single amino acids. Prior to the application of chromatography in the early 1940s, the isolation of amino acid depended on slight differences in the solubilities of amino acid salts in various solvents and at different pH values. For example, the isolation of aspartic acid was accomplished by adding an excess of calcium hydroxide to an aqueous solution of amino acids and then precipitating the calcium aspartate with alcohol.

Such methods, although used successfully to isolate each of the common amino acids of protein, are difficult as well as tedious, and require relatively large amounts of starting material. Chromatography, on the other hand, is simple, rapid, and capable of isolating amino acids even when they are present in microgram quantities. Thus chromatography has been the method of choice for amino acid isolation ever since its first application by A. J. P. Martin and R. L. M. Synge in 1941. *See* CHROMATOGRAPHY.

Chromatography is carried out by using either cylindrical glass tubes (columns) packed with a porous solid or by using sheets of filter paper. In the former method, the column is packed with any of a variety of substances, such as starch, powdered cellulose, or cation-exchange resin, and is saturated with the chosen solvent. A solution of amino acids is allowed to percolate into the top of the column,

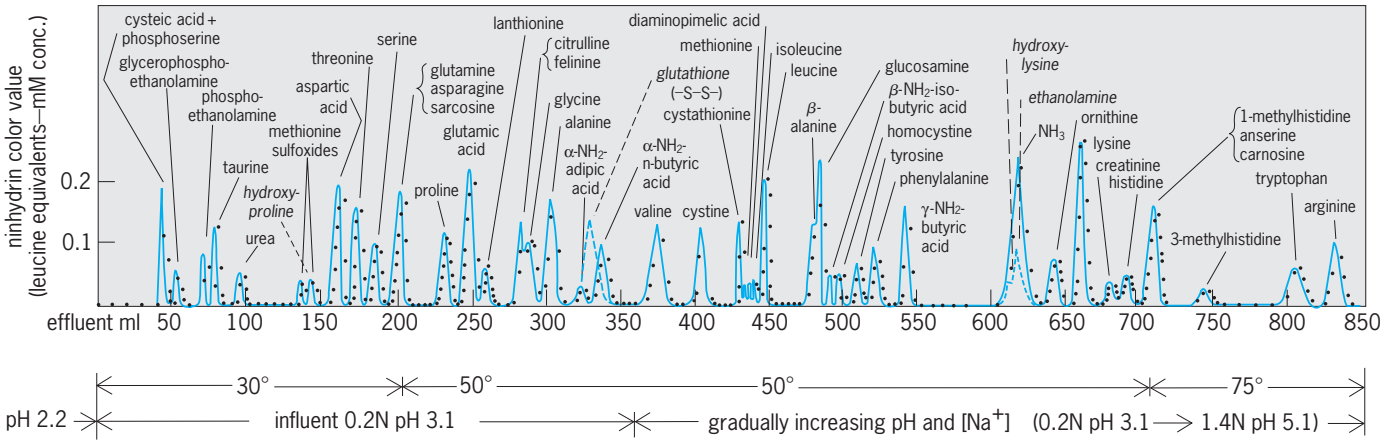


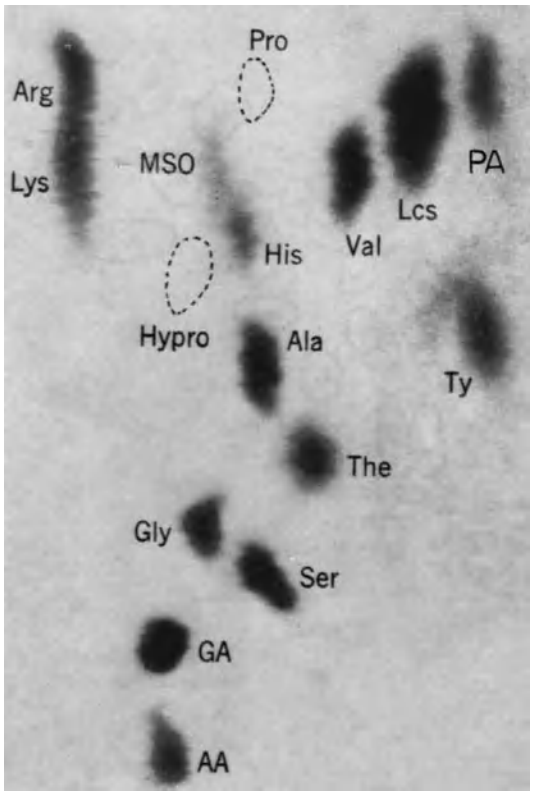
Fig. 4. Typical protein fractionation into amino acids shown in a cation-exchange resin column chromatogram. (*After S. Moore and W. H. Stein, J. Biol. Chem., 211:893–906, 1954*)

and then solvent is forced through the column at a controlled rate. The solvent is allowed to flow out through an opening at the bottom of the column, and the eluate is caught in a series of test tubes.

A given amino acid will have been eluted from the column at a time depending on its own characteristic rate of movement, and will have been caught in one or a few tubes; a different amino acid will have been eluted into a separate set of tubes. To detect their presence, as well as to determine their exact quantity, a substance which reacts with amino acids to give a visible color is then added to each tube. The best such reagent is ninhydrin, which reacts with amino acids to produce carbon dioxide, ammonia, and aldehyde, and forms a purple compound with the liberated ammonia. The amount of color which develops under standard conditions can then be measured in a photoelectric colorimeter, and the precise amount of amino acid determined by comparison with a standard curve based on reactions with known quantities. **Figure 4** shows a typical separation by passing water adjusted to different pH values and temperatures through a column of cation-exchange resin, collecting fractions, and determining the amount of amino acid in each with ninhydrin.

The column method described above is capable of giving the most precise quantitative data, and can handle relatively large quantities of amino acids. For the analysis of mixtures containing only a few micrograms of each amino acid, however, paper chromatography is the most simple and rapid procedure, and can also be made quantitative. In this procedure, the mixture of amino acids is applied as a drop of solution to a spot close to one corner of a sheet of paper. The sheet is then placed in a vapor-tight chamber with one edge of the paper dipping into a chosen, water-saturated organic solvent. The solvent flows through the paper by capillarity, water becoming bound to the paper and the organic solvent flowing past it. The amino acids travel through the paper as discrete spots, exactly as described above for their travel through a column. When the solvent has traveled to the opposite edge of the paper, the sheet is removed, dried, and sprayed with a reagent such as ninhydrin. The position of each amino acid is then revealed by a colored spot which appears when the sheet is heated.

If the solvent used brings two or more amino acids to the same position, two-dimensional chromatography is employed. The dried sheet is not sprayed, but is rotated 90° and is placed with the edge along which the amino acids are located in a second solvent. This solvent flows through the paper at right angles to the direction taken by the first solvent, and if correctly chosen, will separate those amino acids which stayed together in the first solvent. The sheet is then dried and sprayed to locate the amino acids, and then the spots can be cut out with scissors and eluted with water in separate tubes for colorimetric determination. A two-dimensional chromatogram of the amino acids of a protein is shown in **Fig. 5**.



KEY:

Arg = arginine	His = histidine
Lys = lysine	Ala = alanine
MSO = methionine sulfoxide	The = threonine
	Se = serine
Hypro = hydroxyproline	Val = valine
Gly = glycine	Lcs = leucines
GA = glutamic acid	Ty = tyrosine
AA = aspartic acid	PA = phenylalanine
Pro = proline	

Fig. 5. Two-dimensional chromatogram of a protein; the solvents were phenol and lutidine. (*From R. J. Block, E. L. Durrum, and G. Zweig, A Manual of Paper Chromatography and Paper Electrophoresis, 2d ed., Academic Press, 1958*)

The determination of total α-amino acids in extracts of natural materials can be carried out by using the ninhydrin method, described above, or the Van Slyke method, which measures the amount of nitrogen gas given off on treatment with nitrous acid. Specific determinations of a given amino acid were previously carried by means of sensitive microbiological assays which measured the growth of a microorganism dependent on the amino acid as a function of the concentration of the unknown material. Nowadays this rather difficult and cumbersome method has been largely replaced by ion-exchange chromatography.

There are relatively specific color reactions for many amino acids; while still used to some extent as quantitative assays, these now find most widespread

use as qualitative tests for the presence of the amino acids.

Amino Acid Metabolism

Although amino acids and some other charged molecules can enter a cell passively by means of simple diffusion, there are present in all cells so far examined special systems for concentrating such small molecules inside the cell. These systems, called permeases, are localized in the cell membrane. They are protein complexes, probably containing at least two parts, which utilize metabolic energy to transport small molecules against the concentration gradient. Bacterial amino acid permeases are capable of achieving a concentration inside a cell 1000 times greater than that outside the cell.

The criteria for a permease, or active transport system, are (1) it must require energy, (2) it must be relatively specific, and (3) it must concentrate the transported substance against a gradient. Most permeases obey the classical Michaelis-Henry enzyme kinetics. Portions of the system which concentrates valine, isoleucine, and leucine in *Escherichia coli* have been purified; thus the mechanism of action of this permease may soon be clear.

Biosynthesis. Since amino acids, as precursors of proteins, are essential to all organisms, all cells must be able to synthesize those they cannot obtain from their environment. The selective advantage of being able rapidly to shift from endogenous to exogenous sources of these compounds has led to the evolution in bacteria and many other organisms of very complex and precise methods of adjusting the rate of synthesis to the available level of the compound. These regulatory mechanisms can be divided according to whether they require a short or a long time to take effect.

The immediately effective control is that of feedback inhibition. As Figs. 9–14 show, the biosynthesis of amino acids is relatively complicated and usually requires at least three enzymatic steps. In most cases so far examined, the amino acid end product of the biosynthetic pathway inhibits the first enzyme to catalyze a reaction specific to the biosynthesis of that amino acid. This inhibition is extremely specific; the enzymes involved have special sites for binding the inhibitor. This inhibition functions to shut off the pathway in the presence of transient high levels of the product, thus saving both carbon and energy for other biosynthetic reactions. When the level of the product decreases, the pathway begins to function once more.

The one exception to this general rule is the growth factor requirement which results from the presence in the environment of a metabolic inhibitor. For example, a certain strain of bacterium is very sensitive to inhibition by valine, and this inhibition can be overcome by isoleucine. In the presence of valine, then, this strain requires isoleucine for growth. There are many such antagonisms between amino acids, with the result that organisms which require several amino acids must receive them in balanced amounts; any one in excess may prove inhibitory.

If a microorganism is grown for several generations in the presence of an amino acid, the levels of the enzymes of the biosynthetic pathway decrease considerably. This phenomenon is called enzyme repression, and it comes about because the synthesis of the enzymes is decreased in the presence of the end product. However, the enzyme already present in the cell is stable; therefore several generations are required before it is diluted to its lowest level by being apportioned among daughter cells at each division. (The level of such enzymes never reaches zero.) If at any time during this process the amino acid ceases to be available to the microorganism, synthesis of the enzyme immediately begins and continues until the proper intracellular concentration of the amino acid is reached. At this point the biosynthesis of the enzyme slows down and an equilibrium is reached such that the level of the enzyme remains constant until there is another alteration in the exogenous level of the amino acid. In contrast to feedback inhibition, which requires only milliseconds to act, repression and derepression require from one-half to several generations to reach a new equilibrium.

The actual metabolic pathways by which amino acids are synthesized are presented in diagrammatic form in Figs. 9–14. These pathways generally are found to be the same in all living cells investigated, whether microbial or animal. Biosynthetic mechanisms thus appear to have developed soon after the origin of life and to have remained unchanged through the divergent evolution of modern organisms. The major exception is lysine, which is formed from aspartic acid via diaminopimelic acid in bacteria, but from α-ketoglutaric acid in the fungi. Indeed, the occurrence of diaminopimelic acid as a precursor of lysine, or as a constituent of proteins, or both, is a major taxonomic property of the bacteria and the related blue-green algae.

Formation and transfer of amino groups. The biosynthetic pathway diagrams reveal only one quantitatively important reaction by which organic nitrogen enters the amino groups of amino acids: the reductive amination of α-ketoglutaric acid to glutamic acid by the enzyme glutamic acid dehydrogenase. All other amino acids are formed either by transamination (transfer of an amino group, ultimately from glutamic acid) or by a modification of an existing amino acid. An example of the former is the formation of valine by transfer of the amino group from glutamic acid to α-ketoisovaleric acid; an example of the latter is the reduction and cyclization of glutamic acid to form proline.

Two other direct conversions of inorganic nitrogen to amino acid nitrogen are known: the reductive amination of pyruvic acid to alanine and the addition of ammonia to fumaric acid to form aspartic acid. However, there is no evidence that either of these reactions is quantitatively important in amino nitrogen formation. In any case, ammonia is the only form

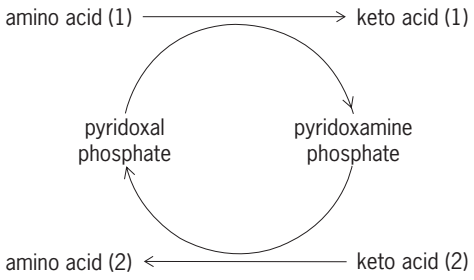


Fig. 6. Graphic representation of transamination mechanism by true transaminase.

of inorganic nitrogen which has been clearly shown to enter organic compounds directly; nitrate (NO_3^-), nitrite (NO_2^-), and nitrogen gas (N_2) are probably reduced to free intracellular ammonia before being converted to organic form in those plants and microorganisms which can use them as a nutritional source of nitrogen.

The principal mechanism of amino group transfer, transamination, is extremely important in many phases of nitrogen metabolism. Although many transamination reactions are known (all or most naturally occurring amino compounds probably participate in transamination in one tissue or another), the actual number of transaminases involved is uncertain. The few transaminases which have been highly purified all catalyze amino group exchange between more than just one pair of amino acids.

A true transaminase uses pyridoxal phosphate or pyridoxamine phosphate as coenzyme; the amino group is transferred to the former, which then gives it up to the keto acid acceptor (**Fig. 6**). A different mechanism of amino group transfer occurs in the biosyntheses of arginine and of adenylic acid; here, aspartic acid is added to a keto group to form a stable intermediate; a second enzyme then cleaves the intermediate to fumaric acid plus the new amino compound (see Fig. 10, showing the pathway of arginine biosynthesis).

One other important route by which ammonia enters organic compounds is by way of the amide group of glutamine. This group is formed by the direct addition of ammonia to glutamic acid, the necessary energy coming from the breakdown of adenosine triphosphate (ATP), first to adenosine diphosphate (ADP), then to inorganic phosphate. Once formed, amide nitrogen can be transferred to suitable acceptors to form precursors of the purine and histidine rings, as well as to hexose 6-phosphate to form glucosamine 6-phosphate.

Asparagine is another important amide, but the mechanism of asparagine formation is still in some doubt, and the only known product of the asparagine amide group is free ammonia. Glutamine is also readily deamidated to ammonia; both glutamine and asparagine serve as important storage forms of ammonia in higher plants and animals, as well as being constituent amino acids of proteins.

Degradation. Most organisms are capable of degrading some amino acids, and metabolic pathways leading to degradation to CO_2 and H_2O are known for each of the common amino acids. These pathways are detailed in the articles on the individual amino acids. There are, however, general features (**Fig. 7**) characteristic of degradative pathways which are discussed below.

The first step in the degradation of all amino acids, with the exception of tyrosine and phenylalanine, is the labilization of one of the four groups on the α-carbon atom. The labilization always involves an enzyme containing pyridoxal phosphate as a cofactor, except in the case of oxidative deamination where a flavoprotein is involved. Pyridoxal phosphate acts by forming a Schiff's base (shown as **1** in Fig. 7) with the amino acid, as shown in Fig. 7. By a rearrangement of electrons and protons (a tautomerization) the double bond moves to the position shown in **2**, with a concomitant alteration in the electron distribution in the pyridine ring of pyridoxal phosphate and the loss of the hydrogen on the α-carbon to the solvent. Intermediate **2** may now be hydrolyzed to yield the α-keto acid (**3**) corresponding to the amino acid, or it may return to intermediate **1**. In the latter case there is often a racemization, since **2** is a symmetric compound and the reversal of the tautomerization is not always carried out in an asymmetric manner.

If, instead of the hydrogen atom, the carboxyl group is labilized by donating its electrons to form the new double bond, intermediate **2***a* is formed. A reversal of the tautomerization cannot regenerate an amino acid in this case; the reaction is a decarboxylation and the product is an amine. Similarly, the organic radical R may be lost (**2***b*) whereupon the reversal of the tautomerization yields glycine. The last reaction is possible only when the radical is substituted with a hydroxyl group on the β-carbon (adjacent to the α-carbon). A final variation on the basic reactions of pyridoxal phosphate takes place when the R group contains an electronegative substituent on the β-carbon (**4**). In this case it is possible to expel the substituent, leading to an allyl amino acid, intermediate **5**, analogous to **1**. Intermediate **5** is now hydrolyzed to yield the extremely unstable intermediate **6** which rearranges to give the keto acid, ammonia, and X^- from the amino acid.

Transamination. Transamination is accomplished by hydrolysis of intermediate **2** to yield the keto acid and pyridoxamine phosphate. The latter compound reacts with another keto acid, yielding pyridoxal phosphate and the new amino acid. Transaminase enzymes are relatively nonspecific, reacting with groups of amino acids with similar R groups (for example, the aromatic amino acids or the branched-chain amino acids). The equilibrium constant of a transaminase reaction is usually very close to 1, reflecting the similarity in the free energies of formation of keto acids from amino acids. Transamination is much commoner than deamination in nature, since a deamination without a subsequent transfer to the amino group to another keto acid renders the coenzyme inactive. Most transaminase reactions involve

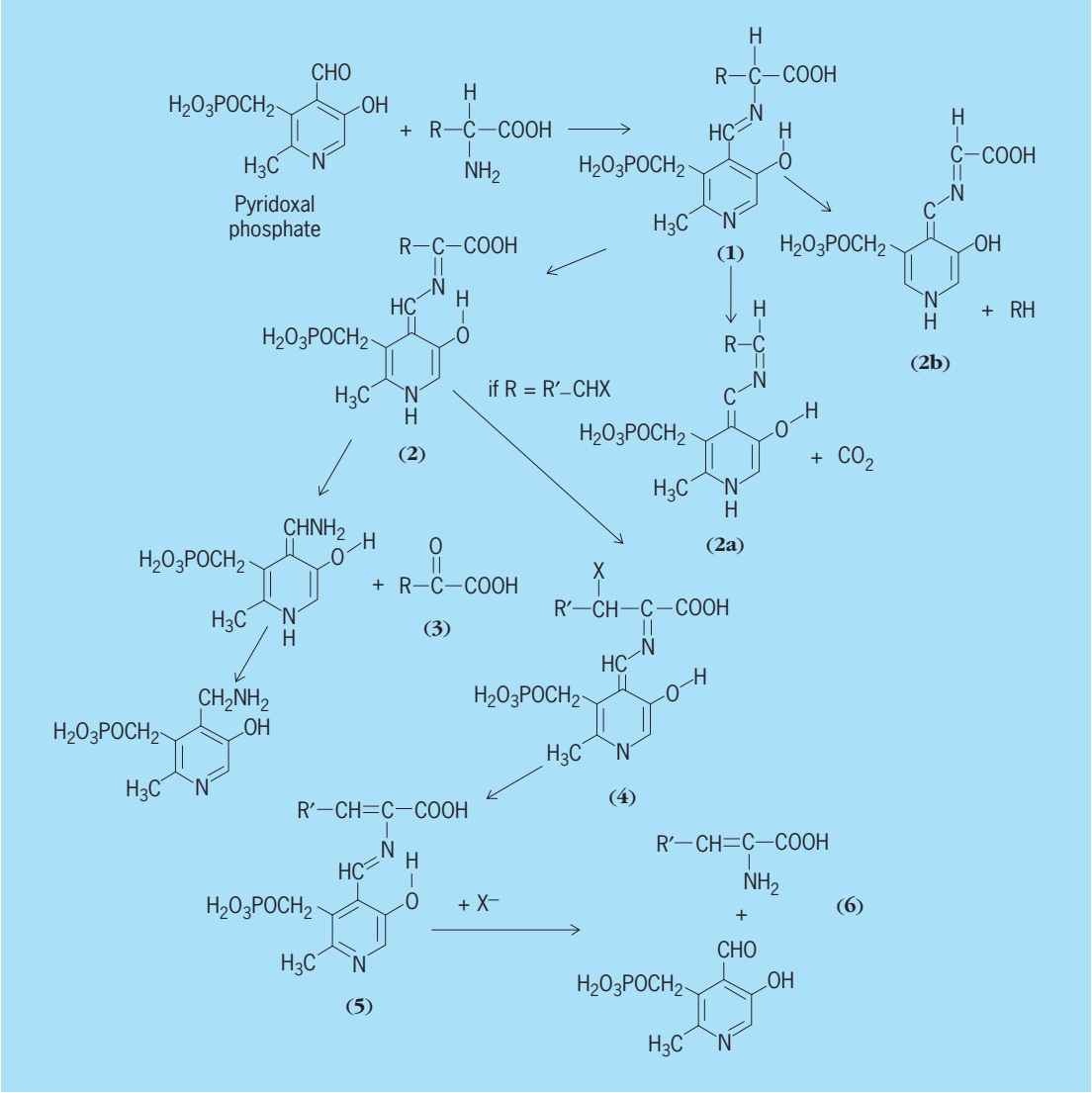


Fig. 7. General features of degradative pathways for amino acids.

aspartate or glutamate as one partner, with alanine also being rather common.

Decarboxylation. Decarboxylations are carried out by rather specific enzymes, and their products are often biologically active substances. Histamine, tryptamine (serotonin), and dopamine are all examples of decarboxylation products which are very active in animal tissues.

β Elimination. These reactions result in the loss of an electronegative group, such as OH^- or SH^-, from the carbon adjacent to the α carbon. Usually the net result is a deamination as well, since α,β unsaturated amino acids rapidly tautomerize to α,β saturated imino acids, which hydrolyze to yield the corresponding keto acid. Sometimes, however, there is addition to the α,β double bond of the pyridoxal phosphate–amino acid complex, so that the net result is a β substitution. This type of reaction is important in the synthesis of tryptophan.

β Cleavage. Instead of the α hydrogen or the carboxyl group, the β carbon can be eliminated, giving rise to glycine. These reactions only occur when the β substituent is electronegative, as in serine or threonine.

Conjugation. Until 1956 the mechanism of protein synthesis was totally unknown. Since then most of the details have been elucidated. A brief outline of the steps involved follows. The brevity of the outline precludes mention of many of the details.

The structure of a protein is determined by the arrangement of bases in the portion of the deoxyribonucleic acid (DNA) making up the gene for that protein. This sequence of DNA is transcribed into a complementary molecule of ribonucleic acid (RNA). The specific RNA, called messenger RNA (mRNA), binds to ribosomes, which are complex subcellular particles composed of a different sort of RNA, ribosomal RNA (rRNA), and protein. The ribosomes serve as the sites where conjugation of the amino acids takes place.

The amino acids are activated by reaction with a molecule of ATP to form an amino-acyl adenylate (**Fig. 8**) that remains temporarily bound to the enzyme which catalyzes the activation, an

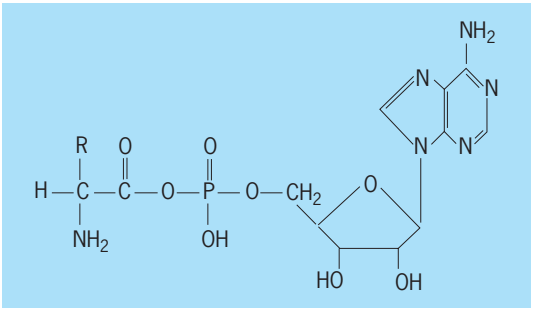


Fig. 8. Structural formula for amino-acyl adenylate.

amino-acyl-tRNA synthetase (tRNA indicates transfer RNA). The amino-acyl moiety is then transferred to a specific molecule of RNA, that is, to tRNA. The linkage of the amino acid to the tRNA takes place by esterification to the 2′-hydroxyl of the ribose moiety of the 3′-terminal end. The tRNA molecule has a complex secondary structure which exposes a sequence of three bases in the interior of the molecule. These three bases are called the anticodon. Their nature and order are specific for the amino acid involved. The amino-acyl-tRNA synthetase contributes the specificity which ensures, with very high certainty, that the amino acid is attached to a tRNA with the proper anticodon.

The amino-acyl-tRNA interacts with the ribosome messenger complex when there is a sequence of three bases (the codon) on the messenger complementary, in the Watson-Crick sense, to the anticodon. The tRNA is aligned on the ribosome in such a way that the nascent protein chain, held by a C-terminal amino acid still linked to its tRNA, is brought into proximity with the free α-amino group of the amino acid. A series of enzymes effects formation of the peptide bond, and the nascent protein chain has been elongated by one amino acid. The chain is now held to the ribosome by the tRNA of the latest residue added, and it is ready to accept the amino acid specified by the next codon. Thus the ribosome progresses along the mRNA, "reading" it, until all the amino acids specified by the original DNA gene have been added. It then reaches a codon which signifies termination of the chain, and a soluble protein is released.

The nascent protein chain begins in bacteria with a special methionyl-tRNA complex, in which the amino group of methionine is blocked with a formyl group. This is the first amino-acyl-tRNA to bind to the ribosome, and apparently all genes begin with the codon specific for this tRNA. In the first peptide bond formed, the *N*-formyl-methionyl-tRNA plays the part of the nascent protein chain.

For smaller peptides, such as glutathione and the mucopeptide of bacterial cell walls, the conjugation process is catalyzed by specific enzymes and involves different intermediates. Necessarily this means separate enzymes for each such peptide, in contrast to protein synthesis where the same machinery (except for the mRNA) serves for all chains of whatever sequence.

Amino Acids in Nutrition

The nutritional requirement for the amino acids of protein can vary from zero, in the case of an organism which synthesizes them all, to the complete list, in the case of an organism in which all the biosynthetic pathways are blocked. There are 8 or 10 amino acids required by certain mammals; most plants synthesize all of their amino acids, while microorganisms vary from types which synthesize all, to others (such as certain lactic acid bacteria) which require as many as 18 different amino acids. *See* NUTRITION; PROTEIN METABOLISM.

It seems likely that, when life originated, amino acids were taken from the rich organic medium which the oceans then offered, and biosynthetic abilities evolved only slowly as the supply of exogenous materials became depleted. A stage must have eventually been reached, however, at which all the amino acids were being synthesized metabolically, and none was required nutritionally. As evolution progressed, food chains developed, and some forms of life became adapted to obtain many of their organic nutrients at the expense of other living forms, either directly or indirectly. In these dependent types, mutations had occurred, causing the loss of specific biosynthetic enzymes and hence the gain of nutritional requirements. It is easy to duplicate this process in the laboratory: A microorganism with full biosynthetic ability can be induced to undergo random mutations, and selective methods can then be used to isolate mutants requiring amino acids, vitamins, or other normal metabolites. In every case, it is found that a given mutation deprives the cell of a single biosynthetic enzyme, blocking the reaction which that enzyme catalyzes and thus the entire pathway of which that reaction is a part. *See* PREBIOTIC ORGANIC SYNTHESIS.

In summary, the nutrition of many organisms must include the provision of growth factors, which are defined as organic compounds which the organism requires for its growth but which it cannot synthesize. Growth factor requirements reflect the heritable loss of biosynthetic enzymes, as the result of gene mutations. Amino acids are typical growth factors for many organisms.

Graphic Presentation of Amino Acid Biosynthesis

The amino acids are grouped into families on the basis of their common biosynthetic origins (Figs. 9–13). Lysine is shown in two families, because its biosynthesis in bacteria differs from that in fungi. Intermediates which are hypothetical are shown in brackets. The notation −2H or +2H refers to the removal or addition of two electrons and two hydrogen ions with the aid of either diphosphopyridine nucleotide (DPN) or triphosphopyridine nucleotide (TPN), both of which are coenzymes of hydrogen transfer. Symbols used are ~Ac, coenzyme A–bound acetate; PRRP, phosphoribosyl pyrophosphate; CAP, carbamyl phosphate; and ATP, adenosine triphosphate. An arrow between two compounds in the diagrams does not necessarily imply a single

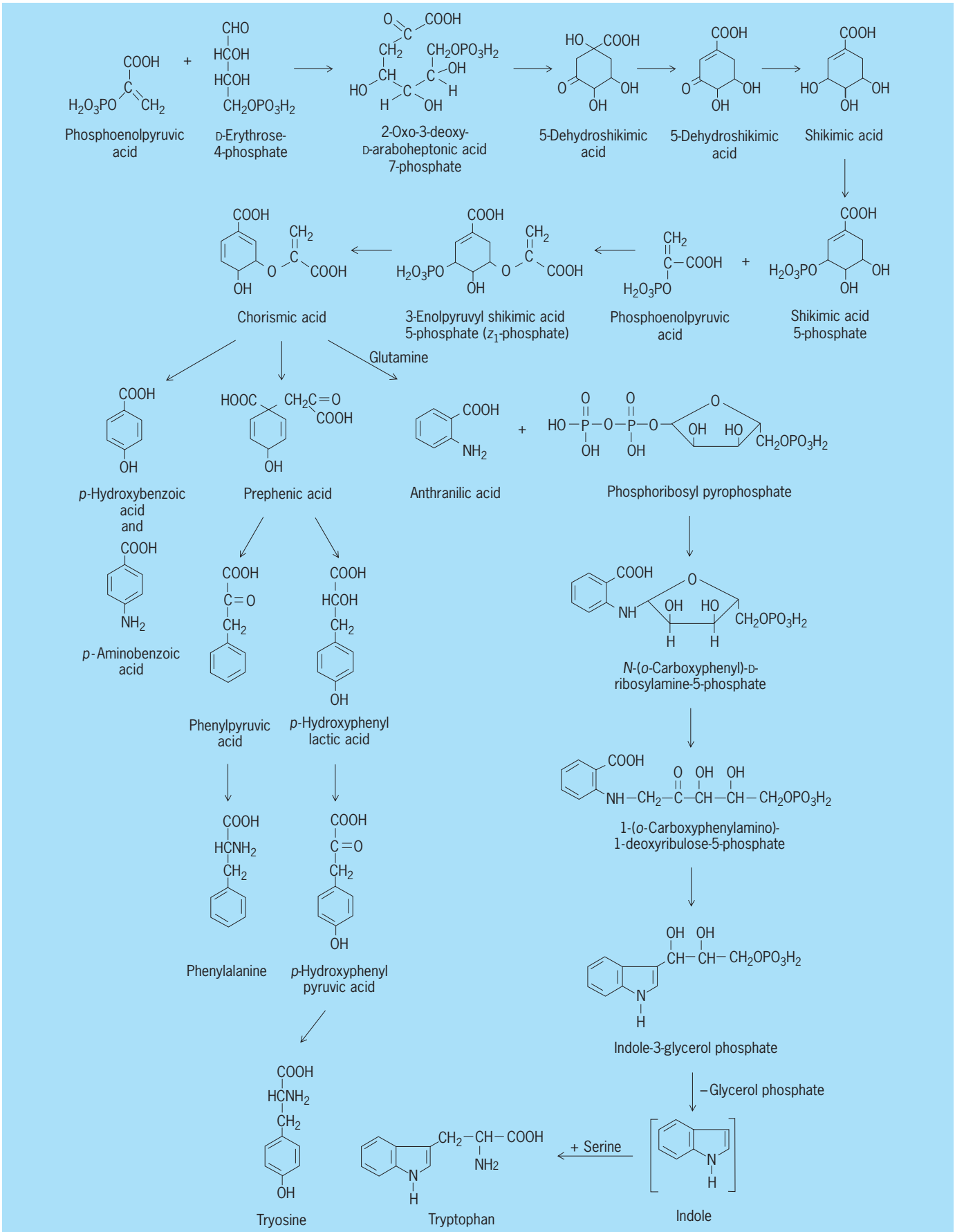


Fig. 9. Metabolic pathways for the aromatic family of amino acids.

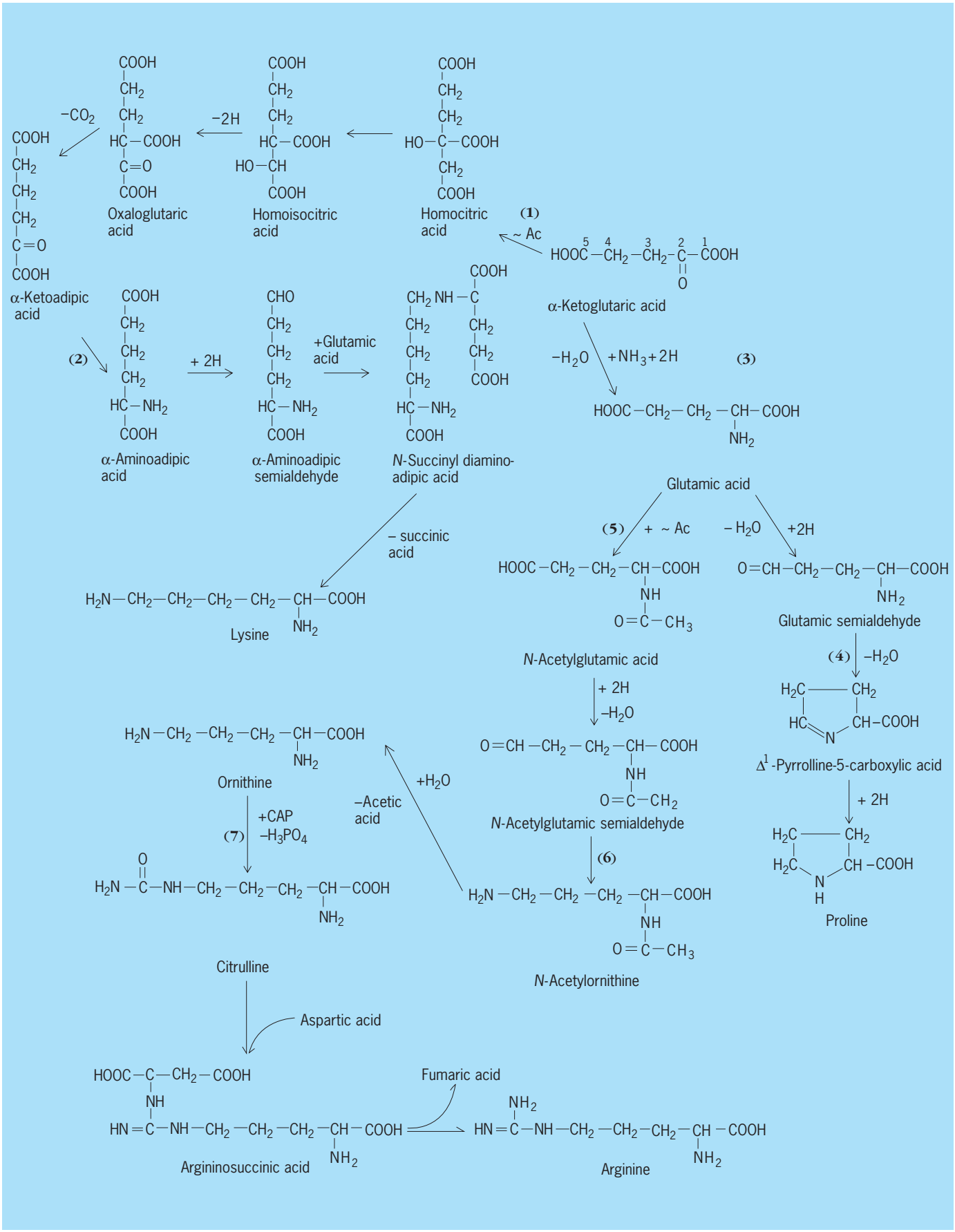


Fig. 10. Metabolic pathways for the α-ketoglutaric family of amino acids.

Fig. 11. Metabolic pathways for the aspartic acid family of amino acids.

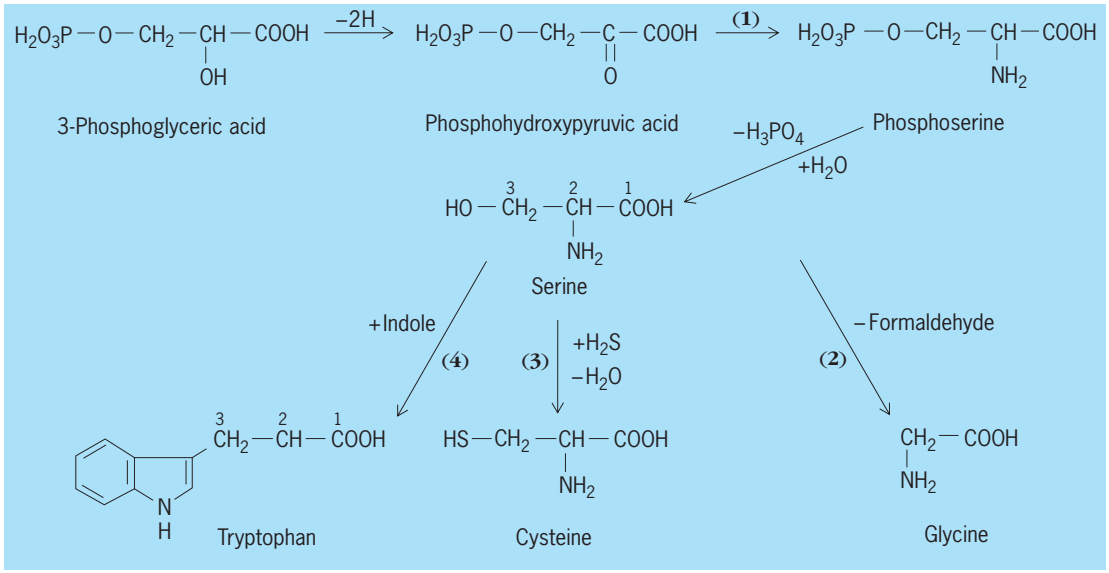


Fig. 12. Metabolic pathways for the serine family of amino acids.

enzymatic reaction. In many cases, the arrow represents a sequence of reactions for which the intermediates are unknown.

Edward A. Adelberg; Paul T. Magee

Aromatic family. This family is composed of phenylalanine, tyrosine, tryptophan, and two other important metabolites, *p*-aminobenzoic acid and *p*-hydroxybenzoic acid. The initial precursors for the biosynthesis of these amino acids, phosphoenol pyruvate and D-erythrose 4-phosphate, are metabolites of glucose catabolism.

Although the intermediates shown in **Fig. 9** have all been isolated and identified, not all of the enzymes (particularly those involved in chorismic acid metabolism) have been studied. It is probable that *p*-hydroxybenzoic acid and prephenic acid are synthesized directly from chorismic acid by a single enzymatic reaction, and *p*-aminobenzoic

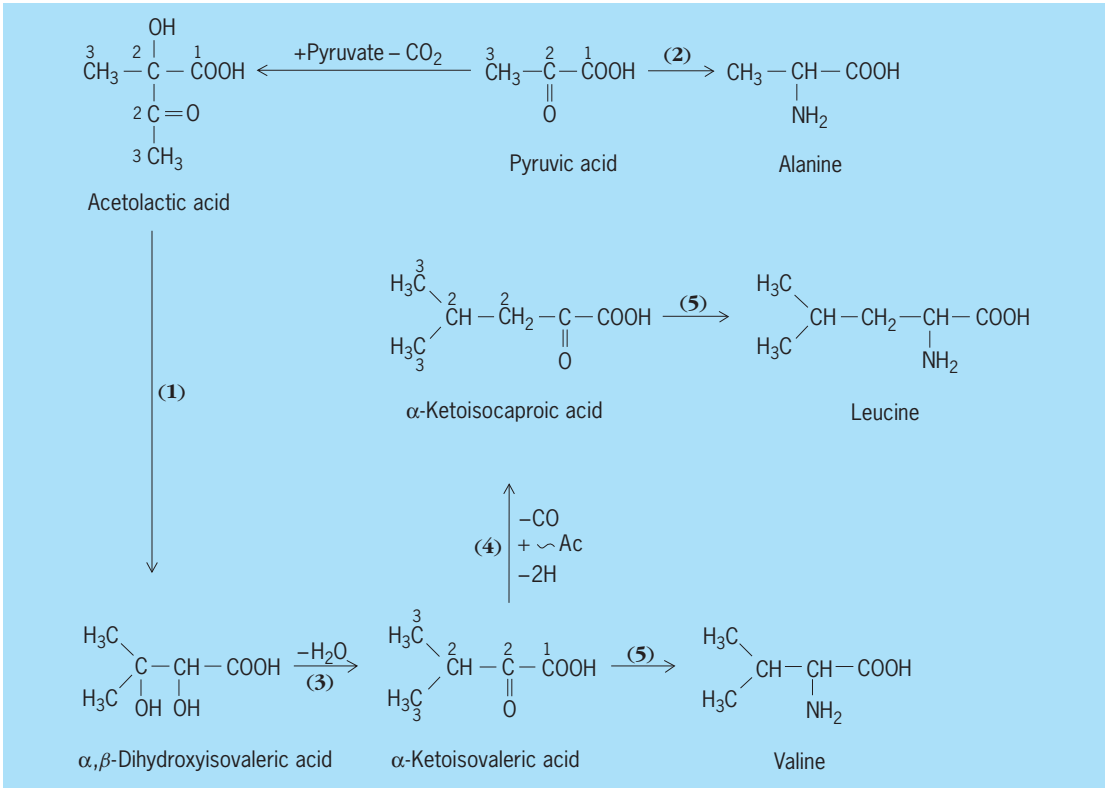


Fig. 13. Metabolic pathways for the pyruvic acid family of amino acids.

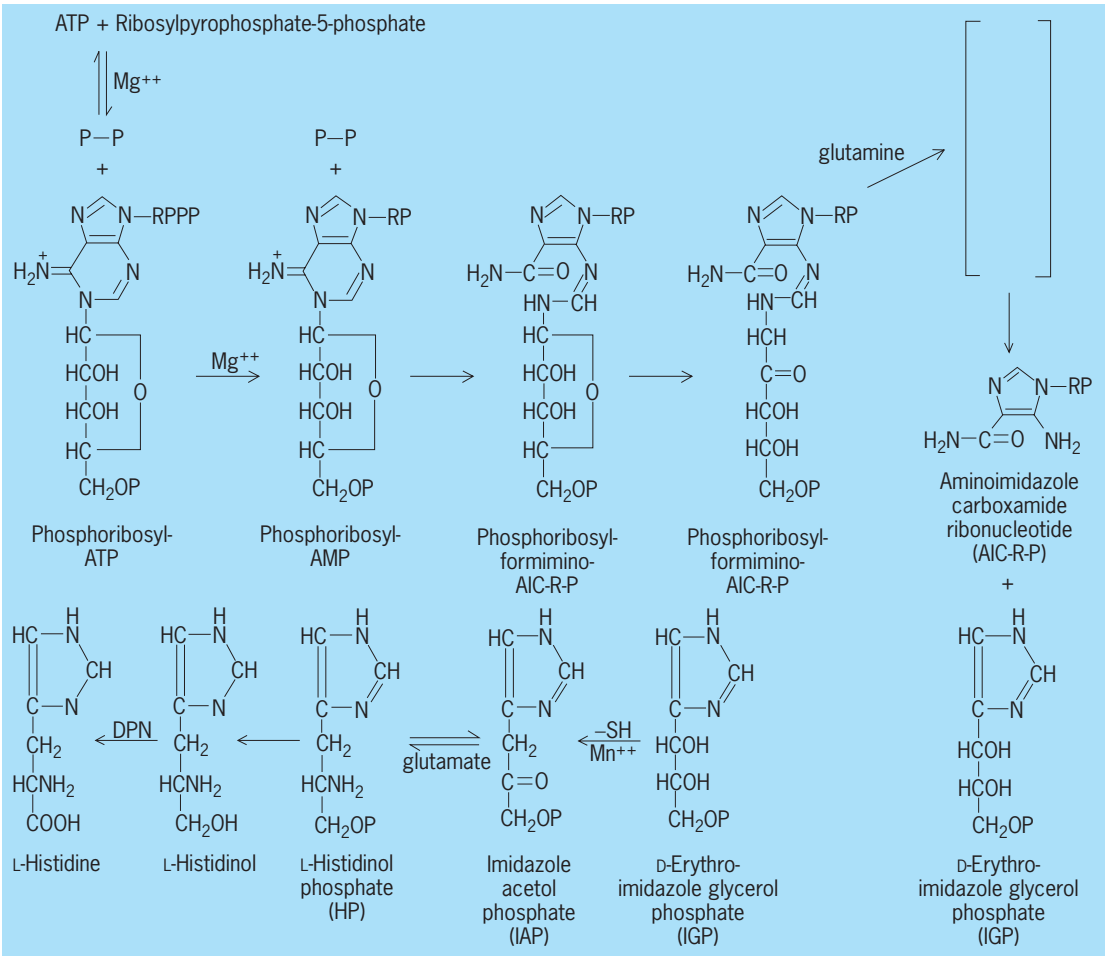


Fig. 14. Metabolic pathways in histidine biosynthesis.

acid may also be a direct metabolite. It is certain that other unidentified intermediates exist in the pathway between chorismic acid and anthranilic acid.

Two different enzymes for the conversion of chorismic acid to prephenic acid have been demonstrated in microorganisms. One of these enzymes is controlled by the pool size of tyrosine, while the other is controlled by phenylalanine.

The glycerol phosphate side chain of indoleglycerol phosphate derived from phosphoribosyl pyrophosphate can be exchanged directly for serine without the formation of free indole as an intermediate. In the absence of serine, the enzyme liberates free indole from indoleglycerol phosphate, and the same enzyme will condense indole with serine to form tryptophan.

There is some evidence for the existence of an anthranilic acid–tryptophan cycle in microorganisms. Formylkynurenine produced by the action of tryptophan pyrrolase on tryptophan can regenerate anthranilic acid by the combined action of kynureninase and kynureinine formamidase. R. G. Martin

α-Ketoglutaric acid family. This family is composed of glutamic acid, proline, lysine, and arginine. The numbered items refer to **Fig. 10**.

1. In yeast and other fungi, lysine is formed from α-ketoglutaric acid plus a C_2 fragment derivable from acetate. Lysine is formed by a different pathway in bacteria (see following section on aspartic acid).

2. Presumably by transamination.

3. This reductive amination is the main source of organic nitrogen for most microorganisms.

4. The cyclization takes place spontaneously.

5. The acetylation of glutamic acid prevents cyclization at the next step and permits the eventual formation of ornithine. This mechanism has been demonstrated in *Escherichia coli*, but does not take place in the fungus *Neurospora*; the fungus appears able to form ornithine via the nonacetylated intermediates.

6. Transamination.

7. Carbamyl phosphate (CAP).

Aspartic acid family. This family is composed of aspartic acid, lysine, threonine, methionine, and isoleucine. The numbered items refer to **Fig. 11**.

1. Aspartate arises principally by the transamination of oxaloacetate. In plants and in some microorganisms, it is formed by the direct amination of fumaric acid.

2. The compound formed by the transmination at carbon 6 and the subsequent desuccinylation is

L,L-diaminopimelic acid. For the decarboxylase to function, the compound must be racemized to form *meso*-diaminopimelic acid.

3. In bacteria, and presumably in blue-green algae, lysine is formed by decarboxylation of diaminopimelic acid. In fungi and in higher animals, lysine is formed by a different route as seen in the α-ketoglutaric acid family (Fig. 10).

4. A series of reactions probably involves transfer of an active formaldehyde group from serine, followed by reduction.

5. Intramolecular rearrangement and reduction in α-aceto-α-hydroxybutyric acid take place in one step. The same enzyme catalyzes the analogous step in valine biosynthesis (Fig. 13).

6. Transamination from glutamic acid. The same transaminase functions for the keto acids of both isoleucine and valine.

Serine family. The serine family is composed of serine, glycine, cysteine, and tryptophan. The numbered items refer to **Fig. 12**.

1. Transamination from alanine. There is equal evidence for a second pathway in which dephosphorylation precedes transamination.

2. The terminal group of serine is transferred to tetrahydrofolic acid (THFA) to form *N*(10)-hydroxymethyl-THFA. In this form, it can be transferred at various levels of oxidation for biosynthesis of compounds methionine, purine, and thymine.

3. This reaction, inferred to occur in microorganisms, is yet to be directly demonstrated. In animal tissues, serine receives the sulfhydryl group by transsulfuration from homocysteine, which is formed in animal tissues from dietary methionine.

4. See the aromatic family for considerations of this reaction.

Pyruvic acid family. This family is composed of valine, leucine, and alanine. The numbered items refer to **Fig. 13**.

1. Intramolecular rearrangement and reduction take place in one step.

2. Generally by transamination from glutamate, although cases of direct reductive amination with ammonia have been cited.

3. A single enzyme, dihydroxy acid dehydrase, catalyzes the dehydration of both the isoleucine and valine dihydroxy acid precursors.

4. The complex series of reactions involved in the formation of α-ketoisocaproic acid is exactly analogous to the steps in the citric acid cycle leading to formation of α-ketoglutarate, with the keto acid taking the place of oxaloacetate.

5. Transamination from glutamate. The valine transaminase also functions in isoleucine biosynthesis.

Edward A. Adelberg; Paul T. Magee

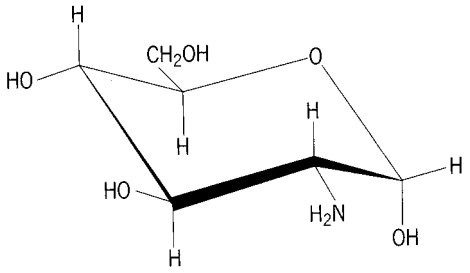
Histidine biosynthesis. The pathway of histidine biosynthesis shown in **Fig. 14** is known to occur in the mold *Neurospora* and in coliform bacteria. It should be noted that the imidazole ring of histidine is formed by the pathway and is not derived from the five-membered ring of adenosine triphosphate (ATP).

David W. E. Smith

Bibliography. G. C. Barrett et al. (eds.), *Chemistry and Biochemistry of the Amino Acids*, 1983; H. D. Jakubke and H. Jeshkeit, *Amino Acids, Peptides and Proteins*, 1978; A. Meister, *Biochemistry of the Amino Acids*, 2 vols., 2d ed., 1965.

Amino sugar

A sugar in which one or more nonglycosidic hydroxyl groups is replaced by an amino or substituted amino group. The most abundant example is D-glucosamine (2-amino-2-deoxy-D-glucose), whose structure is shown below in the α-pyranose ring form.



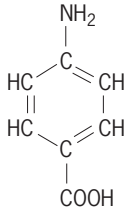
A linear polymer of *N*-acetyl-D-glucosamine is widely distributed as chitin, the exoskeletal material of arthropods. The glycoproteins of higher animals, which are components of the proteoglycans of cartilage and skin, consist of polysaccharides that are generally sulfated and have *N*-acetylated glucosamine or galactosamine alternating with a uronic acid.

Amino sugars are important constituents of glycoproteins and oligosaccharides involved in biological recognition. The 9-carbon amino sugar *N*-acetylneuraminic acid occurs widely. Biosynthetic incorporation of amino sugars commonly involves the nucleotide-sugar pathway. Amino sugars of the greatest structural diversity are found in microorganisms as constituents of cell walls, in antigenic carbohydrates produced at the cell surface, and as antibiotic substances secreted from the cell. Streptomycin is the first demonstrated example of numerous amino-sugar-containing antibiotics produced notably by Actinomycetes (bacteria). *See* GLYCOPROTEIN; OLIGOSACCHARIDE; POLYSACCHARIDE.

Derek Horton

para-Aminobenzoic acid

A compound also known as PABA, often considered to be a water-soluble vitamin, with the structure below. *p*-Aminobenzoic acid is widely distributed in



foods and has been isolated from liver, yeast, and other sources rich in vitamin B. There is doubt, however, that it is significant as a nutrient. In the early 1940s a considerable literature concerning the role of *p*-aminobenzoic acid in curing deficiency disease appeared, but it is now known that this effect was due in great part to other vitamins, particularly folic acid, about which little was then known. *p*-Aminobenzoic acid is a part of the folic acid molecule, and its presence in a folic acid–deficient diet results in increased intestinal synthesis of the folic acid.

p-Aminobenzoic acid antagonizes the bacteriostatic action of sulfonamides. Because of the similar chemical compositions of these substances, it is probable that the sulfonamides function at least in part by displacing *p*-aminobenzoic acid in bacterial enzyme systems. *p*-Aminobenzoic acid is an effective antirickettsial agent and has been used to treat typhus, scrub typhus, and spotted fever. There is no evidence that humans have a dietary requirement for this vitamin. The compound is also used as a sun-screen agent in suntan lotions. *See* FOLIC ACID; RICKETTSIOSES; VITAMIN. Stanley N. Gershoff

Ammeter

An electrical instrument for measuring electric current. Currents are usually either unidirectional and steady (direct current or dc) or alternating in direction at a relatively low frequency (alternating current or ac). A current that is unidirectional but regularly fluctuating is a superposition of dc and ac. Higher-frequency ac is often referred to as radio-frequency or RF current. At frequencies above about 10 MHz, where the wavelength of the signal becomes comparable with the dimensions of the measuring instrument, current measurements become inaccurate and finally meaningless, since the value obtained depends on the position where the measurement is made. In these circumstances, power measurements are usually used. *See* CURRENT MEASUREMENT; MICROWAVE MEASUREMENTS.

Magnetic fields, thermal (heating) effects, and chemical effects resulting from the passage of electric current may be employed to create measuring instruments. Current may also be measured in terms of the voltage that appears across a resistive shunt through which the current passes. This method has become the most common basis for ammeters, primarily because of the very wide range of current measurement that it makes possible, and more recently through its compatibility with digital techniques. *See* MULTIMETER; VOLTMETER.

The unit of current, the ampere, is the base unit on which rest the International System (SI) definitions of all the electrical units. The ampere itself is defined in terms of the mechanical forces that are produced by the magnetic interaction of two currents. Primary instruments in national standards laboratories measure currents in accordance with this definition and are at the head of a hierarchy of everyday calibrated current-measuring instruments. *See* CURRENT BALANCE; ELECTRICAL UNITS AND STANDARDS; WATT BALANCE.

Direct-current ammeters. The moving-coil, permanent-magnet (d'Arsonval) ammeter remains an important type. Their mechanical construction is identical with that of a d'Arsonval moving-coil instrument. *See* VOLTMETER.

These ammeters continue to be in large-scale production and are available in many versions, from small indicators to large instruments suitable for control panels. Generally they are of modest accuracy, no better than 1%. Digital instruments have taken over all measurements of greater precision because of the greater ease of reading their indications where high resolution is required.

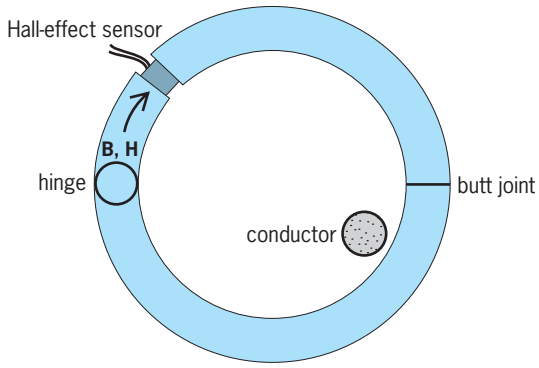
By suitable choice of coil winding and strength of magnet, moving-coil meters of this basic design can give full deflection at currents in the range from 10 microamperes to about 10 milliamperes. Increased sensitivity can be obtained by the use of electronic amplifiers. Ammeters intended for higher currents use similar instruments, modified to require some 50 mV for full-scale deflection, in conjunction with shunts. *See* AMPLIFIER.

Low-frequency ac ammeters. Moving-iron instruments are widely used as ammeters for low-frequency ac applications and their mechanical construction is identical with a moving-iron repulsion voltmeter. *See* VOLTMETER.

Although the design will also operate with direct current, hysteresis in the magnetic material limits accuracy. Eddy currents limit the usefulness to low frequencies.

High-frequency ammeters. High-frequency currents are measured by the heating effect of the current passing through a physically small resistance element. One early design, the hot-wire ammeter, used the increase in length of a wire that was heated by the current. The increase was mechanically amplified in order to move the pointer. In modern instruments the temperature of the center of the wire is sensed by a thermocouple, the output of which is used to drive a moving-coil indicator. A wide range of full-scale sensitivities can be provided, from about 50 A with a heavy-gage heater in air to 1 mA by using a fine winding on an insulating bead around the thermocouple, working in a small evacuated glass bulb. *See* THERMOCOUPLE.

Digital ammeters. Although most forms of digital ammeter operate by balancing currents derived from the signal and a reference source, analog-to-digital converters are restricted to full-range currents in the region of 1 mA. Higher values require inconveniently large circuit currents; lower currents bring problems of accuracy because of leakage. Digital ammeters are therefore almost invariably based on a conventional digital voltmeter which measures the voltage drop caused by the current flowing through a shunt resistor. *See* VOLTMETER.



Magnetic circuit of a clip-on ammeter or current probe. B = magnetic flux density, H = magnetic field strength.

Alternating-current measurement. Alternating currents may be measured by the addition of an ac converter. This converter transforms the ac signal at the output of the scaling amplifier to an equivalent dc signal for processing by the analog-to-digital conversion circuits. Shunts and current transformers can be used for ranges up to tens of thousands of amperes; electronic amplifiers and electrometers can extend the current measurement capability to the femtoampere (10^{-15} A) region. More refined analog-to-digital converters can be used to provide improved resolution and accuracy. *See* ANALOG-TO-DIGITAL CONVERTER; TRANSFORMER.

Clip-on ammeters (clamp meters). These hand-held devices possess the exceedingly useful attribute of being able to measure a current in a particular conductor without interrupting the circuit to insert, for example, a current-measuring resistor. They accomplish this by clamping a hinged high-permeability magnetic core around the conductor, which thus becomes a single-turn primary winding of a current transformer. If alternating current is to be measured, a secondary winding around the core which is connected to an ammeter suffices to complete the instrument. If direct current is to be measured, a Hall-effect flux sensor is inserted transversely in a narrow gap in the core to measure the magnetic flux (see **illustration**). The flux density in the magnetic circuit resides perdominantly in the gap whose dominant and constant reluctance ensures that this flux, and therefore the voltage output of the Hall sensor, is proportional to the current. This direct-current instrument is best calibrated with known currents. *See* HALL EFFECT.

Both the alternating-current and direct-current instruments possess the convenient property that the measurement is, in principle, independent of the position of the current-carrying conductor within the core. This is because of Ampère's theorem that the line integral of flux around a closed magnetic path is independent of the position within the magnetic path of the current which is the source of the flux.

The accuracy of these instruments is usually only a few percent, but this is adequate for many applications. Connecting the secondary winding or Hall-sensor output to an oscilloscope produces the particularly useful fault-tracing tool called a current probe. It is often the current in one particular conductor of a network rather than the voltage differences between junction points that is important.

Sampling techniques. A different digital solution to the problem of measuring alternating or varying currents is the use of sampling. The signal, after suitable scaling and conversion to a voltage, is sampled repetitively. These samples are digitized, using a fast analog-to-digital converter. The resulting stream of numerical values is then processed mathematically to yield the information required, such as the root-mean-square (rms) value, mean value, and dc component. In order to accurately reflect the properties of the input waveform, the sampling rate needs to be at least twice the highest frequency present in the signal. *See* ELECTRICAL MEASUREMENTS; INFORMATION THEORY; ROOT-MEAN-SQUARE.

R. B. D. Knight; Bryan P. Kibble

Bibliography. H. M. Berlin and F. C. Getz, Jr., *Principles of Electronic Instrumentation and Measurement,* 1988; H. H. Chiang, *Electrical and Electronic Instrumentation,* 1984; S. Geczy, *Basic Electrical Measurements,* 1984; A. D. Helfrick and W. D. Cooper, *Modern Electronic Instrumentation and Measurement Techniques,* 1990; L. Jones, *Electronic Instruments and Measurements,* 1991; B. Kibble et al., *A Guide To Measuring Direct and Alternating Current and Voltage Below 1 MHz*, Institute of Measurement and Control, 2003; M. V. Reissland, *Electrical Measurements: Fundamentals, Concepts, Applications,* 1990.

Ammine

One of a group of complex compounds formed by the coordination of ammonia molecules with metal ions and, in a few instances, such as calcium, strontium, and barium, with metal atoms. Some typical examples of ammines include $[Co(NH_3)_6]Cl_2$ (rose), $[Cu(NH_3)_4]Cl_2$ (blue), $[Cr(NH_3)_6]Cl_3$ (yellow), $[Cr(NH_3)_4Cl_2]Cl$ (cis is violet, trans is green), $[Ni(NH_3)_6]Cl_2$ (blue), $[Pt(NH_3)_4]Cl_2{\cdot}H_2O$ (colorless), and $[Hg(NH_3)_2]Br_2$ (colorless). Although these ammines are formally analogous to many salt hydrates, the general characteristics of the group of ammines differ considerably from those of the hydrates. For example, hydrated Co(III) salts are strong oxidizing agents whereas Co(II) ammines are strong reducing agents. The ammines of principal interest are those of the transition metals and of the zinc family, but even here there is wide variation in stability or rate of decomposition. For example, iron ammines are unstable in aqueous solution; Cu(II) and Co(II) ammines exist in aqueous solution but are decomposed by aqueous acids; Co(III) and Pt(IV) ammines can be recrystallized from strong acids. Ammines are prepared by treating aqueous solutions of the metal salt with ammonia or, in some instances, by the action of dry gaseous or liquid ammonia on the anhydrous salt. These, and similar, differences have motivated much

theoretical study of the bonding in such compounds. *See* AMMONIA; CHEMICAL BONDING; COORDINATION CHEMISTRY.

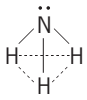
Harry H. Sisler

Ammonia

The most familiar compound composed of the elements nitrogen and hydrogen, NH_3. It is formed as a result of the decomposition of most nitrogenous organic material, and its presence is indicated by its pungent and irritating order.

Uses. Because of the wide range of industrial and agricultural applications, ammonia is produced in tremendous quantities. Examples of its use are the production of nitric acid and ammonium salts, particularly the sulfate, nitrate, carbonate, and chloride, and the synthesis of hundreds of organic compounds including many drugs, plastics, and dyes. Its dilute aqueous solution finds use as a household cleansing agent. Anhydrous ammonia and ammonium salts are used as fertilizers, and anhydrous ammonia also serves as a refrigerant, because of its high heat of vaporization and relative ease of liquefaction. *See* FERTILIZER.

Molecular structure. The NH_3 molecule has a pyramidal structure of the type shown in the diagram below, in which the nitrogen atom has achieved a



stable electronic configuration by forming three electron pair bonds with the three hydrogen atoms. The HNH bond angle in the pyramid is 106.75°, which is best explained as resulting from the use of sp^3 hybrid bonding orbitals by the nitrogen atom. This should yield tetrahedral bond angles with the result that one sp^3 orbital is occupied by the unshared pair of electrons. The repulsive effect of this unshared pair, which is concentrated relatively near to the nitrogen nucleus on the shared pairs of electrons forming the N-H bonds, produces a slight compression of the HNH bond angles, thus accounting for the fact that they are slightly less than tetrahedral (109.5°). The dipole moment of the ammonia molecule, 1.5 debyes, is a resultant of the combined polarities of the three N-H bonds and of the unshared electron pair in the highly directional sp^3 orbital. The pyramidal ammonia molecule turns inside out readily, and it oscillates between the two extreme positions at the precisely determined frequency of 2.387013×10^{10} Hz. This property has been used in the highly accurate time-measuring device known as the ammonia clock.

Physical characteristics. The physical properties of ammonia are analogous to those of water and hydrogen fluoride in that the physical constants are abnormal with respect to those of the binary hydrogen compounds of the other members of the respective periodic families. This is particularly true of the boiling point, freezing point, heat of fusion, heat of vaporization, and dielectric constant. These abnormalities may be related to the association of molecules through intermolecular hydrogen bonding. The principal physical constants for ammonia are summarized in **Table 1**. Ammonia is highly mobile in the liquid state and has a high thermal coefficient of expansion.

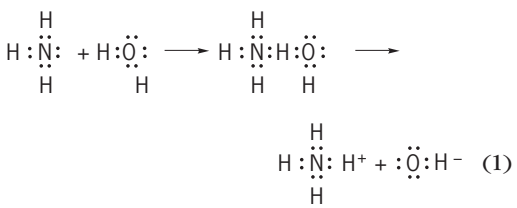
TABLE 1. Physical properties of ammonia

Property	Value
Melting point	−77.74°C (−107.9°F)
Boiling point	−33.35°C (−28.03°F)
$\Delta H_{fus.}$ at mp	5657 J mole^{-1}
$\Delta H_{vap.}$ at bp	23,350 J mole^{-1}
Critical temperature	132.4°C (270.3°F)
Critical pressure	112.5 atm (11.40 MPa)
Dielectric constant (−77.7°C)	25
Density (−70°C)	0.7253 g cm^{-3}
Density (−30°C)	0.6777 g cm^{-3}
$\Delta H°_{form.}$ (25°C)	46.19 kJ mole^{-1}
$\Delta G°_{form.}$ (25°C)	16.64 kJ mole^{-1}
$S°$ (298.1 K) (exptl.)	192.5 J deg^{-1} mole^{-1}
$C°_p$	35.66 J deg^{-1} mole^{-1}
Viscosity (liquid, 25°C)	0.01350 Pa · s
Vapor pressure (−20°C)	1426.8 torr (190.22 kPa)
Vapor pressure (0°C)	3221.0 torr (429.43 kPa)
Vapor pressure (20°C)	6428.5 torr (857.06 kPa)
Solubility in H_2O	
(1 atm,* 0°C)	42.8% by wt
(1 atm, 20°C)	33.1% by wt
(1 atm, 40°C)	23.4% by wt

*1 atm = 101.325 kilopascals.

Chemical properties. Most of the chemical reactions of ammonia may be classified under three chief groups: (1) addition reactions, commonly referred to as ammonation; (2) substitution reactions, commonly referred to as ammonolysis; and (3) oxidation-reduction reactions.

Ammonation. Ammonation reactions include those in which ammonia molecules add to other molecules or ions either through the mechanism of covalent bond formation using the unshared pair of electrons on the nitrogen atom, through ion-dipole electrostatic interactions, or through hydrogen bonding. Most familiar of the ammonation reactions is the reaction with water, which may be represented schematically as reaction (1). The strong tendency of water

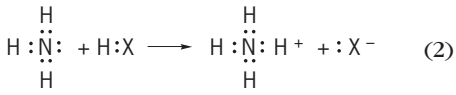


(1)

and ammonia to combine is evidenced by the very high solubility of ammonia in water (700 vol of ammonia gas in 1 vol of water at 20°C and 1 atm ammonia pressure). The ammonia hydrate (ammonium hydroxide) is a weak electrolyte in aqueous solution as is indicated by an ionization constant of 1.77×10^{-5} at 25°C. Phase diagrams for the NH_3-H_2O system indicate the existence of $NH_3 \cdot H_2O$ (mp −79.0°C

or −110.0°F) in the solid state at low temperatures. Under these conditions, the compound $2NH_3 \cdot H_2O$ (mp −78.8°C or −109.8°F) also exists. Ammonia reacts readily with strong acids to form ammonium salts, reaction (2). Ammonium salts of weak acids in the solid state dissociate readily into ammonia and the free acid.

Included among ammonation reactions is the formation of complexes (called ammines) with many

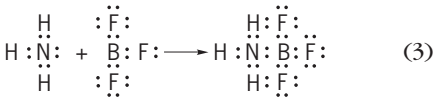


(2)

($:X^-$ = an anion)

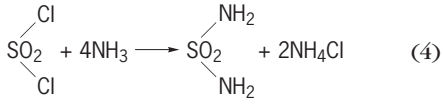
metal ions, particularly transition-metal ions, for example, $Hg(NH_3)_2^{+}$, $Cr(NH_3)_6^{3+}$, $Zn(NH_3)_4^{2+}$, and $Co(NH_3)_6^{3+}$. *See* AMMINE.

Ammonation occurs with a variety of molecules capable of acting as electron acceptors (Lewis acids). The reaction of ammonia with such substances as sulfur trioxide, sulfur dioxide, silicon tetrafluoride, and boron trifluoride are typical and illustrated by reaction (3).

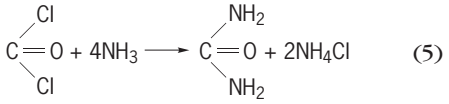


(3)

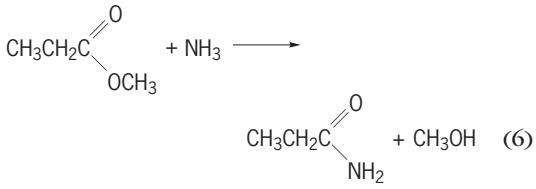
Ammonolysis. Ammonolytic reactions include reactions of ammonia in which an amide group (—NH_2), an imide group (=NH), or a nitride group (≡N) replaces one or more atoms or groups in the reacting molecule. Examples include reactions (4) to (8).



(4)

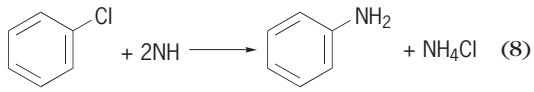


(5)



(6)

$$HgCl_2 + 2NH_3 \longrightarrow Hg(NH_2)Cl + NH_4Cl \qquad (7)$$



(8)

Oxidation-reduction. These reactions may be subdivided into those which involve a change in the oxidation state of the nitrogen atom and those in which elemental hydrogen is liberated. An example of the first group is the catalytic oxidation of ammonia in air to form nitric oxide, shown in reaction (9). In the

$$4NH_3 + 5O_2 \xrightarrow{Pt} 4NO + 6H_2O \qquad (9)$$

absence of a catalyst, ammonia burns in oxygen to yield nitrogen. This is shown by reaction (10). Another example, reaction (11) is the reduction with ammonia of hot metal oxides such as cupric oxide. Still another example is the reaction of ammonia with chlorine (12).

$$4NH_3 + 3O_2 \longrightarrow N_2 + 6H_2O \qquad (10)$$

$$3CuO + 2NH_3 \longrightarrow 3Cu + 3H_2O + N_2 \qquad (11)$$

$$2NH_3 + Cl_2 \longrightarrow NH_2Cl + NH_4Cl \qquad (12)$$

Oxidation-reduction reactions of ammonia of the second type are exemplified by reactions of active metals with ammonia, shown in reactions (13) and (14).

$$2Na + 2NH_3 \rightarrow 2NaNH_2 + H_2 \qquad (13)$$

$$3Mg + 2NH_3 \xrightarrow{\Delta} Mg_3N_2 + 3H_2 \qquad (14)$$

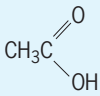
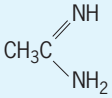
Liquid ammonia as a solvent. The physical and chemical properties of liquid ammonia make it appropriate for use as a solvent in certain types of chemical reactions. The solvent properties of liquid ammonia are, in many ways, qualitatively intermediate between those of water and of ethyl alcohol. This is particularly true with respect to dielectric constant; therefore, ammonia is generally superior to ethyl alcohol as a solvent for ionic substances but is inferior to water in this respect. On the other hand, ammonia is generally a better solvent for covalent substances than is water. The chemical properties of ammonia, for example, its ability to undergo ammonation, ammonolysis, and oxidation-reduction reactions, are roughly analogous to reactions of water (hydration, hydrolysis, and oxidation-reduction). Both water and liquid ammonia undergo autoionization; liquid ammonia, reaction (15), undergoes the process to a lesser extent, as its very low ion product constant indicates, reaction (16).

$$2NH_3 \rightleftharpoons NH_4^+ + NH_2^- \qquad (15)$$

$$[NH_4^+][NH_2^-] = 1.9 \times 10^{-33} \text{ at } -50°C \qquad (16)$$

Among the particularly interesting aspects of chemistry in liquid ammonia is the fact that metals of sufficiently low lattice energy, high cation solvation energy, and low ionization energy are reversibly soluble in liquid ammonia. The solutions obtained are highly colored—blue if dilute, bronze if concentrated. Metals which form such solutions include the alkali metals, the heavier alkaline earth metals, and the divalent lanthanides. Concentrations as high as 10 to 20 molal are obtainable for the alkali metals. Dilute solutions exhibit the phenomenon of strong electrolytic conduction, whereas the concentrated solutions act as metallic conductors. At very high

TABLE 2. Analogous compounds in the water and ammonia systems

Aquo	Ammono
$(H_3O)Cl$	$(NH_4)Cl$
	
KOH	KNH_2
$Na_2[Zn(OH)_4]$	$Na_2[Zn(NH_2)_4]$
$Cu(H_2O)_4^{2+}$	$Cu(NH_3)_4^{2+}$
MgO	$MgNH$, Mg_3N_2
	
C_2H_5OH	$C_2H_5NH_2$
$(CH_3)_2O$	$(CH_3)_2NH$, $(CH_3)_3N$
$Hg(OH)Cl$	$HgNH_2Cl$
$HOCl$	H_2NCl

dilutions magnetic measurements indicate the presence of 1 mole of unpaired electrons per mole of dissolved metal, but as the concentrations of the solutions are increased, their paramagnetism decreases. Much remains to be learned concerning the nature of these solutions, but presently available data may be interpreted in terms of the five solute species M, $M_{2(am)}$, $M^-_{(am)}$, $M^+_{(am)}$, and $e^-_{(am)}$ participating in equilibria (17)–(19). The ammoniated electron, $e_{(am)}$,

$$M \rightleftharpoons M^+_{(am)} + e^-_{(am)} \tag{17}$$

$$2M \rightleftharpoons M_{2(am)} \tag{18}$$

$$e^-_{(am)} + M \rightleftharpoons M^-_{(am)} \tag{19}$$

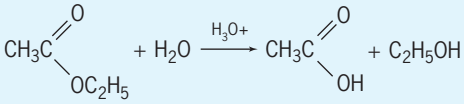
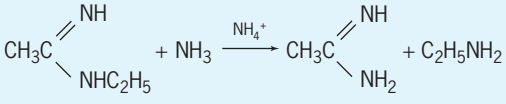
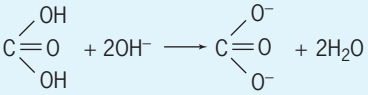
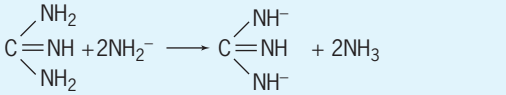
appears to consist of an electron trapped in a cavity (approximately 0.3 nanometer in diameter) of NH_3 molecules. These solutions are unstable but, in the absence of catalysts, decompose only very slowly to yield the metal amide plus hydrogen gas [reaction (20)]. Solutions of metals in liquid ammonia are excellent reducing agents and are particularly useful for the reduction of organic compounds which are miscible with liquid ammonia.

$$e^-_{(am)} + NH_3 \rightarrow {}^1/_2H_2 + NH_2^- \tag{20}$$

The usefulness of liquid ammonia as a solvent is based on the differences in the chemical properties of liquid ammonia and of other common solvents, notably water. Principal among these differences (compared with water) are (1) the lesser tendency of ammonia to release protons, (2) the greater electron donor tendency (or proton affinity) of ammonia, and (3) the stronger reducing character of ammonia may be used as a solvent for very strong bases (such as NH_2^- or $C_2H_5O^-$) which would undergo complete protolysis in aqueous solution. Liquid ammonia may also be used as a solvent for very strong reducing agents (such as solvated electrons) which would immediately displace hydrogen from water. Because of the second difference, liquid ammonia solutions do not provide very strong acids, since all strong acids are converted immediately to ammonium ion, NH_4^+, which is a much weaker acid than hydronium ion, H_3O^+, its counterpart in aqueous systems. In summary, it may be said that as a solvent for chemical reactions, liquid ammonia affords much stronger bases and stronger reducing agents but much weaker acids and weaker oxidizing agents than does water. By application of these differences, a number of interesting synthetic procedures may be carried out in liquid ammonia.

Ammonia system of compounds. Many of the familiar compounds which contain oxygen may be considered to be derived from water as the parent solvent. In an analogous way it is sometimes useful to consider many nitrogen-containing compounds to be derived from ammonia as parent solvent. These latter compounds are sometimes considered to constitute the nitrogen or ammonia system of compounds. This analogy is useful in understanding the chemistry of various nitrogen compounds (**Tables 2** and **3**).

TABLE 3. Analogous compounds in the water and ammonia systems

Aquo	Ammonia
$KOH + (H_3O)Cl \longrightarrow KCl + 2H_2O$	$KNH_2 + (NH_4)Cl \longrightarrow KCl + 2NH_3$
$Zn + 2H_3O^+ \longrightarrow Zn^{2+} + H_2 + 2H_2O$	$Zn + 2NH_4^+ \longrightarrow Zn^{2+} + H_2 + 2NH_3$
	
	
$Zn(OH)_2 + 2OH^- \longrightarrow Zn(OH)_4^{2-}$	$Zn(NH_2)_2 + 2NH_2^- \longrightarrow Zn(NH_2)_4^{2-}$

Synthesis of ammonia. The Haber-Bosch synthesis is the major source of industrial ammonia. In a typical process, water gas (CO, H_2, CO_2) mixed with nitrogen is passed through a scrubber cooler to remove dust and undecomposed material. The CO_2 and CO are removed by a CO_2 purifier and ammoniacal cuprous solution, respectively. The remaining H_2 and N_2 gases are passed over a catalyst at high pressures (up to 1000 atm or 100 megapascals) and high temperatures (approx. 700°C or 1300°F) [reaction (21)].

$$3H_{2(g)} + N_{2(g)} \rightleftharpoons 2NH_{3(g)} \quad (21)$$

The ammonia is separated by absorption in water. Processes used vary widely in the sources of N_2 and H_2, treatment of the catalysts, temperature, pressure, and methods of ammonia separation.

Other industrial sources of ammonia include its formation as a by-product of the destructive distillation of coal, and its synthesis through the cyanamide process, which is indicated by reactions (22) and (23).

$$CaC_2 + N_2 \xrightarrow{1000^\circ C} CaCN_2 + C \quad (22)$$

$$CaCN_2 + 3H_2O\ (\text{steam}) \longrightarrow CaCO_3 + 2NH_3 \quad (23)$$

In the laboratory, ammonia is usually formed by its displacement from ammonium salts (either dry or in solution) by strong bases, as indicated by reaction (24). Another source is the hydrolysis of metal nitrides, as indicated reaction (25).

$$NH_4^+ + OH^- \xrightarrow{\text{heat}} NH_3 + H_2O \quad (24)$$

$$Mg_3N_2 + 6H_2O \longrightarrow 3Mg(OH)_2 + 2NH_3 \quad (25)$$

See AMIDE; AMINE; HIGH-PRESSURE PROCESSES; HYDRAZINE; NITROGEN.

Harry H. Sisler

Bibliography. J. R. Jennings, *Catalytic Ammonia Synthesis*: *Fundamentals and Practice*, 1991; H. H. Sisler, R. D. Dresdner, and W. T. Mooney, Jr., *Chemistry*: *A Systematic Approach*, 1980; S. Strelzoff, *Technology and Manufacture of Ammonia*, 1981, reprint 1988.

Ammonium salt

A product of a reaction between ammonia, NH_3, and various acids. The general reaction for formation is (1). Examples of ammonium salts are ammonium chloride, NH_4Cl, ammonium nitrate, NH_4NO_3, and ammonium carbonate, $(NH_4)_2CO_3$. These compounds are addition products of ammonia and the acid. For this reason, their formulas are sometimes written as $[H(NH_3)]X$.

$$NH_3 + HX \longrightarrow NH_2X \quad (1)$$

All ammonium salts decompose into ammonia and the acid when heated. Their stability, however, varies according to the nature of the acid. Salts of weak acids decompose at lower temperatures than do salts of strong acids. Ammonium chloride, the salt of the strong acid hydrogen chloride, HCl, decomposes at 320°C (608°F), whereas ammonium sulfide, $(NH_4)_2S$, the salt of the weak acid hydrogen sulfide, H_2S, decomposes at 32°C (90°F). If the salt is heated in a closed vessel, a definite pressure of ammonia is established in the presence of the solid salt. This pressure is determined solely by the temperature and, if the acid is nonvolatile, is called the dissociation pressure at that temperature. For a detailed discussion of such equilibra *see* CHEMICAL EQUILIBRIUM.

If anhydrous ammonia is added to many of the ammonium salts at very low temperatures, salts containing several molecules of ammonia are formed. Ammonium chloride, for example, can add three or six molecules of ammonia to form complex salts. The reaction for the formation of the tetra compound is (2). The number of such complexes which maybe

$$[H(NH_3)]Cl + 3(NH_3) \longrightarrow [H(NH_3)_4]Cl \quad (2)$$

formed depends upon the nature of the acid radical. When warmed, they lose ammonia; all are unstable above 0°C (32°F).

Ammonium chloride is made by absorbing ammonia in hydrochloric acid. It crystallizes from the solution in feathery crystals of the regular crystal system. It is a colorless solid with a density of 1.52. This salt, sometimes called sal ammoniac, is used in galvanizing iron, in textile dyeing, and in manufacturing dry cell batteries.

Ammonium nitrate, NH_4NO_3, a colorless salt with a density of 1.73, is prepared from ammonia and nitric acid. The solid salt deliquesces, or absorbs water from moist air, thus appearing to melt. It is used as a source of nitrous oxide, N_2O, or laughing gas, and in the manufacture of explosives. A mixture of ammonium nitrate and trinitrotoluene is referred to as amatol.

Ammonium sulfate, $(NH_4)_2SO_4$, obtained from ammonia and sulfuric acid, is a colorless solid with a density of 1.77. It is prepared commercially by passing ammonia and carbon dioxide, CO_2, into a suspension of finely ground calcium sulfate, $CaSO_4$, as shown in reaction (3). Large quantities are also produced as a by-product of coke ovens and coal-gas works. The chief use of ammonium sulfate is as a fertilizer.

$$CaSO_4 + 2NH_3 + CO_2 + H_2O \longrightarrow CaCO_3 + (NH_4)_2SO_4 \quad (3)$$

Ammonium carbonate, $(NH_4)_2CO_3$, may be prepared by bringing ammonia and carbon dioxide together in aqueous solution, shown in reaction (4). It

$$2NH_3 + CO_2 + H_2O \longrightarrow (NH_4)_2CO_3 \quad (4)$$

is also obtained by heating a mixture of ammonium sulfate and a fine suspension of calcium carbonate, shown in reaction (5).

$$(NH_4)_2SO_4 + CaCO_3 \longrightarrow (NH_4)_2CO_3 + CaSO_4 \quad (5)$$

Ammonium thiocyanate, NH_4SCN, a colorless solid with a density of 1.31, is prepared by the reaction of ammonia and carbon disulfide, CS_2, as in reaction (6). It is used as a protective agent in dyeing.

$$2NH_3 + CS_2 \longrightarrow NH_4SCN + H_2S \quad (6)$$

Except for several complex species such as ammonium chloroplatinate, $(NH_4)_2PtCl_6$, ammonium salts are very soluble in water. In aqueous solutions, they ionize to produce the ammonium ion, NH_4^+, and an acid anion. Solutions of ammonium salts of strong or moderately strong acids are acidic as a result of hydrolysis of the ammonium ion. The reaction involving a molecule of water and producing a hydrogen ion, H^+, is shown as reaction (7).

$$NH_4^+ + H_2O \rightleftharpoons NH_3 + H_3O^+ \quad (7)$$

When a strong base is added to a solution of an ammonium salt, ammonia is evolved. This is a test for the presence of NH_4^+ ion and occurs because the reaction of reaction (8) is driven toward the left.

$$NH_3 + H_2O \rightleftharpoons NH_4^+ + OH^- \quad (8)$$

See AMMONIA; FERTILIZER; HYDROLYSIS.

Francis J. Johnston

Amnesia

A significant but relatively selective inability to remember. Amnesia can be characterized along two dimensions with respect to its onset: an inability to remember events that occurred after the onset of amnesia is referred to as anterograde amnesia, and a deficit in remembering events that occurred prior to the onset of amnesia is referred to as retrograde amnesia. Amnesia can be due to a variety of causes, or etiologies, and it can be classified according to whether the cause is primarily neurological or psychological in origin.

Neurological amnesia. Neurological amnesias are the result of brain dysfunction and can be transient or permanent. They are usually characterized by a severe anterograde amnesia and a relatively less severe retrograde amnesia.

Transient amnesias are temporary memory disturbances and can range in duration from hours to months, depending on the cause and severity. They can be caused by epilepsy, head injury, and electroconvulsive therapy (most frequently used for the treatment of depression). In cases of transient global amnesia, an extensive amnesia that is usually sudden in onset and resolves within a day, the etiology is still not known, although many believe that it is vascular in origin. *See* SEIZURE DISORDERS.

Temporal lobe damage. Permanent amnesia can occur following brain damage to the medial temporal lobe. This discovery was made in 1953 following an experimental bilateral temporal lobe resection in an individual suffering from epilepsy. Although this procedure was moderately successful in treating the epilepsy, the patient developed severe anterograde amnesia. Although this individual had an average intelligence quotient (I.Q.) and normal short-term memory, he had great difficulty remembering recent events in his life. For example, he did not know where he lived, what he ate during his last meal, what year it was or how old he was; and he found it very difficult to learn new words and people's names. A temporally graded retrograde amnesia was also exhibited, with more remote memories being better preserved. The length of his retrograde amnesia was somewhere between 2 and 11 years. Bilateral temporal lobectomies are now avoided because of the resulting amnesia.

On occasion it is necessary to do a unilateral temporal lobectomy, which results in material specific memory deficits. Left temporal lobectomies result in verbal memory deficits and right temporal lobectomies result in nonverbal, visual memory deficits. Amnesia resulting from impairment to the medial temporal lobe can also occur following anoxia, cerebrovascular accidents, head injury, and viral infections to the brain. The primary structures involved in the processing of memory within the medial temporal lobe are the hippocampus and the amygdala.

Wernicke-Korsakoff syndrome. Neurological amnesia can also occur following damage to the diencephalon. One of the most common causes of diencephalic amnesia is Wernicke-Korsakoff syndrome, a disorder caused by a thiamine deficiency, usually related to chronic alcoholism. It is characterized by confusion, ataxia, and an oculomotor disturbance. Following thiamine replacement therapy, these symptoms reverse, but the individual is left with an amnesia. The major characteristics in diencephalic amnesia are very similar to medial temporal lobe amnesia. Damage to the mediodorsal thalamus and the mamillary bodies is largely responsible for the amnesia. *See* ALCOHOLISM.

Preserved memory function. Amnesics have provided important insights into how the diencephalon and medial temporal lobe are involved in preserved memory function. Amnesics have been shown to demonstrate normal learning on tasks involving motor, perceptual, and cognitive skills. They generally show a normal pattern of performance on tasks measuring how previously presented information indirectly influences subsequent responses. For example, amnesics will show normal rates of decreasing response times for identifying previously shown stimuli. This process is referred to as priming. Other examples include normal tendencies to complete word stems (such as MOT___) and word fragments (such as _O_EL) with previously presented words (Motel) when asked to complete the words with the first word that comes to mind. *See* COGNITION.

Multiple memory systems. Studies of preserved memory functions in amnesics supported the view that there are multiple memory systems represented in the brain. Damage to the medial temporal lobe area or the diencephalon appears to disrupt memory requiring conscious recollection of previously experienced information or events, frequently referred

to as explicit memory; however, amnesics' ability to show evidence of memory when conscious recollection is not required (referred to as implicit memory) is generally intact. Furthermore, since the retrograde amnesia is limited to events closer to the onset of the amnesia, with more remote memories better preserved, it appears that the diencephalon and medial temporal lobe play a role in the initial formation of explicit memory, as well as for some period of time, perhaps even years, following formation of the memory.

Memory improvements. Attempts to improve amnesics' ability to remember have generally met with little success. The most common approach to retraining has involved practicing remembering. Since damage to the medial temporal lobes and diencephalon is relatively permanent, and regeneration of damaged areas in the brain is limited, there is little hope that memory practice alone will improve the impaired functions, at least in adults. More encouraging approaches depend on using preserved implicit memory function in ways that will allow the patient to learn new information. *See* BRAIN; NEUROBIOLOGY.

Functional amnesia. Memory impairment that is not associated with brain damage is referred to as functional amnesia. Functional amnesia can be classified according to whether the amnesia is nonpathological or pathological. Nonpathological functional amnesia is a normal memory loss for events occurring during infancy and early childhood, sleep, hypnosis, and anesthesia. Pathological functional amnesia is an abnormal memory loss found in cases of functional retrograde amnesia and multiple personality. In contrast to neurological amnesia, pathological functional amnesia is usually associated with more severe retrograde than anterograde amnesia. *See* ANESTHESIA; HYPNOSIS; SLEEP AND DREAMING.

Functional retrograde amnesia. Functional retrograde amnesia usually occurs in response to a severe emotional trauma. Initially, individuals typically enter a fugue state in which they may wander around for an indefinite period of time. This stage is characterized by loss of personal identity, loss of memory of their past, and a lack of awareness of their amnesia. This stage usually ends when the individual is challenged by questions about his or her identity or past. Individuals who do not fully recover from this stage will enter a period characterized by awareness of the amnesia. The amnesia may spontaneously remit in response to an external trigger, such as the appearance of a familiar person, an occurrence that reminds them of their identity or past, or as the result of hypnosis or sodium amytal therapy. Recovered patients typically remain amnestic for the events during the fugue state, but usually remember the events during the amnestic period following the fugue state. Various explanations of functional retrograde amnesia have been proposed, including a dissociation of traumatic information from the ego, repression of some threatening event or information by the ego, and a selective failure of episodic memory, memory characterized by events in one's life.

Multiple personality disorder. A common characteristic in multiple personality disorders is amnesia among different personalities. The amnesia may be unidirectional, in which one personality will be amnestic for another personality, but the latter personality will not be amnestic for the former; or bidirectional, in which two personalities will be amnestic for each other. The amnesia can be quite problematic for an individual with a multiple personality disorder when the present personality cannot account for a predominant action of another personality (for example, when one personality spends the money earned by another personality). Amnesia in individuals with multiple personality disorders has been explained in terms of a dissociation from the ego of other personalities. Additional explanations include difficulty in retrieving stored information associated with the different states related to different personalities, and difficulty in retrieving analogous to posthypnotic states. Just as in the neurological amnesias, there have been demonstrations of intact implicit memory in functional amnesia with relatively severe impairment of explicit memories. *See* MEMORY; SCHIZOPHRENIA.

Richard S. Lewis

Bibliography. A. J. Parkin and N. R. C. Leng, *Neuropsychology of the Amnesic Syndrome,* 1992; L. Squire (ed.), *Memory and Its Disorders,* 1991; L. R. Squire and N. Butters (eds.), *Neuropsychology of Memory,* 2d ed., 1992.

Amnion

A thin, cellular, extraembryonic membrane forming a closed sac which surrounds the embryo in all reptiles, birds, and mammals and is present only in these forms; hence the collective term amniotes is applied to these animals. The amnion contains a serous fluid in which the embryo is immersed. *See* AMNIOTA.

Typically, the amnion wall is a tough, transparent, nerve-free, and nonvascular membrane consisting of two layers of cells; an inner, single-cell-thick layer of ectodermal epithelium, continuous with that covering the body of the embryo as the outer layer of its skin, and an outer covering of mesodermal, connective, and specialized smooth muscular tissue, also continuous with the mesodermal germ layer of the embryo. Early after the formation of the amnion, waves of contraction of the muscles pass over the amniotic sac and produce a characteristic rocking of the embryo. *See* GERM LAYERS.

In reptiles, birds, and some mammals, the amnion arises by a process of folding over the embryo body of the extraembryonic ectoderm along with its underlying, closely applied mesoderm (collectively, the somatopleure). Head, tail, and lateral folds of this sheet of tissue meet and fuse over the back of the embryo. Only the inner limb of the fold forms the true amnion, the outer limb of the fold becoming part of another fetal membrane, the chorion. In other mammals, including humans, the amnion arises by a process of

cavitation in a mass of cells in which embryonic and extraembryonic cells become separated. The cavity forms above those cells destined to form the embryo body and eventually spreads over and around the embryo up to the region of the developing umbilical cord.

The major function of the amnion and its fluid is to protect the delicate embryo. Thus, developmental stages of terrestrial animals are provided with the same type of cushioning against mechanical shock as is provided by the water environment of aquatic forms. *See* FETAL MEMBRANE.

Nelson T. Spratt, Jr.

Amniota

A collective term for the classes Reptilia (reptiles), Aves (birds), and Mammalia (mammals) of the subphylum Vertebrata. The remaining vertebrates, including the several classes of fishes and the amphibians, are grouped together as the Anamnia. Members of the Amniota are characterized by having a series of specialized protective extraembryonic membranes during development. Three of the membranes—amnion, chorion or serosa, and allantois—occur only in this group, but a fourth, the yolk sac, is sometimes present and is found in many anamniotes. The presence of the extraembryonic membranes makes it possible for the embryonic development of the amniotes to take place out of the water. In the most primitive forms the early stages of development take place inside a shell-covered egg that is deposited on land. This pattern is typical of most reptiles, all birds, and some mammals (**Figs. 1** and **2**). In these animals the amnion and chorion form fluid-filled sacs which protect the embryo from desiccation and shock. The allantois usually acts as a storage place for digestive and nitrogenous wastes and, in conjunction with the chorion, as a respiratory structure. In viviparous reptiles and mammals the chorion and allantois generally fuse and become more or less intimately associated with the uterine lining of the mother. Nutritive, excretory and respiratory exchanges take place across the chorioallantoic membrane between the allantoic circulation of the embryo and the uterine circulatory vessels of the mother. *See* ALLANTOIS; AMNION; ANAMNIA; CHORION; VERTEBRATA; YOLK SAC.

Jay M. Savage

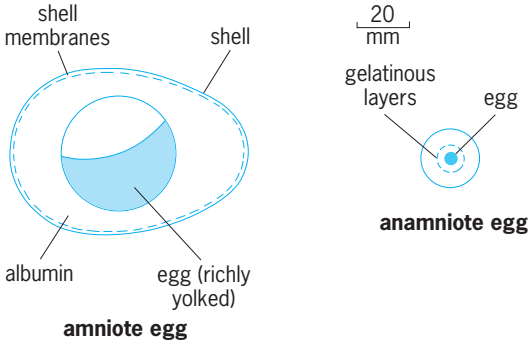


Fig. 1. Basic types of vertebrate eggs.

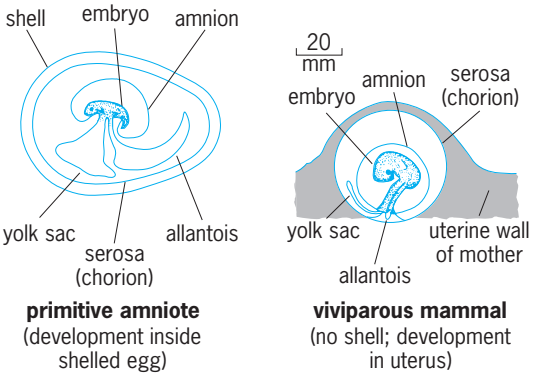


Fig. 2. Embryo developing in amniote eggs.

Amoebida

An order of Lobosia without protective coverings (tests). These protozoa range in size from about 4 micrometers to 0.08–0.12 in. (2–3 mm). Pellicles may be thin, as in *Amoeba proteus*, or thicker and less flexible, as in *Thecamoeba verrucosa* (**Fig. 1***a*). Pellicular folds may develop during locomotion, particularly in species with thick pellicles. Both flagellate and ameboid stages occur in certain soil amebas. In *Naegleria*, for example, the cycle includes an ameba, a flagellate, and a cyst. Most Amoebida have no flagellate stage.

Locomotion. The term ameboid movement is rather loose because locomotion of Amoebida varies somewhat from genus to genus. In some cases movement involves protoplasmic flow of the body as a whole, without typical pseudopodia. In *A. proteus* (Fig. 1*c* and *d*) there may be several ridged indeterminate pseudopodia, into one of which the organism appears to flow in locomotion. In other species (Fig. 1*b*) determinate pseudopodia never become large enough to direct locomotion. In some cases (Fig. 1*l*–*n*) the form of the pseudopodia may vary in a single species. Certain amebas commonly have a relatively inert posterior mass, the uroid, which may or may not be partially constricted, is sometimes covered with projections, and often contains food vacuoles. The uroid may be discarded by autotomy.

Locomotion involving protoplasmic flow depends upon sol-gel reversibility. According to the contractile-hydraulic theory, endoplasm of *A. proteus* is under constant pressure from the ectoplasm (plasmagel). The plasmagel contracts posteriorly to drive the endoplasm (plasmasol) forward to a point at which pressure from the plasmagel is relaxed momentarily. The result is a bulge, a developing pseudopodium (Fig. 1*c* and *m*). The plasmasol reaching the pseudopodial tip is diverted peripherally to make contact with, and become part of, the plasmasol at the posterior end of the body. The new plasmasol flows anteriorly and the process is repeated. In

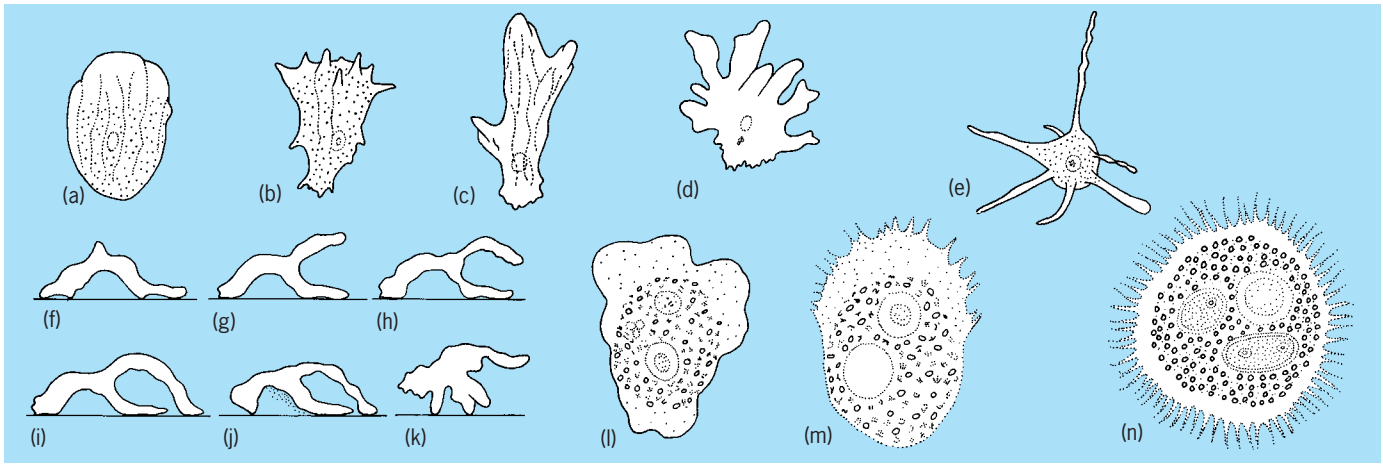


Fig. 1. Typical ameboid locomotion. (*a*) *Thecamoeba verucosa*, locomotion without formation of distinct pseudopodia. (*b*) *Mayorella bigemia*, formation of conical pseudopodia. (*c*) *Amoeba proteus*, formation of large pseudopodia. (*d*) *Amoeba dubia*, formation of a number of large pseudopodia. (*e*) *Astramoeba flagellipodia*, floating form with slender and sometimes spiral pseudopodia. (*f–k*) *Chaos* (*Pelomyxa*) *carolinensis*, locomotion of walking type, as seen in thriving cultures. (*l–n*) *Acanthamoeba castelanii*, specimens with different forms of pseudopodia, 12–30 μm. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

Thecamoeba there may be a rolling type of locomotion. A peculiar kind of walking has been described in one of the giant amebas (Fig. 1*f–k*). *See* CELL MOTILITY.

Cellular inclusions. Amebas, normally phagotrophic, usually contain food vacuoles. Certain species contain crystals of apparently differing chemical nature. Other inclusions are globules of different sizes, mitochondria, and stored food reserves. In addition, bacteria or algae may occur in the cytoplasm, changing the color to a gray or green. Nuclei range in number from one to several hundred, as in *Chaos carolinensis*. The giant amebas are visible without a microscope. *See* CELL (BIOLOGY).

Endoparasitic species. Those species found in the digestive tract of invertebrates and vertebrates include relatively harmless species and a few pathogens, such as *Entamoeba histolytica* of humans and *E. invadens* of reptiles.

Amebiasis. Entamoeba histolytica causes amebiasis. In primary cases the amebas are localized in the colon. Cases range from mild amebiasis to acute amebic dysentery. The amebas (**Fig. 2***f–k*) invade the wall of the colon, causing ulceration in symptomatic infections. Complications may include perforations of the appendix or colon, adhesions involving the colon, or other visceral damage. In secondary amebiasis the amebas may invade the ileum or liver (causing hepatic abscess) or, rarely, the lungs, brain, spleen, lymph glands, urinary bladder, uterus, penis, vagina, and skin. The amebas may migrate to the ileum; the usual route to most other organs is the blood, but contamination is probably involved in invasion of the urogenital tract. Severe infections may be correlated with vitamin deficiencies or generally substandard diets. Intestinal amebiasis has been treated with various drugs, such as arsenicals (carbarsone), quinoline derivatives (diodoquin, vioform), alkaloid derivatives such as emetine, and antibiotics such as terramycin.

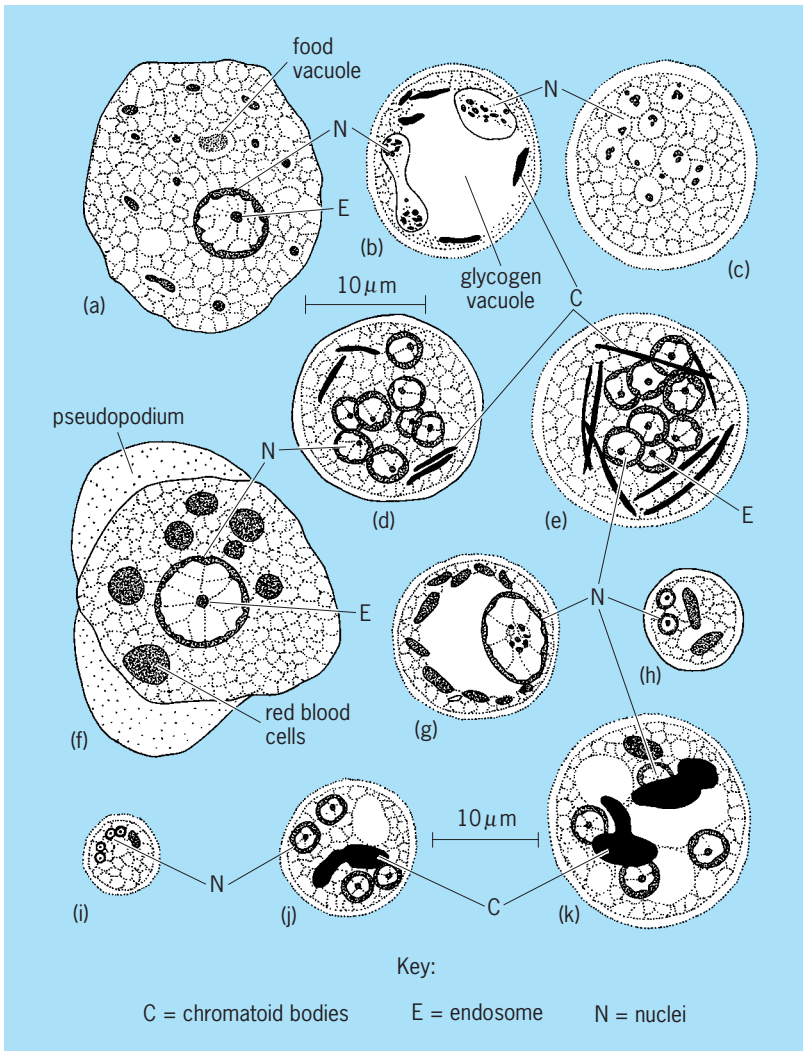


Fig. 2. Commensal and parasitic amebas. (*a–e*) *Entamoeba coli*: (*a*) rounded form; (*b*) cyst; (*c*)multinucleate cyst; (*d*) typical cyst from naturally infected monkey; (*e*) octunucleate cyst. (*f–k*) *E. histolytica*: (*f*) ameboid form; (*g*) cyst with glycogen vacuole; (*h*) binucleate cyst; (*i, j, k*) quadrinucleate cysts. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

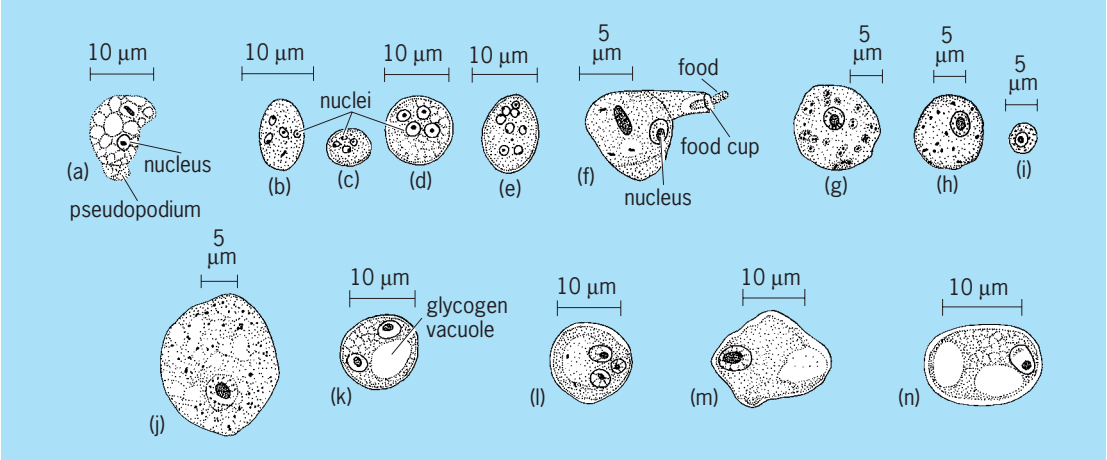


Fig. 3. Nonpathogenic amebas. (*a–e*) *Endolimax nana*: (*a*) ameboid stage; (*b*) cyst from monkey; (*c–e*) cyst from human. (*f–n*) *Iodamoeba bütschlii*: (*f*) ameboid stage; (*g,h*) ameboid stage, medium size; (*i*) ameboid stage, small size; (*j*) ameboid stage, large size; (*k*) binucleate cyst; (*l*) cyst showing three cysts; (*m*) uninucleate cyst; (*n*) uninucleate cyst from monkey. (*After R. P. Hall, Protozoology, Prentice-Hall, 1953*)

Nonpathogenic species. *Entamoeba coli* (Fig. 2*a–e*), a similar ameba, does not invade human tissues. Also limited to the lumen of the colon are *Endolimax nana, Iodamoeba bütschlii* (**Fig. 3**), and *Dientamoeba fragilis*. These four are relatively harmless although sometimes associated with digestive disturbances. In laboratory identification, number and structure of nuclei, shape of chromatoid bodies (Fig. 2) if present, and size and shape of cysts are important criteria. Intestinal amebas are transferred usually as cysts which are passed in the feces. Voided cysts remain viable for several days if kept moist and shaded. *Dientamoeba fragilis*, for which cysts are unknown, is an apparent exception to usual methods of transfer. Cysts are distributed widely by an infected individual and a new host becomes infected by swallowing cysts. Uncooked vegetables from soil fertilized with human feces are a potential source of infection. Standard methods of water purification seem reasonably protective, but it is difficult to control spread of cysts by food handlers. *See* LOBOSIA; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA.

Richard P. Hall

Amorphous solid

A rigid material whose structure lacks crystalline periodicity; that is, the pattern of its constituent atoms or molecules does not repeat periodically in three dimensions. In the present terminology amorphous and noncrystalline are synonymous. A solid is distinguished from its other amorphous counterparts (liquids and gases) by its viscosity: a material is considered solid (rigid) if its shear viscosity exceeds $10^{14.6}$ poise ($10^{13.6}$ pascal · second). *See* CRYSTAL; SOLID-STATE PHYSICS; VISCOSITY.

Preparation. Techniques commonly used to prepare amorphous solids include vapor deposition, electrodeposition, anodization, evaporation of a solvent (gel, glue), and chemical reaction (often oxidation) of a crystalline solid. None of these techniques involves the liquid state of the material. A distinctive class of amorphous solids consists of glasses, which are defined as amorphous solids obtained by cooling of the melt. Upon continued cooling below the crystalline melting point, a liquid either crystallizes with a discontinuous change in volume, viscosity, entropy, and internal energy, or (if the crystallization kinetics are slow enough and the quenching rate is fast enough) forms a glass with a continuous change in these properties. The glass transition temperature is defined as the temperature at which the fluid becomes solid (that is, the viscosity = $10^{14.6}$ poise = $10^{13.6}$ Pa · s) and is generally marked by a change in the thermal expansion coefficient and heat capacity. [Silicon dioxide (SiO_2) and germanium dioxide (GeO_2) are exceptions.] It is intuitively appealing to consider a glass to be both structurally and thermodynamically related to its liquid; such a connection is more tenuous for amorphous solids prepared by the other techniques. *See* GLASS.

Types of solids. Oxide glasses, generally the silicates, are the most familiar amorphous solids. However, as a state of matter, amorphous solids are much more widespread than just the oxide glasses. There are both organic (for example, polyethylene and some hard candies) and inorganic (for example, the silicates) amorphous solids. Examples of glass formers exist for each of the bonding types: covalent [As_2S_3], ionic [$KNO_3{-}Ca(NO_3)_2$], metallic [Pd_4Si], van der Waals [*o*-terphenyl], and hydrogen [$KHSO_4$]. Glasses can be prepared which span a broad range of physical properties. Dielectrics (for example, SiO_2) have very low electrical conductivity and are optically transparent, hard, and brittle. Semiconductors (for example, As_2SeTe_2) have intermediate electrical conductivities and are optically opaque and brittle. Metallic glasses (for example, Pd_4Si) have high electrical and thermal conductivities, have metallic luster, and are ductile and strong.

Uses. The obvious uses for amorphous solids are as window glass, container glass, and the glassy

polymers (plastics). Less widely recognized but nevertheless established technological uses include the dielectrics and protective coatings used in integrated circuits, and the active element in photocopying by xerography, which depends for its action upon photoconduction in an amorphous semiconductor. In optical communications a highly transparent dielectric glass in the form of a fiber is used as the transmission medium. In addition, metallic amorphous solids have been considered for uses that take advantage of their high strength, excellent corrosion resistance, extreme hardness and wear resistance, and unique magnetic properties. *See* OPTICAL COMMUNICATIONS.

Semiconductors. It is the changes in short-range order (on the scale of a localized electron), rather than the loss of long-range order alone, that have a profound effect on the properties of amorphous semiconductors. For example, the difference in resistivity between the crystalline and amorphous states for dielectrics and metals is always less than an order of magnitude and is generally less than a factor of 3. For semiconductors, however, resistivity changes at a given temperature of 10 orders of magnitude between the crystalline and amorphous states are not uncommon, and accompanying changes in optical properties can also be large.

Electronic structure. The model that has evolved for the electronic structure of an amorphous semiconductor is that the forbidden energy gap characteristic of the electronic states of a crystalline material is replaced in an amorphous semiconductor by a pseudogap. Within this pseudogap the density of states of the valence and conduction bands is sharply lower but tails off gradually and remains finite due to structural disorder (**Fig. 1**). The states in the tail region are localized; that is, their wave functions extend over small distances in contrast to the extended states that exist elsewhere in the energy spectrum. Because the localized states have low mobility (velocity per unit electric field), the extended states are separated by a mobility gap (Fig. 1) within which charge transport is markedly impeded. In each band, the energy at which the extended states meet the localized states is called the mobility edge. *See* BAND THEORY OF SOLIDS.

An ideal amorphous solid can be conceptually defined as having no unsatisfied bonds, a minimum of bond distortions (bond angles and lengths), and no internal surfaces associated with voids. Deviations from this ideality introduce localized states in the gap in addition to those in the band edge tails due to disorder alone. One important defect is called an unsatisfied, broken, or dangling bond. These dangling bonds create states deep in the gap which can act as recombination centers and markedly limit carrier lifetime and mobility. A large number of such states introduced, for example, during the deposition process will dominate the electrical properties.

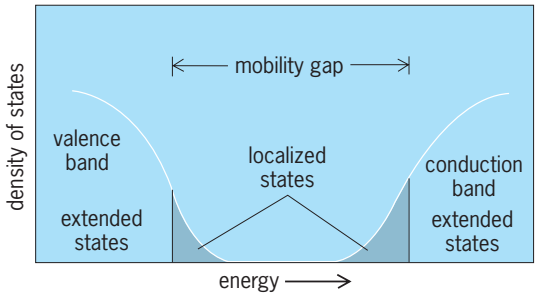


Fig. 1. Density of states versus energy for an amorphous semiconductor.

Charge transport can occur by two mechanisms. The first is conduction of mobile extended-state carriers (analogous to that which occurs in crystalline semiconductors), for which the conductivity is proportional to $\exp(-E_g/2kT)$, where E_g is the gap width, T is the absolute temperature, and k is Boltzmann's constant. The second mechanism is hopping of the localized carriers, for which the conductivity is proportional to $\exp[-(T_o/T)^{1/4}]$, where T_o is a constant (Mott's law). At low temperatures carriers hop from one localized trap to another, whereas at high temperatures they can be excited to the mobility edge.

Glassy chalcogenides. One class of amorphous semiconductors is the glassy chalcogenides, which contain one (or more) of the chalcogens sulfur, selenium, or tellurium as major constituents. These amorphous solids behave like intrinsic semiconductors, show no detectable unpaired spin states, and exhibit no doping effects. It is thought that essentially all atoms in these glasses assume a bonding configuration such that bonding requirements are satisfied; that is, the structure accommodates the coordination of any atom. These materials have application in switching and memory devices. *See* GLASS SWITCH.

Tetrahedrally bonded solids. Another group is the tetrahedrally bonded amorphous solids, such as amorphous silicon and germanium. These materials cannot be easily formed by quenching from the melt (that is, as glasses) but can be prepared by one of the deposition techniques mentioned above.

When amorphous silicon (or germanium) is prepared by evaporation, not all bonding requirements are satisfied, so a large number of dangling bonds are introduced into the material. These dangling bonds are easily detected by spin resonance or low-temperature magnetic susceptibility and create states deep in the gap which limit the transport properties. The number of dangling bonds can be reduced by a thermal anneal below the crystallization temperature, but the number cannot be reduced sufficiently to permit doping.

Amorphous silicon prepared by the decomposition of silane (SiH_4) in a plasma has been found to have a significantly lower density of defect states within the gap, and consequently the carrier lifetimes are expected to be longer. This material can be doped p- or n-type with boron or phosphorus (as examples) by the addition of B_2H_6 or PH_3 to the SiH_4 during deposition. This permits exploration of possible devices based on doping, which are analogous to devices based on doping of crystalline silicon.

One reason plasma-deposited silicon has a

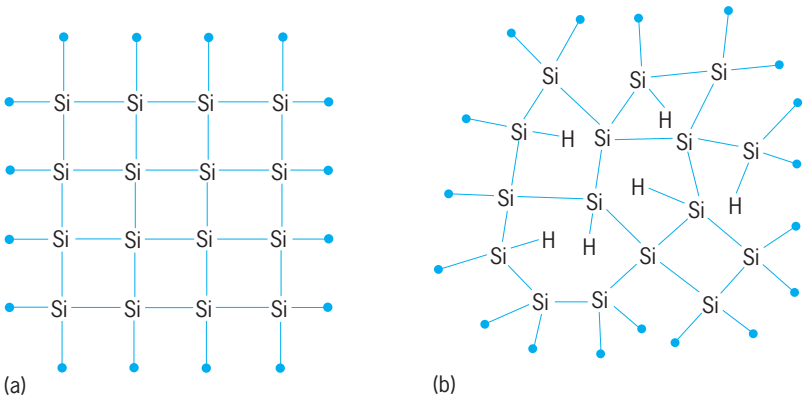


Fig. 2. Bonding of silicon. Bonds which continue the network are shown terminated by a dot. (*a*) Crystalline arrangement. (*b*) Amorphous structure with dangling (or unsatisfied) bonds terminated by hydrogen.

significantly lower density of defect states within the gap is that the process codeposits large amounts of hydrogen (typically 5–30% of the atoms, depending upon deposition conditions), and this hydrogen is very effective at terminating dangling bonds (**Fig. 2**). Other possible dangling-bond terminators (for example, fluorine) have been explored.

The ability to reduce the number of states deep in the gap and to dope amorphous silicon led directly to the development of an amorphous silicon photovoltaic solar cell. Solar conversion efficiencies of 10% have been achieved in the laboratory, leading to a development effort aimed at large-scale power applications. The appeal of amorphous silicon is that it holds promise for low-cost, easily fabricated, large-area cells.

Amorphous silicon solar cells have been constructed in heterojunction, *pin*-junction, and Schottky-barrier device configurations, and are widely used in consumer products such as calculators and watches. The optical properties of amorphous silicon provide a better match to the solar spectrum than do those for crystalline silicon, but the transport properties of the crystalline material are better. Experiments indicate that hole transport in the amorphous material is the limiting factor in the conversion efficiency. *See* SEMICONDUCTOR; SEMICONDUCTOR DIODE; SEMICONDUCTOR HETEROSTRUCTURES; SOLAR CELL; SOLAR ENERGY.

Brian G. Bagley

Bibliography. S. Elliott, *Physics of Amorphous Materials*, 2d ed., 1990; N. F. Mott, *Conduction in Non-Crystalline Materials*, 2d ed., 1993; K. Morigaki, *Physics of Amorphous Semiconductors*, 1994; R. Zallen, *The Physics of Amorphous Solids*, 1983.

Ampère's law

A law of electromagnetism which expresses the contribution of a current element of length dl to the magnetic induction (flux density) B at a point near the current. Ampère's law, sometimes called Laplace's law, was derived by A. M. Ampère after a series of experiments during 1820–1825. It plays a fundamental role in the International System (SI) definition of the ampere. *See* PHYSICAL MEASUREMENT.

Whenever an electric charge is in motion, there is a magnetic field associated with that motion. The flow of charges through a conductor sets up a magnetic field in the surrounding region. Any current may be considered to be broken up into infinitesimal elements of length dl, and each such element contributes to the magnetic induction at every point in the neighborhood. The contribution dB of the element is found to depend upon the current I, the length dl of the element, the distance r of the point P from the current element, and the angle θ between the current element and the line joining the element to the point P (see **illus.**). Ampère's law expresses the manner of the dependence by Eq. (1).

$$dB = k\frac{I\,dl\,\sin\theta}{r^2} \tag{1}$$

Choice of units. The proportionality factor k depends upon the units used in Eq. (1) and upon the properties of the medium surrounding the current. As in other equations expressing observed relationships, there is an arbitrary choice as to which of the units is to be defined from Ampère's law, or by a relationship derived from the law. If values are assigned to B, I, and l, the factor k can be found by experiment. If, however, an arbitrary value is assigned to k, Ampère's law or any equation derived from Ampère's law may be used to define the unit of current since units of B are otherwise defined. In SI, the latter choice is made, a value is assigned to k, and the ampere as a unit of current is defined from an equation derived from Ampère's law. When the current is in empty space, the factor k is assigned a value of 10^{-7} weber/A-m.

As in other equations associated with electric and magnetic fields, for example Coulomb's law, it is convenient to replace k by a new factor μ_0 related to k as in Eq. (2). This substitution removes the factor

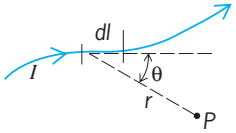
$$\mu_0 = 4\pi k \tag{2}$$

4π from many derived equations in which it would otherwise appear. With this substitution Ampère's law becomes Eq. (3). The factor μ_0 is called the

$$dB = \frac{\mu_0}{4\pi}\frac{I\,dl\,\sin\theta}{r^2} \tag{3}$$

permeability of empty space: it has the value $4\pi \times 10^{-7}$ henry/m. *See* ELECTRICAL UNITS AND STANDARDS.

Right-hand rule. The direction of dB is always perpendicular to the plane determined by the line



Graphic representation of Ampère's law.

tangent to the current element *dl* and the line joining *dl* to *P*. The sense of the lines is clockwise when looking in the direction of the current. The direction of each point may also be described in terms of a right-hand rule. If the current element is grasped by the right hand with the thumb pointing in the direction of the current, the fingers encircle the current in the direction of the magnetic induction. (If electron flow is used as the convention for current direction, the left hand is used in place of the right.)

Since the magnetic induction is everywhere perpendicular to the current element, it follows that the lines of induction or flux always form closed paths.

Field near a current. From Ampère's law, the field near a current may be calculated by finding the vector sum of the contributions of all the various elements that make up the current. This sum can be found (provided the integration can be carried out) by integrating over the whole length of the current, as in Eq. (4), where the limits are taken so that all

$$B = \int dB = \frac{\mu_0}{4\pi} \int \frac{I\, dl\, \sin\theta}{r^2} \qquad (4)$$

current elements are included and the integral represents the vector sum.

The experimental test of the validity of Ampère's law is not direct since experiments are made not upon the flux density due to individual elements, but upon the flux density resulting from the current as a whole. Thus, the applications of Ampère's law are in the computation of the flux density for known geometrical arrangements of the current. For simple current paths, the summation is easily carried out, as for the flux density at the center of a single circular conductor of radius *a*. For this conductor, $r = a$, and *r* is always perpendicular to *dl*, so that $\sin\theta = 1$. Furthermore, all contributions are perpendicular to the plane of the coil, so that the vector sum is the arithmetic sum as in Eq. (5). The flux density anywhere

$$B = \frac{\mu_0}{4\pi} \int_0^{2\pi a} \frac{I\, dl}{a^2} = \frac{\mu_0}{4\pi} \frac{I}{a^2} \int_0^{2\pi a} dl = \frac{\mu_0 I}{2a} \qquad (5)$$

on the axis of a flat coil, on the axis of a solenoid, or near a long straight conductor may be computed readily, and the result compared with experimental values. In general, the field of any current may be evaluated if *dB* can be expressed in Ampère's law, and the integration can be carried out. For the case of a long straight conductor *see* BIOT-SAVART LAW.

Ampère's law can be used as an alternative definition of magnetic induction (flux density). *See* MAGNETISM.

Kenneth V. Manning

Bibliography. B. I. Bleaney and B. Bleaney, *Electricity and Magnetism*, 3d ed., 1976, paper 1989; W. J. Duffin *Electricity and Magnetism*, 4th ed., 1990; D. Halliday, R. Resnick, and K. Krane, *Physics*, 5th ed., 2002; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1985; H. D. Young and R. A. Freedman, *Sears and Zemansky's University Physics*, 11th ed., 2003.

Amphetamine

A stimulating drug that affects the brain and the body in a variety of ways, also known by the trade name Benzedrine. Chemically, amphetamine is a racemic mixture of the *l* and *d* isomers of α-methyl-β-phenethylamine. The *l* isomer has more pronounced effects on the body, while the *d* isomer (commonly referred to as Dexedrine) has a greater effect on the brain. On the whole, the pharmacological effects of amphetamine are to produce an increase in blood pressure, a relaxation of bronchial smooth muscle, a constriction of the blood vessels supplying the skin and mucous membranes, and a variety of alterations in behavior.

Pharmacology. The mechanisms by which amphetamine produces its effects are not precisely defined. The effects on the body seem to be mediated predominantly through an increase in the activity of the neurons in the sympathetic nervous system via the transmitter norepinephrine. Amphetamine has the ability to release norepinephrine from nerve terminals. The consequence of release is that amphetamine has a spectrum of activity similar to the normal physiological effects of norepinephrine on the peripheral nervous system. For example, amphetamine and norepinephrine increase blood pressure by increasing the strength with which the muscles of the heart contract. Both drugs cause a constriction of blood vessels in the skin and mucous membranes, and they both cause the smooth muscles in the lung to relax.

Similarly, the major effects of amphetamine on the brain have been related to its ability to release norepinephrine in the hypothalamus, the reticular activating system, and the cerebral cortex. However, amphetamine may also release another neurotransmitter, dopamine, in the mesolimbic system and the cerebral cortex, and some of the effects of amphetamine on the brain are due to this other chemical.

The effects of amphetamine on the brain are generally stimulating. The drug causes an increase in motor activity, a decrease in appetite, an increase in arousal and ability to concentrate, an elevation of mood, insomnia, an enhanced sense of well-being, and hyperventilation. *See* CENTRAL NERVOUS SYSTEM; SYMPATHETIC NERVOUS SYSTEM.

Indications. The pharmacological effects noted above are the basis for the use of this drug in medicine. For example, the ability of amphetamine to contract blood vessels in the mucous membranes and to relax smooth muscles in the lung make it an efficacious nasal decongestant. Indeed, the original use of amphetamine was in Benzedrine inhalers for the purpose of clearing the nasal passages.

In a like manner, the pronounced excitatory effects of amphetamine on the brain have led to several other medical uses. For example, its ability to cause hyperventilation (by stimulation of the respiratory centers in the medulla) has resulted in its use as an analeptic. The arousing and insomnia-producing

effects (via stimulation of the reticular activating system and the cerebral cortex) have been exploited to treat narcolepsy, a disease characterized by an inability to stay awake. Furthermore, the ability of amphetamine to cause a loss of appetite (via the hypothalamus) has promoted its use in diet programs as an anorexic; however, tolerance to this effect limits its usefulness. Finally, the enhanced sense of well-being and mild euphoria that are seen after taking amphetamine have led to its use in certain forms of psychiatric depression.

One important use of amphetamine is in the treatment of hyperkinetic children. Hyperkinesis is a behavioral disorder in children that is characterized by increased motor activity, short attention span, lack of concentration, impulsiveness, emotional lability, and low tolerance for frustration. Consequently the child is difficult to control and disrupts discipline in the home and classroom. Amphetamine has been used successfully to treat this syndrome, although the paradox of how a stimulant can reduce hyperkinetic behavior is not understood.

Side effects. Just as the therapeutic uses of amphetamine are a consequence of its pharmacology, so are its side effects. Thus, the typical side effects on the body include dry mouth, heart rhythm alterations (palpitations and arrhythmias), hypertension, stomach cramps, and decreased urinary frequency. The central nervous system side effects include dizziness, dysphoria, headache, tremor, restlessness, insomnia, decreased appetite, increased aggressiveness, anxiety, and paranoid panic states. Extreme overdosage can result in convulsions, cerebral hemorrhaging, coma, and death.

Illegal uses. The stimulant and euphorogenic side effects of amphetamine have made this drug subject to widespread abuse. The patterns of this abuse, however, vary greatly. For example, in the early 1930s, when amphetamine was in Benzedrine inhalers, the drug was taken by dunking the inside paper liner of the inhaler in a glass of beer and then drinking the mixture. Today, most self-administration is with pills and capsules that are illegally diverted from legitimate sources. There also developed growing abuse of methamphetamine (a chemical derivative of amphetamine). This drug is taken in crystal form, either intranasally ("snorted") or intravenously ("mainlined") after being dissolved in water. Another problem concerns multiple drug use in a single individual (polydrug abuse). For example, amphetamine is often taken in "binges" which are terminated by taking large doses of depressants (such as secobarbital). A cycle then develops of alternating "uppers" and "downers." Amphetamine is also used as an adulterant of illegal cocaine; the mixture purportedly has a more pleasant and longer-lasting effect. Finally, amphetamine has been mixed with heroin, with the resulting concoction (a "speedball") supposedly having a more potent and euphorogenic effect. However, an implicit danger in taking illegal amphetamine, as with taking any illegal drug, is that the user has no idea as to purity or the source of the drug. Thus, the safety of the user is jeopardized.

The diversity of people taking amphetamine without proper medical supervision makes it difficult to characterize a typical user. Amphetamine abuse by students has been observed. There is also evidence that it is used by truck drivers to fend off fatigue and boredom so that they can drive longer trips. The drug is used by housewives to give them a "lift" and help them through their day. It is thought to be used by athletes to increase endurance and improve performance. It is also taken by diet-conscious people who wish to use amphetamine's anorexic effects to help lose weight. Amphetamine abuse by physicians and nurses has also been reported; they use the drug to stay alert during their long working hours.

More serious abuse occurs in some individuals who take the drug over long periods. As with other drugs of abuse (such as morphine, alcohol, or barbiturates), tolerance will occur after repeated dosing. Thus, whereas in the past one pill might be sufficient to induce the appropriate high, it becomes necessary to increase that amount in the future to achieve the same feeling. Over time, the users may become "psychologically dependent" on amphetamine, and feel that they cannot live without it. Subsequent drug usage and the acquisition of more drugs may then dominate and deleteriously affect their lives. *See* ALCOHOLISM; BARBITURATES; MORPHINE ALKALOIDS.

Chronic use of large amounts of amphetamine can have severe effects on personality. It has been shown that large doses of amphetamine can induce a behavioral state in humans that is nearly indistinguishable from paranoid schizophrenia, but can be reversed upon cessation of the drug. This drug-induced panic state is dangerous not only to the user (whose irrational acts may be self-destructive) but also to those in contact with the user. This is because, in these fearful, illogical, and agitated states, the user may become deluded that other persons have conspiratorial or hurtful intentions; this may lead to impulsive, "self-defensive" acts by the user. The potential for injury then becomes very high, and acute psychiatric intervention is critical.

The question of whether or not physical dependence occurs with amphetamine is unresolved. It is known that after long-term treatment with amphetamine, removal of the drug results in feelings of anxiety, stomach cramps, headache, and extreme fatigue. However, the mild severity of these symptoms makes it unclear if this is a true withdrawal response. Regardless, the other effects of long-term abuse of amphetamine (weight loss, insomnia, behavioral disorders) make unsupervised self-administration highly risky and unadvisable. *See* ADDICTIVE DISORDERS; NARCOTIC.

Richard E. Chipkin

Bibliography. L. S. Goodman and A. Gillman (eds.), *The Pharmacological Basis of Therapeutics*, 9th ed., 1996; L. E. Hollister, *Clinical Use of Psychotherapeutic Drugs*, 1973; R. M. Julien, *A Primer of Drug Action*, 7th ed., 1995; O. S. Ray, *Drugs, Society and Human Behavior*, 8th ed., 1998.

Amphibia

One of four classes (the others are Reptilia, Aves, and Mammalia) composing the superclass Tetrapoda of the subphylum Vertebrata. The living amphibians number over 4700 species and are classified in three orders: Salientia or Anura (frogs and toads, over 4200 species); Urodela or Caudata (salamanders, approximately 420 species); and Gymnophiona or Apoda (caecilians, about 165 species). The orders in the subclasses Labyrinthodontia and Lepospondyli are now extinct. A classification scheme for the Amphibia follows. See separate articles on each group listed.

Class Amphibia
- Subclass Labyrinthodontia
  - Order: Ichthyostegalia
  - Temnospondyli
  - Anthracosauria
- Subclass Lepospondyli
  - Order: Nectridea
  - Aistopoda
  - Microsauria
  - Lysorophia
- Subclass Lissamphibia
  - Order: Anura
  - Caudata
  - Apoda

General characteristics. A typical amphibian is characterized by a moist glandular skin, the possession of gills at some point in its life history, four limbs, and an egg lacking the embryonic membrane called the amnion. The closest evolutionary relatives of the amphibians are the fishes, from which they evolved, and the reptiles, to which they gave rise (**Fig. 1**). Present-day amphibians, however, are highly specialized animals, rather different from the primitive forms that probably arose from crossopterygian fishes and far removed from those that gave rise to the earliest reptiles. *See* AMNION; ANAMNIA.

Fish and amphibians. In general, modern amphibians as adults differ from fishes in lacking scales, breathing by means of the skin and lungs instead of gills, and having limbs in place of fins. There are many exceptions to these generalizations, however. Some fishes lack scales, and some members of one group of amphibians, the Apoda, have scales buried in the skin. Some salamanders retain the gills throughout life, and there are air-breathing lungfish. There are no amphibians with paired fins, but the caecilians lack limbs entirely; presumably these animals have evolved from ancestors that possessed typical tetrapod limbs. The significant differences that distinguish typical adult amphibians from fishes are those closely associated with adaptation to life on land. Respiration is by means of skin and lungs rather than

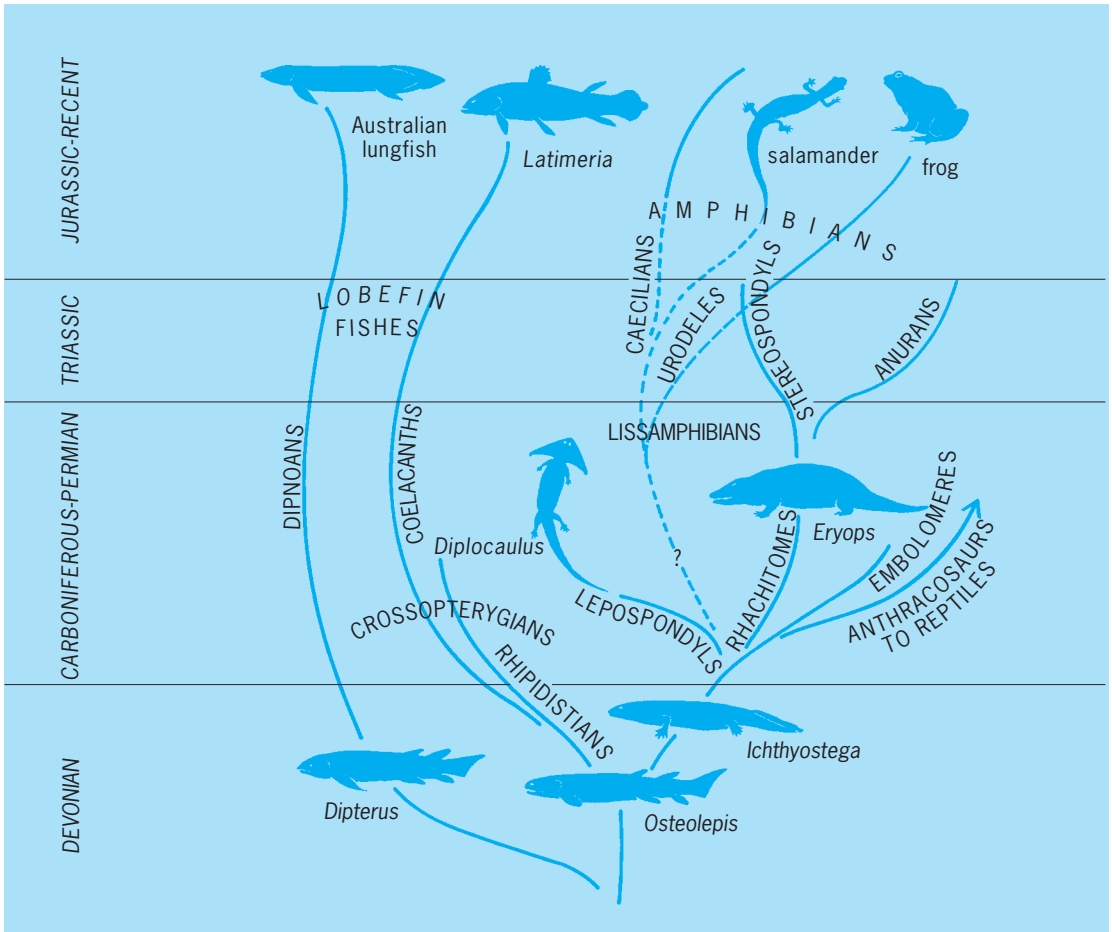


Fig. 1. Vertebrate phylogeny, including tetrapod ancestors, stem reptile groups, and amphibian groups. (*After E. H. Colbert, Evolution of the Vertebrates, Wiley, 1955*)

gills, and tetrapod limbs are for terrestrial locomotion. *See* OSTEICHTHYES; RESPIRATION.

Reptiles and amphibians. Reptiles usually have a dry scaly skin that greatly reduces water loss, and so are very different from the amphibians with their moist skin that permits much evaporation. Young (larval) amphibians have gills, but there is no comparable gill-breathing, larval stage in the life history of a reptile. An important difference between the two groups is the absence of the amniotic egg in the Amphibia and its presence in the Reptilia. The amniotic egg is far more complex than its anamniotic predecessor (represented today by the eggs of fish and modern amphibians). The amniotic egg has three layers: the chorion, the allantois, and the amnion. These layers are involved in protection, nutrient supply, and waste removal, all important functions when moving away from a dependence on the aquatic environment, a predominant condition in all fishes and amphibians today. Lacking these membranes, amphibian eggs must be laid in water or in very moist places. The shell of the reptile egg also makes it more able to resist desiccation, and the eggs can be laid in relatively dry places. The ability to resist water loss through the skin and the development of a land egg are major evolutionary steps for reptiles. *See* REPTILIA.

Ecology. The all-important factor in amphibian life is water. Most species must return to the water to breed, and all must have access to water (even if only in the form of rain or dew) or die of desiccation in a short time. An important consequence of this basic fact of physiology is that vast arid and semiarid areas of the Earth are inhabited by a relatively few specialized amphibians. The majority of amphibian species are found in moist, tropical regions.

Amphibians typically are absent from highly saline waters. A notable exception is the crab-eating frog (*Rana cancrivora*) of southeastern Asia, which can tolerate full-strength seawater. A high level of urea in this frog's blood (as in the blood of sharks and their allies) produces an osmotic pressure high enough to help withstand the otherwise dehydrating effect of salt water.

Physiology. Amphibians are poikilothermic animals; that is, the temperature of the body of an amphibian is largely regulated by the temperature of the surrounding environment. An amphibian in the water will have a temperature close to, if not the same as, that of the water; an amphibian out of the water will often be somewhat cooler than the air as a result of evaporative cooling.

An animal such as an amphibian that burns little of its food energy in keeping warm is able to survive on less food than a bird or mammal of similar size. This advantage is offset by the inability of amphibians to be active under cold conditions. Thus the far northern and southern parts of the world which support large populations of birds and mammals have few amphibian species.

Evolution. The amphibians mark a significant point in the evolution of the vertebrates, the transition from aquatic to terrestrial life. Although restricted to aquatic or at least moist habitats, modern amphibians are an important component of many ecosystems. In the South American rainforest, for example, frogs are a dominant component of the biome. There are as many as 80 or more species of anurans in a small geographic area alone. Also, in certain eastern deciduous forests of the United States, terrestrial salamanders have been calculated to be one of the highest-biomass components of the forest. *See* TETRAPODA; THERMOREGULATION.

W. Ben Cash; Richard G. Zweifel

History and classification. Traditionally, the living vertebrates have been divided into five main groups: fishes, amphibians, reptiles, birds, and mammals. Recent studies have shown, however, that only three of these groups (amphibians, birds, and mammals) are monophyletic, with each including all the descendants of a single common ancestor. Fishes are paraphyletic, as they include the ancestors of all other vertebrate groups. Reptiles are also paraphyletic, as they include the ancestors of birds and mammals. Fortunately, reptiles, birds, and mammals constitute a monophyletic group, the Amniota.

The principal feature which distinguishes living amphibians from amniotes is their mode of reproduction. All amniotes have extraembryonic membranes, including the amnion, which surround the developing embryo within a fluid-filled sac, making it independent of an external aquatic environment. Amphibians, however, still depend on water as the environment into which eggs are normally laid.

Living amphibians, sometimes called the Lissamphibia, are divided into three main groups which are anatomically distinct (Fig. 1). The Caudata (salamanders, including newts) are the least derived with fore- and hindlimbs of equal length and a distinct tail. They are largely aquatic and found mainly in the Northern Hemisphere. The Gymnophionia, apodans and caecilians, are rare, limbless burrowers restricted mainly to the tropics. The Anura, frogs and toads, have large hindlimbs which are used in swimming and hopping; they have well-developed hearing; and many have the ability to produce a characteristic croak. They are truly amphibious and have a worldwide distribution.

Despite these obvious anatomical differences, the three groups of living amphibians are united by the common possession of some unique features: (1) a specialized structure in the inner ear used in hearing, the papilla amphibiorum; (2) green rods in the retina; (3) specialized mucous glands in the skin; (4) specialized glands in the mouth; and (5) bicuspid pedicellate teeth. These features have not been observed in any other living vertebrate, and they are clear evidence that all three groups must have descended from a single common ancestor.

Fossils. The fossil record of the three groups of living amphibians is extensive, and the earliest member of each has been found in Mesozoic rocks. The earliest salamanders are from the Middle Jurassic of England and Kirghizia (Kyrgyzstan), the earliest gymnophionan is from the Lower Jurassic of Arizona;

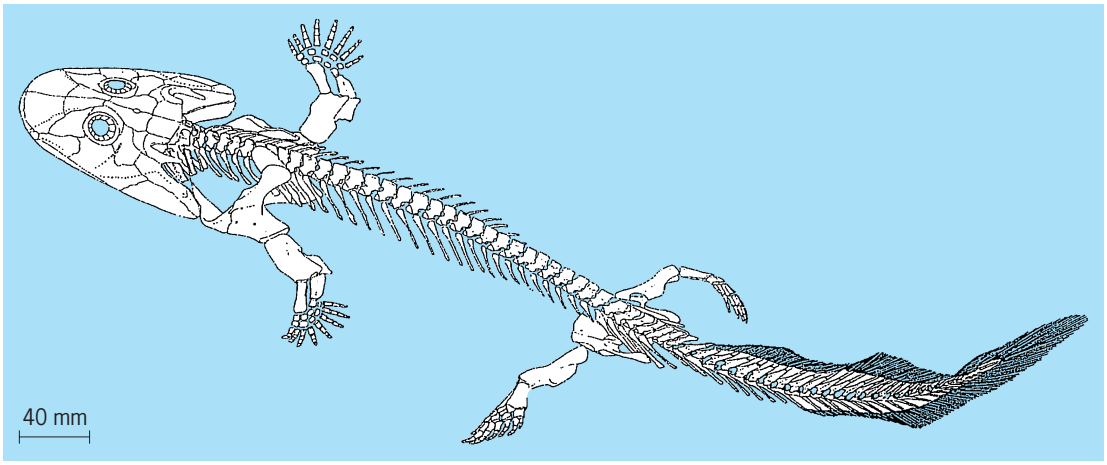


Fig. 2. Reconstruction of the skeleton of the Devonian amphibian *Acanthostega*. (*From M. I. Coates, The Devonian tetrapod Acanthostega gunnari Jarvik: Postcranial anatomy, basal tetrapod interrelationships and patterns of skeletal evolution, Trans. Roy. Soc. Edinburgh Earth Sci., 87:363–421, 1996*)

and the earliest anuran is from the Lower Jurassic of South America. However, no intermediary forms which link the three groups together have been found in the Mesozoic, and it is necessary to look in the Paleozoic, some 100 million years earlier, for the common ancestor of modern amphibians.

The earliest known amphibians have been found in the Upper Devonian rocks of Greenland. Animals such as *Acanthostega* (**Fig. 2**) retain a large number of fishlike features, including internal gills and a caudal fin supported by fin rays. However, they lack an anal fin, and the paired appendages bear digits and form distinct flexible limbs rather than fins. The earliest amphibians had as many as eight digits on each foot, which may have been used as flexible paddles to assist aquatic locomotion. The first amphibians with the characteristic pentadactyl limb, with five digits on each foot, are recorded in the early Carboniferous.

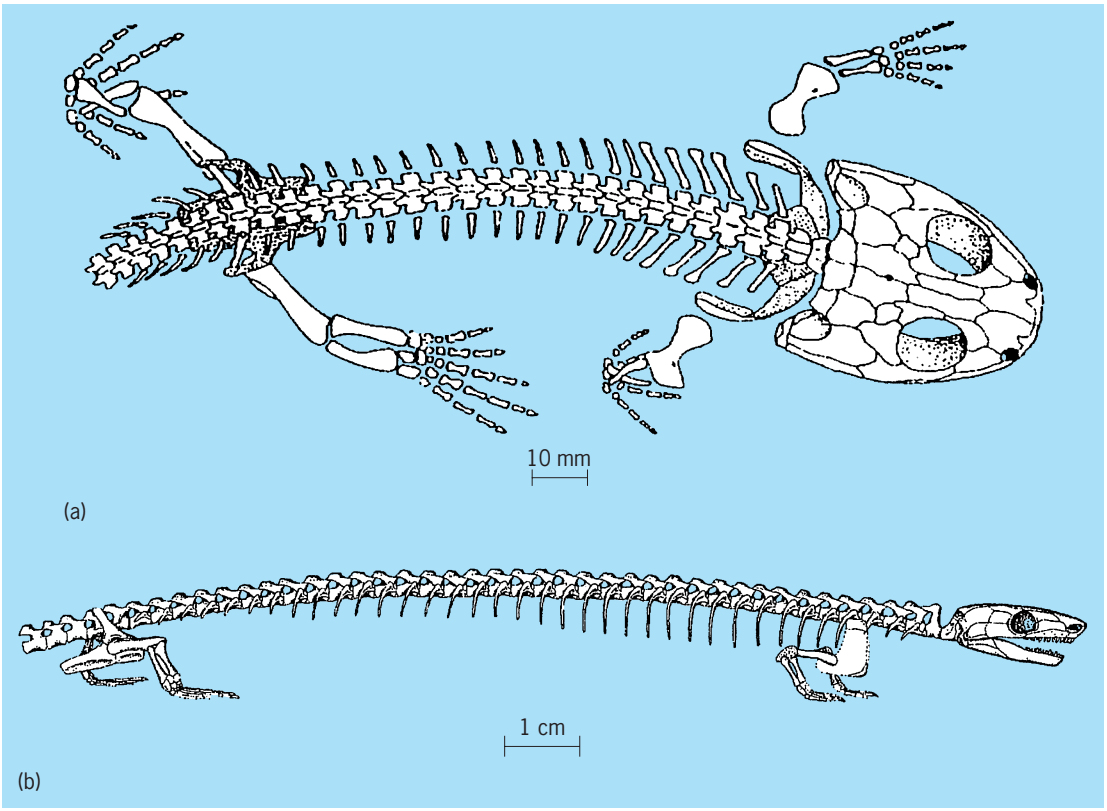


Fig. 3. Reconstructions of amphibian skeletons. (*a*) Carboniferous temnospondyl *Balanerpeton* (*from A. R. Milner and S. E. K. Sequeira, The temnospondyl amphibians from the Visean of East Kirkton, West Lothian, Scotland, Trans. Roy. Soc. Edinburgh Earth Sci., 84:331-362, 1994*). (*b*) Permian microsaur *Rhynchonkos* (*as Goniorhynchus, from R. L. Carroll and P. Gaskill, The Order Microsauria, Mem. Amer. Phil. Soc., 126:1–211, 1978*).

The development of the pentadactyl limb coincided with a rapid radiation of amphibians, and by the end of the Mississippian (325 million years ago) all the main groups had evolved, including those from which both the modern amphibians and the amniotes were to evolve.

Much of the evidence of this rapid radiation is based on single, isolated fossil specimens or is from fossil localities where species diversity is very low. In contrast, the East Kirkton Quarry, a site found recently in the Mississippian of Scotland, has an extensive fauna of terrestrial amphibians, and includes some of the earliest known members of groups of amphibians which were to dominate the later Carboniferous and Permian periods. Among these is the temnospondyl *Balanerpeton* (**Fig. 3**). This is the earliest known member of the group from which all modern amphibians may have evolved. Temnospondyls have a salamanderlike postcranium, with a four-digit manus (hand) and five-digit pes (foot), but a froglike skull with large vacuities in the palate and an otic notch which almost certainly supported a tympanum. Some temnospondyl species also have bicuspid pedicellate teeth. However, the gymnophionans are so different from temnospondyls that a close relationship has been disputed. Instead, a separate origin has been proposed among the microsaurs (Fig. 3). The earliest record of this group is from rocks similar in age to these at East Kirkton, in the Upper Mississippian of Illinois.

This dispute clearly influences the date of the most recent common ancestor of the three groups of modern amphibians. If all three descended from temnospondyls, their most recent common ancestor is possibly represented by the dissorophid temnospondyls of the late Carboniferous and early Permian. However, if the gymnophionans have a separate origin among the microsaurs, then their most recent common ancestor was a much earlier amphibian from the Mississippian, when the microsaurs and temnospondyls had not yet diverged into distinct groups.

Whichever hypothesis is accepted, it is clear that modern amphibians have a very long history extending back almost to the time of the origin and radiation of land vertebrates 340 million years ago. Their unique sensory biology and specialized glands must have evolved at that time and have remained unchanged to the present day. *See* REPTILIA.

T. R. Smithson

Bibliography. W. F. Blair et al., *Vertebrates of the United States*, 2d ed., 1968; J. F. Breen, *Encyclopedia of Reptiles and Amphibians*, 1974; R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987; W. E. Duellman and L. Trueb, *Biology of Amphibians*, 1986; M. J. Lannoo (ed.), *Amphibian Declines: The Conservation Status of United States Species*, University of California Press, Berkeley, 2005; D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, New York, 2001; A. R. Milner, The Paleozoic relatives of lissamphibians, *Herp. Monogr.*, 7:8–27, 1993; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; F. H. Pough et al., *Herpetology*, 3d ed., Prentice Hall, Upper Saddle River, NJ, 2004; W. D. I. Rolfe, E. N. K. Clarkson, and A. L. Panchen, Volcanism and early terrestrial biotas, *Trans. Roy. Soc. Edinburgh Earth Sci.*, 84:175–464, 1994; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966; R. D. Semlitsch (ed.), *Amphibian Conservation*, Smithsonian Institution Press, Washington, DC, 2003; R. C. Stebbins and N. W. Cohen, *A Natural History of Amphibians*, Princeton University Press, 1995; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001 .

Amphibole

A group of common ferromagnesian silicate minerals that occur as major or minor constituents in a wide variety of rocks. The crystal structure of the amphiboles is very flexible and, as a result, the amphiboles show a larger range of chemical composition than any other group of minerals. The structural and chemical complexity of the amphiboles reveals considerable information on the geological processes that have affected the rocks in which they occur. *See* MINERAL; SILICATE MINERALS.

Chemistry. A general formula for amphiboles may be written as $A_{0-1}B_2C_5T_8O_{22}W_2$, where

A = Na, K, Ca
B = Ca, Na, Mn^{2+}, Fe^{2+}, Li, Mg
C = Mg, Fe^{2+}, Al, Fe^{3+}, Ti^{4+}, Mn, Li
T = Si, Al, Ti^{4+}
O = oxygen
W = OH, F, O^{2-}, Cl

and the chemical species are written in order of their importance. Amphiboles are divided into four main groups, according to the type of chemical species in the B group:

Iron-magnesium-manganese amphiboles	$B = (Fe^{2+}, Mg, Mn^{2+}, Li)_2$
Calcic amphiboles	$B = Ca_2$
Sodic-calcic amphiboles	$B = NaCa$
Sodic amphiboles	$B = Na_2$

Iron-magnesium-manganese amphiboles. These amphiboles may be orthorhombic or monoclinic, and the names and formulas of the common amphiboles are as follows:

Orthorhombic amphiboles

Anthophyllite	$(Mg,Fe^{2+})_7Si_8O_{22}(OH)_2$
Gedrite	$Na_{0-1}(Mg, Fe^{2+},Al)_7(Si,Al)_8O_{22}(OH)_2$
Holmquistite	$Li_2(Mg,Fe^{2+})_3(Fe^{3+},Al)_2Si_8O_{22}(OH)_2$

Monoclinic amphiboles

Cummingtonite	$(Mg,Fe^{2+})_7Si_8O_{22}(OH)_2$
Grunerite	$(Fe^{2+},Mg)_7Si_8O_{22}(OH)_2$
Tirodite	$Mn_2^{2+}(Mg,Fe^{2+})_5Si_8O_{22}(OH)_2$
Dannemorite	$Mn_2^{2+}(Fe^{2+},Mg)_5Si_8O_{22}(OH)_2$

Calcic amphiboles. These amphiboles are monoclinic; the more important species are as follows:

Tremolite	$Ca_2Mg_5Si_8O_{22}(OH)_2$
Actinolite	$Ca_2(Mg,Fe^{2+})_5Si_8O_{22}(OH)_2$
Edenite	$NaCa_2Mg_5(Si_7Al)O_{22}(OH)_2$
Pargasite	$NaCa_2(Mg_4Al)(Si_6Al_2)O_{22}(OH)_2$
Hastingsite	$NaCa_2(Fe_4^{2+}Fe^{3+})(Si_6Al_2)O_{22}(OH)_2$
Tschermakite	$Ca_2(Mg_3Al_2)(Si_6Al_2)O_{22}(OH)_2$
Hornblende	$Na_{0-1}Ca_2(Mg,Fe^{2+},Fe^{3+},Al)_5(Si,Al)_8O_{22}(OH,F)_2$
Kaersutite	$NaCa_2(Mg_4Ti)(Si_6Al_2)O_{22}(O,OH)_2$
Cannilloite	$CaCa_2(Mg_4Al)(Si_5Al_3)O_{22}(OH)_2$

Sodic-calcic amphiboles. The sodic-calcic amphiboles are monoclinic; the more important species are as follows:

Richterite	$NaCaNaMg_5Si_8O_{22}(OH)_2$
Winchite	$CaNa(Mg_4Al)Si_8O_{22}(OH)_2$
Barroisite	$CaNa(Mg_3Al_2)(Si_7Al)O_{22}(OH)_2$
Katophorite	$NaCaNa(Mg_4Al)(Si_7Al)O_{22}(OH)_2$
Taramite	$NaCaNa(Fe_3^{2+}Al_2)(Si_6Al_2)O_{22}(OH)_2$

Sodic amphiboles. The sodic amphiboles are monoclinic; the most important species are as follows:

Glaucophane	$Na_2(Mg_3Al_2)Si_8O_{22}(OH)_2$
Riebeckite	$Na_2(Fe_2^{3+}Fe_2^{3+})Si_8O_{22}(OH)_2$
Arfvedsonite	$NaNa_2(Fe_4^{2+}Fe^{3+})Si_8O_{22}(OH)_2$
Ungarettiite	$NaNa_2(Mn_2^{2+}Mn_3^{3+})Si_8O_{22}O_2$
Obertiite	$NaNa_2(Mg_3Fe^{3+}Ti^{4+})Si_8O_{22}O_2$

Other amphiboles. Solid solution is extremely common in amphiboles and gives rise to a large number of additional distinct mineral species. These species are named by attaching prefixes to the names listed above, with the prefix indicating the substituent element differentiating the two minerals. For example, ferropargasite is related to pargasite by substitution of Fe^{2+} for Mg, and ferriwinchite is related to winchite by substitution of Fe^{3+} for Al:

Pargasite	$NaCa_2(Mg_4Al)(Si_6Al_2)O_{22}(OH)_2$
Ferropargasite	$NaCa_2(Fe_4^{2+}Al)(Si_6Al_2)O_{22}(OH)_2$
Winchite	$CaNa(Mg_4Al)Si_8O_{22}(OH)_2$
Ferriwinchite	$CaNa(Mg_4Fe^{3+})Si_8O_{22}(OH)_2$

In this way, a large number of distinct amphibole minerals can be named in a rational way, without generating a large number of unrelated names.

Structure. Amphiboles can have orthorhombic and monoclinic symmetries; these can be distinguished either by x-ray crystallography or by the optical properties of the mineral in polarized light. *See* X-RAY CRYSTALLOGRAPHY.

Monoclinic amphiboles are by far the most common. The characteristic feature of the amphibole structure is the chain of corner-sharing tetrahedrally coordinated groups, labeled T(1) and T(2) [**Fig. 1**].

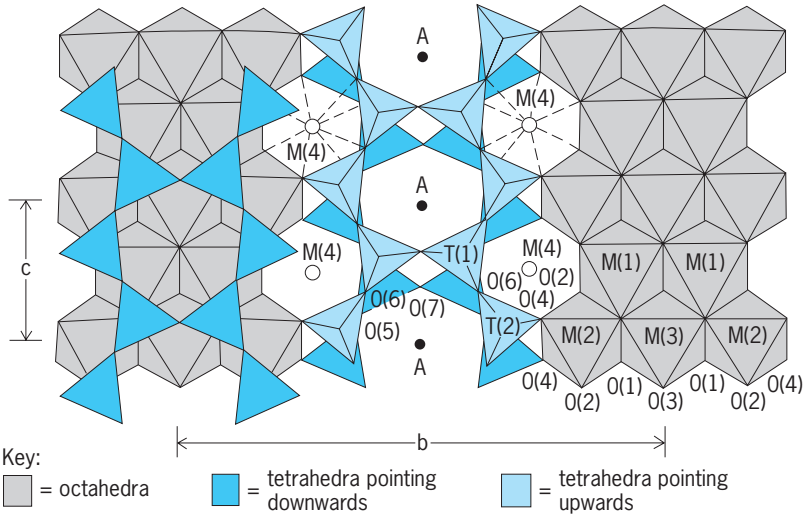


Fig. 1. Crystal structure of monoclinic (*C2/m*) amphibole viewed perpendicular to the tetrahedral double chains; the corner-sharing tetrahedral chains [T(1) and T(2)] and the edge-sharing octahedral strips [M(1) to M(3)] are shaded. O(1) to O(7) = distinct oxygen atoms. (*After F. C. Hawthorne, The crystal chemistry of the amphiboles, Can. Mineral., 21:173–480, 1983*)

These tetrahedra are occupied by the T cations (commonly Si and Al) of the formula, and their repeat distance along the z axis defines the unit-cell length, c, in this direction. The C cations of the formula (commonly Mg, Fe^{2+}, Al, and Fe^{3+}) occupy the M(1), M(2), and M(3) octahedra. These octahedra share edges to form strips that extend along the z axis, and link to the tetrahedral double chains in the x and y directions to form a continuous three-dimensional framework. The B cations occupy the M(4) site (Fig. 1) and are surrounded by eight anions; they provide additional linkage between the tetrahedral double chains and the octahedral strips, and are important in controlling the miscibility and stability behavior of the amphiboles. The A cations occupy the A site, which occurs in a large cavity between adjacent tetrahedral double chains (Fig. 1) and is surrounded by 12 oxygen atoms.

At the center of the octahedral strip is the O(3) site (Fig. 1) that is occupied by the W anions of the chemical formula. This site is bonded to the cations at the adjacent two M(1) sites and one M(3) site, and is usually occupied by monovalent anions (OH^-, F^-, and Cl^-). In the calcic amphibole kaersutite, it has long been recognized that O^{2-} (oxygen) can occur at the O(3) site, associated with the presence of high Fe^{3+} and Ti^{4+} that are characteristic of a highly oxidizing environment. In sodic amphiboles, it has been shown that major amounts of Ti^{4+} can couple to major amounts of O^{2-} at O(3), to the extent that the new anhydrous amphibole species obertiite [with O^{2-} at O(3)] has been recognized.

Inspection of the structure down the y axis (**Fig. 2**) shows that the amphibole structure consists of sheets of tetrahedra interleaved with octahedrally coordinated C-group cations. It is here that the difference between the orthorhombic and monoclinic structures becomes apparent. In the monoclinic structure (Fig. 2*a*), the tetrahedral layers all stack in

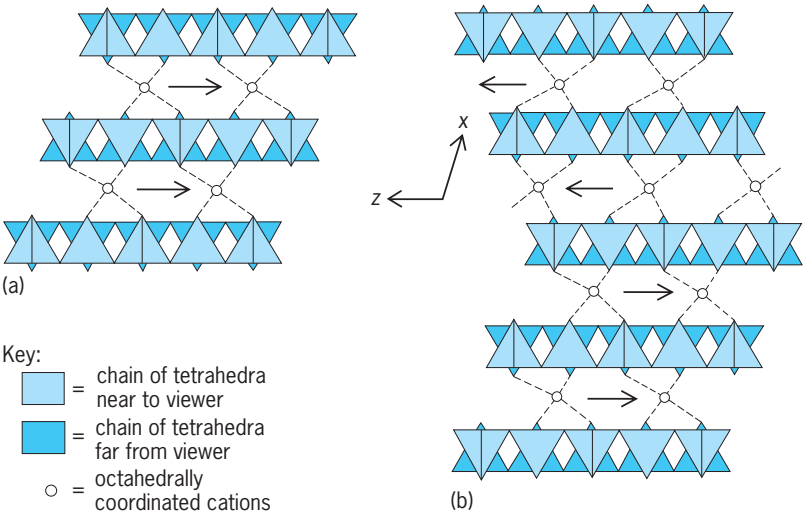


Fig. 2. Crystal structures of (*a*) monoclinic (*C2/m*) and (*b*) orthorhombic (*Pnma*) amphiboles, shown schematically as a stacking of layers. (*After F. C. Hawthorne, The crystal chemistry of the amphiboles, Can. Mineral., 21:173–480, 1983*)

the x direction with the same sense of displacement (along the z direction) relative to the underlying layer, and hence the x axis is inclined to the z axis, producing a monoclinic structure. In the orthorhombic structure (Fig. 2*b*), the displacement of the tetrahedral layers reverses every third layer, and hence the x axis is orthogonal to the z axis, producing an orthorhombic structure. The different stacking of layers in the orthorhombic structure results in a much smaller cavity around the M(4) site. Consequently, the M(4) site can accept only small B-group cations (Mn^{2+}, Fe^{2+}, Mg, and Li), and only iron-magnesium-manganese amphiboles can be orthorhombic. The greater flexibility of the monoclinic structure allows all B-group cations to occupy the M(4) site; as a result, all of the principal chemical groups of amphiboles can have the monoclinic structure. *See* CRYSTAL; CRYSTALLOGRAPHY; CRYSTAL STRUCTURE; SYMMETRY.

Immiscibility and exsolution. Amphiboles do not show the complete range of possible compositions suggested by the general chemical formula and common idealized compositions described above. In particular, there is not a continuous range of chemical composition between the four main amphibole groups: the iron-magnesium-manganese amphiboles, the calcic amphiboles, the sodic-calcic amphiboles, and the alkali amphiboles. This lack of so-called solid solution is a result of the structure not being able to accommodate two types of cations (positively charged atoms) of very different size (or charge) at the same set of sites in the structure of a single crystal. This effect is very prominent between the iron-magnesium-manganese amphiboles and the calcic amphiboles, in which the M(4) site (Fig. 1*a*) is occupied by the B-group cations (Fe^{2+}, Mg, Mn) and Ca, respectively. There is a wide range of chemical compositions common in simple calcic and iron-magnesium-manganese amphiboles (**Fig. 3**). For some chemical compositions (within the shaded areas of Fig. 3), a single amphibole crystallizes. For other chemical compositions (the unshaded areas), a single amphibole is not stable. Instead, two coexisting amphiboles form: a calcic amphibole and an iron-magnesium-manganese amphibole. Thus actinolite and cummingtonite-grunerite are immiscible.

The degree to which two amphiboles are immiscible often varies as a function of temperature and pressure; at high temperatures or pressures, miscibility is usually enhanced. The immiscible region, which is known as a miscibility gap, is narrow at high temperature but widens at lower temperature. When the amphibole composition is within the miscibility gap, a single amphibole is no longer stable. It is here that the process of exsolution (or unmixing) occurs. Actinolite exsolves cummingtonite-grunerite to form lamellae of cummingtonite within the actinolite. Coexisting amphiboles and exsolution textures are very informative about temperatures of crystallization and cooling history, particularly in metamorphic rocks. *See* PHASE EQUILIBRIUM; SOLID SOLUTION; SOLID-STATE CHEMISTRY.

Occurrence. Amphiboles are common minerals in many types of igneous rocks, and the composition of the amphibole reflects the silica content of the rock. The high melting temperature of many rocks often precludes crystallization of amphiboles from the melt. However, high-temperature minerals such as pyroxenes often react with hydrothermal fluids during cooling to produce secondary amphiboles. At higher temperatures such secondary amphiboles can form an integral part of the original rock, whereas at lower temperatures they often occur as veins infilling fractures formed during cooling. Calcic amphiboles, particularly pargasite, are characteristic of ultramafic and metabasaltic rocks and are usually quite rich in magnesium. Titanium-rich hornblendes and kaersutites occur in intermediate rocks and are often strongly oxidized—an unusual feature in amphiboles from any other environment. Acidic rocks, particularly granites, contain a wide range of amphiboles, from hastingsite to riebeckite and arfvedsonite; these are often iron-rich and can contain significant amounts of more unusual elements such as Li, Zn, and Mn. Understanding the variation in amphibole composition as a function of the progressive crystallization of a suite of rocks is important for understanding the petrological processes in an igneous environment.

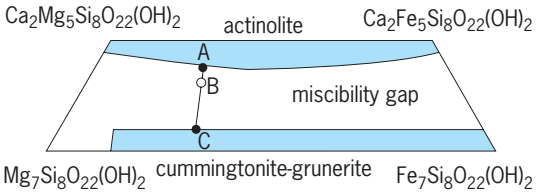


Fig. 3. Compositional variations in calcic and iron-magnesium-manganese amphiboles at low temperature. Shaded areas denote compositional regions within which single amphiboles are stable; compositions within the unshaded area (for example, composition B) cannot form a single amphibole but form two coexisting amphiboles (compositions A and C); composition A represents a calcic amphibole.

Amphiboles usually weather very easily and hence are not important in sedimentary rocks, although they can be significant components of soil. Iron-rich alkali amphiboles can form at essentially ambient conditions in the sedimentary environment, but this occurrence is rare. Amphiboles are common and important rock-forming minerals in many types of metamorphic rocks. They are particularly abundant in rocks of basaltic composition at most grades of metamorphism. In the greenschist facies, the first amphiboles to form are tremolite-actinolite at low to high pressures. With increasing grade, the amphibole composition becomes more complex, changing in the general direction of pargasite. The specific compositional changes are sensitive to differences in temperature and pressure, and are useful in following the detailed history of metamorphic rocks. Blueschists are characteristic of very high pressures and relatively low temperatures. At low grade, riebeckite and crossite form, giving the blueschists their characteristic color. With increasing temperature and pressure, the amphibole composition evolves toward glaucophane, which can coexist with actinolite in transitional facies. At very high pressures and temperatures, the miscibility gap between the alkali and calcic amphiboles disappears with the crystallization of sodic-calcic amphiboles such as winchite. Understanding these processes allows derivation of the tectonic history of a rock from the compositional variations recorded in its constituent amphiboles. *See* BASALT; BLUESCHIST; FACIES (GEOLOGY); GRANITE; IGNEOUS ROCK; METAMORPHIC ROCK; METAMORPHISM; SCHIST.

Industrial uses. Amphiboles are economically important as commercial asbestos minerals and as semiprecious gem materials. World asbestos production is dominated by the serpentine-group mineral chrysotile; but the amphibole minerals anthophyllite, cummingtonite-grunerite (amosite), actinolite, and riebeckite (crocidolite) are also important, particularly in Australia and South Africa. It is now well established that the mineralogical characteristics of asbestos have a major influence on its degree of carcinogenicity; amphibole asbestos seems to be far more harmful in this respect than chrysotile. *See* ASBESTOS.

Some amphiboles with attractive physical properties are marketed as semiprecious gem material. Most important is nephrite, a dense compact form of fibrous actinolite that is a principal variety of jade. Fibrous riebeckite is marketed as one of the less common varieties of tiger's eye. Iridescent gedrite and gem-quality pargasite are used as semiprecious gems in contemporary jewelry. *See* GEM; JADE; MINERALOGY.

Frank C. Hawthorne

Bibliography. W. A. Deer et al., *An Introduction to the Rock-Forming Minerals*, 2d ed., 1992; F. C. Hawthorne, The crystal chemistry of the amphiboles, *Can. Mineral.*, 21:173–480, 1983; F. C. Hawthorne et al., The role of Ti in hydrogen-deficient amphiboles, *Can. Mineral.*, 36:1253–1265, 1998; B. E. Leake et al., Nomenclature of amphiboles: Report of the Subcommittee on Amphiboles of the International Mineralogical Association Commission on New Minerals and Mineral Names, *Can. Mineral.*, 35:219–246, 1997; D. R. Veblen (ed.), *Reviews in Mineralogy*, vol. 9A: *Amphiboles and Other Hydrous Pyriboles: Mineralogy*, 1981; D. R. Veblen and P. H. Ribbe (eds.), *Reviews in Mineralogy*, vol. 9B: *Amphiboles: Petrology and Experimental Phase Relations*, 1982.

Amphibolite

A class of metamorphic rocks with one of the amphibole minerals as the dominant constituent. Most of the amphibolites are dark green to black crystalline rocks that occur as extensive layers widely distributed in mountain belts and deeply eroded shield areas of the continental crust. Amphibolite is the main country rock that has been intruded by the large granite masses found in most mountain ranges, with small and large masses of amphibolite present also as inclusions in granites. *See* AMPHIBOLE.

Amphibolites are the products of regional metamorphism and crustal deformation of older materials of appropriate composition. The deformation that accompanied the formation of the amphibolites is expressed by the segregation of the minerals into layers and the orientation of the needles and prisms of the amphibole crystals along a direction in the layer. The temperature range of formation is roughly 400–500°C (750–930°F) of the epidote-amphibolite and amphibolite metamorphic facies. The amphibolites are of remarkably simple mineralogies with most consisting of only common hornblende and plagioclase, although epidote, chlorite, diopside, garnet, biotite, magnetite, sphene, and quartz may be present as minor minerals.

The features of the original rock are obliterated; thus it is difficult and sometimes impossible to determine the premetamorphic rock. The bulk composition of the amphibolites could result by the addition of water and recrystallization of original intrusive masses of gabbro, diabase, or piles of basaltic lava flows and, if so, are referred to as ortho amphibolites. The bulk composition could result by the dehydration and recrystallization of lower-temperature-of-formation serpentine and chlorite schists, which also could represent altered original gabbro or basaltic lavas, but which also could be original clay-rich marine sediments (and then the rocks are referred to as para amphibolites). Compositional relations between the minor elements titanium, chromium, and nickel have been used to distinguish the ortho from para amphibolites in some occurrences.

Amphibolite layers commonly are associated with layers of various biotite schists, granite gneisses, quartzite, and marble masses that formed during the same metamorphic episode. Extensive folding, shearing, and plastic deformation such as pinch and swell structures of the amphibolite layers are characteristic features of most occurrences. Multiple episodes of deformation and recrystallization are

common in many localities, and imply a long and complex history of evolution of the rock mass.

The amphibolites are seldom sites of ore deposits, although mineralized quartz veins that crosscut the amphibolites occur in some localities.

The plate tectonic model of crustal evolution interprets the large masses of amphibolites as the alterations of basaltic crustal materials in subduction zones. *See* METAMORPHIC ROCKS; SCHIST.

George De Vore

Amphidiscosa

An order of the subclass Amphidiscophora in the class Hexactinellida. These sponges are distinguished from the order Hemidiscosa in that the birotulates are amphidiscs. Examples of this order are *Pheronema, Monorhaphis*, and *Hyalonema*. *See* HEMIDISCOSA.

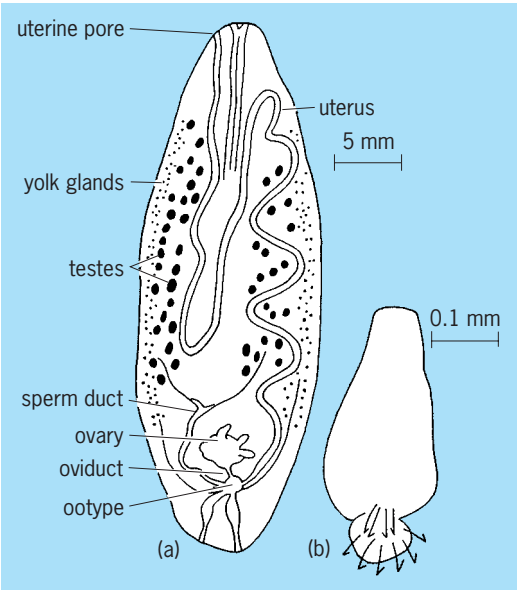
The record of fossil Amphidiscosa is poor, but goes back to the Carboniferous *Uralonema*.

Willard D. Hartman; Robert M. Finks

Amphilinidea

An order of tapeworms of the subclass Cestodaria. These worms have a protrusible proboscis and frontal glands at the anterior end. No holdfast organ is evident. All members of the order inhabit the coelom of sturgeon and other fish. The only life history which is completely known is that of *Amphilina* (**illus.** *a*). The 10-hooked embryos leave the parental uterus through a pore, and if upon escaping into the water they are eaten by an amphipod crustacean, they undergo further development to the procercoid larva (illus. *b*). When the parasitized amphipod is eaten by a sturgeon, the larval worm enters the coelom of the fish and develops to sexual maturity.

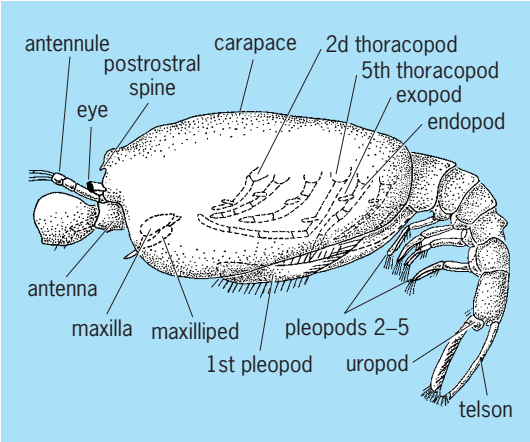
Clark P. Read



***Amphilina*. (*a*) Adult. (*b*) Procercoid larva.**

Amphionidacea

An order of the Eucarida comprising a single species, *Amphionides reynaydii* (see **illus.**). Because of the similarity of its early larval stages to those found among the caridean shrimps, for more than a century it was classified as an aberrant member of the Caridea. However, the fact that the pleuron of the second abdominal somite never overlaps that of the first in any stage immediately distinguishes this species from all true Caridea.



Amphionides reynaudii. (After P. A. McLaughlin, Comparative Morphology of Recent Crustacea, W. H. Freeman, 1980)

The cephalothorax is covered by a large but very thin carapace. Only the first pair of thoracic appendages is modified as maxillipeds. The following six pairs of limbs are biramous, with the outer rami short and virtually functionless; the inner rami are segmented and somewhat sticklike. There are no appendages on the last thoracic somite in females, but a four-segmented uniramous pair of appendages has been seen in juvenile males. The eyestalk has a large tubercle on the inner side, which has been postulated to be a photophore used for attracting prey. In addition to the single pair of maxillipeds, the pronounced development of the first pair of pleopods in adult females clearly sets *Amphionides* apart from other eucarids. When extended, these pleopods close off the ventral carapace region and appear to form a unique type of brood pouch beneath the thorax. Although no brooding females have ever been captured, the fact that the genital pores on the coxae of the sixth thoracopods open into this chamber gives support to the brood pouch concept. The very long fifth thoracopods could be used for cleaning and egg tending.

Amphionides may have up to 12 or 13 free-swimming larval stages followed by a metamorphic

molt to the juvenile form. It is these larval stages that are most frequently captured in plankton samples, and in the nineteenth century several of the larval stages were described as distinct species. Adult males have never been collected; however, sexual dimorphism in the last juvenile stage has been described. It is presumed that adult males are much more efficient swimmers, and thus able to avoid capture. It is also possible that their only function is one of reproduction, and that adult males are relatively short-lived.

Amphionides reynaudii has worldwide distribution, primarily in equatorial regions. Adult females have been collected at depths between 21,000 and 111,000 ft (700 and 3700 m), while most larvae occur in depths of only 90–300 ft (30–100 m). *See* EUCARIDA.

Patsy A. McLaughlin

Bibliography. P. Heegaard, *Larvae of Decapod Crustacea: The Amphionidae, Dana*, Rep. 77, pp. 1–67, 1969; D. I. Williamson, *Amphionides reynaudii* (H. Milne Edwards), representative of a proposed new order of eucaridan Malacostraca, *Crustaceana*, 25:35–50, 1973.

Amphipoda

An order of crustaceans in the subclass Malacostraca, which lack a carapace, bear unstalked eyes, and respire by thoracic branchiae, or gills. The abdomen usually bears three pairs of biramous swimmerets (pleopods), three pairs of rather rigid uropods, and a telson which may be lobed or entire. The body is usually flattened laterally, and the pereiopods (walking legs), unlike those of the Isopoda, are elongated so that walking is difficult. In contrast to the Isopoda, the maxillipeds lack epipodites. The sexes are separate, but reproductive and copulatory organs are very simple. The eggs are extruded by the female into a ventral brood pouch composed of setose lamellae attached to the medial bases of the legs. The young hatch as miniature adults, growing usually to a length of 0.12–0.48 in. (3–12 mm), and in exceptional cases to 5.6 in. (140 mm). *See* ISOPODA.

Amphipods are very abundant in the oceans, being represented by 3200 species. More than 600 other species occur in streams, lakes and subterranean waters and in terrestrial leaf molds and mosses. Many are excellent swimmers. Nonpelagic species of the suborders Gammaridea and Caprellidea live on aquatic bottoms, plants, and epifaunal growths. Predation by amphipods is occasional. Their mouthparts are well adapted for chewing (**Fig. 1**); they either eat aquatic plants, debris, and detritus or swallow mud containing food particles. A few are filter feeders; others feed by sucking animal tissues.

Four suborders are known, the Gammaridea, Hyperiidea, Caprellidea, and Ingolfiellidea.

Gammaridea. The pleon, or abdomen, is well developed and the maxilliped bears a palp **(Fig. 2*a*)**. This is the largest suborder, comprising 3200 species. They are principally marine, but the group contains

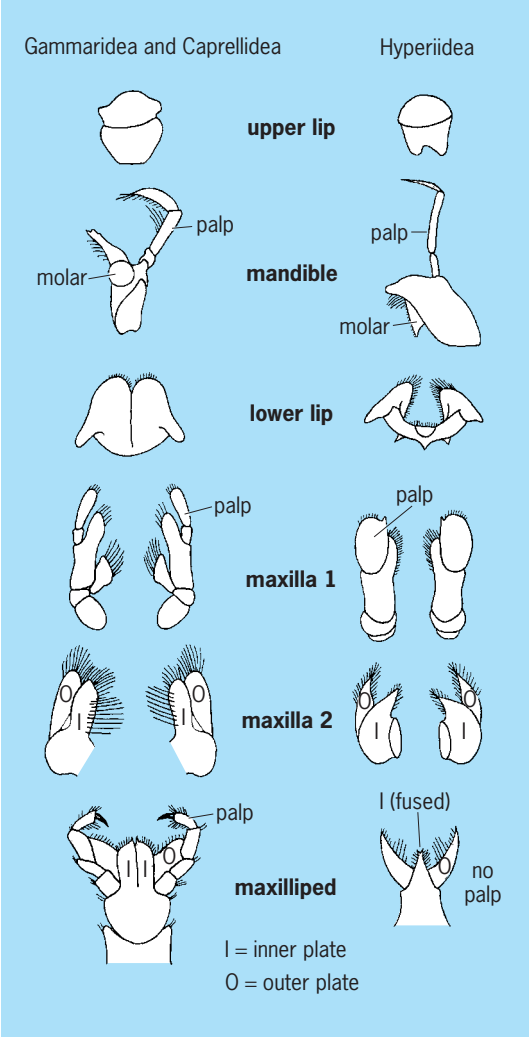


Fig. 1. Generalized mouthparts of Amphipoda. (*After G. O. Sars, An Account of the Crustacea of Norway, vol. 1, 1895*)

the only nonmarine forms, of which 500 are limnetic and 80 terrestrial. About 200 of the marine species are pelagic, the remainder being benthic or intertidal. Gammarideans have been discovered in the deepest oceanic trenches and at altitudes of 13,000 ft (4000 m) on tropical islands.

Hyperiidea. The abdomen is well developed (Fig. 2*b*), but the maxilliped lacks a palp. The suborder comprises about 300 species which are entirely marine and pelagic. Adaptations of pelagic life include suspensory mechanisms such as oil storage, broadened body surfaces, and elongated appendages. The eyes are conspicuous, because of the proliferation of ommatidia to cover the entire cephalic area (Fig. 2*b*) or to their absence in bathypelagic species. Some hyperiids spend much of their lives encased within medusae and salps.

Caprellidea. The abdomen is vestigial, and the maxilliped bears a palp. The body is very slender, resulting in the name "skeleton shrimp" (Fig. 2*c*). Caprellids are entirely marine. This group contains

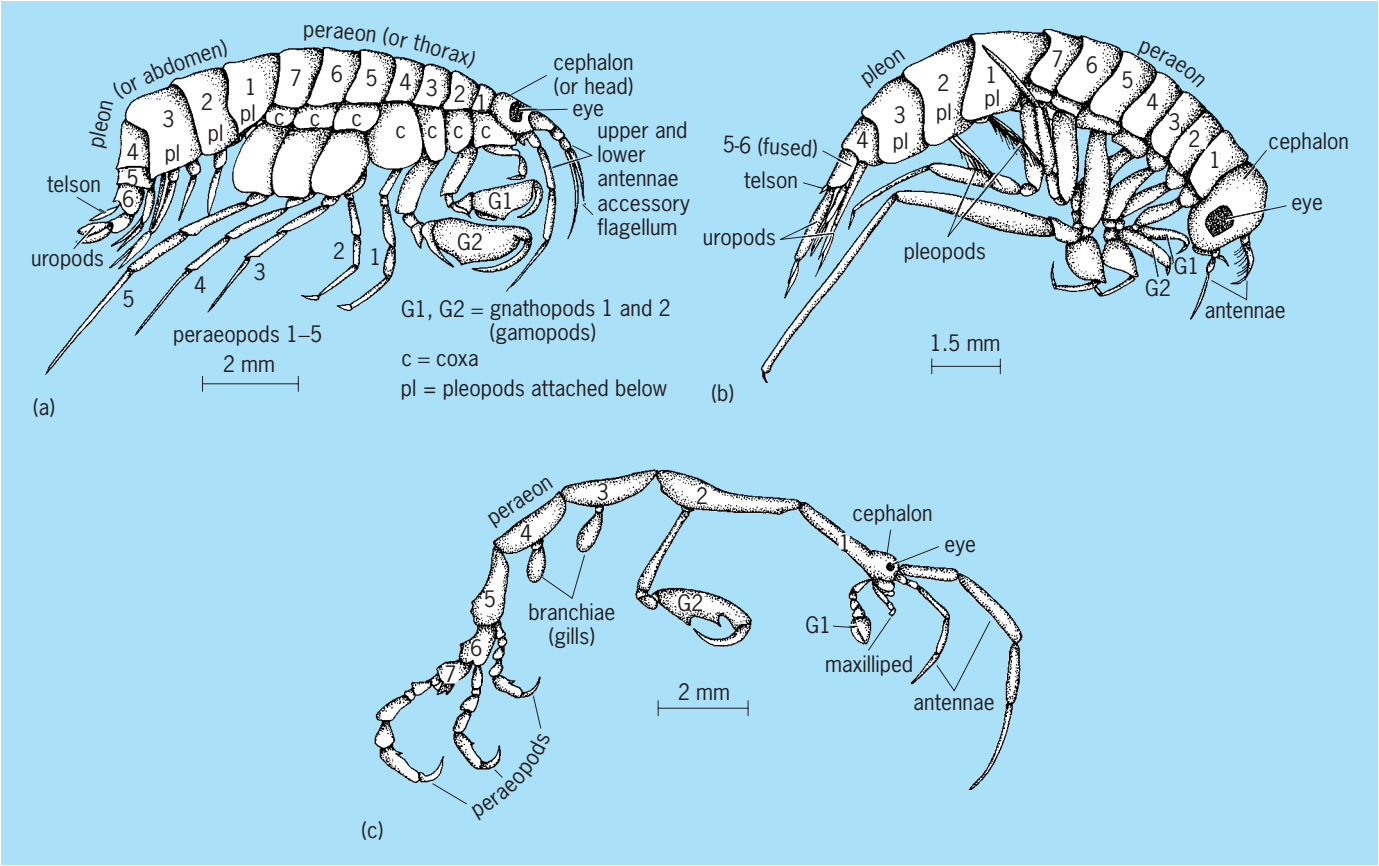


Fig. 2. Examples of the suborders of the Amphipoda. (*a*) Gammaridea, *Lilljeborgia pallida,* male; (*b*) Hyperiidea, *Parathemisto bispinosa*, female; (c) Caprellidea, *Caprella linearis*, male. (*After G. O. Sars, An Account of the Crustacea of Norway, vol. 1, 1895*)

200 species of skeleton shrimps and about 30 species of whale lice, or cyamids. The cyamids are dorsoventrally flattened animals, which are epibionts in external orifices of cetaceans (**Fig. 3**). The first thoracic segment bearing the first gnathopod is usually fused solidly to the head in caprellids, leaving only six free thoracic segments. Most species lack the first two pairs of pereiopods, while a few bear remnants of these legs. Branchial lamellae occur on two or three free segments, and female brood lamellae occupy the second and third free segments. Caprellids commonly occur on epifaunal growths in shallow water; some of them feed on hydrozoan coelenterates. A behavioral motion similar to that of the praying mantids (insects) has been described for them.

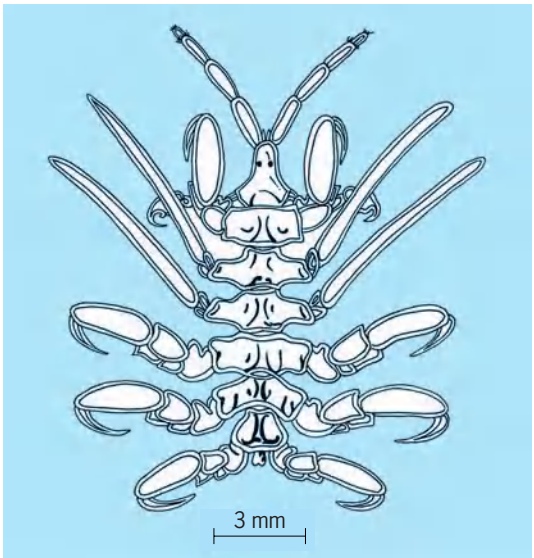


Fig. 3. *Cyanis boopis*, male; a whale louse (Caprellidea) which is modified as an epibiont on cetaceans. (*After G. O. Sars, An Account of the Crustacea of Norway, vol. 1, 1895*)

Ingolfiellidea. Both abdomen and maxilliped are well developed as in the Gammaridea, but the head often bears a separate ocular lobe lacking eyes. The pleopods are simple and leaflike, and the body is thin, resembling a tanaid. The dactyls of the gnathopods are composed of two articles. The suborder comprises a dozen species, in the abyssal sea, in the shallow sea, and in subterranean localities.

Economic importance. Marine species are important food for various stages of many commercial fishes. Hyperiids are the principal food of seals at certain seasons, and also of balaenoid whales at times. One gammaridean genus, *Chelura*, is a minor wood borer, associated with the isopod *Limnoria*.

A few fossil species are known in Tertiary amber deposits. *See* HYPERIIDEA; MALACOSTRACA; MARINE ECOLOGY.

J. Laurens Barnard

Amplifier

A device capable of increasing the magnitude of a physical quantity. This article discusses electronic amplifiers whose operation depends on transistors. Some amplifiers are magnetic, while others may take the form of rotating electrical machinery. Forms of nonelectrical amplifiers are hydraulic actuators and levers which are amplifiers of mechanical forces. *See* DIRECT-CURRENT MOTOR; HYDRAULIC ACTUATOR; LEVER.

The operation of an amplifier can be explained with a model (**Fig. 1**). A controlled voltage source of

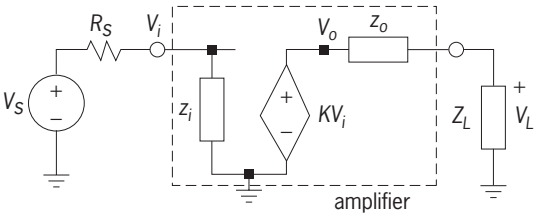


Fig. 1. Amplifier model with source, load, and input and output impedances.

gain K generates an output voltage $V_o = KV_i$ from an input voltage V_i. This input voltage is obtained from a source voltage V_S with source resistance R_S via voltage division with the amplifier's input impedance z_i. The load voltage V_L across the load impedance Z_L is obtained from V_o by voltage division with the amplifier's output impedance z_o. The input voltage and load voltage are given by Eqs. (1), where k_i

$$V_i = \frac{z_i}{z_i + R_S} V_S = k_i V_S$$
$$V_L = \frac{Z_L}{Z_L + z_o} V_o = k_o V_o \tag{1}$$

and k_o, respectively, express the effects of the amplifier loading the source and of the load impedance loading the amplifier. The impedances z_i and z_o mostly consist of a resistor in parallel with a capacitor; often they may be assumed to be purely resistive: $z_i = r_i$ and $z_o = r_o$. From Eq. (1), the amplifier's operation is given by Eq. (2). Thus, the amplification

$$V_L = k_i K k_o V_S = \frac{z_i}{z_i + R_S} K \frac{Z_L}{Z_L + z_o} V_S \tag{2}$$

is decreased from the ideal value K by the two load factors k_i and k_o. The reduction in gain is avoided if the two factors equal unity, that is, if $z_i = \infty$ and $z_o = 0$. Thus, in addition to the required gain K, a good amplifier has a very large input impedance z_i and a very small output impedance z_o. *See* GAIN.

Operational Amplifiers

The operational amplifier (op amp) is a commonly used general-purpose amplifier. It is implemented as an integrated circuit on a semiconductor chip, and functions as a voltage amplifier whose essential characteristics at low frequencies are very high voltage amplification, very high input resistance, and very low output resistance. *See* INTEGRATED CIRCUITS; VOLTAGE AMPLIFIER.

The operational amplifier has three signal terminals (**Fig. 2***a*): an inverting and a noninverting input and one output. There are also two terminals to apply dc power, which are normally not drawn in circuit diagrams. Further terminals may be present for other functions, such as offset control. Because the electronic circuitry inside the operational amplifier does not permit excursions of the output voltage to exceed the supply voltages of either polarity, both positive and negative supply voltages, $\pm V_B$, with respect to ground are required if both positive and negative excursions of output voltage with respect to ground are desired. All signal voltages are referenced to ground.

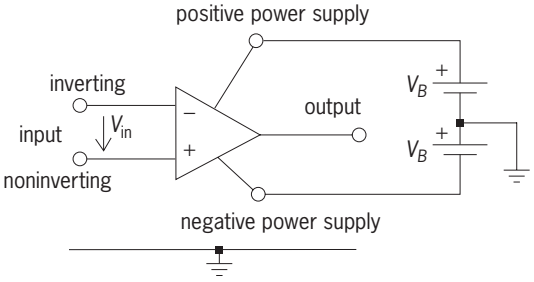
If the voltages at the inverting and noninverting input terminals are called V^- and V^+, respectively, the operational amplifier input voltage V_{in} equals $V^+ - V^-$, and the output voltage V_o is given by Eq. (3).

$$V_o = AV_{\text{in}} = A(V^+ - V^-) \tag{3}$$

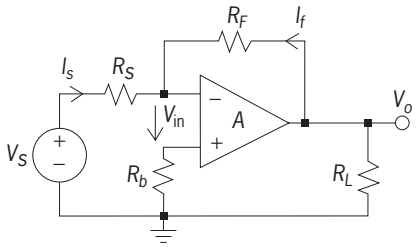
Because the magnitude of the output voltage V_o cannot become larger than the power supply voltage V_B, and the operational amplifier gain A is very large, the magnitude of the voltage V_{in} across the input terminals of an op amp is very small, being limited as in Eq. (4). With $A = 10^5$ and, typically, $V_B = 12$

$$V_{\text{in}} = V^+ - V^- = \frac{V_o}{A} < \frac{V_B}{A} \tag{4}$$

V, it follows that $V_{\text{in}} = 12 \times 10^{-5}$ V $= 120\ \mu$V. This voltage is normally much smaller than any other signal in the circuit, and V_{in} is therefore approximated as zero. Expressed differently, because the gain of the operational amplifier is so large, even very small



(a)



(b)

Fig. 2. Operational amplifier. (*a*) Circuit symbol, showing signal terminals and power supply terminals. (*b*) Simple operational amplifier circuit to amplify the signal voltage V_S by the gain $-R_F/R_S$, and apply it to load resistor R_L.

input voltages would be amplified to levels larger than the power supplies permit. Thus, to obtain practical amplifiers with gain less than $A = V_o/V_{\text{in}}$, operational amplifiers are always used with negative feedback (Fig. 2*b*). Feedback also reduces distortion and increases gain stability by the same factor as the reduction in voltage gain. *See* ANALOG COMPUTER; DISTORTION (ELECTRONIC CIRCUITS); FEEDBACK CIRCUIT; OPERATIONAL AMPLIFIER.

Transconductance Amplifiers

The transconductance amplifier has become widely used. In contrast to operational amplifiers, which convert an input voltage to an output voltage, transconductors are voltage-to-current converters described by the transconductance parameter g_m, which satisfies Eq. (5). Thus, the output current I_{out}

$$I_{\text{out}} = g_m V_{\text{in}} \tag{5}$$

is proportional to the input voltage V_{in}. Whereas operational amplifier circuits are mostly found with single-ended output terminals (Fig. 2*a*), transconductors have either single-ended (**Fig. 3***a*) or (mostly)

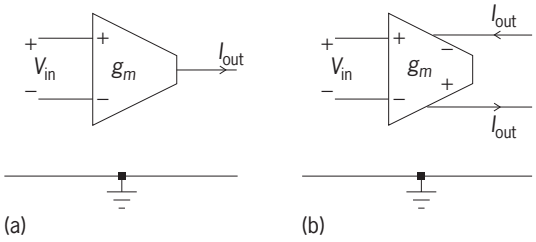


Fig. 3. Circuit symbols for transconductance amplifiers with (*a*) single-ended, output and (*b*) differential output.

differential outputs (Fig. 3*b*). As do operational amplifiers, ideal transconductors have an infinite input resistance, but they also have an infinite output resistance so that the output is an ideal current source (see **table**). One of the attractive properties of transconductance amplifiers is their wide bandwidth. Very simple transconductance circuits can be designed which maintain their nominal g_m values over bandwidths of several hundred megahertz (10^8 Hz), whereas operational amplifiers often have high gain only over a frequency range of less than 100 Hz, after which the gain falls off rapidly. Consequently, in high-frequency communications applications, circuits built with transconductance amplifiers

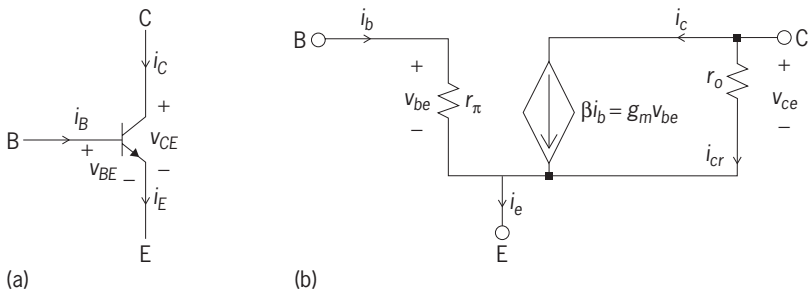


Fig. 4. An *npn* transistor. (*a*) Circuit symbol. (*b*) Small-signal model.

Transconductance amplifier parameters

Parameter	Symbol	Typical value
Transconductance	g_m	10–1000 μA/V
Input resistance	r_i	1,000,000 Ω
Output resistance	r_o	100,000 Ω

generally give much better performance than those with operational amplifiers.

Bipolar Junction Transistor Amplifiers

A bipolar junction transistor has three terminals named the base (*B*), emitter (*E*), and collector (*C*) (**Fig. 4***a*). It consists of two *pn* junctions, the emitter-base junction, and the collector-base junction. In normal amplifier operation, the emitter-base junction is forward biased and the collector-base junction is reverse biased; that is, for an *npn* transistor (Fig. 4*a*), the base voltage is more positive than the emitter voltage by $V_{BE} \approx 0.65$ V, and the collector is more positive than the base. These polarities are inverted in a *pnp* transistor.

Because it is necessary to distinguish between constant currents and voltages (dc or bias conditions), time-varying signals (ac), and the total quantities in the analysis of electronic circuits, the following standard notation is used: dc is denoted as uppercase sub uppercase, ac is lowercase sub lowercase, and the total quantity is lowercase sub uppercase. Thus, the total collector current i_C, consisting of the sum of the dc bias current I_C and the time-varying ac signal current i_c, would be written as Eq. (6). Other

$$i_C(t) = I_C + i_c(t) \tag{6}$$

quantities are expressed similarly; for example, $v_{BE} = V_{BE} + v_{be}$ for the base-emitter voltage.

Transistor operation. The operation of the bipolar transistor can be described by the simple equations (7) and (8), where V_T is the strongly

$$i_C \cong I_S[\exp(v_{BE}/V_T) - 1] \approx I_S \exp(v_{BE}/V_T) \tag{7}$$

$$i_B = \frac{i_C}{\beta} \qquad i_E = \frac{i_C}{\alpha} \tag{8}$$

temperature-dependent thermal voltage (≈ 0.026 V at room temperature), and the approximation in Eq. (7) is valid in the forward-biased region, where $v_{BE} \approx 0.65$ V is much larger than V_T. The quantity I_S is the reverse saturation current (the current that flows in the collector when $v_{BE} < 0$; a few nanoamperes), β is the base-to-collector current gain, and α is the emitter-to-collector current gain. It follows from the transistor configuration (Fig. 4*a*) that $i_E = i_C + i_B$, and this result, along with Eq. (8), leads to Eqs. (9).

$$\alpha = \frac{\beta}{\beta + 1} \qquad \beta = \frac{\alpha}{1 - \alpha} \tag{9}$$

Typical numbers are $\alpha \approx 0.99$ and $\beta \approx 100$. Thus, i_B is much smaller than both i_C and i_E, so that $i_C \approx i_E$. *See* TRANSISTOR.

Linear performance. To achieve linear amplifier performance for small ac signals, the highly nonlinear bipolar junction transistor is operated at a dc bias point (I_C, V_{CE}) by choosing a dc base current I_B as discussed below. The ac signals are then superimposed on the dc bias points. For small-signal operation, the bipolar junction transistor can be modeled by a simple linear circuit (Fig. 4*b*). To derive this circuit, the small-signal approximation, $v_{be} \ll V_T$, is used with Eq. (7). The small-signal current i_c is found to be linearly related to the small-signal voltage v_{be} by Eq. (10), where the transconductance parameter g_m, given by Eq. (11), is proportional to the dc bias

$$i_c = g_m v_{be} = \beta i_b \tag{10}$$

$$g_m = \frac{I_C}{V_T} \tag{11}$$

current I_C that is selected by the designer. The second equality in Eq. (10) follows from Eq. (8) Dividing both sides of Eq. (10) by $i_b g_m$ yields Eq. (12), where

$$r_\pi = \frac{v_{be}}{i_b} = \frac{\beta}{g_m} \tag{12}$$

r_π is the small-signal input resistance of the transistor. In some amplifier configurations the input resistance r_e, given by Eq. (13), is also of interest.

$$r_e = \frac{v_{be}}{i_e} = \frac{\alpha}{g_m} \approx \frac{1}{g_m} \tag{13}$$

An additional component in the small-signal circuit is the output resistor r_o. It cannot be obtained from Eqs. (7) and (8) but arises from the Early effect, which causes the collector current to increase with increasing collector-emitter voltage. Thus, a current i_{cr} is added to i_c such that Eq. (14) is satisfied.

$$r_o = \frac{v_{ce}}{i_{cr}} \tag{14}$$

With the small-signal transistor parameters identified, the gain and input and output resistances of bipolar junction transistor amplifiers can now be readily derived. First, however, it is necessary to deal with the problem of bias design, which will be explained on the case of a common-emitter amplifier. *See* BIAS (ELECTRONICS).

Common-emitter amplifier. Because bipolar junction transistor parameters are strong functions of temperature and β varies widely from device to device, amplifier performance must be made fairly independent of the transistor parameters and the bias circuitry must be designed such that the collector bias current I_C is constant (**Fig. 5***a*). The bias circuitry consists of four resistors, labeled R_1, R_2, R_E, and R_C, and a battery, which provides voltage V_{CC}. Two coupling capacitors, labeled C_C, conduct no dc current and serve to isolate the dc bias quantities from the load resistor R_L and the source v_S with source resistor R_S. The resistor R_E, between the emitter terminal and ground, provides feedback between the output and the input of the amplifier. This feedback causes the desired stabilization of the dc operating point as discussed below, but also reduces the gain. Thus, to achieve a stable operating point, R_E must be included in the dc circuit but removed from the ac circuit. This step is accomplished by a bypass capacitor C_B which shunts ac currents away from R_E and places the emitter essentially at ac ground. It makes the emitter terminal common to both input and output and leads to the name common-emitter amplifier.

The base bias current I_B is derived by replacing the bias circuit seen from the base by a Thévenin equivalent: The voltage V_B is obtained by voltage division as Eq. (15), and the Thévenin resistance at

$$V_B = \frac{R_2}{R_1 + R_2} V_{CC} \tag{15}$$

the base equals the parallel connection of R_1 and R_2, given by Eq. (16). For sufficiently large values of β,

$$R_B = \frac{R_1 R_2}{R_1 + R_2} \tag{16}$$

the bias point is given approximately by Eqs. (17) and (18). Thus, the bias point is almost independent

$$I_C = \frac{(V_B - V_{BE})}{R_E} \tag{17}$$

$$V_{CE} \approx V_{CC} - I_C(R_C + R_E) \tag{18}$$

of the transistor parameters such as β, and is essentially fixed by the choice of V_{CC} and of the resistors. *See* THÉVENIN'S THEOREM (ELECTRIC NETWORKS).

The operating point is the dc bias point on a graph of I_C versus V_{CE} (Fig. 5*b*). It is specified by the intersection of the straight load line, given by Eq. (18), and the transistor characteristics given by the manufacturer. The operating point is chosen approximately in the center of the linear range, at a suitable current I_C and at $V_{CE} \approx V_{CC}/2$. Placing the operating point at $V_{CC}/2$ permits the largest possible undistorted signal amplitude of approximately $V_{CC}/2$ at the output. The dc voltage across R_E, V_E, is chosen approximately as $V_E \approx 0.2\ V_{CC}$.

With the operating point set as prescribed, the transistor can be replaced in the common-emitter amplifier (Fig. 5*a*) by its small-signal model (Fig. 4*b*) to compute the gain of the amplifier and its input and output resistances. In the resulting small-signal circuit (Fig. 5*c*), all capacitor values are assumed to be infinite, that is, the capacitors are short circuits for ac signals. Furthermore, the power supply V_{CC} is ac signal ground.

The small-signal circuit shows that the input resistance seen by the source is R_B in parallel with r_π; that is, with Eqs. (11) and (12) and the chosen bias conditions, it is given by Eq. (19). (The resistor R_B

$$r_{\text{in}} = \frac{r_\pi R_B}{r_\pi + R_B} \approx r_\pi \tag{19}$$

is usually much greater than r_π.) Similarly, the output resistance seen by the load resistor is given by

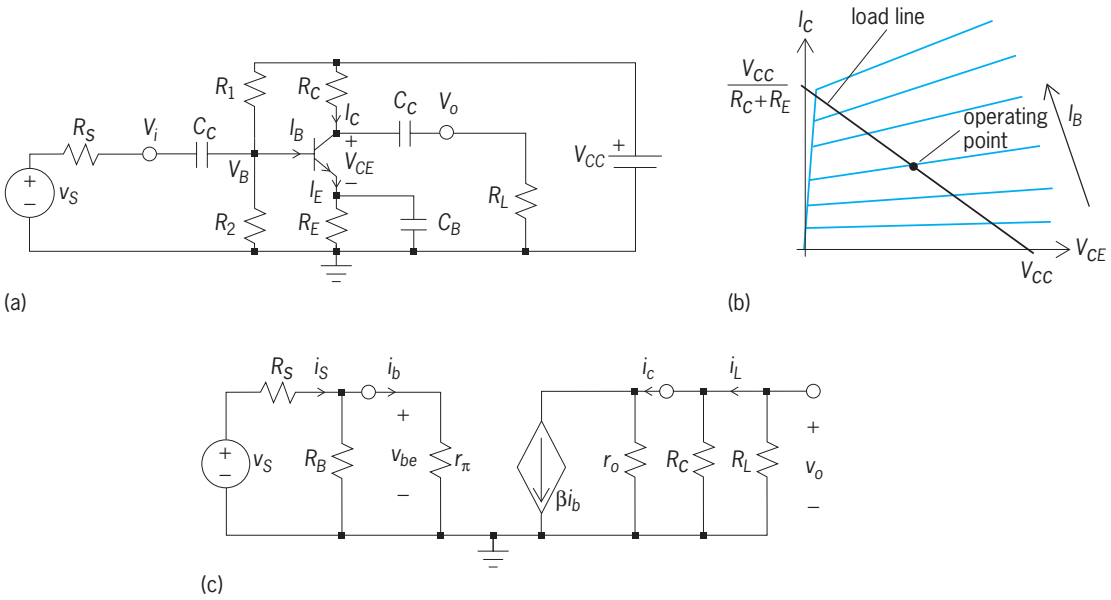


Fig. 5. Bipolar junction transistor common-emitter amplifier. (*a*) Circuit diagram. (*b*) Transistor dc characteristics and load line, showing placement of operating point. Arrow indicates increasing values of I_B. (*c*) Small-signal model of amplifier.

Eq. (20). (The resistor R_C is usually much smaller than r_o.)

$$r_{\text{out}} = \frac{r_o R_C}{r_o + R_C} \approx R_C \tag{20}$$

Finally, the voltage gain can be computed from the small-signal circuit, neglecting R_B and r_o, as Eq. (21),

$$A_v = \frac{v_o}{v_S} = -\frac{r_\pi}{R_S + r_\pi}\left(\frac{\beta}{r_\pi}R_C\right)\frac{R_L}{R_C + R_L} \tag{21}$$

which compares directly with Eq. (2). The quantity $(\beta/r_\pi)R_C = g_m R_C$ is the intrinsic gain of the unloaded amplifier, labeled K in Eq. (2). The amplifier's current gain is readily obtained from the small-signal circuit (Fig. 5*c*), neglecting r_o, as Eq. (22).

$$A_i = \frac{i_L}{i_S} \approx -\beta\frac{R_C}{R_C + R_L} \tag{22}$$

Thus, the common-emitter amplifier provides substantial voltage and current gains. It has low input resistance r_{in} and somewhat larger, but still small, output resistance r_{out}.

Common-base amplifiers. In a common-base amplifier (**Fig. 6***a*), the input voltage v_S with source resistor R_S is applied through a coupling capacitor C_C to the emitter-base junction which is forward-biased by a dc bias source $-V_{EE}$. Bias current flows from the power supply V_{CC} to $-V_{EE}$. The load R_L is connected through a second coupling capacitor C_C. In the small-signal model (Fig. 6*b*), the controlled current source is labeled $\alpha\ i_e$ because the emitter current i_e is the controlling input current into the bipolar junction transistor. The model uses the input resistance r_e defined in Eq. (13).

According to the small-signal model (Fig. 6*b*), the input resistance seen by the source to the right of R_S is given by Eq. (23). (The resistance r_e is usually much

$$r_{\text{in}} = \left(\frac{1}{R_E} + \frac{1}{r_e}\right)^{-1} \approx r_e \tag{23}$$

smaller than R_E, the resistor between the emitter and $-V_{EE}$.) The resistance r_{in} is bias dependent and by Eqs. (11) and (13) is typically quite small. Thus, the common-base amplifier has a very low input resistance. The output resistance r_{out} seen from the output voltage terminal v_o back into the circuit is found by setting $v_S = 0$, so that $i_e = 0$. Thus, this resistance

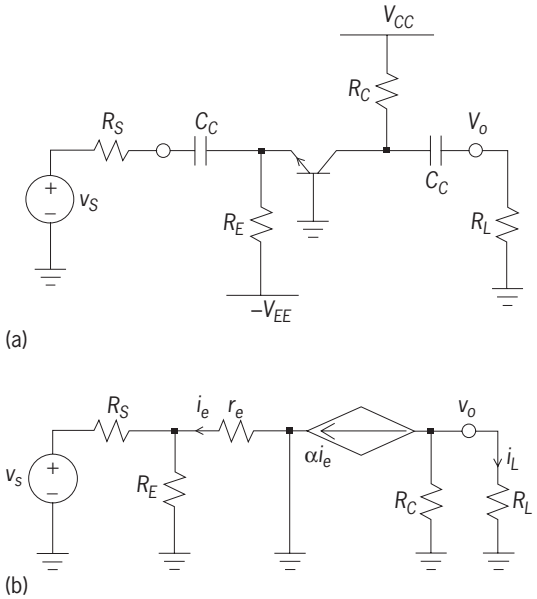


Fig. 6. Common-base amplifier. (*a*) Circuit diagram. (*b*) Small-signal model.

is given by Eq. (24), where R_C is the resistor between

$$r_{\text{out}} = R_C \tag{24}$$

the collector and V_{CC}.The voltage and current gains are derived from the small-signal model (Fig. 6*b*) as Eqs. (25) and (26).

$$A_v = \frac{V_o}{V_S} \approx \alpha \frac{1}{r_e + R_S} \frac{R_C R_L}{R_C + R_L} \tag{25}$$

$$A_i = \frac{i_L}{i_e} = \alpha \frac{R_C}{R_C + R_L} \tag{26}$$

The common-base amplifier is used as a current buffer which accepts an input current into a very small input resistance r_e and delivers that current (multiplied by $\alpha \approx 1$) with a much larger output resistance into the load. If R_C can be used as load resistor and R_L is absent, the current gain is almost equal to unity. Another advantage of the common-base amplifier is that its bandwidth is much wider than that of the common-emitter amplifier.

Common-collector amplifiers. The common-collector configuration, (**Fig. 7***a*), also called an emitter follower, is the third amplifier connection of a bipolar junction transistor. Here the collector is connected directly to the power supply, that is, to ac signal ground, so that it is common to input and output. The output is now taken from the emitter terminal. The input resistance r_{in} seen from the source with resistor R_S is computed from the small-signal model (Fig. 7*b*), neglecting r_o, as in Eq. (27), where R_B is the

$$r_{\text{in}} = R_B || R_{\text{in}} = R_B || [r_\pi + (\beta + 1)(R_E || R_L)] \tag{27}$$

resistor between the base and ground and R_E is the resistor between the emitter and a dc bias source $-V_{EE}$. (The notation || means "in parallel with.") Here, the resistance connected to the emitter is multiplied by $\beta + 1$ and adds to the resistance seen into the base of the amplifier, so that $R_{\text{in}} = v_b/i_b \approx (\beta + 1)(R_E || R_L)$ is very large. This is one of the advantages of the emitter-follower: it does not load the signal source. The output resistance r_{out} seen from the load R_L into the circuit, which is calculated from the small-signal model by setting $v_S = 0$, is approximated by Eq. (28). Thus, into the emitter is seen the total re-

$$r_{\text{out}} \approx \frac{r_\pi + R_S}{\beta + 1} \tag{28}$$

sistance of the base circuit divided by $\beta + 1$; that is, the output resistance is very small.

From the small-signal model, neglecting R_B, it follows that the voltage and current gains are given approximately by Eqs. (29) and (30).

$$A_v = \frac{v_o}{v_S} \approx 1 \tag{29}$$

$$A_i = \frac{v_o/R_L}{v_S/r_{\text{in}}} \approx \frac{r_{\text{in}}}{R_L} \tag{30}$$

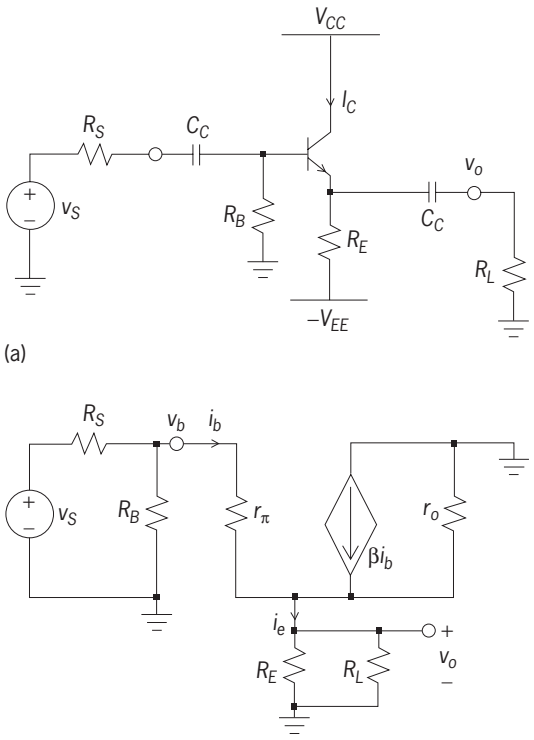


Fig. 7. Common-collector amplifier (emitter follower). (*a*) Circuit diagram. (*b*) Small-signal model.

These results show that the emitter follower is a very useful circuit; it is a buffer: without loading the source because of its large input resistance, it has a voltage gain of (somewhat less than) unity combined with a large current gain and a very small output resistance. *See* EMITTER FOLLOWER.

Field-Effect Transistor Amplifiers

In a field-effect transistor (FET) the currents are controlled by an applied electric field, generated by a voltage at a control electrode (the gate). There are two types of field-effect transistors, the junction field-effect transistor (JFET) and the insulated-gate or metal-oxide-semiconductor FET (MOSFET), each of which is available in two polarities (*n* channel and *p* channel). Further, there are depletion- and enhancement-mode MOSFETs, depending on how the conducting channel is established. The field-effect transistor differs from the bipolar junction transistor in the following characteristics: it is simpler to fabricate and occupies less space; in normal operation, the gate draws essentially no current, resulting in a very high input resistance; it exhibits no offset voltage at zero drain current; but, unfortunately, it generally has a significantly lower transconductance g_m at normal bias current levels. The operation of JFET and MOSFET amplifiers is similar in most respects. Differences arise from the fact that in a MOSFET the gate is insulated from the semiconducting channel by an oxide, whereas in a JFET this oxide insulator is replaced

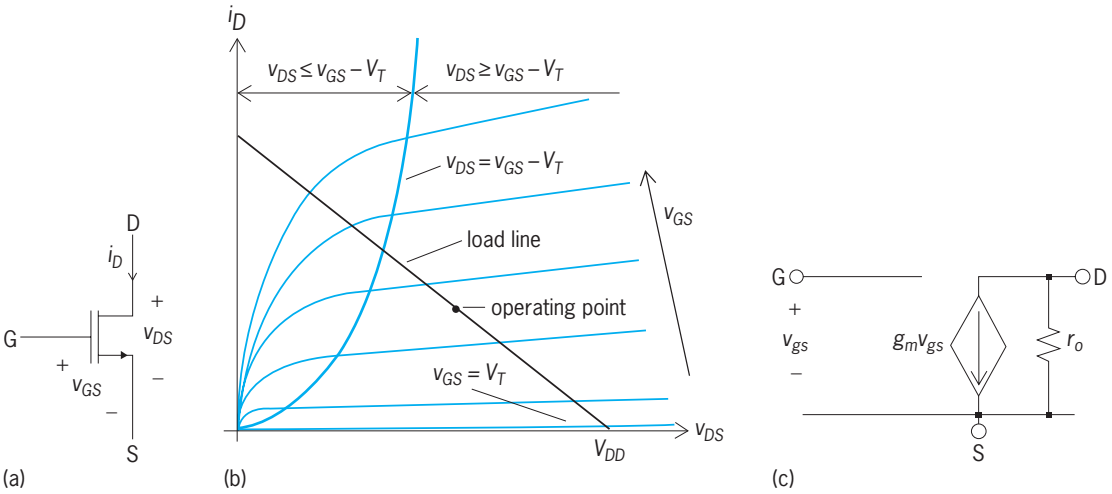


Fig. 8. An *n*-channel MOSFET. (*a*) Symbol. (*b*) Transistor dc characteristics. (*c*) Small-signal model.

by the depletion region of a reverse-biased *pn* junction. Only *n*-channel MOSFET amplifiers will be discussed.

MOSFET operation. A MOSFET has source (*S*), drain (*D*), and gate (*G*) terminals (**Fig. 8***a*). In an *n*-channel MOSFET, electrons flow from the source through the channel to the drain by applying a positive bias voltage V_{DS} to the drain. In an enhancement-type MOSFET, the channel is established by applying a positive voltage V_{GS} to the gate. The voltage V_{GS} must at least equal the threshold voltage V_T; increasing V_{GS} widens the channel and increases the current (Fig. 8*b*).

In the region $v_{DS} \ge v_{GS} - V_T$, the operation of the MOSFET is described by Eq. (31), where K is

$$i_D = K(v_{GS} - V_T)^2(1 + \lambda v_{DS}) \qquad (31)$$

a parameter that depends on the electron mobility, the oxide capacitance per unit area, and the channel width and length; and λ is the channel-length modulation parameter that accounts for the increase in i_D with increasing v_{DS} (Fig. 8*b*). As in bipolar junction transistors, to achieve linear amplification from the nonlinear MOSFET it is necessary to establish an operating point, that is, a dc bias point on a graph of I_D versus V_{DS} (Fig. 8*b*), and superimpose the ac signals on the dc conditions. The bias voltage V_{DS} must lie in the range $V_{GS} - V_T < V_{DS} < V_{DD}$, where V_{DD} is the power supply voltage. The small-signal operation is then described by a small-signal model (Fig. 8*c*) with the output resistance given by Eq. (32) and the transconductance given by Eq. (33). The derivatives

$$r_o = \left[\frac{di_D}{dv_{DS}}\right]^{-1} = \frac{1}{\lambda K(V_{GS} - V_T)^2} \simeq \frac{1}{\lambda I_D} \qquad (32)$$

$$g_m = \frac{di_D}{dv_{GS}} = 2K(V_{GS} - V_T) \qquad (33)$$

are evaluated at the dc operating point. The gate is an open circuit so that no gate current flows.

MOSFET amplifiers. Analogous to bipolar junction transistor circuits, there are common-source, common-gate, and common-drain amplifiers. Biasing methods use the same principles as in bipolar junction transistors, but the situation is simpler here because no gate current flows. For example, the common-source amplifier (**Fig. 9***a*) can be constructed in a manner analogous to the bipolar junction transistor common-emitter circuit with bias circuitry consisting of resistors labeled R_1, R_2, R_B, and R_D, and battery V_{DD}. The resistors R_1 and R_2 are chosen large to keep the input resistance $R_G = R_1 || R_2$ of the amplifier high. The gain obtained from the

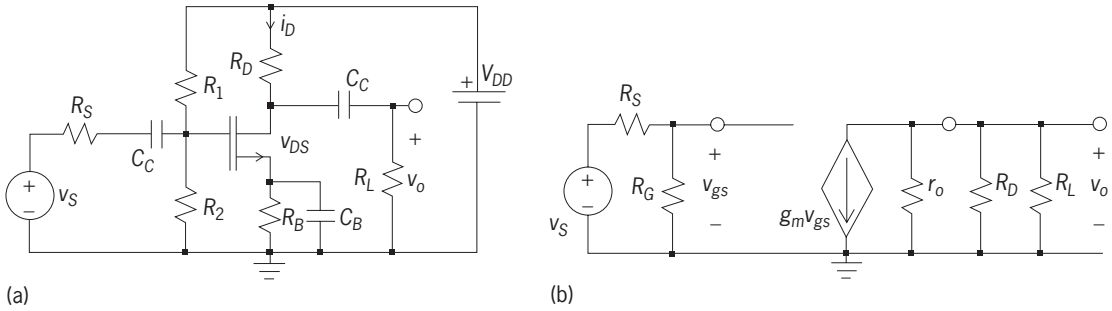


Fig. 9. Common-source amplifier. (*a*) Circuit diagram. (*b*) Small-signal model.

small-signal model (Fig. 9*b*) is given by Eq. (34),

$$A_V = \frac{v_o}{v_S} = -g_m \frac{R_G}{R_G + R_S} \frac{R_D R_L}{R_D + R_L} \qquad (34)$$

where R_S has been neglected compared to R_G, and also r_o. (Typical values for R_G and r_o are 5.4 MΩ and 50 kΩ, respectively.) This gain is much smaller than the one obtainable with a bipolar junction transistor amplifier. The output resistance equals $r_{\text{out}} = R_D \,||\, r_o \approx R_D$.

Common-gate and common-drain (source-follower) amplifiers can be constructed in a way analogous to bipolar junction transistor common-base and common-collector circuits. The source follower is found to be a good buffer circuit which does not load the signal source, provides a voltage gain of unity, and has a very low output resistance.

Differential Amplifiers

Single-ended (one input and one output) amplifiers have a number of disadvantages. They require large-valued bias resistors to achieve bias stability, which in turn restricts the dc gain. Also, they have low noise immunity, and are generally not useful for implementation in integrated circuits. The differential amplifier avoids many of these problems, but uses two or more active devices, which are either bipolar junction transistors or MOSFETs (**Fig. 10**). It is used as the input stage for almost all integrated amplifiers, including operational amplifiers.

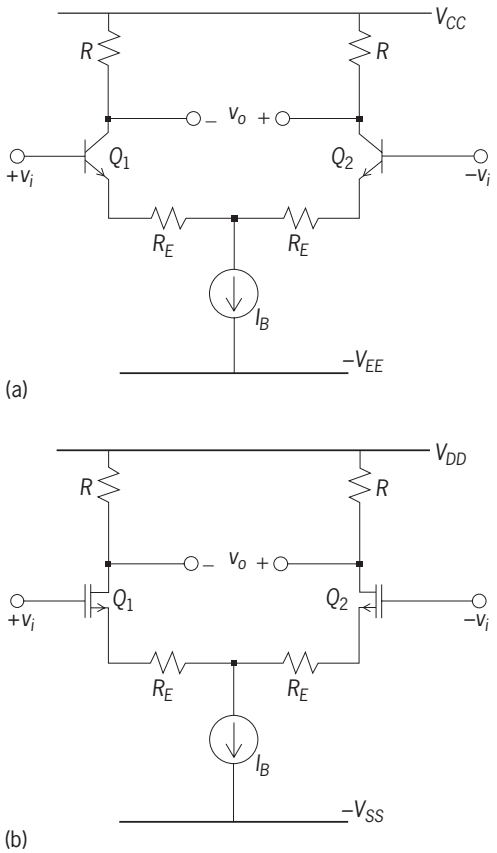


Fig. 10. Differential amplifiers. (*a*) Bipolar junction transistor amplifier. (*b*) MOSFET amplifier.

In the bipolar junction transistor stage (Fig. 10*a*), the differential gain is given by Eq. (35*a*), with the

$$A_{vd} = \frac{v_o}{v_d} = g_m(R||r_o) \approx g_m R \qquad (35a)$$

$$A_{vd} \approx \frac{g_m R}{1 + g_m R_E} \qquad (35b)$$

transconductance g_m defined in Eq. (11), where v_d is the differential voltage applied between the bases and R represents the load resistors. This simple differential transistor pair structure has the very small linear input range of only $|v_d| \approx V_T/2 \approx 13$ mV at room temperature. To increase the range, the so-called emitter-degeneration resistors R_E are included so that $v_d/2$ does not entirely appear across the base-emitter junctions of the bipolar junction transistors but also across R_E. The presence of R_E increases the linear range approximately to $|v_d| \approx R_E I_B/4$, but at the cost of a reduced gain given by Eq. (35*b*). For large R_E, the gain is essentially determined by a ratio of resistors, R/R_E. The differential input resistance between the bases of the amplifier can be derived from small-signal analysis as Eq. (36).

$$r_{d,\text{in}} = 2[r_\pi + (\beta + 1)R_E] \qquad (36)$$

The analysis of the MOSFET differential amplifier (Fig. 10*b*) is largely identical to that of the bipolar junction transistor circuit: The gain is given by Eq. (35) with g_m in Eq. (33).

The current source I_B can be realized in principle by a large resistor R_B; this approach, however, requires large power supply voltages because the voltage drop $R_B I_B$ across the bias "source" subtracts from the voltage available on the transistors. Large power supplies are normally not available in integrated circuits, and large-valued resistors may be difficult to implement. The solution is to establish a reference bias current, $I_{\text{ref}} = I_B$, possibly external to the integrated circuit chip, and use current mirrors to obtain bias currents throughout the integrated circuit. At the same time, current mirrors are used to avoid the load resistors R. *See* CURRENT SOURCES AND MIRRORS; DIFFERENTIAL AMPLIFIER.

Output Stages

Typically, amplifiers increase the power or signal levels from low-power sources, such as microphones, strain gages, magnetic disks, or antennas. After the small signals have been amplified, the amplifier's output stage must deliver the amplified signal efficiently, with minimal loss, and with no distortion to a load, such as a loudspeaker. Because the signal at the output stage is large, the small-signal approximations used earlier are normally not useful and the large-signal behavior must be investigated. The following discussion will concentrate on bipolar junction transistors, but it could equally well be based on MOSFETs; the circuit principles and concepts are the same.

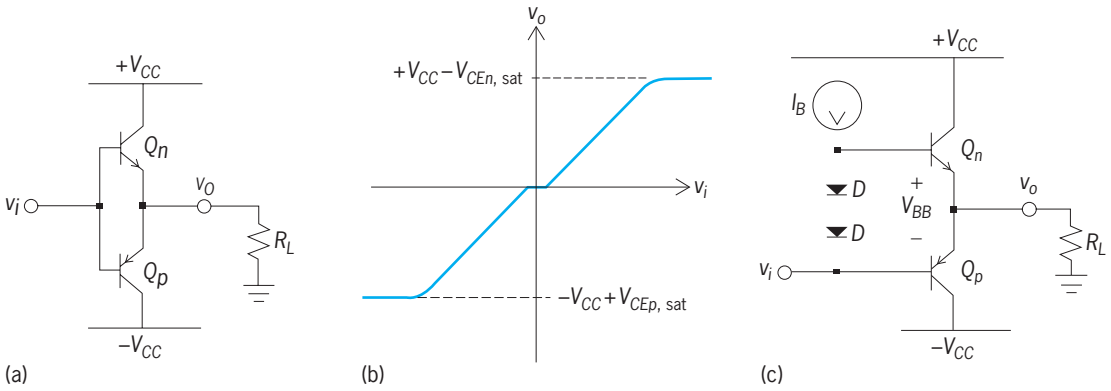


Fig. 11. Complementary bipolar junction transistor output stage. (*a*) Basic circuit, illustrating principle of operation. (*b*) Transfer curve. (*c*) Practical class AB stage.

To transfer power efficiently implies that the power dissipated in the output stage should be as small as possible. A serious problem arises when the power P dissipated in the transistors, given by Eq. (37), rises above a safe level specified by the

$$P = V_{CE} I_C \tag{37}$$

manufacturer, because an undue rise in the transistor's junction temperature will reduce the device's useful life or destroy it.

Buffer stage. An output stage should take a signal with minimal loading from an amplifier and deliver it with the least possible attenuation to the load. A buffer circuit will perform that job, such as the emitter follower (Fig. 7), where I_C must always flow and must be at least as large as the peak expected load current because I_C must be positive. This arrangement is referred to as class A operation. The average power dissipated in the transistor equals $V_{CC}I_C$, because the full power supply voltage V_{CC} appears across the device when the output voltage $v_o = 0$, that is, in the center of the signal range. The instantaneous power rises to 2 $V_{CC}I_C$ (assuming symmetrical supplies, $V_{CC} = V_{EE}$) when v_o reaches $-V_{CC}$. The power in the transistor depends on the value of the load resistor R_L; it can become infinite when $R_L = 0$, that is, when the load is short circuited, because then the collector current becomes arbitrarily large (flowing theoretically without resistance from V_{CC} through the transistor to ground) and the voltage V_{CE} equals V_{CC}. This dangerous situation must be avoided by clever circuit techniques that make an output stage short-circuit proof.

Since the power into a buffer is almost zero (because $i_b \approx 0$), the output power P_L must be drawn from the power supply and the conversion efficiency is a parameter of interest. It is defined by Eq. (38).

$$\eta = \frac{\text{load power}}{\text{supply power}} = \frac{P_L}{P_S} = \frac{0.5V_{o-\text{peak}}^2/R_L}{2V_{CC}I_C}$$

$$= \frac{1}{4}\frac{V_{o-\text{peak}}}{V_{CC}}\frac{V_{o-\text{peak}}}{R_L I_C} \le \frac{1}{4} \tag{38}$$

The factor $2V_{CC}$ appears in the denominator because (for $-V_{EE} = -V_{CC}$) the voltage across R_E is also equal to V_{CC} on average and R_E dissipates the same average power as the transistor. The last two factors in Eq. (38) are less than unity, so that the efficiency reaches only 25% at best. Much better performance can be obtained by using complementary devices.

Complementary (class AB) stage. This output stage has a complementary transistor pair, consisting of an *npn* transistor Q_n and a *pnp* transistor Q_p, connected between symmetrical power supplies $\pm V_{CC}$ (**Fig. 11***a*). When the input voltage $v_i = 0$, no current flows because neither of the transistors is forward biased ($V_{BE} = 0$), and therefore $v_o = 0$. If $v_i > V_{BEn} \approx 0.65$ V, Q_p is off, that is, does not conduct ($V_{EBp} < 0$), but Q_n conducts and current flows from $+V_{CC}$ through Q_n and the load resistance R_L to ground: $v_o > 0$. If $v_i < V_{EBp} \approx 0.65$ V, the situation is reversed: Q_n is off ($V_{BEn} < 0$), and Q_p conducts so that current flows from ground through R_L and Q_p to $-V_{CC}$: $v_o < 0$. Both transistors act as emitter followers, so that $v_o \approx v_i$, but only one at a time conducts and no dc current flows from $+V_{CC}$ to $-V_{CC}$. Apart from the small saturation voltage, $V_{CE,\text{sat}} \approx 0.2$ V, v_o can swing from $+ V_{CC}$ to $-V_{CC}$ with a gain of approximately unity and very little distortion. This behavior is depicted in the graph of v_o versus v_i, called the transfer curve (Fig. 11*b*), which also shows the so-called crossover distortion at the origin. This distortion arises because v_i must at least equal the base-emitter voltages of the transistors before the emitter followers turn on and become active.

The problem is avoided by arranging the circuit such that a dc voltage V_{BB} equal to two forward-biased *pn*-junction voltages, that is, ≈1.3 V always appears between the bases of the transistors (Fig. 11*c*). Two diodes D (in practice, two diode-connected transistors) are forward biased by a current source I_B, which also supplies base currents to Q_n and Q_p. For ac signals, the forward-biased diodes are simply very small resistors. The operation is then as just described for the basic circuit (Fig. 11*a*), but the crossover distortion is avoided.

The power conversion efficiency for this circuit, neglecting the small power dissipated in the diodes, can be shown to equal $\eta = 0.785V_{o-\text{peak}}/V_{CC}$, that is, maximally 78.5%. By using the earlier results from the emitter follower, the gain is found to equal

approximately unity, and the input and output resistances are given by Eq. (39), where i_n and i_p are the

$$r_{\text{in}} \approx (\beta + 1)R_L \qquad r_{\text{out}} \approx r_{en} || r_{ep} = \frac{V_T}{i_n + i_p} \quad (39)$$

currents in the *npn* and *pnp* devices, respectively, and r_{en} and r_{ep} are their emitter resistances as defined in Eq. (13). Thus, the input resistance is very large and the output resistance very small, as required for a good follower or buffer circuit. *See* POWER AMPLIFIER.

Rolf Schaumann

Bibliography. P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3d ed., 1993; R. Gregorian and G. C. Temes, *Analog MOS Integrated Circuits for Signal Processing*, 1986; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Amplitude (wave motion)

The maximum magnitude (value without regard to sign) of the disturbance of a wave. The term "disturbance" refers to that property of a wave which perturbs or alters its surroundings. It may mean, for example, the displacement of mechanical waves, the pressure variations of a sound wave, or the electric or magnetic field of light waves. Sometimes in older texts the word "amplitude" is used for the disturbance itself; in that case, amplitude as meant there is called peak amplitude. This is no longer common usage. *See* LIGHT; SOUND.

As an example, consider one-dimensional traveling waves in a linear, homogeneous medium. The wave disturbance y is a function of both a space coordinate x and time t. Frequently the disturbance may be expressed as $y(x, t) = f(x \pm vt)$, where v denotes the wave velocity. The plus or minus sign indicates the direction in which the wave moves, and the shape of the wave dictates the functional form symbolized by f. Then, the amplitude of the disturbance at some point x_0 is the maximum magnitude (that is, the maximum absolute value) achieved by f as time changes over the duration required for the wave to pass point x_0. A special case of this is the one-dimensional, simple harmonic wave $y(x, t) = A \sin [k(x \pm vt)]$, where k is a constant. The amplitude is A since the absolute maximum of the sine function is $+1$. The amplitude for such a wave is a constant. *See* HARMONIC MOTION; SINE WAVE.

If the medium which a wave disturbs dissipates the wave by some nonlinear behavior or other means, then the amplitude will, in general, depend upon position. *See* WAVE MOTION.

S. A. Williams

Amplitude modulation

The process or result of the process whereby the amplitude of a carrier wave is changed in accordance with a modulating wave. This broad definition includes applications using sinusoidal carriers, pulse carriers, or any other form of carrier, the amplitude factor of which changes in accordance with the modulating wave in any unique manner. *See* MODULATION.

Amplitude modulation (AM) is also defined in a more restrictive sense to mean modulation in which the amplitude factor of a sine-wave carrier is linearly proportional to the modulating wave. AM radio broadcasting is a familiar example. At the radio transmitter the modulating wave is the audio-frequency program signal to be communicated; the modulated wave that is broadcast is a radio-frequency, amplitude-modulated sinusoid. *See* RADIO BROADCASTING.

In AM the modulated wave is composed of the transmitted carrier, which conveys no information, plus the upper and lower sidebands, which (assuming the carrier frequency exceeds twice the top audio frequency) convey identical and therefore mutually redundant information. J. R. Carson in 1915 was the first to recognize that, under these conditions and assuming adequate knowledge of the carrier, either sideband alone would uniquely define the message. Apart from a scale factor, the spectrum of the upper sideband and lower sideband is the spectrum of the modulating wave displaced, respectively, without and with inversion by an amount equal to the carrier frequency.

For example, suppose the audio-frequency signal is a single-frequency tone, such as 1000 Hz, and the carrier frequency is 1,000,000 Hz; then the lower and upper sidebands will be a pair of single-frequency waves. The lower-sideband frequency will be 999,000 Hz, corresponding to the difference between the carrier and audio-signal frequencies; and the upper-sideband frequency will be 1,001,000 Hz, corresponding to the sum of the carrier and audio-signal frequencies. The amplitude of the signal appears in the amplitude of the sidebands. In practice, the modulating waveform will be more complex. A typical program signal might occupy a frequency band range of perhaps 100 Hz to 5000 Hz.

An important characteristic of AM, as illustrated by **Fig. 1**, is that, apart from a scale factor and constant term, either the upper or lower envelope of the modulated wave is an exact replica of the modulating wave, provided two conditions are satisfied: first, that the carrier frequency exceeds twice the highest speech frequency to be transmitted; and second, that the carrier is not overmodulated.

Single-sideband (SSB) modulation. This is a form of linear modulation that takes advantage of the fact that in an AM signal the carrier conveys no information and each of the two sidebands (that is, the upper sideband and the lower sideband) contains the same information. Hence it is not necessary to transmit both of them, and it is possible to achieve a 50% savings in bandwidth relative to normal AM or double-sideband (DSB) modulation.

The price for this bandwidth efficiency is increased complexity. The generation of SSB is typically accomplished either by filtering out the unwanted sideband of a DSB signal or by using a phase-shift modulator. This latter system involves use of a Hilbert

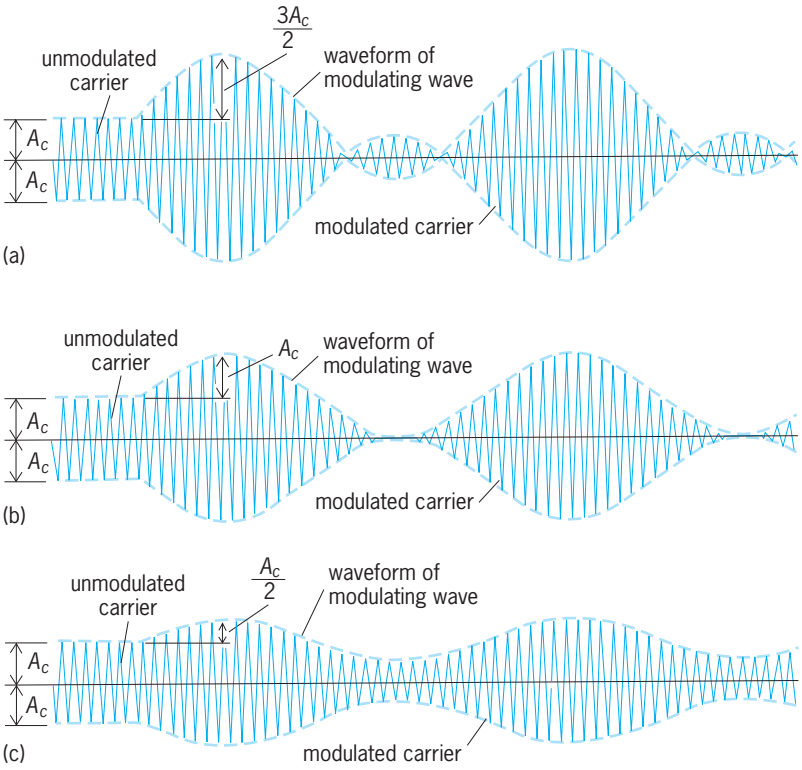


Fig. 1. Amplitude modulation of a sine-wave carrier by a sine-wave signal, with (*a*) 50% overmodulation, (*b*) 100% modulation, and (*c*) 50% modulation. (*After H. S. Black, Modulation Theory, Van Nostrand, 1953*)

transformer to shift the phase of the frequency components of the message by -90°. Specifically, if the message is denoted by $m(t)$, the transmitted waveform $s(t)$ is given by Eq. (1), where ω_0 is the angular

$$s(t) = m(t)\cos\omega_0 t \pm \hat{m}(t)\sin\omega_0 t \qquad (1)$$

frequency of the carrier and $\hat{m}(t)$ is the Hilbert transform of $m(t)$. If the plus sign is used in the above equation, the lower sideband is retained, whereas if the minus sign is used, the upper sideband is retained.

Either a coherent demodulator or a carrier reinsertion receiver can be used to demodulate an SSB signal. In either case, there are certain limits on phase and frequency errors that can be tolerated. Also of concern in SSB transmission are linearity requirements, since the envelope of an SSB signal can undergo significant variations.

SSB signaling is extensively used in voice transmission. Indeed, for applications such as telephone transmission over satellite links, one of the most common forms of modulation is the use of SSB subcarriers, which in turn frequency-modulate a carrier. The technique works as follows: Each voice message occupies roughly a 3-kHz bandwidth. The various voice messages are SSB-modulated onto individual subcarriers spaced 4 kHz apart. This composite baseband signal then frequency-modulates a carrier which is transmitted over the radio channel.

SSB modulation is also used for voice transmission in radio relay systems and coaxial cable systems. Indeed, this type of transmission has become so well accepted on a worldwide basis that standard frequency allocation plans have been set up by the Consultative Committee in International Telegraphy and Telephony (CCITT).

The choice of whether or not to use SSB depends upon the particular conditions under which the system is operating. SSB requires both less power and less bandwidth than does AM, but requires a more complex transmitter and receiver. Also, SSB systems have been shown to be less sensitive to the effects of selective fading than are both AM and DSB. *See* SINGLE SIDEBAND.

Vestigial-sideband (VSB) modulation. This is modulation whereby in effect the modulated wave to be transmitted is composed of one sideband plus a portion of the other adjoining the carrier. The carrier may or may not be transmitted. VSB is like SSB except in a restricted region around the carrier. The overall frequency response to the wanted sideband and to the vestigial sideband is so proportioned by networks that upon demodulation, preferably but not necessarily by a product demodulator, the original modulating wave will be recovered with adequate accuracy.

By thus transmitting a linearly distorted copy of both sidebands in a restricted frequency region above and below the carrier, the original modulating wave can now be permitted to contain significant components at extremely low frequencies, even approaching zero in the limit. By this means, at the cost of a modest increase in bandwidth occupancy, network requirements are greatly simplified. Furthermore, the low-frequency limitation and the inherent delay associated with SSB are avoided.

In standard television broadcasting in the continental United States the carrier is transmitted, envelope detection is used, and the vestigial sideband possesses a bandwidth one-sixth that of a full sideband. *See* TELEVISION.

Comparison of modulation spectra. In order to better understand the differences between AM, DSB, SSB, and VSB, it is helpful to consider the resulting spectrum of each waveform when the signal given by Eq. (2) is used as the modulation, where A and B

$$m(t) = A\cos\omega_1 t + B\cos\omega_2 t \qquad (2)$$

are constant amplitudes. **Figure 2***a* shows the single-sided spectrum of the modulation itself, where $f_1 = \omega_1/2\pi$ and $f_2 = \omega_2/2\pi$; and Fig. 2*b*–*e* shows the spectra of the AM, DSB, SSB, and VSB waveforms, respectively, when modulated by $m(t)$. In Fig. 2*b*, f_0 is the carrier frequency and a is the modulation index. In Fig. 2*e*, ϵ represents the fraction of the lower sideband at frequency $f_0 - f_1$ which is retained, that is, that is, it represents a vestige of the lower sideband of the DSB waveform of Fig. 2*c*.

Uses of AM in multiplexing. Multiplexing is the process of transmitting a number of independent messages over a common medium simultaneously. To multiplex, it is necessary to modulate. Two or more communicating channels sharing a common propagation path may be multiplexed by arranging them

along the frequency scale as in frequency division, by arranging them in time sequence as in time division, or by a process known as phase discrimination, in which there need be no separation of channels in either frequency or time.

Frequency-division multiplexing. When communication channels are multiplexed by frequency division, a different frequency band is allocated to each channel. Single-sideband carrier telephone systems are a good example. At the sending end, the spectrum of each channel input is translated by SSB to a different frequency band. For example, speech signals occupying a band of 300–3300 Hz might be translated to occupy a band range of 12,300–15,300 Hz corresponding to the upper sideband of a sinusoidal carrier, the frequency of which is 12,000 Hz. Another message channel might be transmitted as the upper sideband of a different carrier, the frequency of which might be 16,000 Hz.

At the receiving end, individual channels are separated by electric networks called filters, and each original message is recovered by demodulation. The modulated wave produced by SSB at the sending end becomes the modulating wave applied to the receiving demodulator.

When communication channels are multiplexed by frequency division, all channels may be busy simultaneously and continuously, but each uses only its allocated fraction of the total available frequency range.

Time-division multiplexing. When communication channels are multiplexed by time division, a number of messages is propagated over a common transmitting medium by allocating different time intervals in sequence for the transmission of each message. **Figure 3** depicts a particularly simple example of a two-channel, time-division system. Ordinarily, the number of channels to be multiplexed would be considerably greater. Transmitting and receiving switches must be synchronized; time is of the essence in this system, and the problem is to switch at the right time.

On the theoretical side there is a certain basic, fundamental question which must always be answered about any time-division system. The question is: At what rate must each communication channel be connected to its common transmitting path? Today it is known from the sampling principle that for successful communication each channel must be momentarily connected to the common path at a rate that is in excess of twice the highest message frequency conveyed by that channel. *See* PULSE MODULATION.

Viewed broadly, amplitude modulation of pulse carriers generates the desired amplitude and phase relationships essential to time-division multiplexing. In addition, whereas each communication channel may use the entire available frequency band for transmitting its message, it may transmit only during its allocated fraction of the total time.

Phase-discrimination multiplexing. This type of multiplexing, like SSB, saves bandwidth, may save signal power, and, like AM and VSB, has the important advantage of freely transmitting extremely low modulating frequencies. Furthermore, each communication channel may utilize all of the available frequency range all of the time.

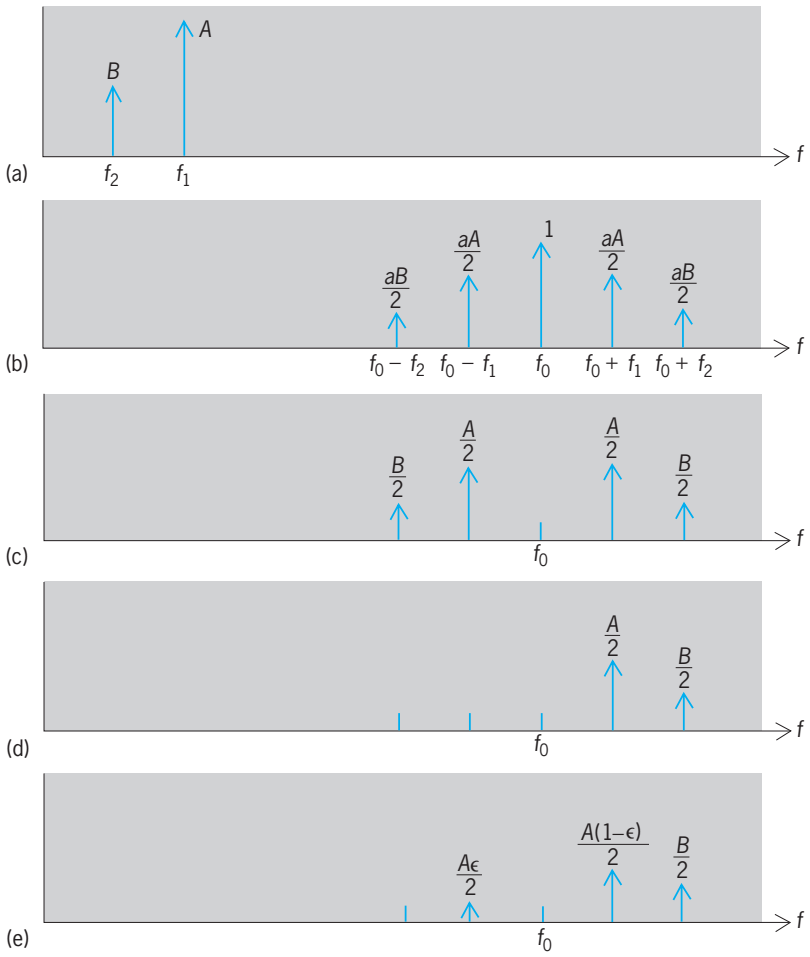


Fig. 2. Comparison of spectra of modulation waveforms. Signal amplitude is plotted as a function of frequency *f*. (*a*) Spectrum of modulation itself. (*b*) Normal amplitude modulation (AM). (*c*) Double-sideband (DSB). (*d*) Single-sideband (SSB). (*e*) Vestigial-sideband (VSB).

When n channels are multiplexed by phase discrimination, the modulating wave associated with each channel simultaneously amplitude-modulates $n/2$ carriers, with a different set of carrier phases provided for each channel. All sidebands are transmitted; $n/2$ carriers may or may not be transmitted. At the receiving end, with the aid of locally supplied carriers and an appropriate demodulation process, the n channels can be separated, assuming distortionless transmission, ideal modulators, and so on. Systems with odd numbers of channels can also be devised. Equality of sidebands and their exact phases account for the suppression of interchannel interference.

Day's system. This is a simple example of phase-discrimination multiplexing. Two sine-wave carriers of the same frequency but differing in phase by 90° are amplitude-modulated, each by a different message wave. The spectrum of each modulated sinusoid occupies the same frequency band. These modulated sinusoids are then added and propagated without distortion to a pair of demodulators. Quadrature carriers of correct frequency and phase are applied

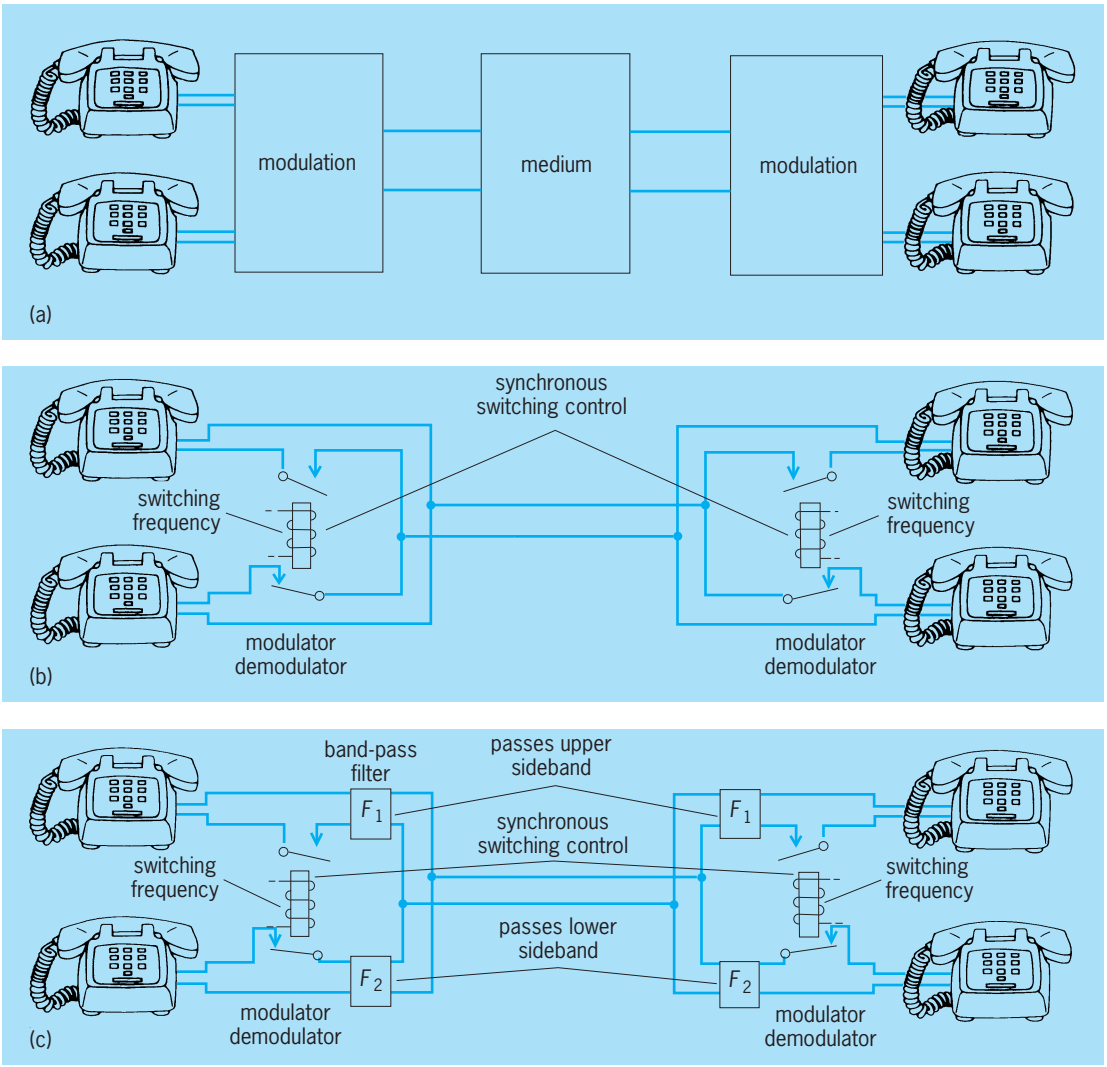


Fig. 3. Two-channel, time-division carrier system and two-channel, frequency-division carrier system. (*a*) General diagram of two-channel carrier system. (*b*) Amplitude modulation, time division. (*c*) Single-sideband modulation, frequency division. (*After H. S. Black, Modulation Theory, Van Nostrand, 1953*)

locally, one to each demodulator. Theoretically, a faithful copy of each message can be recovered.

Within the continental United States, for purposes of saving bandwidth, Day's system is used for multiplexing the two so-called color components associated with color television broadcasting. *See* MULTIPLEXING AND MULTIPLE ACCESS.

Modulator and demodulator. Many methods of modulating and demodulating are possible, and many kinds of modulators and demodulators are available for each method. *See* MODULATOR.

Fundamentally, since the sidebands of an amplitude-modulated sinusoid are generated by a multiplication of wave components which produces frequency components corresponding to the products, it is natural to envisage a product modulator having an output proportional to the product of two inputs: modulating wave and carrier. An ideal product modulator suppresses both modulating wave and carrier, transmitting only upper and lower sidebands. *See* AMPLITUDE MODULATOR.

At the receiving end, the original message may be recovered from either sideband or from both sidebands by a repetition of the original modulating process using a product modulator, commonly referred to as a product demodulator, followed by a low-pass filter. Perfect recovery requires a locally applied demodulator carrier of the correct frequency and phase. For example, a reduced carrier system creates its correct demodulator carrier supply by transmitting only enough carrier to control the frequency and phase of a strong, locally generated carrier at the receiver. *See* AMPLITUDE-MODULATION DETECTOR.

SSB systems commonly generate their modulator and demodulator carriers locally with independent oscillators. For the high-quality reproduction of music, the permissible frequency difference between modulator and demodulator carriers associated with a particular channel is limited to about 1–2 Hz. For monaural telephony, frequency differences approaching 10 Hz are permissible.

Unbalanced square-law demodulators, rectifier-type demodulators, and envelope detectors are often used to demodulate AM. However, even though

overmodulation (Fig. 1) is avoided, significant distortion may be introduced. In general, the amount of distortion introduced in this manner will depend upon the kind of demodulator or detector used, the amount of noise and distortion introduced prior to reception, and the percentage modulation. Harold S. Black; Laurence B. Milstein

Bibliography. L. W. Couch, *Digital and Analog Communication Systems*, 5th ed., 1996; M. Schwartz, *Information, Transmission, Modulation and Noise*, 4th ed., 1990; H. Stark, F. B. Tuteur, and J. B. Anderson, *Modern Electrical Communications*, 2d ed., 1988; F. G. Stremlers, *Introduction to Communication Systems*, 3d ed., 1990; H. Taub, *Principles of Communication Systems*, 2d ed., 1986; R. E. Ziemer and W. H. Tranter, *Principles of Communications*, 4th ed., 1994.

Amplitude-modulation detector

A device for recovering information from an amplitude-modulated (AM) electrical signal. Such a signal is received, usually at radio frequency, with information impressed in one of several forms. The carrier signal may be modulated by the information signal as double-sideband (DSB) suppressed-carrier (DSSC or DSBSC), double-sideband transmitted-carrier (DSTC or DSBTC), single-sideband suppressed-carrier (SSBSC or SSB), vestigial sideband (VSB), or quadrature-amplitude modulated (QAM).

The field of amplitude-modulation detector requirements splits by application, complexity, and cost into the two categories of synchronous and asynchronous detection. Analog implementation of nonlinear asynchronous detection, which is typically carried out with a diode circuit, is favored for consumer applications, AM-broadcast radio receivers, minimum-cost products, and less critical performance requirements. Synchronous detectors, in which the received signal is multiplied by a replica of the carrier signal, are implemented directly according to their mathematics and block diagrams, and the same general detector satisfies the detection requirements of all SSB, DSB, and VSB signals. Although synchronous detectors may operate in the analog domain by using integrated circuits, more commonly digital circuits are used because of cost, performance, reliability, and power advantages. A quadrature-amplitude-modulation detector is simply a two-channel version of a synchronous detector.

Synchronous detection. The idea behind synchronous detection is simple. There are two conceptual approaches: to reverse the modulation process (which is rather difficult), or to remodulate the signal from the passband (at or near the transmitter's carrier frequency) to the baseband (centered at dc or zero frequency). The remodulation approach is routine. Unfortunately, a nearly exact replica of the transmitter's carrier signal is needed at the receiver in order to synchronously demodulate a transmitted signal. Synchronous means that the carrier-signal reference in the receiver has the same frequency and phase as the carrier signal in the transmitter. There are three means available to obtain a carrier-signal reference. First, the carrier signal may actually be available via a second channel. There are no difficulties with this method because a perfect replica is in hand. Second, the carrier signal may be transmitted with the modulated signal. It then must be recovered by a circuit known as a phase-lock loop with potential phase and frequency errors. Third, the carrier signal may be synthesized by a local oscillator at the receiver, with great potential for errors. Unless otherwise stated, it will be assumed that a perfect replica of the carrier signal is available. *See* OSCILLATOR; PHASE-LOCKED LOOPS.

SSB, VSB, and DSB signals. Demodulation of an SSB-modulated (upper or lower), VSB-modulated, or DSB-modulated signal is as simple as multiplying it by the transmitter carrier-signal replica and low-pass filtering to remove unwanted high-frequency terms (sometimes called images) from the product in order to obtain the base-band signal (**Fig. 1***a*). Sometimes the multiplication is carried out in a single step, sometimes in two steps called conversion stages. If the type of transmitted-signal modulation is transmitted carrier (TC), the consequent dc component must be removed by a filter. If the information signal itself has a dc component, an amplitude-modulated TC method cannot be used to transmit the data because of the difficulty in separating the dc component of the signal from the carrier which is shifted to dc by the demodulation process. *See* ELECTRIC FILTER.

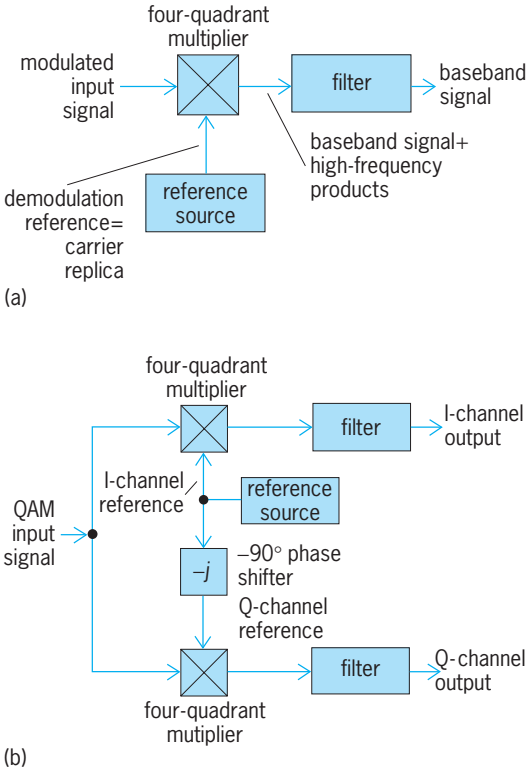


Fig. 1. Synchronous (coherent) detectors for (*a*) double-sideband (DSB), single-sideband (SSB), and vestigial-sideband (VSB) signals, and (*b*) for quadrature-amplitude-modulated (QAM) signals.

If the ith information component in a modulated signal is $A_i \cos(w_i t)$ and the modulating carrier is $\cos(w_c t)$, then the ith modulated element is given by Eq. (1). The ideal local oscillator signal is $y(t) =$

$$x(t) = A_i \cos(w_i t) \cos(w_c t) \quad (1)$$

$2 \cos(w_c t)$, so the demodulated signal is given by Eq. (2). The last term is removed by the filter, so that

$$\begin{aligned} z(t) = x(t)y(t) &= 2A_i \cos(w_i t) \cos^2(w_c t) \\ &= 2(A_i/2) \cos(w_i t)[1 + \cos(2w_c t)] \\ &= A_i \cos(w_i t) + [A_i \cos(w_i t) \cos(2w_c t)] \end{aligned} \quad (2)$$

the information signal is completely recovered. *See* TRIGONOMETRY.

The center frequency of the signal is often shifted down in two conversion stages. First an analog conversion by an oscillator translates the frequency of the signal to a value convenient for the analog-to-digital converter to operate; then the four-quadrant digital multiplier serves as a second converter and completes the frequency translation. The increasing performance and decreasing costs of analog-to-digital converters favor detectors in which analog conversion is eliminated and the analog-to-digital converter operates directly on the higher-frequency signal.

The importance of the accuracy of the reference in synchronous demodulation cannot be overemphasized. An average frequency error in the demodulating reference is intolerable because it will translate the modulated signals to the wrong frequencies. The cosine of the phase angle in the demodulating reference is the demodulator gain. A phase error can therefore seriously reduce the gain. A small frequency error integrates with time to become a phase error, causing objectionable periodic amplitude modulation of the received signal.

However, it is not necessary for the reference waveform to be a pure sinusoid, as long as it does have the correct periodicity and phase. An appeal to Fourier theory shows that this reference waveform can be represented by a sum of sinusoids harmonically related to the fundamental frequency of the reference. Multiplying the signal to be detected by this nonsinusoidal reference causes so-called images to appear after the multiplication process. These images, which exist at multiples of the fundamental frequency, can be removed by suitably filtering the multiplier output signal. *See* FOURIER SERIES AND TRANSFORMS; NONSINUSOIDAL WAVEFORM.

QAM signals. A synchronous detector of quadrature-amplitude-modulated signals consists of two synchronous demodulators operating in parallel (Fig. 1*b*). The two demodulating carrier-reference signals are separated in phase by precisely 90°. The accuracy of the phase angle of the carrier-reference signal limits the degree of separation (after demodulation) that can be maintained between the two QAM signals that have been transmitted on top of one another. If the ith transmitted signal element is given by Eq. (3), and the I and Q demodulation carrier references are given by Eqs. (4) and (5), then the filtered

$$x(t) = A_i \cos(w_i t) \cos(w_c t) + B_i \cos(w_i t) \sin(w_c t) \quad (3)$$

$$\begin{aligned} y_I(t) &= 2 \cos(w_c t + \Theta) \\ &= 2[\cos(w_c t) \cos(\Theta) - \sin(w_c t) \sin(\Theta)] \end{aligned} \quad (4)$$

$$\begin{aligned} y_Q(t) &= 2 \sin(w_c t + \Theta) \\ &= 2[\sin(w_c t) \cos(\Theta) + \cos(w_c t) \sin(\Theta)] \end{aligned} \quad (5)$$

(that is, the double-frequency products have been

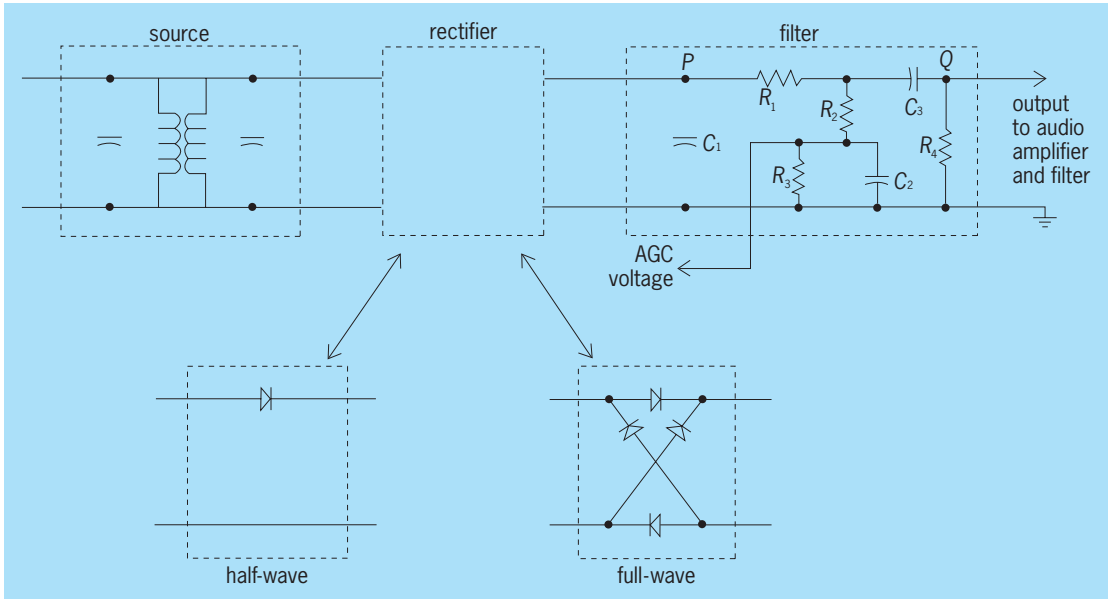


Fig. 2. Diode detector circuit. *P* = input point of filter; *Q* = output point of filter.

removed) demodulated outputs are given by Eqs. (6) and (7). Separation is complete for $\Theta = 0$.

$$z_I(t) = A_i \cos(w_i t) \cos(\Theta) - B_i \cos(w_i t) \sin(\Theta) \quad (6)$$

$$z_Q(t) = B_i \cos(w_i t) \cos(\Theta) + A_i \cos(w_i t) \sin(\Theta) \quad (7)$$

Asynchronous detection. Asynchronous detection applies to DSTC signals whose modulation index is less than 1, and is quite simple. First the received signal is full-wave rectified, then the result is low-pass filtered to eliminate the carrier frequency and its products, and finally the average value is removed (by dc blocking, that is, ac coupling). This result is identical to that which is obtained by synchronous detection with a reference that has been amplitude distorted to a square wave. The cheaper but less efficient half-wave rectifier can also be used, in which case the demodulation process is called envelope detection.

Diode detector. A diode detector circuit in a radio receiver has three stages (**Fig. 2**). The first stage is the signal source, which consists of a pair of tuned circuits that represent the last intermediate-frequency (i.f.) transformer which couples the signal energy out of the i.f.-amplifier stage into the detector. The second stage is a diode rectifier, which may be either full wave or half wave. Finally, the signal is passed through the third stage, a filter, to smooth high-frequency noise artifacts and remove the average value. *See* DIODE; INTERMEDIATE-FREQUENCY AMPLIFIER; RADIO RECEIVER; RECTIFIER.

The waveform shaping at the input point of the filter (Fig. 2) is determined by a capacitor C_1 in parallel with an equivalent resistance of R_1, the latter in series with a parallel combination of resistors, R_2 and R_4. The filter also has a capacitor C_3 between R_2 and R_4, and a parallel combination of a resistor R_3 and a capacitor C_2 in series with R_2. The reactances of both capacitors C_2 and C_3 are quite small at the information frequency, so the capacitors can be viewed as short circuits. Meanwhile, the C_3–R_4

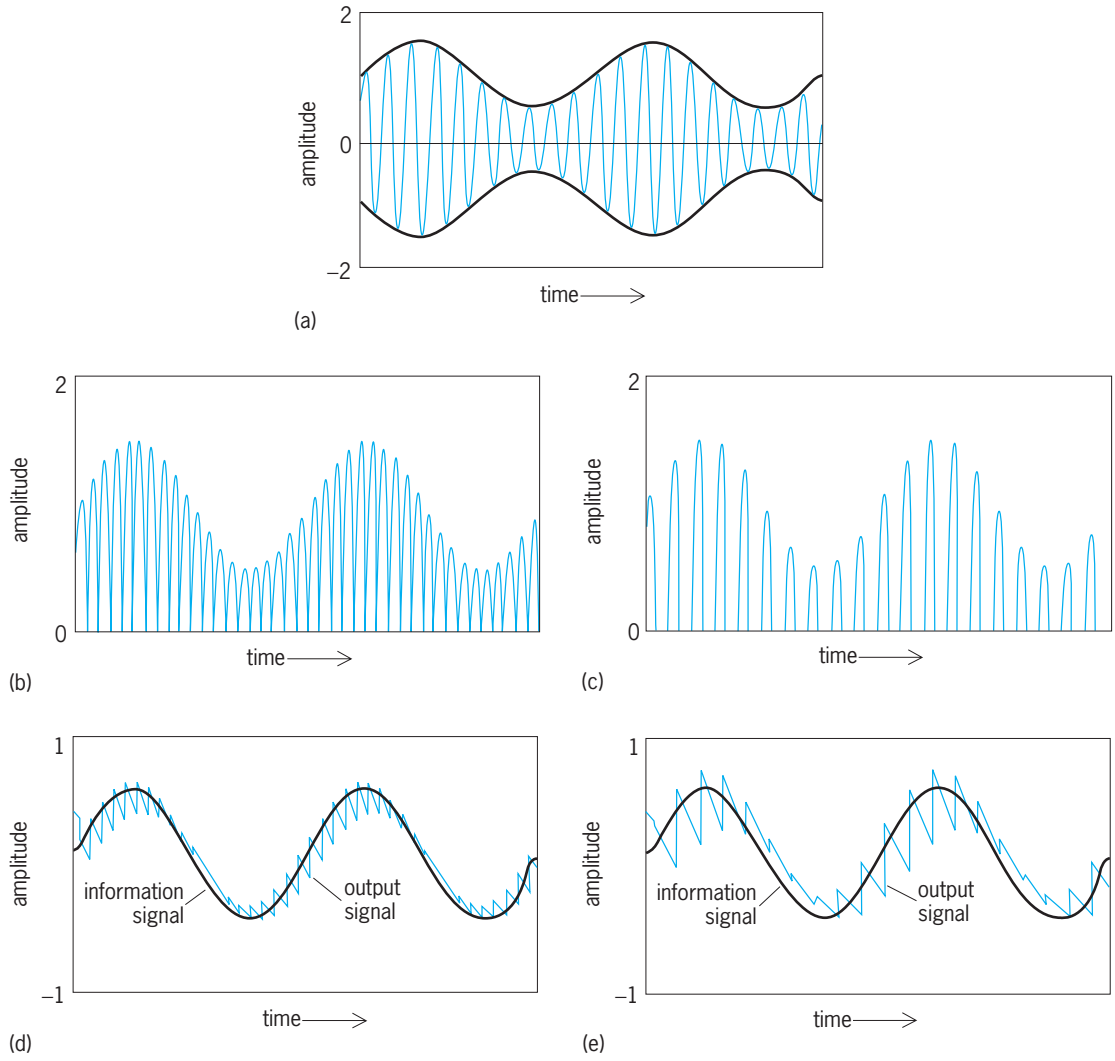


Fig. 3. Detection of a 50%-modulated double-sideband transmitted-carrier (DSTC) signal. An artificially low frequency ratio of 10:1 is used to illustrate raggedness of the output. (*a*) Received signal and modulating information-signal envelope. (*b*) Full-wave rectified received signal. (*c*) Half-wave rectified signal. (*d*) Information signal and output signal after full-wave rectifier and filter. (*e*) Information signal and output signal after half-wave rectifier and filter.

combination serves as a dc-blocking circuit to eliminate the constant or dc component at the output point of the filter. In order to bias the output point properly for the next amplifier stage, R_4 is replaced by a biasing resistor network in a real filter; R_4 is the single-resistor equivalent. *See* ELECTRICAL IMPEDANCE.

The strength of the signal arriving at the detector (**Fig. 3***a*) is proportional to the mean value of the signal at the input point of the filter and is sensed as the automatic-gain-control (AGC) voltage. Changes in this voltage level are used to adjust the amplification before the detector so that the signal strength at the detector input can remain relatively constant, although the signal strength at the receiver's antenna may fluctuate. Since capacitor C_2 shunts all signal energy and any surviving carrier energy to ground, the AGC voltage is roughly the average value at the input point of the filter scaled by $R_3/(R_1 + R_2 + R_3)$ because R_1 is much smaller than R_4. *See* AUTOMATIC GAIN CONTROL (AGC).

Waveforms. The received signal (Fig. 3*a*) is delivered from the i.f. stage to the diode rectifier. If the filter were simply a resistor, then the output from the diode rectifier would simply be the full-wave rectification (Fig. 3*b*) or half-wave rectification (Fig. 3*c*) of the received signal. (The presence of the capacitors in the filter provides the waveform shaping.) In both cases the modulating (information) signal shapes the envelopes of the rectified waveforms. In an amplitude-modulation radio, the output from the filter after full-wave rectification (Fig. 3*d*) would be the audio signal (plus some noise) that is delivered to the audio amplifier. Additional filtering can be provided as necessary to reduce the noise to an acceptable level. This amplified and filtered signal is finally delivered to an output device, such as a loudspeaker. The ragged waveform of the filter output contrasts with the smooth waveform of the information signal (Fig. 3*d*). The raggedness vanishes as the ratio of the i.f. frequency to the information frequency increases. While a synthetic example with a low ratio can be used to clearly show the effects within the demodulator (Fig. 3*d*), the amplitude of this raggedness noise decreases in almost direct proportion to the increase in the frequency ratios. The raggedness of the output signal from the filter after half-wave rectification (Fig. 3*e*) is much greater than that of the full-wave-rectifier case.

The actual worst-case ratio for standard-broadcast amplitude-modulation radio is 46.5:1. In this case the raggedness on the filtered outputs is reduced to a fuzz that can be seen in graphs of the waveforms but is well outside the frequency range of audio circuits, loudspeakers, and human hearing. *See* AMPLITUDE MODULATION; AMPLITUDE MODULATOR; MODULATION; WAVEFORM.

Stanley A. White

Bibliography. R. C. Dorf (ed.), *The Electrical Engineering Handbook*, 2d ed., 1997; M. S. Roden, *Analog and Digital Communication Systems*, 3d ed., 1991; M. Valkenburg (ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 8th ed., 1996.

Amplitude-modulation radio

Radio communication employing amplitude modulation of a radio-frequency carrier wave as the means of conveying the desired intelligence. In amplitude modulation the amplitude of the carrier wave is made to vary in response to the fluctuations of a sound wave, television image, or other information to be conveyed. *See* AMPLITUDE MODULATION; RADIO.

Amplitude modulation (AM), the oldest and simplest form of modulation, is widely used for radio services. The most familiar of these is broadcasting; others include radiotelephony and radiotelegraphy, television picture transmission, and navigational aids. The essentials of these radio systems are discussed in this article.

Low frequency (long wave). European and Asian countries use frequencies in the range 150–255 kHz for some broadcast services. An advantage of these frequencies is stable and relatively low-attenuation wave propagation. When not limited by atmospheric noise, large areas may be served by one station. In the United States these frequencies are reserved for navigational systems and so are not available for broadcasting. *See* RADIO SPECTRUM ALLOCATION; RADIOWAVE PROPAGATION.

Low-frequency (lf) broadcast antennas are omnidirective and radiate vertically polarized waves. Unless special means are used to reduce antenna selectivity, the highest program frequencies transmitted are substantially below 10,000 Hz.

Medium frequency. The frequencies in the range from 535 to 1705 kHz are reserved for AM (standard) broadcasting. In the Western Hemisphere this band is divided into channels at 10-kHz intervals, certain channels being known as clear, regional, and local, according to the licensed coverage and class of service. The local channels are occupied by stations, usually of 250-W output, servicing smaller localities. Many stations occupy a channel, but they are situated far enough apart to permit interference-free coverage in the local area. Fewer stations of higher power, but with greater distances between them, share the regional channels. A few clear channels are occupied by high-power stations (50,000-W maximum output in the United States). These stations may have exclusive use of a channel, or may share it with another distant station. *See* RADIO BROADCASTING.

Interference between co-channel regional stations and clear-channel stations is minimized by use of directive antennas, which suppress radiation toward other stations and direct it to main populated areas from a properly located station.

European medium-frequency (mf) broadcasting channels are assigned at 9-kHz intervals rather than the 10-kHz intervals used in the Western Hemisphere. This reduced spacing provides more channels within the mf band. The technique of directive antennas, which also provides more channels within a band by minimizing interference between stations, has not been used extensively in Europe.

Vertically polarized radiation is used at medium and low frequencies propagated over the Earth's

surface. There is also propagation of high-angle radiation via reflection from the ionosphere, a phenomenon that predominates at night but is relatively absent during daylight. This sky-wave propagation accounts for the familiar long-distance reception at night. At distances where the downcoming sky waves overlap the ground wave, fading and distortion of the signal occurs. Receivers in the groundwave zone get stable signals day and night. In the overlap zone, daylight reception may be stable but night reception disturbed by fading. In the sky-wave zone at night, assuming no interference from other stations, satisfactory service may be received over long distances. Reception in this zone depends on atmospheric noise, which varies seasonally, and the state of the ionosphere, which varies greatly from night to night depending upon astronomical conditions that affect the upper atmosphere.

Individual AM broadcast stations transmit program frequencies ranging from 30 to 10,000 Hz with excellent fidelity. To obtain suitable tuning selectivity between channels, AM broadcast receivers may reproduce less extensive program frequencies, according to make, cost, type, and condition of the receiver.

High frequency (shortwave). Small bands of frequencies between 3000 and 7500 kHz are used in tropical areas of high atmospheric noise for regional broadcasting. This takes advantage of the lower atmospheric noise at these frequencies and permits service under conditions where medium frequencies have only severely limited coverage. Wave propagation day and night is by sky wave. Groundwave coverage from such stations is usually negligible, since high-angle horizontally polarized radiation is used. Short-distance coverage by this mode is by ionospheric reflection of waves radiated almost vertically.

Long-distance international broadcasting uses high-power transmitters and directive (beam) antennas operating in the bands 5950–6200 kHz, 9500–9775 kHz, 11,700–11,975 kHz, 15,100–15,450 kHz, 17,700–17,900 kHz, 21,450–21,750 kHz, and 25,600–26,100 kHz. These bands are allocated throughout the world for this purpose, and the band used for any path depends upon ionospheric conditions which vary with direction, distance, hour, season, and the phase of the 11-year sunspot cycle. Typically, waves are propagated in low-angle beams by multiple reflections between ionosphere and Earth to cover transoceanic distances, and signals are often distorted during transmission. These bands are so crowded that a signal is seldom received without interference for any prolonged period. Reception from particular stations can be improved by the use of special directive receiving antennas. Propagation partially through the ionosphere also causes rotation of the transmission's plane of polarization due to the Faraday effect. *See* FARADAY EFFECT.

The technical performance of high-frequency (hf) broadcast transmission systems is usually to the same standards employed for mf broadcasting, although propagation and interference conditions seldom make this evident to a listener.

AM telephony and telegraphy. The first radiotelephony was by means of amplitude modulation, and its use has continued with increasing importance. Radiotelephony refers to two-way voice communication. Amplitude modulation and a modified form called single-sideband are used almost exclusively for radiotelephony on frequencies below 30 MHz. Above 30 MHz, frequency or phase modulation is used almost exclusively, a notable exception being 118–132 MHz, where amplitude modulation is used for all two-way very high frequency (vhf) radiotelephony in aviation operations.

The least expensive method known for communicating by telephony over distances longer than a few tens of miles is by using the high frequencies of 3–30 MHz. Furthermore, since radio is the only way to communicate with ships and aircraft, hf AM radiotelephony has remained essential to these operations, except for short distances that can be covered from land stations using the very high frequencies. Therefore, hf radiotelephony has become established for a great variety of pioneering and exploration enterprises and private, public, and government services needing telephone communication, fixed or mobile, over substantial distances where there are no other ways to telephone. Because AM techniques are simple and inexpensive, this form of modulation has predominated. The economic development of distant hinterland areas depends greatly on hf AM radio telephony to the outside world, although satellite transmission and reception has generally replaced hf AM radio telephony for all purposes except broadcasting.

Widespread use has led to serious crowding of the hf band. All governments are suspending the use of the hf band wherever it is technically and economically feasible to employ frequencies above 30 MHz using either direct transmission or radio repeater stations. The trend to single-sideband modulation also alleviates the pressure of congestion in the hf band.

Many of the radiotelephone systems in use operate two ways on one frequency in simplex fashion, that is, all stations on the frequency are normally in a receiving status. Transmission, by press-to-talk (manual) or voice-operated carrier (automatic) switching from reception to transmission, occurs only while the sender is talking. Many stations can thus occupy one frequency provided the volume of traffic by each of the stations is small. Two-way telephony must be strictly sequential. This system is not adapted for connection to a normal two-wire telephone.

Full duplex radiotelephony, for interconnection with wire telephone systems, is essential for most public correspondence. This requires two frequencies, each available full time in one direction. Even so, typical fading of signals during propagation requires that voice-operated antiregeneration devices be used to maintain circuit stability. Talking must be sequential between speakers as there can be no interrupting, but the system will interconnect with conventional business or home telephones.

Amplitude-modulated telegraphy consists of

interrupting a carrier wave in accordance with the Morse dot-dash code or codes used for the printing telegraph. Much of the radiotelegraphic traffic of the world uses AM telegraphy, although there has been extensive conversion to frequency-shift (frequency-modulation) telegraphy since 1944, the latter being better adapted to automatic teleprinting operations. Radiotelegraph operations have been refined, speeded, and mechanized, but, under adverse noise and fading conditions, AM manual telegraphy between experienced operators is still more reliable. Most aviation and marine radiotelegraphy uses AM manual methods. *See* TELEGRAPHY; TELEPHONE SERVICE.

Single-sideband (SSB) hf telephony. This is a modified form of amplitude modulation in which only one of the modulation sidebands is transmitted. In some systems the carrier is transmitted at a low level to act as a pilot frequency for the regeneration of a replacement carrier at the receiver. Where other means are available for this purpose, the carrier may be completely suppressed. Where intercommunication between AM and SSB systems is desired, the full carrier may be transmitted with one sideband.

Since 1954 the use of SSB has expanded rapidly in replacing common AM telephony for military and many nonpublic radio services. In time, SSB will gradually displace AM radiotelephony to reduce serious interference due to overcrowding of the radio spectrum. SSB transmission also is less affected by selective fading in propagation.

Multiplexing, both multiple-voice channels or teleprinter channels included with voice, is applied to SSB transmission. Teleprinting and data transmission by frequency multiplex using SSB radiotelephone equipment is increasing and displacing older methods of radiotelegraphy for fixed point-to-point government and common-carrier services. *See* SINGLE SIDEBAND.

Aviation and marine navigation aids. Amplitude-modulated radio has a dominant role in guidance and position location, especially in aviation, which is almost wholly under radio guidance. Radio is used in traffic control from point to point, in holding a position in a traffic pattern, and in approach and landing at airports. Distance measuring from a known point, radio position markers, runway localizers, and glidepath directors all use AM in some form, if only for coded identification of a facility.

Marine operations are not so dependent on radio facilities as are those of aviation, but almost every oceangoing ship makes use of direction finding, at least to determine its bearing from a radio station of known location. Special marine coastal beacon stations emit identified signals solely for direction-finding purposes. Certain navigational systems (Decca, loran) for aviation are also used by ships for continuous position indication and guidance. *See* ELECTRONIC NAVIGATION SYSTEMS.

Television broadcasting. Amplitude modulation is used for the broadcasting of the picture (video) portion of television. In England, France, and a few other places amplitude modulation is also used for the sound channel associated with the television picture, but frequency modulation is more commonly used for sound.

Countries of the Western Hemisphere, Japan, Philippines, Thailand, and Iran broadcast television video in an emission band of 4.25 MHz; the English video bandwidth is 3 MHz; the French system, 10 MHz. The rest of continental Europe (except Russia) use a bandwidth of 5.25 MHz. The carrier frequencies employed are between 40 and 216 MHz, and 470 to 890 MHz. A channel allocation includes the spectrum needed for both sound and picture. Japan and Australia also use 88–108 MHz for television broadcasting. Reception of such transmissions is often impaired by multipath (the same signal arriving out-of-phase at the receiver from different propagation paths due to reflections by large objects and terrain), causing effects such as ghosting.

The English and French systems employ positive video modulation, in which white corresponds to higher amplitudes of the modulation envelope and black corresponds to lower amplitudes. All other established video broadcasting uses negative modulation, working in the opposite sense. Synchronizing pulses in the negative system are at maximum carrier amplitude. The dynamic range from black to white in the picture varies from 75 to 25% of maximum amplitude.

The upper-frequency portion of one video sideband is suppressed by filters so that its remaining vestige, together with the other complete sideband, is transmitted. This is called vestigial sideband transmission and avoids unnecessary spectrum usage. *See* TELEVISION TRANSMITTER.

Edmund A. Laport; Michael C. Rau

Stereophonic broadcasting. A number of systems for stereophonic AM radio broadcasting have existed since the late 1970s. These systems allow AM radio stations to transmit two channels of information in the same spectrum space where only one could exist previously. The result is similar to that of stereophonic FM, audio cassette, and other binaural entertainment media.

To transmit in AM stereo, a broadcast station employs a device known as an exciter to adapt its existing transmitter. The exciter has left- and right-channel audio inputs as well as summed audio- and radio-frequency outputs. Stereophonic audio program material is connected to the audio inputs, while the outputs attach to the AM transmitter. The summed audio (left channel plus right channel) information is used to amplitude-modulate the transmitter, to assure that compatibility with existing monophonic receivers is maintained. The radio-frequency exciter output, however, contains the stereophonic information to be transmitted. This output is connected in place of the transmitter's crystal oscillator, and although the signal present is at the exact carrier frequency of the radio station, it is also phase-modulated by the difference (left channel minus right channel) audio information. A subaudible and low-level identification tone is usually added to this difference audio information. This allows AM radio

receivers to identify stereophonic transmissions and activate the appropriate decoding circuitry, as well as a stereo indicator lamp commonly found on the radio front panel. *See* PHASE MODULATION; RADIO RECEIVER.

Stereophonic AM broadcasting methods all have their roots in a system known as quadrature multiplexing. This system allows the transmission of two channels of information on a single carrier frequency, but true quadrature transmissions are inherently incompatible with the majority of radios available to consumers. The differences between various systems employed to transmit stereophonic AM broadcasts all relate to the methods used to overcome this incompatibility. *See* AMPLITUDE MODULATION; RADIO; STEREOPHONIC RADIO TRANSMISSION.

Stanley Salek

Bibliography. National Association of Broadcasters, *NAB Guide to Advanced Television Service,* 2d ed., 1991; J. Whitaker (ed.), *NAB Engineering Handbook*, 9th ed., National Association of Broadcasters, 1999.

Amplitude modulator

A device for moving the frequency of an information signal, which is generally at baseband (such as an audio or instrumentation signal), to a higher frequency, by varying the amplitude of a mediating (carrier) signal. The motivation to modulate may be to shift the signal of interest from a frequency band (for example, the baseband, corresponding to zero frequency or dc) where electrical disturbances exist to another frequency band where the information signal will be subject to less electrical interference; to isolate the signal from shifts in the dc value, due to bias shifts with temperature or time of the characteristics of amplifiers, or other electronic circuits; or to prepare the information signal for transmission.

Familiar applications of amplitude modulators are standard-broadcast or amplitude-modulation (AM) radio; data modems (modem = modulator + demodulator); and remote sensing, where the information signal detected by a remote sensor is modulated by the remote signal-conditioning circuitry for transmission to an electrically quiet data-processing location. *See* AMPLITUDE-MODULATION RADIO; MODEM; REMOTE SENSING.

The primary divisions among amplitude modulators are double sideband (DSB); single sideband (SSB); vestigial sideband (VSB); and quadrature amplitude (QAM), where two DSB signals share the same frequency and time slots simultaneously. Each of these schemes can be additionally tagged as suppressed carrier (SC) or transmitted carrier (TC).

To amplitude-modulate an information signal is (in its simplest form) to multiply it by a second signal, known as the carrier signal (because it then carries the information). A real (as opposed to complex) information signal at baseband, or one whose spectrum is centered about zero frequency, has an amplitude spectrum which is symmetric (an even function; **Fig. 1***a*) and a phase spectrum which is asymmetric (an odd function; Fig. 1*b*) in frequency about zero. Because of this symmetry, the information in the upper or positive sideband (positive frequencies) replicates the information in the lower or negative sideband (negative frequencies). Balanced modulation or linear multiplication (that is, four-quadrant multiplication where each of the two inputs is free to take on positive or negative values without disturbing the validity of the product) of this information signal by a sinusoidal carrier of single frequency, f_c, moves the information signal from baseband and replicates it about the carrier frequencies f_c and $-f_c$ (Fig. 1*c*). *See* FOURIER SERIES AND TRANSFORMS.

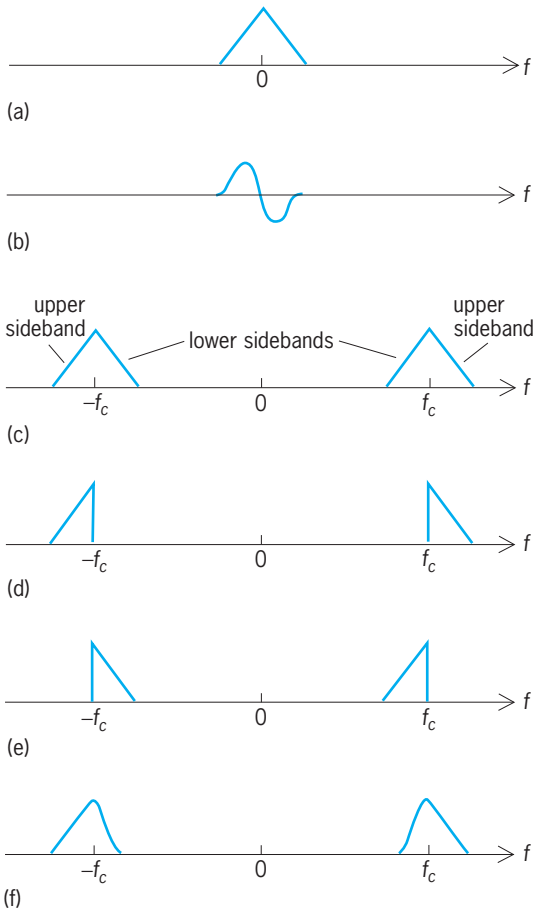


Fig. 1. Baseband spectra and spectra of outputs from suppressed-carrier modulators. Amplitudes and phases are shown as functions of frequency *f*. (*a*) Baseband amplitude spectrum. (*b*) Baseband phase spectrum. (*c*) Amplitude spectrum of output from double-sideband (DSB) modulator. (*d*) Amplitude spectrum of output from upper single-sideband (SSB) modulator. (*e*) Amplitude spectrum of output from lower single-sideband (SSB) modulator. (*f*) Amplitude spectrum of output from vestigial-sideband (VSB) modulator.

Suppressed-carrier modulators. Linear multiplication (described above) produces double-sideband suppressed-carrier (DSSC or DSBSC) modulation of the carrier by the information. One of the redundant sidebands of information may be eliminated by filtering (which can be difficult and costly to do adequately) or by phase cancellation (which is usually a much more reasonable process) to produce the more

efficient single-sideband suppressed-carrier (SSBSC or SSB) modulation (Fig. 1*d, e*). The process of only partially removing the redundant sideband (usually by deliberately imperfect filtering) and leaving only a vestige of it is called vestigial-sideband (VSB) modulation (Fig. 1*f*). *See* ELECTRIC FILTER.

The DSB and SSB generation mechanisms may be illustrated by considering one sinusoid, $A_i \cos (w_i t)$, of a weighted sum of sinusoids that may be assumed to constitute the information signal, $u(t)$. The ith information signal component can be multiplied by the carrier signal, $\cos (w_c t)$, to produce the DSSC product, $x(t)$, as in Eq. (1). This expres-

$$x(t) = A_i \cos (w_i t) \cos (w_c t) = \frac{A_1}{2} \{\cos [(w_c - w_i)t] + \cos [(w_c + w_i)t]\} \quad (1)$$

sion clearly shows the generation of the lower-sideband-frequency signal component at $w_c - w_i$ and the upper-sideband-frequency signal component at $w_c + w_i$. *See* TRIGONOMETRY.

A pair of phase shifters may now be introduced to operate on both the carrier and information signals to produce $\sin (w_c t)$ and $A_i \sin (w_i t)$, respectively. The product of this new signal pair is $y(t)$, given by Eq. (2), which exhibits sideband-frequency signal

$$y(t) = A_i \sin (w_i t) \sin (w_c t) = \frac{A_1}{2} \{\cos [(w_c - w_i)t] - \cos [(w_c + w_i)t]\} \quad (2)$$

components at the same frequencies as $x(t)$. However, the sum of $x(t)$ and $y(t)$ is given by Eq. (3),

$$x(t) + y(t) = A_i \cos [(w_c - w_i)t] \quad (3)$$

which is the lower-sideband-frequency signal only, and the difference of $x(t)$ and $y(t)$ is given by Eq. (4),

$$x(t) - y(t) = A_i \cos [(w_c + w_i)t] \quad (4)$$

which is the upper-sideband-frequency signal only.

Implementation of this phase-shifting method to generate either an upper- or a lower-SSB signal requires a pair of linear modulators (**Fig. 2**), one to generate $x(t)$, the other to generate $y(t)$. The phase shifting of the information signal is generally performed by a wideband Hilbert transformer, a filter whose gain is $-j$ for positive frequencies and j for negative frequencies, where j is the square root of -1. The narrowband phase shifter for the carrier signal can be a simple second-order filter. However, such a filter may not be necessary because the carrier-frequency generator often provides both sine and cosine outputs simultaneously. Because the required transmission bandwidth is 50% less than that of DSB, SSB is an exemplary modulation system for bandwidth preservation. *See* SINGLE SIDEBAND.

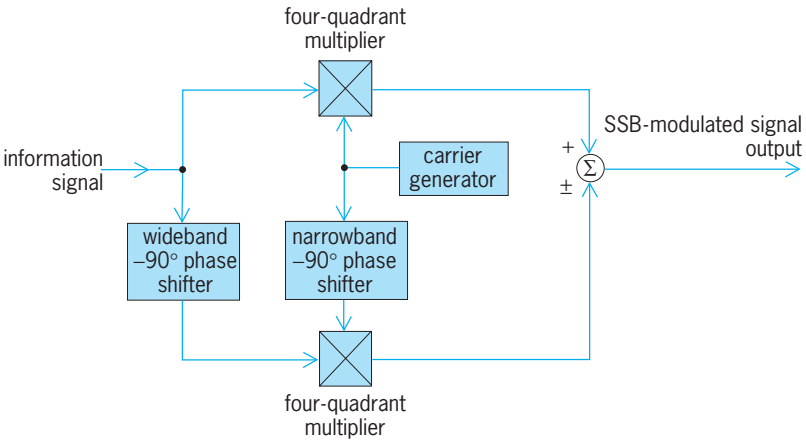


Fig. 2. Single-sideband (SSB) amplitude modulator.

Since the redundant sideband has been reduced (but not eliminated) in the generation of a VSB signal, its transmission efficiency is less than that of a SSB signal but greater than that of a DSB signal. The VSB signal can be generated in two steps: manufacture of a DSB signal, followed by the partial filtering out of a redundant sideband.

Carrier insertion. Receivers demand a carrier-signal reference to properly demodulate (detect) the transmitted signal. This reference may be provided in one of three ways: it may be transmitted with the modulated signal, for example, in double-sideband transmitted-carrier (DSTC or DSBTC) modulation, which is the method used for standard-broadcast AM radio; transmitted on a separate channel (often the case for instrumentation systems); or generated at the receiver.

Receivers for double-sideband signals can be very inexpensive if the carrier is transmitted with the information-bearing sidebands, as in the case of DSTC signals. In concept, the carrier may be inserted by simply adding it to the DSSC signal. For maximum transmission efficiency in generating the DSTC signal, the minimum amount of carrier should be transmitted. (Transmission efficiency may be defined as the ratio of total sideband power to the sum of sideband-plus-carrier power.) However, in order to avoid distortion in simple low-cost receivers, the amplitude of the inserted carrier must be at least equal to the peak amplitude, A, of the information signal. The ratio of the actual amplitude of the information signal to A is known as the modulation index, which should never exceed unity (**Fig. 3**). *See* AMPLITUDE-MODULATION DETECTOR; DISTORTION (ELECTRONIC CIRCUITS).

The transmitter can also be simplified by adding to the information signal a dc value equal to the peak magnitude of the information signal. The value of this modified information signal is now always positive. No longer is a four-quadrant multiplier required to modulate; rather, a far simpler two-quadrant multiplier generates the DSTC signal directly.

Mechanization. DSSC modulation, no matter how it is disguised, is just ordinary (linear) multiplication, or a reasonable approximation to that multiplication. Furthermore, as discussed above, a DSTC signal can be modeled as a DSSC signal with the carrier added. Any amplitude modulator is therefore simply some sort of embodiment of a linear multiplier

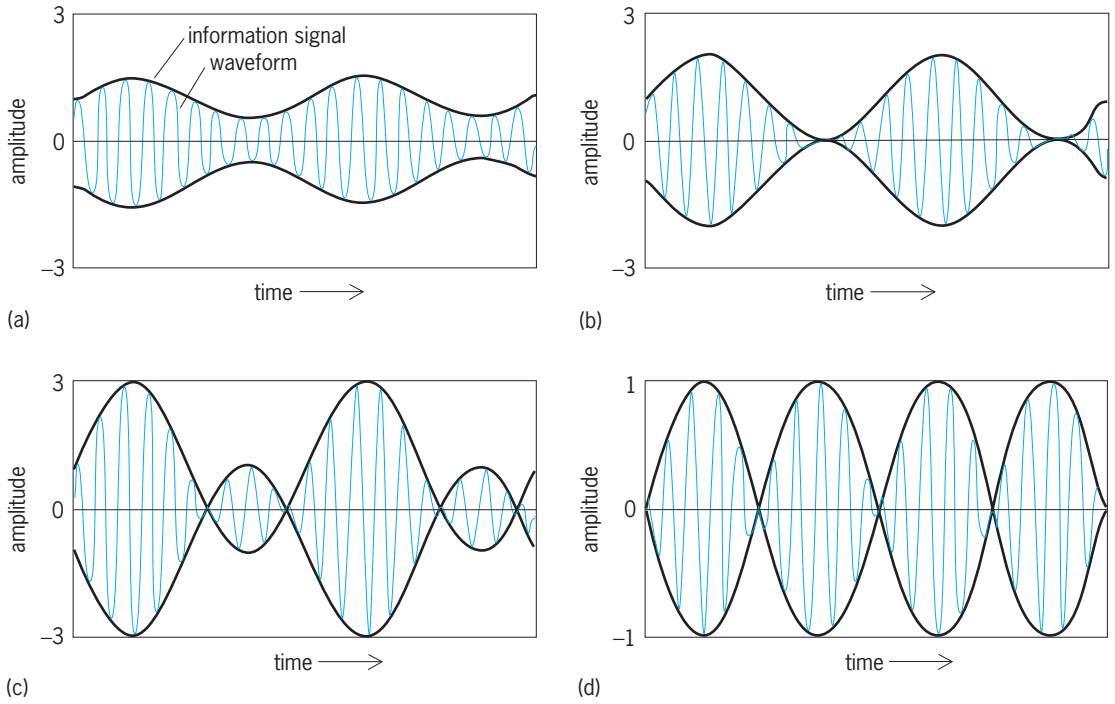


Fig. 3. Waveforms of a signal after various types of double-sideband modulation. Information signal envelopes are shown superimposed on the modulated waveforms. Double-sideband transmitted-carrier (DSTC) modulation with modulation indices of (*a*) 50%, (*b*) 100%, and (*c*) 200% are compared with (*d*) double-sideband suppressed-carrier (DSSC) modulation.

with or without a means to add the carrier.

High-level DSTC modulation is usually done by varying the power-supply voltage to the final high-power amplifier in the radio-frequency transmitter by summing the information signal with the dc output voltage of the power supply. Low-level DSTC modulation is carried out by integrated circuits that approximate the multiplications, followed by a linear amplifier. *See* AMPLIFIER; ELECTRONIC POWER SUPPLY.

Because of the falling costs and improving performance of digital devices, signals are now generally represented by digital data, that is, signal values which are sampled in time and encoded as numbers to represent their sampled values. Digital signal-processing (DSP) functions are carried out by some sequence of addition (or subtraction), multiplication, and delays. Efficient mechanization of each of these functions has been the subject of considerable effort. Digitally, waveform generation and linear multiplication are highly optimized processes. Advances in integrated-circuit fabrication techniques have so lowered the cost of digital circuits for modulation that the overwhelming majority of amplitude modulators in use are now digital. Custom application-specific integrated circuits (ASICs), general-purpose programmable DSP devices, and customizable arrays such as programmable logic arrays (PLAs) and field-programmable arrays (FPAs) are all in widespread use as amplitude modulators and demodulators. Digital modems as data transmission equipment dominate the production of amplitude modulators. *See* INTEGRATED CIRCUITS.

Quadrature amplitude modulation (QAM). It was observed above that the DSSC signal could be reduced to an SSB signal because of the redundant information in the second sideband. This redundancy can be exploited in a different way by quadrature modulation. In addition to the first information signal, $u(t)$, quadrature amplitude modulation acts on a second information signal, $v(t)$. The purpose of the process is to amplitude-modulate the two signals to the same frequency, w_c, with the goal of recovering each signal intact and uncontaminated by the other.

When a first DSSC signal, $u(t)\cos(w_c t)$, is demodulated with a first reference signal, $\cos(w_c t)$, the product signal given by Eq. (5) is obtained. After low-pass

$$u(t)[\cos(w_c t)]^2 = 0.5[1 + \cos(2w_c t)]u(t) \quad (5)$$

filtering to eliminate the double-frequency term, $\cos(2w_c t)$, and scaling by 2 to offset the 0.5 gain, the first information signal, $u(t)$, is recovered. If the first received signal, $u(t)\cos(w_c t)$, had been demodulated with a second reference signal, $\sin(w_c t)$, the product signal given by Eq. (6) would have been ob-

$$u(t)\sin(w_c t)\cos(w_c t) = 0.5\sin(2w_c t)u(t) \quad (6)$$

tained, and low-pass filtering this signal to eliminate the double-frequency term, $\sin(2w_c t)$, would leave nothing.

Similarly, when a second transmitted and received DSSC signal, $v(t)\sin(w_c t)$, is demodulated with the second reference signal, $\sin(w_c t)$, the product signal

given by Eq. (7) is obtained. After low-pass filter-

$$v(t)[\sin(w_c t)]^2 = 0.5[1 - \cos(2w_c t)]v(t) \quad (7)$$

ing to eliminate the first double-frequency term, cos $(2w_c t)$, and scaling by 2 to offset the 0.5 gain, the second information signal, $v(t)$, is obtained. If the second received signal, $v(t)$ sin $(w_c t)$, had been demodulated with the first reference signal, cos $(w_c t)$, the product signal given by Eq. (8) would have been

$$v(t)\sin(w_c t)\cos(w_c t) = 0.5\sin(2w_c t)v(t) \quad (8)$$

obtained, and low-pass filtering this signal to eliminate the double-frequency term, sin $(2w_c t)$, would again leave nothing.

The two carrier signals, sin $(w_c t)$ and cos $(w_c t)$, are 90° apart; they are also orthogonal, or are in quadrature. That is, the average value of their product is zero. An obvious deduction is that, if a third signal is transmitted that is the sum of the first and second transmitted signals of the discussion above, that is, $u(t)\cos(w_c t) + v(t)\sin(w_c t)$, and the received signal is applied to two parallel channels (named by convention I and Q), then it is possible to demodulate in the first (I) channel with the first reference signal, cos $(w_c t)$, and in the second (Q) channel with the second reference signal, sin $(w_c t)$, thereby recovering the first and second information signals, $u(t)$ and $v(t)$, respectively, from the I and Q channels.

The modulation method in the discussion above could be DSBTC if $u(t)$ and $v(t)$ were replaced by $u'(t) = K_1 + u(t)$ and $v'(t) = K_2 + v(t)$, where K_1 and K_2 are equal to or greater than the maximum amplitudes of $u(t)$ and $v(t)$, respectively, in order to keep the modulation indices positive.

In the above discussion, the first and second transmitted signals are both amplitude-modulated signals, but the modulating signals [cos $(w_c t)$ and sin $(w_c t)$] are in quadrature. The pair of information signals, $u(t)$ and $v(t)$, are therefore quadrature-amplitude-modulated (QAM) in the formation of the third transmitted signal. *See* AMPLITUDE MODULATION; MODULATION; MODULATOR.

Stanley A. White

Bibliography. R. C. Dorf (ed.), *The Electrical Engineering Handbook*, 2d ed., 1997; M. S. Roden, *Analog and Digital Communication Systems*, 3d ed., 1991; M. Valkenburg (ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 8th ed., 1996.

Amylase

An enzyme which breaks down (hydrolyzes) starch, the reserve carbohydrate in plants, and glycogen, the reserve carbohydrate in animals, into reducing fermentable sugars, mainly maltose, and reducing nonfermentable or slowly fermentable dextrins. Amylases are classified as saccharifying (β-amylase) and as dextrinizing (α-amylases). The α- and β-amylases are specific for the α- and β-glucosidic bonds which connect the monosaccharide units into large aggregates, the polysaccharides. The α-amylases are found in all types of organs and tissues, whereas β-amylase is found almost exclusively in higher plants. *See* CARBOHYDRATE; ENZYME; GLYCOGEN.

Animals. In animals the highest concentrations of amylase are found in the saliva and in the pancreas. Salivary amylase is also known as ptyalin and is found in humans, the ape, pig, guinea pig, squirrel, mouse, and rat. Pig pancreas is rich in amylase, whereas cattle, sheep, and dog pancreases have lower concentrations.

Starch is one of the most important constituents of human food. The food prepared by the mouth for swallowing (the bolus) is converted by the gastric juices into chyme (a semiliquid paste). Chyme is passed through the pylorus into the duodenum, where intestinal digestion occurs. Part of this digestion is caused by pancreatic amylase, which, like ptyalin, hydrolyzes starch to maltose. A maltase also found in pancreatic juice hydrolyzes maltose to glucose. Glucose is picked up by the bloodstream for use in the tissues for respiration and for conversion to glycogen in the liver for storage.

Plants. Starch is broken down during the germination of seeds (rich in starch) by associated plant enzymes into sugars. These constitute the chief energy source in the early development of the plant. β-Amylase occurs abundantly in seeds and cereals such as malt. It also is found in yeasts, molds, and bacteria.

Industry. Amylase is also used in industry. It is used (1) in brewing and fermentation industries for the conversion of starch to fermentable sugars, (2) in the textile industry for designing textiles, (3) in the laundry industry in a mixture with protease and lipase to launder clothes, (4) in the paper industry for sizing, and (5) in the food industry for preparation of sweet syrups, to increase diastase content of flour, for modification of food for infants, and for the removal of starch in jelly production. The amylase enzyme used in industry comes from many sources: fungi, malt, bacteria, and the pancreas gland of cattle. *See* MALT BEVERAGE.

Daniel N. Lapedes

Amyloidosis

A disorder characterized by the accumulation of an unusual extracellular fibrous protein (amyloid) in the connective tissue of the body. The deposition of amyloid may be widespread, involving major organs and leading to serious clinical consequences, or it may be very limited with little effect on health.

The term amyloid originated with R. Virchow (1854), who mistakenly believed that this substance was akin to starch. Amyloid protein has now been found to have the following unique electron microscopic, x-ray diffraction, and biochemical characteristics. It appears as a homogeneous, eosinophilic extracellular substance when stained with hematoxylin and eosin. When it is stained with Congo red, green birefringence patterns are visible in the polarizing microscope. With the electron microscope, it appears in the form of fine, nonbranching fibrils 7–10 nanometers in diameter, and x-ray diffraction

studies show a cross beta pattern. Amino acid sequence studies demonstrate the presence of unusual proteins (AA, AL transthyretin, beta-2 microglobin, beta/A4 protein, and others).

Amyloidosis has been classified clinically as: (1) primary amyloidosis, with no evidence for preexisting or coexisting disease; (2) amyloidosis associated with multiple myeloma; (3) secondary amyloidosis, associated with chronic infections (such as osteomyelitis, tuberculosis, leprosy), chronic inflammatory disease (such as rheumatoid arthritis, ankylosing spondylitis, regional enteritis), or neoplasms (such as medullary carcinoma of the thyroid); (4) heredofamilial amyloidosis, associated with familial Mediterranean fever and a variety of heritable neuropathic, renal, cardiovascular, and other syndromes; (5) local amyloidosis, with local, often tumorlike, deposits in isolated organs without evidence of systemic involvement; (6) amyloidosis associated with aging in the heart or brain; (7) amyloid associated with hormonal disorders; and (8) amyloid of chronic hemodialysis.

Immunochemical findings. There is an immunochemical classification based on definition of the various types of proteins in amyloidosis: (1) amyloid associated with tissue deposition of protein AA, a tissue protein occurring in secondary amyloidosis and the amyloid of familial Mediterranean fever; and (2) amyloid associated with tissue deposition of protein AL, a protein consisting primarily of the aminoterminal variable segment of the light chain of homogeneous immunoglobulins, or in a few instances the whole light chain, which occurs in primary amyloid and amyloid associated with multiple myeloma. Transthyretin, the major protein of autosomal dominant inherited neuropathic amyloid, is also found in the amyloid lesions in the aged heart. Other chemical types also exist; for example, beta protein is found in the amyloid of Alzheimer's disease, IAPP (amylin) in that of adult onset (type 2) diabetes, and beta-2 microglobulin in the amyloid of chronic hemodialysis.

Antisera to protein AA react with an antigenically related component, serum AA (SAA), that is found in normal human serum, but has a higher molecular weight than AA and is regarded as the precursor of AA. Serum AA has several distinct characteristics: (1) It behaves as an acute phase reactant and is elevated in concentration in infection and inflammation. (2) It migrates as an apolipoprotein in the serum. (3) It is produced in the liver. In the heredofamilial amyloidosis, a variety of single point mutations have been found as the cause of the disorder; that is, the prototype is transthyretin with the substitution of methionine for valine at position 30.

A second component of amyloid, P-component (plasma component or pentagonal unit), with different ultrastructure, x-ray diffraction pattern, and chemical characteristics, has been isolated from amyloid and shown to be identical with a circulating human alpha globulin present in serum in minute amounts. It has certain similarities to C-reactive protein (CRP, an acute-phase protein in humans). However, it does not behave as an acute-phase protein in humans, and has biophysical and immunologic distinctions from CRP.

Clinical features. The different types of amyloid overlap considerably with regard to organ distribution. Since any organ may be involved, the clinical manifestations of amyloidosis are extremely varied. Infiltration of and about peripheral nerves, skin, tongue, joints, and heart is most frequent in the primary forms, while in secondary amyloidosis the liver, spleen, and kidney are the major sites of deposits.

Carpal tunnel syndrome (median neuropathy), resulting from local deposition of amyloid, is a frequent finding in certain types of primary amyloidosis and in the amyloidosis associated with myeloma and some heredofamilial forms. Some individuals with primary amyloidosis develop a coagulation disorder in which there is a deficiency of factor X (a procoagulant in plasma). This condition has been traced to the binding of these proteins to amyloid fibrils.

Amyloid in the heredofamilial (involving multiple siblings of successive generations) forms is responsible for peripheral neuropathy in certain families in Portugal, Japan, and Sweden; upper limb neuropathy in certain families of German or Swiss ancestry in the United States; cranial nerve neuropathy in a group of Finnish ancestry; and kidney disease and cardiovascular disease in others. All of the above conditions are inherited as autosomal dominant traits except for the amyloid associated with familial Mediterranean fever, which is inherited in an autosomal recessive manner.

Clinical features in the diagnosis of primary amyloidosis include unexplained proteinuria, peripheral neuropathy, enlargement of the tongue, enlargement of the heart, intestinal malabsorption, bilateral or familial carpal tunnel syndrome, or low blood pressure when standing upright. Symptoms frequently exist for several years before correct diagnosis. When primary amyloidosis is suspected, evidence of an underlying disease is investigated to rule out possible inflammatory disorders or malignant tumors. In cases of chronic infectious or inflammatory diseases, secondary amyloidosis is suspected if there is unexplained proteinuria, or enlargement of the liver or spleen.

The diagnosis is established by means of microscopic examination of appropriate biopsy specimens. Subcutanous abdominal fat or tissue from the rectum have proved to be particularly useful, and in selected cases, skin, gingival, kidney, and liver biopsy specimens are valuable. Since a biopsy specimen may contain only a small quantity of amyloid, it is important to look for the characteristic green birefringence on polarization microscopy after Congo red staining.

Treatment. There is no specific treatment for amyloidosis, but supportive treatment is very useful. Some individuals with secondary renal amyloidosis have improved after the cure of a predisposing disease, such as active tuberculosis or chronic osteomyelitis. Kidney transplants have been carried out successfully in a few cases of renal failure. Colchicine

has been shown to relieve the acute attacks of familial Mediterranean fever and may decrease the formation of amyloid. Combinations of melphalan, prednisone, and colchicine are used in primary (AL) amyloidosis, and liver transplantation has been introduced in heredofamilial amyloid to remove the site of synthesis of the mutant protein. Alan S. Cohen

Bibliography. A. S. Cohen, Amyloidosis, *Bull. Rheumatic Dis.*, 40:1–12, 1991; A. S. Cohen, Amyloidosis, *N. Engl. J. Med.*, 277:522–530, 1967; J. R. Harris (ed.), *Electron Microscopy of Proteins,* vol. 3, 1982; J. B. Natvig et al., *Amyloid and Amyloidosis*, 1990.

Anaerobic infection

An infection caused by anaerobic bacteria (organisms that are intolerant of oxygen). Most such infections are mixed, involving more than one anaerobe and often aerobic or facultative bacteria as well.

Anaerobes are prevalent throughout the body as indigenous flora, and virtually all anaerobic infections arise endogenously, the principal exception being *Clostridium difficile* colitis. Factors predisposing to anaerobic infection include those disrupting mucosal or other surfaces (trauma, surgery, and malignancy or other disease), those lowering redox potential (impaired blood supply, tissue necrosis, and growth of nonanaerobic bacteria), drugs inactive against anaerobes (such as aminoglycosides), and virulence factors produced by the anaerobes (toxins, capsules, and collagenase, hyaluronidase, and other enzymes).

Anaerobic gram-negative bacilli (*Bacteroides, Prevotella, Porphyromonas, Fusobacterium*) and anaerobic gram-positive cocci (*Peptostreptococcus*) are the most common anaerobic pathogens. *Clostridium* (spore formers) may cause serious infection. The prime pathogen among gram-positive nonsporulating anaerobic bacilli is *Actinomyces*.

Infections commonly involving anaerobes and the incidence of anaerobes in such infections are noted in the **table**. In terms of frequency of occurrence, the oral and dental pleuropulmonary, intraabdominal, obstetric-gynecologic, and skin and soft tissue infections are most important. Foul odor is definitive evidence of involvement of anaerobes in an

Infections commonly involving anaerobic bacteria

Infections	Incidence, %	Proportion of cultures positive for anaerobes yielding only anaerobes
Bacteremia	20	4/5
Central nervous system		
Brain abscess	89	1/2 to 2/3
Extradural or subdural empyema	10	
Head and neck		
Chronic sinusitis	52	4/5*
Chronic otitis media	33 to 56	0 to 1/10
Neck space infections	100	3/4
Wound infection following head and neck surgery	95	0
Dental, oral, facial		
Orofacial, of dental origin	94	4/10
Bite wounds	47	1/34
Thoracic		
Aspiration pneumonia	62 to 100	1/3 to 1/2†
Lung abscess	85 to 93	1/2 to 3/4
Empyema (nonsurgical)	62 to 76	1/3 to 1/2
Abdominal		
Intraabdominal infection (general)	81 to 94	1/10 to 1/3
Appendicitis with peritonitis	96	1/100
Liver abscess	52	1/3
Other intraabdominal infection (postsurgery)	93	1/6
Biliary tract	41 to 45	2/117 to 0
Obstetric-gynecologic		
Miscellaneous types	71 to 100	1/3
Pelvic abscess	88	1/2
Vulvovaginal abscess	75	1/4
Vaginal cuff abscess	98	1/30
Septic abortion, sepsis	63 to 67	
Pelvic inflammatory disease	25 to 48	1/14 to 1/7
Soft tissue and miscellaneous		
Nonclostridial crepitant cellulitis	75	1/12
Pilonidal sinus infection	73+	
Diabetic foot ulcers	95	1/20
Soft-tissue abscesses	60	1/4
Cutaneous abscesses	62	1/5
Decubitus ulcers with bacteremia	63	
Osteomyelitis	40	1/10

* 23/28 cultures (82%) yielding heavy growth of one or more organisms had only anaerobes present.
† Aspiration pneumonia occurring in the community, rather than in the hospital, involves anaerobes to the exclusion of aerobic or facultative forms two-thirds of the time.
SOURCE: S. M. Finegold, Antimicrobial therapy of anaerobic infections: A status report, *Hosp. Pract.,* 14:17–81, October 1979.

infection, but absence of such odor does not exclude the possibility. Other important clues to anaerobic infection, though not specific, are location near a mucosal surface; bite infection; tissue necrosis; gas formation; prior aminoglycoside, quinolone, or trimethoprim/sulfamethoxazole therapy; "sulfur granules" (actinomycosis); certain distinctive clinical presentations (such as gas gangrene); septic thrombophlebitis; unique morphology of certain anaerobic organisms on smear; and failure to grow organisms aerobically from specimens. To document anaerobic infection properly, specimens for culture must be obtained so as to exclude normal flora and must be transported under anaerobic conditions. *See* GANGRENE; GAS.

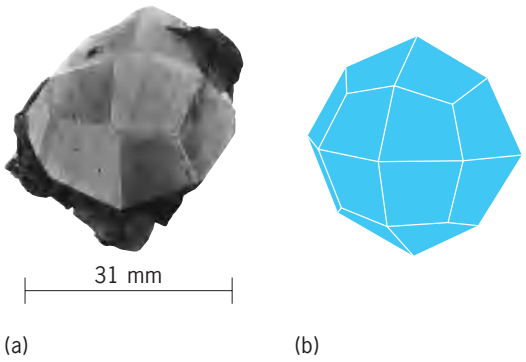
Therapy includes surgery (debridement, drainage) and antimicrobial agents. *Bacteroides fragilis* is the most resistant anaerobe; only chloramphenicol, carbapenems, beta lactam/beta lactamase inhibitor combinations, and metronidazole are essentially active against all strains. Other useful drugs include clindamycin, cefoxitin, ceftizoxime, and cefotetan. *See* ANTIBIOTIC; INFECTION. Sydney M. Finegold

Bibliography. S. M. Finegold and W. L. George (eds.), *Anaerobic Infections in Humans*, 1989; P. N. Levett (ed.), *Anaerobic Microbiology: A Practical Approach*, 1992; T. Willis and L. Phillips, *Anaerobic Infections: Clinical and Laboratory Practice*, 1991.

Analcime

A mineral with a framework structure in which all the aluminosilicate tetrahedral vertices are linked, thus allying it to the feldspars, feldspathoids, and zeolites. The structure is cubic, space group Ia3d, $\alpha = 1.371$ mm. Its formula is $Na(H_2O)[AlSi_2O_6]$; in this sense it is a tectosilicate.

The analcime structure type includes several other mineral species. These include high-temperature leucite, β-$K[AlSi_2O_6]$; pollucite, $Cs(H_2O)$-$[AlSi_2O_6]$; and wairakite, $Ca(H_2O)[AlSi_2O_6]_2$. Temperature-dependent order-disorder relationships between AlO_4 and SiO_4 tetrahedra lead to modifications of lower structural symmetry, birefringent optical properties, and complex composite twinning about a single crystal nucleus. Crystals are most often trapezohedra (see **illus.**); rarely the mineral is massive granular. Hardness is 5–5.5 on Mohs scale; specific gravity is 2.27.



Analcime, (*a*) Specimen from Keweenaw County, Michigan (*American Museum of Natural History*). (*b*) Trapezohedral crystal typical of analcime (*after C. Klein, Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*).

Analcime most frequently occurs as a low-temperature mineral in vesicular cavities in basalts, where it is associated with zeolites (particularly natolite), datolite, prehnite, and calcite. Small grains are frequent constituents of sedimentary rocks and muds in oceanic basins associated with volcanic sources. Noted localities for large crystals include the Bay of Fundy region, Nova Scotia; the Watchung basalts of New Jersey; and the Keweenaw Peninsula in northern Michigan. Analcime also occurs in certain hydrothermal sulfide deposits as a fissure mineral. *See* FELDSPAR; FELDSPATHOID; SILICATE MINERALS; ZEOLITE. Paul B. Moore

Analgesic

Any of a group of drugs of diverse chemical structure and physiological effects which are commonly used for the relief of pain. To qualify as an analgesic a drug must selectively reduce or abolish pain without causing impairment of consciousness, mental confusion, incoordination or paralysis, or other derangements of the nervous system.

Narcotic alkaloids. The oldest and best-known analgesics are opium, a drug obtained by extracting the juice of the poppy seed, and its most active alkaloid, morphine. Morphine and related drugs reduce or block the activation of pain neurons in the gray matter of the spinal cord, and at receptor sites in the brainstem and thalamus. The mechanisms by which this inhibition is effected are complex. Morphine does not act directly on the pain-mediating neurons. Instead, it involves an elaborate pain-modulating system in the posterior horns of the spinal cord, the dorsal-ventral reticular formation of the medulla, the periaqueductal gray matter, and the hypothalamus and thalamus. Separate neurons in these regions have opioid receptors on their surface, and the strength of this affinity between drug and neuron is proportional to the analgesic potency of the opiate (or opioid). All parts of this neuronal system acting locally or by way of descending connections inhibit pain transmission cells. In these same regions there are also neurons that elaborate a naturally occurring (endogenous) group of opioid peptides whose function is to modulate or suppress pain. They are called enkephalins, and the most potent of them is beta-endorphin. Since they are abundant not only in these sites but also in the gut, in sympathetic neurons, and in chromafin cells of the adrenal glands, they are believed to account for the natural inhibition of pain by strong emotion (stress), and expectancy of pain relief by nonanalgesic drugs (placebo effect). *See* ENDORPHINS; MORPHINE ALKALOIDS; OPIATES.

In addition to their use as analgesic drugs, opiates have other biological effects such as sedation,

pupillary constriction, suppression of cough reflex, respiratory depression, reduction of intestinal motility, impairment of segmental flexor reflexes, and decrease in body temperature. This functional diversity is attributed to the activation of other inhibitory systems of neurons. *See* SEDATIVE.

While morphine is the most powerful medical analgesic substance, there are many other naturally occurring alkaloids derived from opium. The best known of these is codeine.

Common to all opiates is the attribute that if they are taken for weeks or months the recipient will need larger doses to obtain the same analgesic and sedative effects. This response is called tolerance. If the drug is stopped, disagreeable withdrawal or abstinence effects are experienced within hours to days. There is severe pain, sweating, salivation, hyperventilation, restlessness, and confusion. These abstinence symptoms, which are marks of habituation, pressure the addicted person to take extreme measures to obtain the narcotic in order to avoid the symptoms. *See* ADDICTIVE DISORDERS; ALKALOID; NARCOTIC.

Synthetics. Because of the strong addictive properties of morphine and related compounds, chemists have synthesized other drugs of similar chemical structure, in the hope of securing analgesia without addiction. This effort has been only partially successful. Methadone, a drug that has been given to addicts as a substitute for morphine, is an effective analgesic when taken orally and is less addictive than morphine. Meperidine (Demerol) is a strong synthetic analgesic but definitely addictive. Other synthetic analgesics are oxycodone (Percodan), levorphanol (levodromoran), propoxyphene (Darvon), and pentazocine (Talwin). The last two of this series cause little or no addiction but, unfortunately, are not strong analgesics. Another synthetic drug, Naloxone, blocks the analgesic effect of all opiate agonists and precipitates withdrawal symptoms in addicted individuals.

Salicylates. Another class of analgesic drugs, which are nonnarcotic (nonaddictive), are the salicylates, the most familiar being acetysalicylic acid (aspirin), and salicylatelike drugs such as phenylbutazone (Butazolidine), indomethacin (Indocin), acetaminophen, and phenacetin. These drugs are most effective in relieving skeletal pain due to inflammation (such as arthritis). Their analgesic properties, which are not nearly as strong as those of morphine and the synthetic opioids, are due to their action on both the peripheral and central nervous system. Peripherally they block the synthesis of prostaglandins in the inflamed tissues and thereby prevent the sensitization of pain receptors. Centrally they act in some obscure way on the hypothalamus. These drugs also have many other effects, such as reducing fever (antipyrexia) and preventing platelet agglutination. They are the most commonly used of all analgesic medications and are often combined with caffeine or a barbiturate sedative under a variety of trade names and sold for the relief of headache, backache, and so forth. *See* ASPIRIN; EICOSANOIDS; NERVOUS SYSTEM (VERTEBRATE); PAIN.

Raymond D. Adams

Analog computer

A computer or computational device in which the problem variables are represented as continuous, varying physical quantities. An analog computer implements a model of the system being studied. The physical form of the analog may be functionally similar to that of the system, but more often the analogy is based solely upon the mathematical equivalence of the interdependence of the computer variables and the variables in the physical system. *See* SIMULATION.

Types. An analog computer is classified either in accordance with its use (general- or specific-purpose) or based on its construction (hydraulic, mechanical, or electronic). General-purpose implies programmability and adaptability to different applications or the ability to solve many kinds of problems. Most electronic analog computers were general-purpose systems, either real-time analog computers in which the results were obtained without any significant time-scale changes, or high-speed repetitive operation computers. The latter utilized time compression typically in the range 1000:1 to 10,000:1, completing a computation or simulation in a few milliseconds and enabling the display of results as flicker-free images.

Special-purpose analog computers contain fixed programs permitting few or no adjustments. They are generally built into physical systems where they serve typically as a subsystem simulator, function controller or results analyzer. For example, the pneumatic computer shown in **Fig. 1** used air bellows and flapper nozzles to generate accurate multiplication, division, squaring, or square-root functions of input signals, encoding data as air pressures.

Use. Large electronic analog computer systems with many hundreds of operational amplifiers were widely used from the early 1960s to the mid-1980s. They solved extremely complex and extensive sets of differential equations (mathematical models) such as six-degree-of freedom space flights, exothermal chemical reaction kinetics, control systems for food processing plants, and the human immunosuppressive system. During the 1970s, they were often paired with a digital computer in hybrid systems

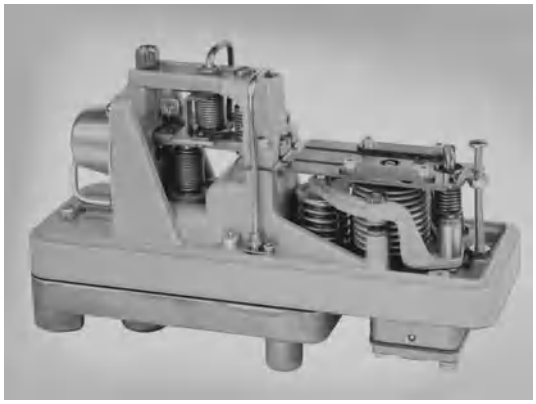


Fig. 1. Special-purpose pneumatic analog computer, model Foxboro 556. (*Foxboro Co.*)

Fig. 2. Analog-hybrid laboratory equiped with a HYSHARE 600 system. This system can include up to six analog computer consoles and one or more digital processors for multiuser, multitask applications. (*Electronic Associates, Inc.*)

(**Fig. 2**). The digital computer then primarily controlled the setting up and running of the analog computers, collecting and checking results and often providing logic and initial values to allow extensive searching through vast multidimensional problem domains. In some instances, hybrid systems also entailed automatic patching (interconnecting) of analog computing units.

Smaller analog computers, with fewer than 100 operational amplifiers, were extensively used for education and training purposes. Models of a theoretical nature, expressed by sets of differential equations, or practical engineering simulations based on empirical data and heuristic models often lent themselves to study by analog computers. The ease by which the user could interact with the modeled system was a major feature, since learning is greatly enhanced when the user safely can explore, experiment with, and even misuse the modeled system.

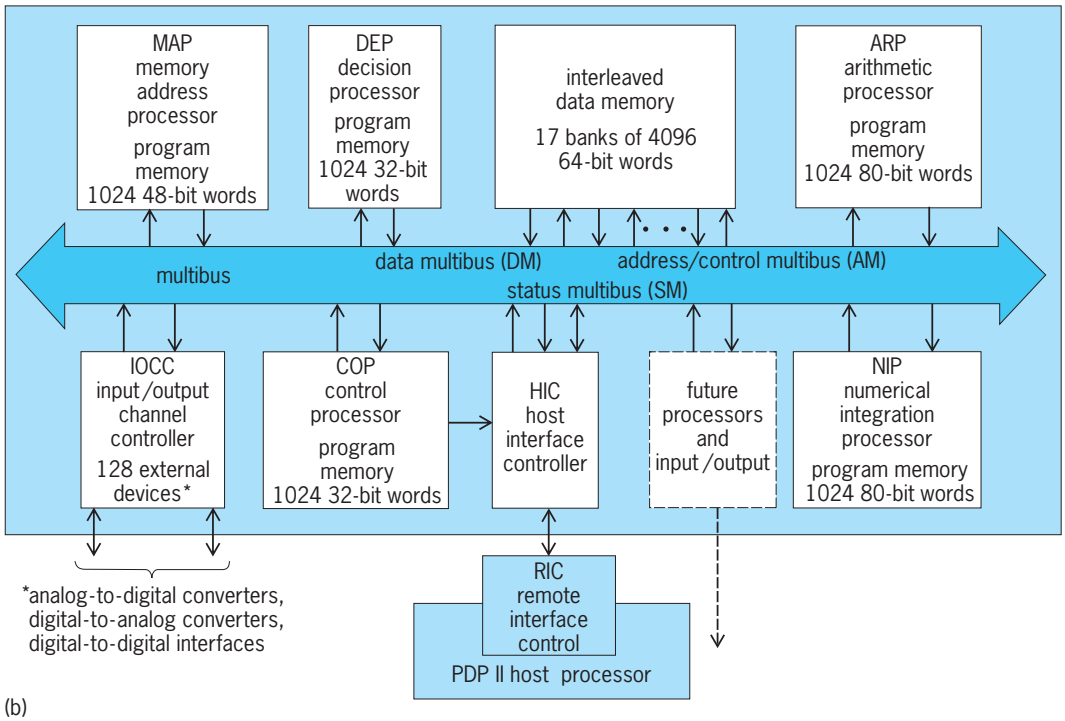
Digital equivalents. Since the 1970s, digital computer programs have been developed which essentially duplicate the functionality of the analog computer. Modern simulation languages, such as ACSL, GASP, GPSS, SLAM, and Simscript, have replaced electronic analog computers. They provide nearly the same highly interactive and parallel solution capabilities of electronic analog computers, but without the technical shortcomings of electronics: accuracy inherently limited to 0.01%, effective bandwidths of 1 MHz, and cumbersome and time-consuming programming. Simulation languages also avoid the large purchase investments and the continual maintenance dependencies of complex electronic systems.

Digital multiprocessor analog system. Another type of analog computer is the digital multiprocessor analog system (**Fig. 3**), in which the relatively slow speeds of sequential digital increment calculations have been radically boosted through parallel processing. In this type of analog computer it is possible to retain the programming convenience and data storage of the digital computer while approximating the speed, interaction potential, and parallel computations of the traditional electronic analogs.

The digital multiprocessor analog computer typically utilizes several specially designed high-speed processors for the numerical integration functions, the data (or variable) memory distributions, the arithmetic functions, and the decision (logic and control) functions. All variables remain as fixed or floating-point digital data, accessible at all times for computational and operational needs.

The digital multiprocessor analog computer achieves an overall problem-solving efficiency comparable to the very best continuous electronic analog computers, at a substantially lower price. An example of such a computer, the model AD10 (Fig. 3*a*), can solve large, complex, and multivariate problems at very high speeds and with the advantages of all-digital hardware. Its computation system (Fig. 3*b*) is based on five parallel processors working in a special computer architecture designed for high-speed operation. The various elements are interconnected by means of a data bus (MULTIBUS). A highly interleaved data memory of up to 10^6 words serves the data and program storage functions. The five processors working in parallel are: the control processor (COP), which controls all operations; the arithmetic processor (ARP), which runs the numerical

(a)



(b)

Fig. 3. Multiprocessor analog computer. (*a*) Exterior. (*b*) Organization. (*Applied Dynamics International*)

calculations; the decision processor (DEP), which executes the logic parts of the program; the memory address processor (MAP), which makes sure that all data are fetched and stored efficiently; and the numerical integration processor (NIP), which carries out the integration functions that are crucially important in an analog computer.

Description. The typical electronic general-purpose analog computer consists of a console containing a collection of operational amplifiers; computing elements, such as summing networks, integrator networks, attenuators, multipliers, and function generators; logic and interface units; control circuits; power supplies; a patch bay; and various meters and display devices. The patch bay is arranged to bring input and output terminals of all programmable devices to one location, where they can be conveniently interconnected by various patch cords and plugs to meet the requirements of a given problem. Prewired problem boards can be exchanged at the patch bay in a few seconds and new coefficients set up typically in less than a half hour. Extensive automatic electronic patching systems have been developed to permit fast setup, as well as remote and time-shared operation.

The analog computer basically represents an instrumentation of calculus, in that it is designed to solve ordinary differential equations. This capability lends itself to the implementation of simulated models of dynamic systems. The computer operates by generating voltages that behave like the physical or mathematical variables in the system under study. Each variable is represented as a continuously varying (or steady) voltage signal at the output of a programmed computational unit. Specific to the analog computer is the fact that individual circuits are used for each feature or equation being represented, so that all variables are generated simultaneously. Thus the analog computer is a parallel computer in which the configuration of the computational units allows direct interactions of the computed variables at all times during the solution of a problem.

Unique features. The unique features of the analog computer which are of value in science and technology are as follows:

1. Within the useful frequency bandwidth of the computational units and components, all programmed computations take place in parallel and are for practical purposes instantaneous. That is, there is no finite execution time associated with each mathematical operation, as is encountered with digital computer methods.

2. The dynamics of an analog model can be programmed for time expansion (slower than real system time), synchronous time (real time), or time compression (faster than real time).

3. The computer has a flexible addressing system so that almost every computed variable can be measured, viewed with vivid display instruments, and recorded at will.

4. One control mode of the computer, usually called HOLD, can freeze the dynamic response of a model to allow detailed examination of interrelationships at that instant in time, and then, after such study, the computing can be made to resume as though no stop had occurred.

5. By means of patch cords, plugs, switches, and adjustment knobs the analog computer program or model can be directly manipulated, especially during dynamic computation, and the resultant changes

in responses observed and interpreted.

6. Because of the fast dynamic response of the analog computer, it is easy to implement implicit computation through the use of problem feedback. (This important and powerful mathematical quality of the analog computer is discussed more fully below.)

7. The computer can be used for on-line model building; that is, a computer model can be constructed in a step-by-step fashion directly at the console by interconnecting computational units on the basis of one-for-one analog representation of the real system elements. Then, by adjusting signal gains and attenuation parameters, dynamic behavior can be generated that corresponds to the desired response or is recognizable as that of the real system. This method allows a skillful person to create models when no rigorous mathematical equations for a system exist.

8. For those applications to which it is well suited, the analog computer operates at relatively low cost, thus affording the analyst ample opportunity to investigate, develop, and experiment within a broad range of parameters and functions.

Components. Manipulations of the signals (voltages) in the analog computer are based upon the fundamental electrical properties associated with circuit components. The interrelation of voltages for representing mathematical functions is derived by combining currents at circuit nodes or junctions. *See* KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS; OHM'S LAW.

The simplest arrangement of components for executing addition would be to impress the voltages to be added across individual resistors (**Fig. 4***a*). The resistors would then be joined (Fig. 4*b*) to allow the currents to combine and to develop the output voltage across a final resistor. Use of this simple configuration of elements for computation is impractical, because of the undesirable interaction between inputs. A change in one input signal (voltage) causes a change in the current that flows through the input resistor; this changes the voltage at the input resistor junction, and the change secondarily causes a different current to flow in the other input resistors. The situation gets more interactive when another computing circuit is attached so that part of the summing current flows away from the summing junction. This interaction effect also prevents exact computing. If, in some way, each voltage to be summed could be made independent of the other voltages connected to the summing junction, and if the required current fed to other circuits could be obtained without loading the summing junction, then precise computation would be possible. *See* DIRECT-CURRENT CIRCUIT THEORY.

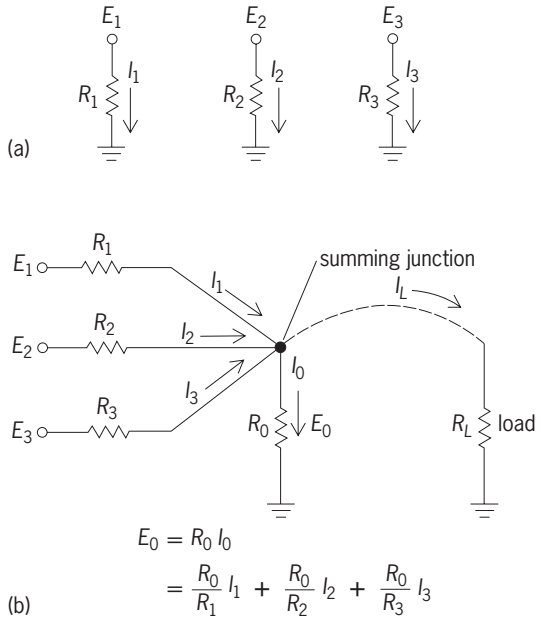


Fig. 4. Addition of electric currents by using a passive network of resistors. (*a*) Individual voltages and resistors. (*b*) Voltages and corresponding currents summed into a common resistor.

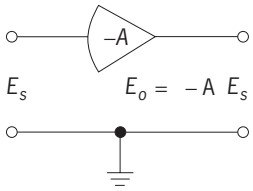


Fig. 5. Symbol for a high-gain direct-current amplifier, with one inverting input referenced to ground.

The electronic analog computer satisfies these needs by using high-gain (high-amplification) dc operational amplifiers. A symbol to represent a dc direct-coupled amplifier is shown in **Fig. 5**. According to convention, the rounded side represents the input to the amplifier, and the pointed end represents the amplifier output. A common reference level or ground exists between the amplifier input and output, and all voltages are measured with respect to it. The ground reference line is understood and is usually omitted from the symbol. The signal input network (consisting of summing resistors) connects to the inverting (or negative) amplifier input terminal. The noninverting (or positive) amplifier input terminal is normally connected to the reference ground. Generally the inverting input is called the summing junction (SJ) of the amplifier. Internal design of the amplifier is such that, if the signal at the summing junction is positive with respect to ground, the amplifier output voltage is negative with respect to ground. The amplifier has an open-loop voltage gain of $-A$; therefore an input voltage of E_s results in an output voltage of $-AE_s$. Gain of a commercial computing amplifier is typically 10^8; thus, an input voltage of less than 1 μV can produce several volts at the amplifier output. *See* AMPLIFIER; DIRECT-COUPLED AMPLIFIER.

Because the operational amplifier thus inverts the signal (changes its sign or polarity), it lends itself to the use of negative feedback, whereby a portion of the output signal is returned to the input. This arrangement has the effect of lowering the net gain of the amplifier signal channel and of improving

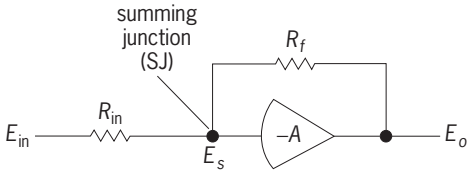
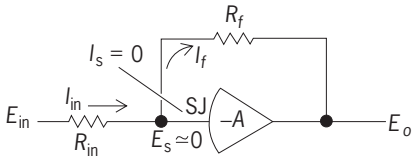


Fig. 6. High-gain amplifier which has been made into an operational amplifier through the inclusion of an input resistor and a feedback resistor tied together at the amplifier's summing junction (SJ).



$$I_{in} = I_f \qquad (1)$$

$$\frac{E_{in}}{R_{in}} = -\frac{E_o}{R_f} \qquad (2)$$

$$\boxed{E_o = -\frac{R_f}{R_{in}} E_{in}} \qquad (3)$$

Fig. 7. Summing junction currents into an operational amplifier create the fixed gain function, determined by resistor values, as indicated in the box.

overall signal-to-noise ratio and increasing computational accuracy.

Circuit operation (**Fig. 6**) can be viewed in the following manner. A dc voltage E_{in} applied to input resistor R_{in} produces a summing junction voltage E_s. The voltage is amplified and appears at the amplifier output as voltage E_o (equal to $-AE_s$, where A is voltage gain of the amplifier). Part of output voltage E_o returns through feedback resistor R_f to the summing junction. Because the returned or feedback voltage is of opposite (negative) polarity to the initial voltage at the summing junction, it tends to reduce the magnitude of E_s, resulting in an overall input-output relationship that may be expressed as Eq. (1). In fact,

$$\frac{E_o}{E_{in}} = \frac{-A}{A+1} = \frac{-1}{1+1/A} \simeq -1 \quad A > 10^8 \qquad (1)$$

the summing junction voltage E_s is so small that it is considered to be at practically zero, a condition called virtual ground.

To illustrate how the operational amplifier serves the needs of the computing network, consider the currents that flow and the corresponding algebraic expressions (**Fig. 7**). The operational amplifier is designed to have high input impedance (high resistance to the flow of current into or out of its input terminal); consequently the amplifier input current I_s can then be considered to be practically zero. The resulting current equation states that input current I_{in} is equal to feedback current I_f. Since the amplifier has a very high gain, the summing junction voltage is virtually zero. Voltage drop across R_{in} is thus equal to E_{in}; voltage drop across R_f is E_o. The equation in the box is the fundamental relationship for the amplifier. As long as the amplifier has such a high gain and requires a negligible current from the summing junction, the amplifier input and output voltages are related by the ratio of the two resistors and are thus not affected by the actual electronic construction of the amplifier. If several input resistors are connected to the same summing junction and voltages are applied to them (**Fig. 8**), then because the summing junction remains at practically zero potential, none of the inputs will interfere with the other inputs. Thus all inputs will exert independent and additive effects on the output.

Because amplifier gain has a negative sign, output voltage equals the negative sum of the input voltages, weighted according to the values of the individual resistors in relation to the feedback resistor, as shown in the box in Fig. 8.

When a computing circuit is connected to the amplifier output, a demand for load current is introduced. Such a required output load current must be supplied by the amplifier without a measurable change in its output voltage. That is, the operational amplifier must act as a voltage controller, supplying whatever current is required within limits, while maintaining the net voltage gain established by the mathematical ratio of its input and feedback elements. The operational amplifier and network is a linear network because once the input-feedback ratios are adjusted, signal channel gains remain constant (a straight-line function) during computation.

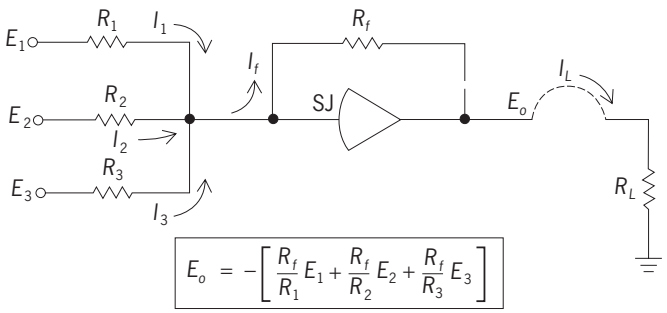


Fig. 8. Operational amplifier which has been made into a summer through the use of several input resistors connected to the summing junction. Output is equal to the inverted weighted sum of the inputs.

Programming. To solve a problem using an analog computer, the problem solver goes through a procedure of general analysis, data preparation, analog circuit development, and patchboard programming. He or she may also make test runs of subprograms to examine partial-system dynamic responses before eventually running the full program to derive specific and final answers. The problem-solving procedure typically involves eight major steps, listed below:

1. The problem under study is described with a set of mathematical equations or, when that is not possible, the system configuration and the interrelations of component influences are defined in block-diagram form, with each block described in terms of black-box input-output relationships.

2. Where necessary, the description of the system (equations or system block diagram) is rearranged

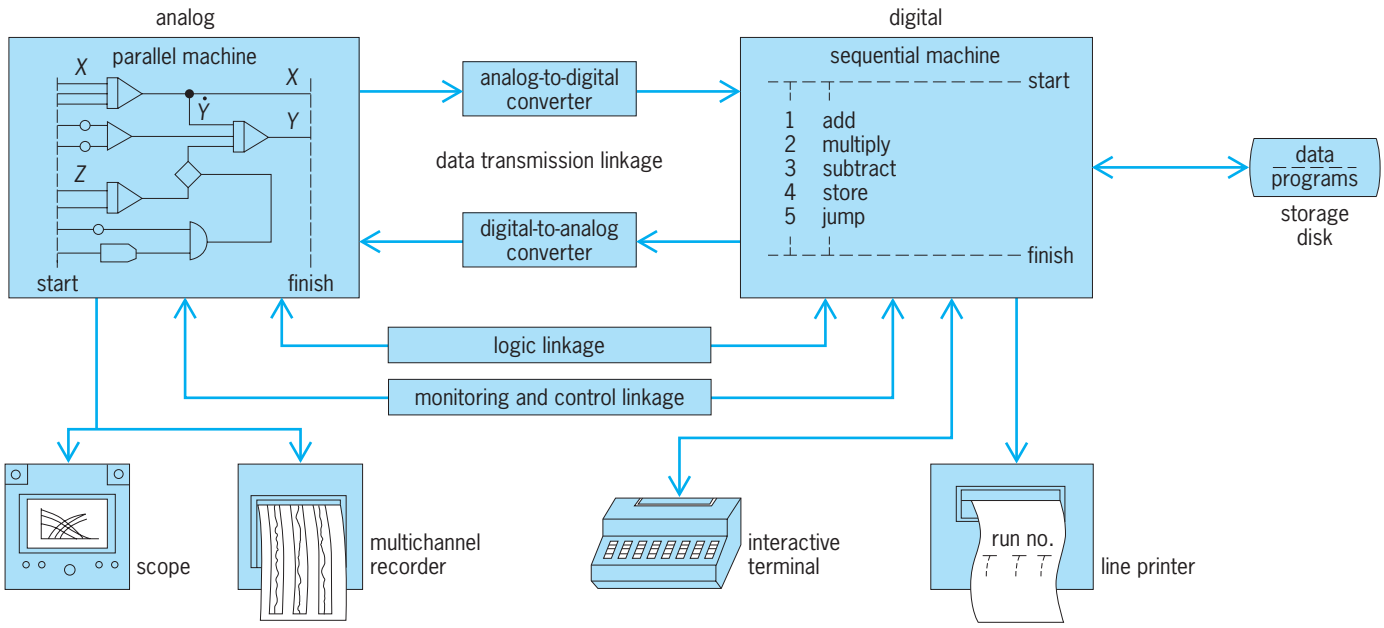


Fig. 9. Block diagram representation of a hybrid analog-to-digital computer.

in a form that may better suit the capabilities of the computer, that is, avoiding duplications or excessive numbers of computational units, or avoiding algebraic (nonintegrational) loops.

3. The assembled information is used to sketch out an analog circuit diagram which shows in detail how the computer could be programmed to handle the problem and achieve the objectives of the study.

4. System variables and parameters are then scaled to fall within the operational ranges of the computer. This may require revisions of the analog circuit diagram and choice of computational units.

5. The finalized circuit arrangement is patched on the computer problem board.

6. Numerical values are set up on the attenuators, the initial conditions of the entire system model established, and test values checked.

7. The computer is run to solve the equations or simulate the black boxes so that the resultant values or system responses can be obtained. This gives the initial answers and the "feel" for the system.

8. Multiple runs are made to check the responses for specific sets of parameters and to explore the influences of problem (system) changes, as well as the behavior which results when the system configuration is driven with different forcing functions.

Hybrid computers. The accuracy of the calculations on a digital computer can often be increased through double precision techniques and more precise algorithms, but at the expense of extended solution time, due to the computer's serial nature of operation. Also, the more computational steps there are to be done, the longer the digital computer will take to do them. On the other hand, the basic solution speed is very rapid on the analog computer because of its parallel nature, but increasing problem complexity demands larger computer size. Thus, for the analog computer the time remains the same regardless of the complexity of the problem, but the size of the computer required grows with the problem.

Interaction between the user and the computer during the course of any calculation, with the ability to vary parameters during computer runs, is a highly desirable and insight-generating part of computer usage. This hands-on interaction with the computed responses is simple to achieve with analog computers. For digital computers, interaction usually takes place through a computer keyboard terminal, between runs, or in an on-line stop-go mode. An often-utilized system combines the speed and interaction possibilities of an analog computer with the accuracy and programming flexibility of a digital computer. This combination is specifically designed into the hybrid computer.

A modern analog-hybrid console (Fig. 2) contains the typical analog components plus a second patchboard area to include programmable, parallel logic circuits using high-speed gates for the functions of AND, NAND, OR, and NOR, as well as flip-flops, registers, and counters. The mode switches in the integrators are interfaced with the digital computer to permit fast iterations of dynamic runs under digital computer control. Data flow in many ways and formats between the analog computer with its fast, parallel circuits and the digital computer with its sequential, logic-controlled program (**Fig. 9**). Special high-speed analog-to-digital and digital-to-analog converters translate between the continuous signal representations of variables in the analog domain and the numerical representations of the digital computer. Control and logic signals are more directly compatible and require only level and timing compatibility. *See* ANALOG-TO-DIGITAL CONVERTER; BOOLEAN ALGEBRA; DIGITAL-TO-ANALOG CONVERTER.

Programming of hybrid models is more complex than described above, requiring the user to consider

the parallel action of the analog computer interlaced with the step-by-step computations progression in the digital computer. For example, in simulating a space mission, the capsule control dynamics will typically be handled on the analog computer in continuous form, but interfaced with the digital computer, where the navigational trajectory is calculated. *See* COMPUTER.

Per A. Holst

Bibliography. R. M. Howe, *Design Fundamentals of Analog Computer Components*, 1961; G. A. Korn, *Electronic Analog and Hybrid Computers*, 1971; D. M. MacKay et al., *Analog Computing at Ultra-High Speed*, 1962; R. Tomovic, *Repetitive Analog Computers*, 1958.

Analog states

States in neighboring nuclear isobars that have the same total angular momentum, parity, and isotopic spin. They also have nearly identical nuclear structure wave functions except for the transformation of one or more neutrons into an equivalent number of protons, which occupy the same single-particle states as the neutrons. Analog states (or isobaric analog states, IAS) have been observed throughout the periodic table, indicating that isotopic spin is a good quantum number. *See* ANGULAR MOMENTUM; I-SPIN; ISOBAR (NUCLEAR PHYSICS); NUCLEAR STRUCTURE; PARITY (QUANTUM MECHANICS).

Since the nucleon–nucleon interaction has been found to be approximately charge-independent, it is possible to consider protons and neutrons as representing different charge states of a single particle, that is, a nucleon. Thus, a level (commonly referred to as a parent state) in a nucleus with Z protons and N neutrons can be expected to have an analog in the neighboring isobar with $Z + 1$ protons and $N - 1$ neutrons (and the same total number of nucleons, $A = Z + N$), where the protons and neutrons occupy the same orbits as those in the parent state. The energy difference between the parent and analog states predominantly arises from the increased contribution from the electrostatic Coulomb interaction to the total energy arising from the extra proton in the analog state. From this amount must be subtracted the neutron-proton mass difference of 0.782 MeV (energies are given on the atomic mass scale). The agreement between such calculated energies of analog states and their measured values is in general fairly precise but not exact. The reason is that small additional factors influence the level energies, such as electromagnetic effects, a small charge-dependent nuclear interaction, isospin mixing, and nuclear structure effects.

Isotopic spin. Since the nucleon has spin $s = {}^1/_2$, it obeys Fermi statistics. Charge independence of nuclear forces then implies that systems of two nucleons that can be described by the same wave functions that are antisymmetric in both space and spin would have identical energies. There are three such combinations of two-nucleon systems that can satisfy these conditions, namely, neutron-neutron, proton-proton, and neutron-proton. The isotopic spin (also called isobaric spin or isospin) quantum number $t = {}^1/_2$ has been introduced to distinguish between the two states of the nucleon, and provides a means to further identify these three states. The neutron is associated with a projection of t along the direction of quantization, that is, the z axis, with $t_z = {}^1/_2$, and the proton with $t_z = -{}^1/_2$. The total z-projection of isospin, T_z, is given by the sum of the t_z of the individual nucleons. Hence, $T_z = 1, 0, -1$ for the neutron-neutron, neutron-proton, and proton-proton systems, respectively, and these represent an isobaric triplet with each system having total isospin $T = 1$. Only one two-nucleon system exists with a wave function that is symmetric in space and spin, that is, the neutron-proton combination for which $T = T_Z = 0$. For a nucleus with N neutrons and Z protons, $T_z = {}^1/_2 (N - Z)$ and, as with angular momentum, the value of isospin T must be equal to or greater than its z-projection, T_z. Analog states that differ solely in the interchange of either one neutron or proton (neighboring isobars) have values of T_z that differ by ± 1. Those that differ by the interchange of two neutrons or protons have T_z that differ by ± 2, and so forth. *See* EXCLUSION PRINCIPLE; FERMI-DIRAC STATISTICS.

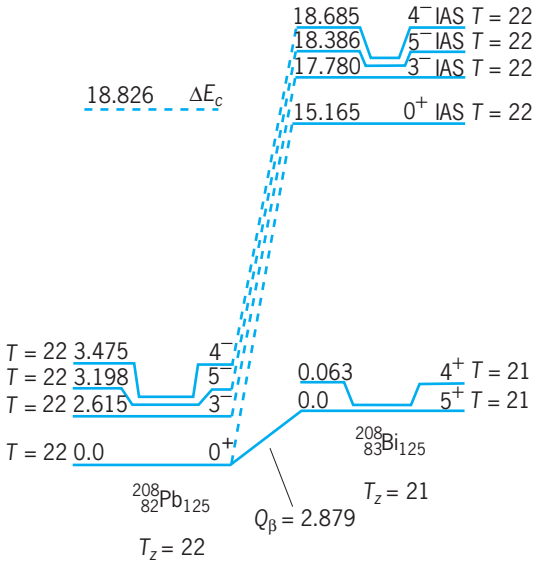
Nuclei with N = Z. Nuclei with $N = Z$ (equal numbers of neutrons and protons) have $T_z = 0$. Hence, there can be no analog for states in such nuclei that also have total isospin $T = 0$. This prohibition follows because, as noted above, T_z changes by ± 1 between neighboring isobars which must then have a total isospin T equal to or greater than 1. As a result, there is no analog of the ground state for these nuclei. However, excited states in $N = Z$ nuclei for which T is greater than zero can have analogs.

Nuclei with N > Z. Nuclei with $N > Z$ (neutron excess) have $T_z = {}^1/_2(N - Z)$, which is equal to or greater than ${}^1/_2$. Therefore, all states in these nuclei have $T \geq {}^1/_2$, and in principle, they can thus have analogs. For heavy nuclei where the Coulomb energy becomes quite large, a question arose as to whether this large energy might destroy the isobaric symmetry. However, this question was resolved in 1961 with the observation of analog states in heavy nuclei by experiments that utilized the (p, n) reaction. *See* NUCLEAR REACTION.

Coulomb displacement energy. The energy required to transform a neutron into a proton in a nucleus with Z protons and N neutrons is called the Coulomb displacement energy, ΔE_c. This energy can be calculated in terms of the energies of the parent and analog states from Eq. (1), where M_{Z+1} is the energy of the

$$\Delta E_c = M_{Z+1} - M_Z + \Delta_{np} \tag{1}$$

analog state, M_Z the energy of the parent state, and $\Delta_{np} = 0.782$ MeV (as noted above) is the neutron-proton mass difference (all energies are given on the atomic mass scale). The difference in the analog energies, that is, $M_{Z+1} - M_Z$, can be expressed in terms of the Q-value for beta decay between the ground states of the isobars, and the energy of excitation of the analog state in the $Z + 1$ nucleus. From analyses of the abundant data on analog states



Partial level diagram of the ^{208}Pb and ^{208}Bi isobar system. Broken lines connect parent–analog state pairs. Energies of excited states are given relative to the ground states, respectively, in megaelectronvolts. The isospin, *T*, and the angular momentum and parity, J^{π}, are shown for each level.

throughout the periodic table, it is found that the Coulomb displacement energy can be represented to a high degree of accuracy by Eq. (2), where $\bar{Z} = Z + {}^1/_2$. The Coulomb displacement energy represents the minimum energy of excitation in the parent nucleus for which states with $T = T_Z + 1$ can exist.

$$\Delta E_c = 1.412(\bar{Z}/A^{1/3}) - 0.861 \text{ (MeV)} \qquad (2)$$

Mass 208 system. A partial level diagram for the ^{208}Pb and ^{208}Bi isobar system demonstrates the energy relationships discussed above (see **illus.**). In such a diagram, only a few of the many observed levels are shown. Below an excitation energy of $\Delta E_c = 18.826$ MeV in ^{208}Pb, all levels have $T = T_z = 22$; the Coulomb displacement energy ΔE_c is the minimum excitation at which $T = 23$ states can occur in ^{208}Pb. The energy difference between the ^{208}Pb and ^{208}Bi ground states is $Q_\beta = 2.879$ MeV. The *z*-component of isospin for ^{208}Bi is $T_z = 21$, which is also the total isospin, *T*, for all levels up to an excitation energy of 15.165 MeV. At this energy, a level (actually a resonance) is observed in ^{208}Bi which possesses characteristics that are similar to those of the ground state of ^{208}Pb, and is identified as the analog. The energy of excitation of the analog state lies in the nuclear continuum of the ^{208}Bi mass system, that is, well above the energy at which ^{208}Bi is unstable to particle decay. As a result, the analog state is accessible only by means of some nuclear reaction. It was first identified in low-energy studies of proton scattering on ^{207}Pb, and later observed in the (*p, n*) reaction on ^{208}Pb. Schematically, in the (*p, n*) reaction an energetic proton impinges upon a ^{208}Pb nucleus and knocks out one of the excess neutrons that occupy the single-particle orbits that lie above those occupied by the protons, and the incident proton is captured into an equivalent orbit. This event is known as a charge-exchange reaction (here, a neutron is transformed into a proton), and is related to the inverse reaction that occurs in positron beta decay. *See* RADIOACTIVITY.

Fermi strength. Analog states are connected by either pure Fermi or mixed Fermi and Gamow-Teller transitions. The selection rules for Fermi transitions are $\Delta J = 0$, $\Delta\pi = 0$, $\Delta T = 0$, and $\Delta T_z = \pm 1$, where *J* represents the total angular momentum and π represents the parity. Those for Gamow-Teller transitions are $\Delta J = 0$ or ± 1 (but transitions between analogs with $J = 0$ are forbidden), $\Delta\pi = 0$, $\Delta T = 0$ or ± 1, and $\Delta T_z = \pm 1$. Both types of transitions require $\Delta L = 0$, where *L* represents the total orbital angular momentum. *See* SELECTION RULES (PHYSICS).

Of special interest in the study of nuclear structure is the distribution of strength for various types of transitions based upon the ground state. For Fermi transitions, all of the strength is concentrated in the transition between the analog states. The total strength is proportional to the isospin matrix element between the analog states, which is given by Eq. (3), where $|TT_z\rangle$ represents the isospin part of the parent wave function, and T_- is an operator which changes a neutron into a proton. Since $T = T_z$, the total Fermi strength is proportional to $(N - Z)$, which is just equal to the number of excess neutrons. This outcome reflects the fact that each of the excess neutrons occupies an orbit (with specific space and spin descriptions) that lies at an energy above those occupied by the protons, and thus there is equal probability for each excess neutron to be transformed into a proton and to participate in the Fermi transition.

$$|\langle TT_z - 1\,|T_-|\,TT_z\rangle|^2 = (T + T_z)(T - T_z + 1) \qquad (3)$$

Since the strength is well defined in practice, Fermi transitions can be used to obtain information pertaining to the reaction mechanisms that effect them. For beta decay, this approach usually involves determination of the vector coupling constant, G_V, as well as fine details about the nuclear wave functions. In charge-exchange studies, knowledge is sought pertaining to details of the effective nucleon–nucleus interaction and, to some extent, details of the overlap between the spatial wave functions of the parent and analog states.

Other applications. The study of analog states provides important information used to test nuclear theories. For example, the double charge-exchange reactions (π^+, π^-) [where π^+ and π^- represent a pion with positive and negative charge, respectively] have been used to identify double isobaric analog states, that is, analogs in isobars removed by 2 charge units with $T_z = T - 2$. Such data have been useful for testing various formulas for predicting the relative masses of isobaric multiplets. Single and double charge-exchange reactions utilizing incident pions have also been used to investigate giant resonances built upon analog states. The single-particle structure of parent states can be studied by

observing the particle decay of the analog state when the decay resides in the nuclear continuum. Measurements of the widths of analog states provide information pertaining to their fragmentation, for example, their mixing with states of the same spin and parity but with total isospin lower by one unit, that is, equal to $T - 1$. *See* GIANT NUCLEAR RESONANCES.

Daniel J. Horen

Bibliography. M. S. Antony, J. Britz, and A. Pape, Coulomb displacement energies between analog levels for $44 \leq A \leq 239$, *Atom. Nucl. Data Tables*, 40:9–56, 1988; M. S. Antony et al., Isobaric mass equation for $A = 1$–45 and systematics of Coulomb displacement energies, *Atom. Nucl. Data Tables*, 33:447–478, 1985; A. Bohr and B. R. Mottelson, *Nuclear Structure*, vol. 1, 1969, vol. 2, 1975; D. H. Wilkinson (ed.), *Isospin in Nuclear Physics*, 1969.

Analog-to-digital converter

A device for converting the information contained in the value or magnitude of some characteristics of an input signal, compared to a standard or reference, to information in the form of discrete states of a signal, usually with numerical values assigned to the various combinations of discrete states of the signal.

Analog-to-digital (A/D) converters are used to translate analog information, such as audio signals or measurements of physical variables (for example, temperature, force, or shaft rotation), into a form suitable for digital handling, which might involve any of these operations: (1) processing by a computer or by logic circuits, including arithmetical operations, comparison, sorting, ordering, and code conversion, (2) storage until ready for further handling, (3) display in numerical or graphical form, and (4) transmission.

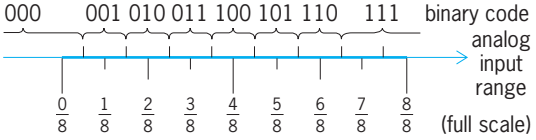


Fig. 1. A three-bit binary representation of a range of input signals.

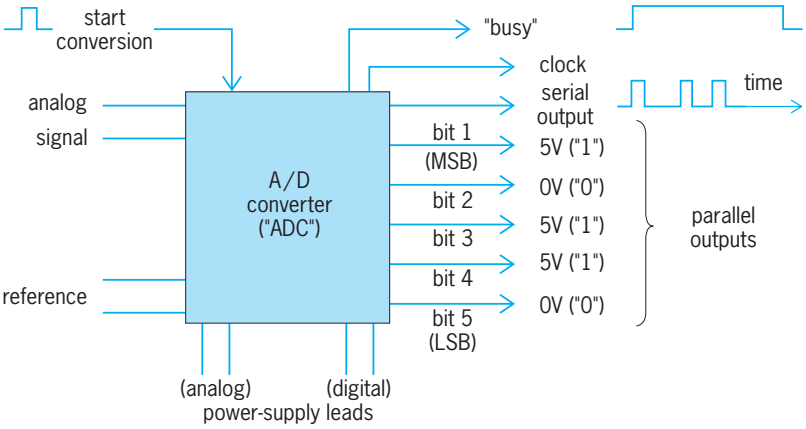


Fig. 2. Basic diagram of an analog-to-digital converter, showing parallel and serial (return-to-zero) output formats for code 10110.

If a wide-range analog signal can be sampled and converted, with sufficient frequency, to an appropriate number of two-level digits, or bits, to represent each sample or snapshot, the digital representation of the signal can be transmitted through a noisy medium without essential degradation of the fine structure of the original signal. *See* COMPUTER GRAPHICS; DATA COMMUNICATIONS; DIGITAL COMPUTER.

Concepts and structure. Conversion involves quantizing and encoding. Quantizing means partitioning the analog signal range into a number of discrete quanta and determining to which quantum the input signal belongs. Encoding means assigning a unique digital code to each quantum and determining the code that corresponds to the input signal. The most common system is binary, in which there are 2^n quanta (where n is some whole number), numbered consecutively; the code is a set of n physical two-valued levels or bits (1 or 0) corresponding to the binary number associated with the signal quantum. [The term binary is used here in two senses: a binary (radix-2) numbering system, and the binary (two-valued) decisions forming each bit.] *See* BIT.

Figure 1 shows a typical three-bit binary representation of a range of input signals, partitioned into eight (that is, 2^3) quanta. For example, a signal in the vicinity of 3/8 full scale (between 5/16 and 7/16) will be coded 011 (binary 3). *See* NUMBERING SYSTEMS.

Conceptually, the conversion can be made to take place in any kind of medium: electrical, mechanical, fluid, optical, and so on (for example, shaft-rotation-to-optical); but by far the most commonly employed form of A/D converters comprises those devices that convert electrical voltages or currents to coded sets of binary electrical levels (for example, +5 V or 0 V) in simultaneous (parallel) or pulse-train (serial) form, as shown in **Fig. 2**. Serial outputs are not always made available in the form of binary numbers.

The converter depicted in Fig. 2 converts the analog input to a five-digit "word." If the coding is binary, the first digit (most significant bit, abbreviated MSB) has a weight of 1/2 full scale, the second 1/4 full scale, and so on, down to the nth digit (least-significant bit, abbreviated LSB), which has a weight of 2^{-n} of full scale (1/32 in this example). Thus, for the output word shown, 10110, the analog input must be given approximately by the equation below.

$$\frac{16}{32} + \frac{0}{32} + \frac{4}{32} + \frac{2}{32} + \frac{0}{32} = \frac{22}{32} = \frac{11}{16}\text{FS (full scale)}$$

The number of bits, n, characterizes the resolution of a converter. The **table** translates bits into other conventional measures of resolution in a binary system.

Figure 2 also shows a commonly used configuration of connections to an A/D converter: the analog signal and reference inputs; the parallel and serial

Binary resolution equivalents*

Bit	2^{-n}	$1/2^n$ (fraction)	dB (decibels)	$1/2^n$ (decimal)	%	Parts per million
FS†	2^0	1	0	1.0	100	1,000,000
MSB‡	2^{-1}	1/2	−6	0.5	50	500,000
2	2^{-2}	1/4	−12	0.25	25	250,000
3	2^{-3}	1/8	−18.1	0.125	12.5	125,000
4	2^{-4}	1/16	−24.1	0.0625	6.2	62,500
5	2^{-5}	1/32	−30.1	0.03125	3.1	31,250
6	2^{-6}	1/64	−36.1	0.015625	1.6	15,625
7	2^{-7}	1/128	−42.1	0.007812	0.8	7,812
8	2^{-8}	1/256	−48.2	0.003906	0.4	3,906
9	2^{-9}	1/512	−54.2	0.001953	0.2	1,953
10	2^{-10}	1/1024	−60.2	0.0009766	0.1	977
11	2^{-11}	1/2048	−66.2	0.00048828	0.05	488
12	2^{-12}	1/4096	−72.2	0.00024414	0.024	244
13	2^{-13}	1/8192	−78.3	0.00012207	0.012	122
14	2^{-14}	1/16,384	−84.3	0.000061035	0.006	61
15	2^{-15}	1/32,768	−90.3	0.0000305176	0.003	31
16	2^{-16}	1/65,536	−96.3	0.0000152588	0.0015	15
17	2^{-17}	1/131,072	−102.3	0.00000762939	0.0008	7.6
18	2^{-18}	1/262,144	−108.4	0.000003814697	0.0004	3.8
19	2^{-19}	1/524,288	−114.4	0.000001907349	0.0002	1.9
20	2^{-20}	1/1,048,576	−120.4	0.0000009536743	0.0001	0.95

* From D. H. Sheingold (ed.), *Analog-Digital Conversion Handbook,* 3d ed., Analog Devices, Inc., 1986.
† Full scale.
‡ Most significant bit.

digital outputs; the leads from the power supply, which provides the required energy for operation; and two control leads—a start-conversion input and a status-indicating output (busy) that indicates when a conversion is in progress. Serial-output converters often furnish a clock output pulse to indicate the time at which a bit is available. The reference voltage or current is often developed within the converter.

Second in importance to the binary code and its many variations is the binary-coded decimal (BCD), which is used when the encoded material is to be displayed or printed in numerical form. In BCD, each digit of a radix-10 number is represented by a four-digit binary subgroup. For example, the BCD code for 379 is 0011 0111 1001. The output of the A/D converter used in digital panel meters is usually BCD.

Techniques. There are many techniques used for A/D conversion, ranging from simple voltage-level comparators to sophisticated closed-loop systems, depending on the input level, output format, control features, and the desired speed, resolution, and accuracy. The three most popular techniques, used individually or in combination, are direct or flash conversion; integrating conversion, exemplified by dual slope; and feedback conversion, exemplified by the method of successive approximations.

Comparator. The most elementary A/D converter is the comparator, a 1-bit (2-value) converter. It has two inputs, the signal and the reference: whenever the signal is greater than the reference, the output is 1; when less, the output is 0. A digitally controlled latch may be included to lock in the value of the comparison at a definite time.

Direct conversion. The fastest, and conceptually simplest, approach to n-bit conversion is via the direct or flash converter. It quantizes by using $2^n - 1$ comparators to compare the input signal simultaneously to a set of reference signals representing all amplitude quanta. Then it employs a logical scheme called a priority encoder to determine the parallel digital output corresponding to a given set of comparator output states.

Figure 3 is a simplified block diagram of a 2-bit (4-level) flash converter. The resistor string acts as a

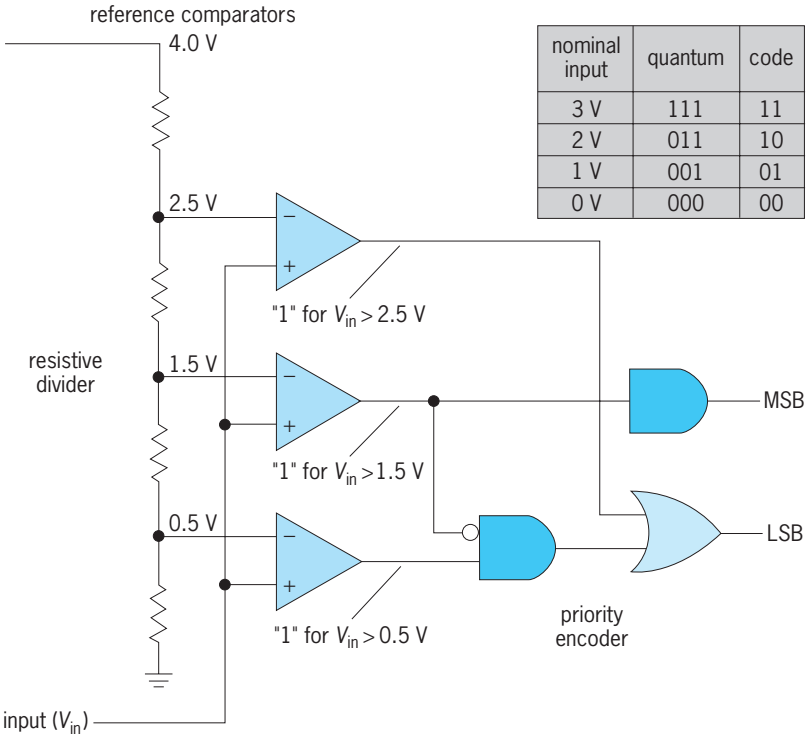


Fig. 3. Diagram of a 2-bit flash converter.

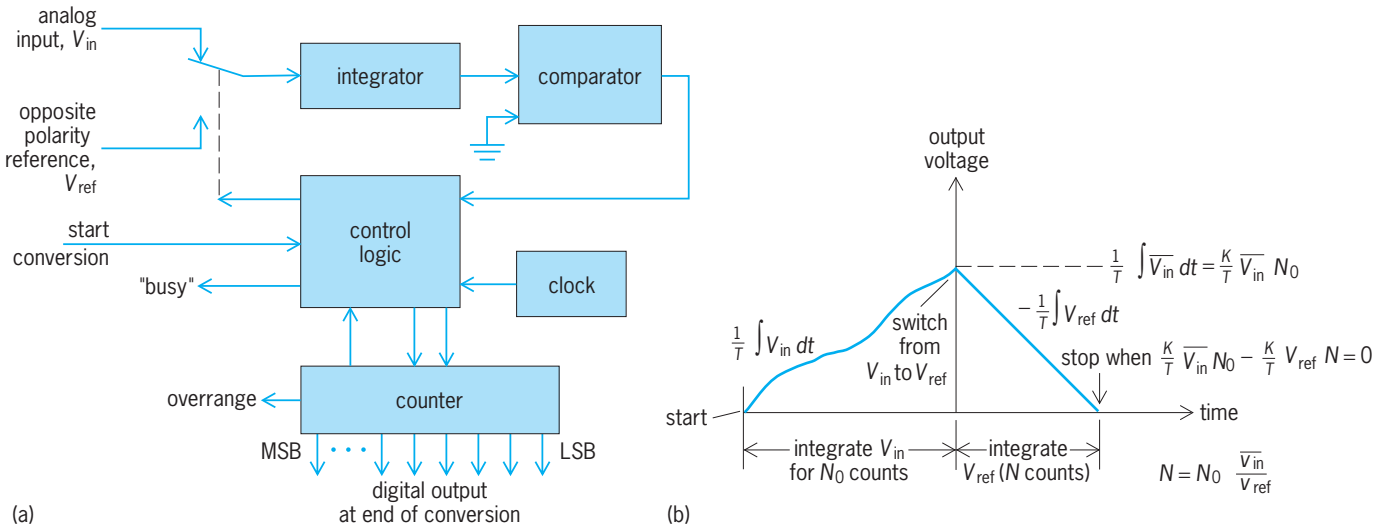


Fig. 4. Example of a dual-slope conversion. (*a*) Block diagram of converter. (*b*) Integrator output. Here, *k* is a constant, and *T* is the *RC* (resistor-capacitor) time constant of the integrator.

voltage divider to provide reference voltages for the three (that is, $2^2 - 1$) comparators. When the common input is below the lowest reference level, all of the comparators have 0 output. When the input is greater than the reference for the lowest comparator, its output switches to 1; when the next comparator's reference is exceeded, its output too switches to 1; and when the third comparator's reference level is exceeded, all three comparators are at 1. Thus, the length of the series of 1's identifies the input voltage's quantum.

The quantum must now be encoded as a 2-bit digital word at the output. In the priority-encoder scheme shown in Fig. 3, the output is a binary word, and the correspondence between quantum and code is given in the table.

Practical flash converters require lengthy resistor strings, large numbers of comparators, and rather complicated encoders. For example, a 10-bit flash converter calls for 1023 comparators and requires encoding 1024 levels into a 10-bit word. Nevertheless, in integrated-circuit form, all of the circuity exists on a single monolithic silicon chip.

Integrating conversion. Unlike flash converters, which measure the input instant by instant, integrating converters use analog integrating circuits in a variety of circuit architectures to convert average values of the input into trains of pulses, which are then encoded by digital counters or processors. Integrating converters have high resolution and low noise sensitivity, and the train of pulses can be transmitted as a high-level noise-insensitive signal for encoding

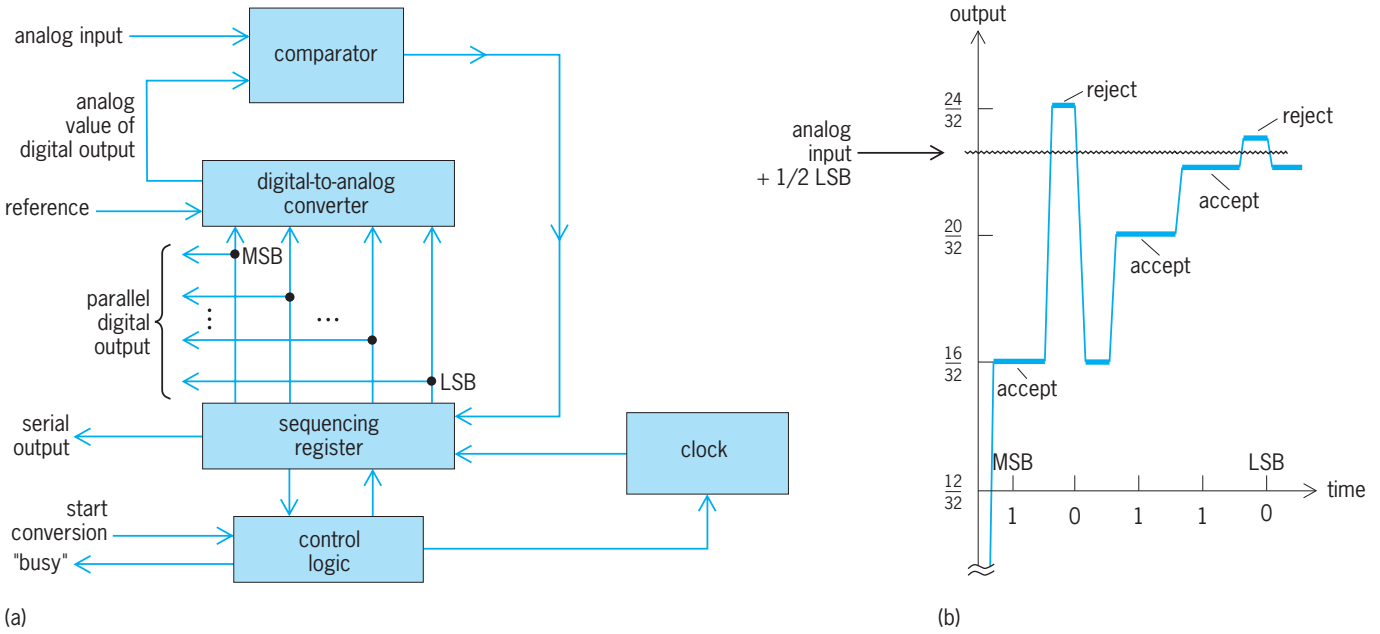


Fig. 5. Successive-approximations conversion. (*a*) Block diagram of converter. (*b*) Digital-to-analog converter output for the example in Fig. 2.

at a remote location. The class includes voltage-to-frequency converters, delta-modulation types, and dual-slope converters.

The dual-slope converter, discussed here as an example of the class, has long been popular in digital voltmeters; it can perform complete conversions with high accuracy at low speeds, generally a few conversions per second. **Figure 4***a* is a simplified block diagram of a dual-slope converter. The input is integrated for a period of time determined by a clock-pulse generator and counter (Fig. 4*b*). The final value of the signal integral becomes the initial condition for integration of the reference in the opposite sense, while the clock output is counted. When the net integral is zero, the count stops. Since the integral "up" of the input over a fixed time (N_0 counts) is equal to the integral "down" of the fixed reference, the ratio of the number of counts of the variable period to that of the fixed period is equal to the ratio of the average value of the signal to the reference.

Feedback conversion. Some A/D converter types employ an *n*-bit digital-to-analog converter to compare an analog version of the digital output with the analog input; the decision (higher or lower) causes a change in the digital output to bring its value closer to the input value. The change may be made by a counter that increases or decreases the digital value in small steps, or by the educated guessing of a successive-approximations converter.

Successive-approximations conversion combines superior speed and accuracy; it is the principal technique used for data-acquisition and computer-interface systems. **Figure 5***a* is a simplified block diagram of a successive-approximations converter. In a manner analogous to the operation of an apothecary's scale with a set of binary weights, the input is weighed against a set of successively smaller fractions of the reference, produced by a digital-to-analog (D/A) converter that reflects the number in the output register. *See* DIGITAL-TO-ANALOG CONVERTER.

First, the MSB is tried (1/2 full scale). If the signal is less than the MSB, the MSB code is returned to zero; if the signal is equal to or greater than the MSB, the MSB code is latched in the output register (Fig. 5*b*). The second bit is tried (1/4 full scale). If the signal is less than 1/4 or 3/4, depending on the previous choice, bit 2 is set to zero; if the signal is equal to or greater than 1/4 or 3/4, bit 2 is retained in the output register. The third bit is tried (1/8 full scale). If the signal is less than 1/8, 3/8, 5/8, or 7/8, depending on previous choices, bit 2 is set to zero; otherwise, it is accepted. The trial continues until the contribution of the least-significant bit has been weighed and either accepted or rejected. The conversion is then complete. The digital code latched in the output register is the digital equivalent of the analog input signal.

Physical electronics. The earliest A/D converters were large rack-panel chassis-type modules using vacuum tubes, requiring about 1.4 ft^3 (0.04 m^3) of space and many watts of power. Since then, they have become smaller in size and cost, evolving through circuit-board, encapsulated-module, and hybrid construction, with improved speed and resolution. Single-chip 12-bit A/D converters with the ability to interface with microprocessors are now available in small integrated-circuit packages. Integrated-circuit A/D converters, with 8-bit and better resolution and conversion rates of hundreds of megahertz, are also commercially available. *See* MICROPROCESSOR.

Daniel H. Sheingold

Bibliography. Analog Devices, Inc. Engineering Staff, *Analog-Digital Conversion Handbook*, ed. by D. H. Sheingold, 3d ed., 1986; M. Demler, *High-Speed Analog to Digital Conversion*, 1991; D. F. Hoeschele, Jr., *Analog-to-Digital and Digital-to-Analog Conversion Techniques*, 2d ed., 1994; R. Van de Plassche, *Integrated Analog-to-Digital and Digital-to-Analog Converters*, 1994.

Analysis of variance

Total variation in experimental data is partitioned into components assignable to specific sources by the analysis of variance. This statistical technique is applicable to data for which (1) effects of sources are additive, (2) uncontrolled or unexplained experimental variations (which are grouped as experimental errors) are independent of other sources of variation, (3) variance of experimental errors is homogeneous, and (4) experimental errors follow a normal distribution. When data depart from these assumptions, one must exercise extreme care in interpreting the results of an analysis of variance. Statistical tests indicate the contribution of the components to the observed variation.

In an illustrative experiment, t methods of treatment are under study, and n samples are measured for each treatment for a total of nt samples. Measurement X_{ij} of the ith sample that received the jth treatment records an overall effect μ, an effect β_j produced by the jth treatment, and an effect ϵ_{ij} produced by experimental error. The three effects are additive, so that Eq. (1) holds, where $i = 1, \ldots, n$;

$$X_{ij} = \mu + \beta_j + \epsilon_{ij} \tag{1}$$

and $j = 1, \ldots, t$. The statistical problem is to test for the existence of these effects.

The analysis of variance in this example is presented in the **table**. Entries in the sum of squares column represent that part of the total variation that is attributable to each source. Total sum of squares Q is the sum over all squared deviations of observations X_j from the grand mean $\overline{X}$, Eq. (2). Similarly,

$$\overline{X} = (\Sigma_i X_{ij})/nt \tag{2}$$

within treatments, sum of squares E is the sum over all squared deviations of observations X_{ij} within a treatment from the mean $\overline{X}_j$ of that treatment, Eq. (3).

$$\overline{X}_j = (\Sigma_i X_{ij})/n \tag{3}$$

Also, between treatments, sum of squares T is n times

Illustrative example of the analysis of variance

Source of variation	Sum of squares	Degrees of freedom	Mean square
Between treatments	$T = n\Sigma_j(\overline{X}_j - \overline{X})^2$	$t - 1$	$T = T/(t - 1)$
Within treatments	$E = \Sigma_{i,j}(X_{ij} - \overline{X}_j)^2$	$t(n - 1)$	$E' = E/t(n - 1)$
Total	$Q = \Sigma_{i,j}(X_{ij} - \overline{X})^2$	$nt - 1$	

the sum over all treatments of the squared deviations of treatment means $\overline{X}_j$ from a grand mean $\overline{X}$ as defined by Eqs. (2) and (3). The sum of squares is generally computed more easily from the equivalent formulas (4)–(6).

$$Q = \Sigma_{i,j}X_{ij}^2 - \frac{(\Sigma_{i,j}X_{ij})^2}{nt} \tag{4}$$

$$T = \Sigma_j\frac{(\Sigma_i X_{ij})^2}{n} - \frac{(\Sigma_{i,j}X_{ij})^2}{nt} \tag{5}$$

$$E = Q - T \tag{6}$$

The entries under degrees of freedom represent the number of independent comparisions upon which the sum of squares for the source of variation is based. In every case the linear restriction imposed by the relationship of the particular mean to the observations results in the loss of one degree of freedom. Therefore the number of degrees of freedom is always one less than the number of deviations used to compute the sum of squares.

The mean squares in the analysis of variance are obtained by dividing the sum of squares by the corresponding degrees of freedom. The within-treatments mean square is an estimate of σ^2, the variance of the error term ϵ_{ij} in the additive model. It represents the random or unexplained variation in the data. The between-treatments mean square is an estimate of $\sigma^2 + \sigma_\beta^2$, where σ_β^2 is the variance of the treatment effects β_j.

If the treatment means differ substantially, the β_j effects estimated by $(\overline{X}_j - \overline{X})$ will differ correspondingly and will have a large variance σ_β^2. If on the other hand the means do not differ, the treatment effects β_j would be zero and σ_β^2 would be zero. In this case the treatment mean square would be equal to the error mean square and both would be independent estimates of σ^2. By comparing the ratio T'/E' of between treatment mean square T' to within-treatment mean square E' with unity, the variation due to treatments is compared with the variation due to random or unexplained factors. If this ratio, called the F ratio, is close to unity, there is no evidence of a treatment effect. However, if ratio T'/E' is substantially greater than unity there may be a significant treatment effect.

To compare the mean squares objectively, one uses the F test of significance in which the statistical hypothesis is that $\sigma_\beta^2 = 0$. Under this hypothesis it can be concluded that the treatment effects are significantly different from zero at the significance level α if the calculated F ratio is greater than the value of F at the α point on the F distribution with $t - 1$ and $t(n - 1)$ degrees of freedom. *See* BIOMETRICS; QUALITY CONTROL; STATISTICS. Robert L. Brickley

Bibliography. D. R. Anderson, D. F. Sweeny, and T. A. Williams, *Introduction to Statistics*: *Concepts and Applications*, 3d ed., 1993; S. Jarrell, *Basic Statistics*, 1994; N. A. Weiss, *Introductory Statistics*, 5th ed., 1999.

Analytic geometry

A branch of mathematics in which algebra is applied to the study of geometry. Because algebraic methods were first systematically applied to geometry in 1637 by the French philosopher-mathematician René Descartes, the subject is also called cartesian geometry. The basis for an algebraic treatment of geometry is provided by the existence of a one-to-one correspondence between the elements, "points" of a directed line g, and the elements, "numbers," that form the set of all real numbers. Such a correspondence establishes a coordinate system on g, and the number corresponding to a point of g is called its coordinate. The point O of g with coordinate zero is the origin of the coordinate system. A coordinate system on g is cartesian provided that for each point P of g, its coordinate is the directed distance $\overline{OP}$. Then all points of g on one side of O have positive coordinates (forming the positive half of g) and all points on the other side have negative coordinates. The point with coordinate 1 is called the unit point. Since the relation $\overline{OP} + \overline{PQ} = \overline{OQ}$ is clearly valid for each two points P, Q, of directed line g, then $\overline{PQ} = \overline{OQ} - \overline{OP} = q - p$, where p and q are the coordinates of P and Q, respectively. Those points of g between P and Q, together with P, Q, form a line segment. In analytic geometry it is convenient to direct segments, writing PQ or QP accordingly as the segment is directed from P to Q or from Q to P, respectively. To find the coordinate of the point P that divides the segment P_1P_2 in a given ratio r, put $\overline{P_1P}/\overline{P_2P} = r$. Then $(x - x_1)/(x - x_2) = r$, where x_1, x_2, x are the coordinates of P_1, P_2, P, respectively, and solving for x gives $x = (x_1 - rx_2)/(1 - r)$. Clearly r is negative for each point between P_1, P_2 and is positive for each point of g external to the segment. The midpoint of the segment divides it in the ratio -1, and hence its coordinate $x = (x_1 + x_2)/2$. *See* MATHEMATICS.

The plane. Choose any two intersecting lines g_1, g_2 of the plane, with cartesian coordinate systems selected on each so that the intersection point has coordinate O in each system. To each point P of the

plane an ordered pair of numbers (x, y) is attached as coordinates, where x is the coordinate of the point of intersection of g_1 with the line through P parallel to g_2, and y is the coordinate of the point of intersection of g_2 with the line through P parallel to g_1. If P lies on g, its coordinates are $(x, 0)$, where x is the coordinate of P in the coordinate system on g_1. Similarly, each point of g_2 has coordinates $(0, y)$. The origin O has coordinates (0, 0). The lines g_1 and g_2 are called the x axis and y axis, respectively. It is usually convenient to take the same scale on each axis; that is, the segments joining the unit points on the two axes to the origin are congruent. The notation $P(x, y)$ denotes a point P with coordinates (x, y). If ω denotes the angle made by the positive halves of the two axes, and d the distance of points $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, application of the law of cosines yields Eq. (1).

$$d = \left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + 2(x_1 - x_2)(y_1 - y_2)\cos\omega\right]^{1/2} \quad (1)$$

Since $\cos 90^\circ = 0$, this important formula is simplified by taking the axes mutually perpendicular. Though it is occasionally useful to employ oblique axes (that is, $\omega \neq 90^\circ$), the simplifications resulting from a rectangular cartesian coordinate system make it the usual choice. Such a cartesian coordinate system is assumed in what follows. Thus, the distance d of $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ is given by Eq. (2).

$$d = [(x_1 - x_2)^2 + (y_1 - y_2)^2]^{1/2} \quad (2)$$

Let θ denote the smaller of the two angles that a line g makes with the positive half of the x axis, measured in the (positive) direction of rotation (that is, from the positive half of the x axis to the positive half of the y axis). Angle θ is called the slope angle of g, and the number $\lambda = \tan\theta$, the slope of g, plays an important role in plane analytic geometry. Slope is not defined for any line perpendicular to the x axis. If $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ are two distinct points of line g with slope angle $\theta \neq 90^\circ$, and Q denotes the intersection of the line through P_1 perpendicular to the y axis with the line through P_2 perpendicular to the x axis, then Eq. (3) holds.

$$\lambda = \tan\theta = \frac{\overline{QP_2}}{\overline{P_1Q}} = \frac{y_2 - y_1}{x_2 - x_1} \quad (3)$$

Loci and equations. The correspondence between the geometric entity "point" and the arithmetic entity "pair of real numbers," upon which plane analytic geometry is based, results in associating with each geometric locus one or more equations that are satisfied by the coordinates of all those (and only those) points forming the locus (equations of the locus), and in associating with each system of equations in the variables x, y the figure (graph of the equations) whose points are determined by the pairs of numbers satisfying the equations. Thus the algebraic method of studying geometry is balanced by a geometric interpretation of algebra. A central problem in analytic geometry is that of finding equations of certain important figures among curves and surfaces. *See* CURVE FITTING.

Equations of lines. A line may be determined by data of various kinds, each yielding a different form for its equation. Thus an equation for a line g through $P_1(x_1, y_1)$ perpendicular to the x axis is evidently $x = x_1$. If g goes through $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ and is not perpendicular to the x axis, let $P(x, y)$ be the coordinates of any other point of g. Then $(y - y_1)/(x - x_1) = \lambda = (y_2 - y_1)/(x_2 - x_1)$, whereas if (x, y) are the coordinates of a point not on g, $(y - y_1)/(x - x_1) \neq \lambda$. Hence an equation for g is shown as Eq. (4), from which follow Eqs. (5) and (6).

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1} \quad (4)$$

$$y - y_1 = \lambda(x - x_1) \quad (5)$$

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} \quad (6)$$

The two-point form is given by Eq. (4), the point-slope form by Eq. (5), and the symmetric form by Eq. (6). The validity of the determinant form,

$$\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0$$

as well as the slope-intercept form, $y = \lambda x + b$, and the intercept for $x/a + y/b = 1$, $a \cdot b \neq 0$, also follow readily from Eq. (1). *See* DETERMINANT.

A line g is determined by its distance p from the origin O ($p \geqq 0$) and the angle β that the perpendicular to g from O makes with the x axis. (The perpendicular is directed from O to g in case g does not go through O, and so as to make $\beta < 180^\circ$ in the contrary case.) The equation of g in terms of p and β (the so-called normal form) is $x\cos\beta + y\sin\beta - p = 0$. The directed distance from g to any point $P(x_0, y_0)$ is $x_0\cos\beta + y_0\sin\beta - p$. It follows that the equations of the bisectors of the angles formed by two lines

$$x\cos\beta_i + y\sin\beta_i - p_i = 0 \quad (i = 1, 2)$$

are

$$x\cos\beta_1 + y\sin\beta_1 - p_1 = \pm(x\cos\beta_2 + y\sin\beta_2 - p_2)$$

The general form of an equation of a line is $Ax + By + C = 0$, where A, B are not both zero. The normal form is obtained from the general form upon dividing the left-hand member by $\pm\sqrt{A^2 + B^2}$, choosing the sign of the radical opposite to that of C in case $C \neq 0$, the same as that of B in case $C = 0$ and $B \neq 0$, and the same as that of A in case $B = 0$ and $C = 0$. If $B = 0$, the general form reduces to $x =$ constant, a line perpendicular to the x axis, and if $B \neq 0$, one obtains $y = -(A/B)x + C/A$, a line with slope $-A/B$ and y intercept C/A. Thus to each line there corresponds a linear equation in x and y, and with each such equation is associated a line.

Angle between two lines. The angle ϕ ($0^\circ < \phi < 180^\circ$) from a line g_1 to another intersecting line g_2 is that through which g_1 must be rotated (about the point of intersection) to coincide with g_2. If θ_1, θ_2 are the slope angles of g_1, g_2, respectively, then

$$\phi = \theta_2 - \theta_1$$

and

$$\tan\phi = \frac{\tan\theta_2 - \tan\theta_1}{1 + \tan\theta_1 \tan\theta_2} = \frac{\lambda_2 - \lambda_1}{1 + \lambda_1 \cdot \lambda_2}$$

provided $\theta_1 \neq 90^\circ \neq \theta_2$. Consequently, g_1 and g_2 are mutually perpendicular if and only if $1 + \lambda_1 \cdot \lambda_2 = 0$. The formula for $\tan\phi$ holds in case g_1, g_2 are parallel. Then $\phi = 0^\circ$ and $\lambda_1 = \lambda_2$. It follows that two distinct lines $A_i x + B_i y + C_i = 0$, with $i = 1, 2$, are mutually perpendicular if and only if $A_1A_2 + B_1B_2 = 0$, and parallel provided $A_1B_2 - A_2B_1 = 0$.

Area of a triangle. Let $P_i(x_i, y_i)$, with $i = 1, 2, 3$, be vertices of a triangle whose area is denoted by A. Then $A = \frac{1}{2}d \cdot D$, where $d = [(x_1 - x_2)^2 + (y_1 - y_2)^2]^{1/2}$ and D is the distance of P_3 from the line joining P_1, P_2. Substituting the coordinates (x_3, y_3) of P_3 for (x, y) in the normal form of the equation of that line gives

$$D = \frac{\pm\left[(y_2 - y_1)x_3 - (x_2 - x_1)y_3 - x_1y_2 + x_2y_1\right]}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}$$

and hence

$$A = \pm(\tfrac{1}{2})(x_1y_2 + x_2y_3 + x_3y_1 - x_3y_2 - x_2y_1 - x_1y_3)$$

or

$$A = \pm(1/2) \cdot \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}$$

The positive sign holds provided the vertices P_1, P_2, P_3 are in counterclockwise order, and the negative sign in the contrary case.

Linear combinations. If $u_1 = 0$, $u_2 = 0$ are equations of lines through a point P, for every choice of constants c_1, c_2 (except $c_1 = c_2 = 0$) the linear combination $c_1u_1 + c_2u_2 = 0$ is an equation of a line through P, and every line through P has an equation of that form. It follows that three lines $u_i = 0$, with $i = 1, 2, 3$, are concurrent provided there exist constants c_1, c_2, c_3 (not all zero) such that the linear combination $c_1u_1 + c_2u_2 + c_3u_3 = 0$ for every pair of numbers (x, y). Putting $u_i = A_ix + B_iy + C_i$, with $i = 1, 2, 3$, then c_1, c_2, c_3 are nontrivial solutions of the system of equations $c_1A_1 + c_2A_2 + c_3A_3 = 0$, $c_1B_1 + c_2B_2 + c_3B_3 = 0$, $c_1C_1 + c_2C_2 + c_3C_3 = 0$. Hence the lines $u_i = 0$, with $i = 1, 2, 3$, are concurrent if and only if

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} = 0$$

Circle. By use of the formula for the distance of two points and the definition of a circle, an equation for a circle with center $C(x_0, y_0)$ and radius r ($r \geqq 0$) is found to be $(x - x_0)^2 + (y - y_0)^2 = r^2$. If $r = 0$, the only (real) point of the locus is the center (x_0, y_0) and the circle is a point circle. The above equation is called the standard form. Expansion yields $x^2 + y^2 - 2x_0x - 2y_0y - x_0^2 + y_0^2 - r^2 = 0$, a quadratic in x, y with equal, nonzero, coefficients of x^2 and y^2, and with the product term xy lacking. Conversely, the locus of each such equation $A(x^2 + y^2) + 2Dx + 2Ey + F = 0$, $A \neq 0$, is a (real) circle with center $(-D/A, -E/A)$ and radius $r = (1/A) \cdot [D^2 + E^2 - AF]^{1/2}$ provided $D^2 + E^2 - AF \geqq 0$, for by "completing the square" the above equation may be put in the standard form. Let $P(x_1, y_1)$ be on the circle $(x - x_0)^2 + (y - y_0)^2 = r^2$. An equation for the tangent to the circle at P is easily seen to be $(y_1 - y_0)(y - y_1) + (x_1 - x_0)(x - x_1) = 0$. Adding $(x_1 - x_0)^2 + (y_1 - y_0)^2$ to the left side of this equation, and its equal r^2 [since $P(x_1, y_1)$ lies on the circle] to the right side, yields $(x_1 - x_0)(x - x_0) + (y_1 - y_0)(y - y_0) = r^2$ as an equation of the tangent. The formal process by which this equation may be obtained from the standard form of the equation of a circle is called polarization. If $P(x_1, y_1)$ is not on the circle, the line that is the locus of that equation is called the polar line of P with respect to the circle. If P is outside the circle, its polar line joins the contact points of the two tangents from P to the circle. *See* CIRCLE.

Polarization of the general equation of a circle gives $A(x_1x + y_1y) + D(x + x_1) + E(y + y_1) + F = 0$, as an equation of the tangent at (x_1, y_1) if that point is on the circle, and of the polar line of (x_1, y_1) otherwise. The tangential distance t from (x_1, y_1) to the circle $(x - x_0)^2 + (y - y_0)^2 = r^2$ is given by $t^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2 - r^2$. Hence $(x - x_0)^2 + (y - y_0)^2 - r_0^2 = (x - x_1)^2 + (y - y_1)^2 - r_1^2$ is an equation of the locus of points with equal tangential distances from the two circles with centers (x_0, y_0), (x_1, y_1) and radii r_0, r_1, respectively. Since the equation is linear, it represents a line which evidently contains the common chord of the circles in case they intersect in two distinct points. If $u_i = 0$, with $i = 1, 2, 3$, denotes equations of three circles that intersect pairwise in two distinct points, then the three common chords are concurrent, since the equation $u_1 - u_2 = 0$ is a linear combination of the equations $u_2 - u_3 = 0$, $u_3 - u_1 = 0$. An equation of the circle through three noncollinear points $P_i(x_i, y_i)$, with $i = 1, 2, 3$, may be written

$$\begin{vmatrix} x^2 + y^2 & x & y & 1 \\ x_1^2 + y_1^2 & x_1 & y_1 & 1 \\ x_2^2 + y_2^2 & x_2 & y_2 & 1 \\ x_3^2 + y_3^2 & x_3 & y_3 & 1 \end{vmatrix}$$

Conic sections. Much of plane analytic geometry deals with a class of curves which (from the way in which they were first studied) are known as conic sections or conics. A conic is the locus of a point P that moves so that its distance from a fixed point F (the focus) is in a constant positive ratio ϵ (the

eccentricity) to its distance from a fixed line (the directrix) not through F. Let $(c, 0)$ be the coordinates of F, $c > 0$, and take the y axis as the directrix. Then $P(x, y)$ satisfies the equation $[(x - c)^2 + y^2]^{1/2} = \epsilon x$; that is, $(1 - \epsilon^2)x^2 + y^2 - 2cx + c^2 = 0$, and it is easily seen that each point whose coordinates satisfy this equation is on the conic. Hence each conic is represented by a second-degree equation in the cartesian coordinates (x, y). A conic is called a parabola, ellipse, or hyperbola accordingly as $\epsilon = 1$, $\epsilon < 1$, $\epsilon > 1$, respectively. *See* CONIC SECTION.

Parabola. When $\epsilon = 1$, the equation for the conic obtained above becomes $y^2 - 2cx + c^2 = 0$ or $y^2 = 2c(x - c/2)$. The curve has an axis of symmetry (the x axis when the focus and directrix are chosen as above). The point $(c/2, 0)$ at which the axis intersects the curve is the vertex. Putting $x = c$ in the equation, it is seen that the chord through the focus that is perpendicular to the axis (the latus rectum) has length $2c$. All quadratics, $x = Ay^2 + By + C$, $y = Ax^2 + Bx + C$, $A \neq 0$, are equations of parabolas (with axes perpendicular to the y axis, or the x axis, respectively). All other parabolas have equations $Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$, with discriminant

$$\begin{vmatrix} A & B & D \\ B & C & E \\ D & E & F \end{vmatrix} \neq 0 \quad \text{and} \quad AC - B^2 = 0$$

(A quadratic with nonvanishing discriminant is called irreducible.) A simple standard form of an equation of a parabola is obtained by taking $(c/2, 0)$ for focus $(c > 0)$ and the line $x = -c/2$ for directrix, resulting in $y^2 = 2cx$. Polarizing gives $y_1y = c(x + x_1)$ for an equation of the tangent at the point $P(x_1, y_1)$ on the parabola. This tangent cuts the axis at the point $Q(-x_1, 0)$ whose distance from the focus is $x_1 + c/2$. But this is the distance from the directrix to P, and consequently the triangle FPQ is isosceles with $\angle FQP = \angle FPQ$. It follows that the line joining F to any point P of the parabola makes the same angle with the tangent at P as does the line through P that is parallel to the axis of the parabola. Thus each light ray emanating from the focus is reflected parallel to the axis. *See* PARABOLA.

Ellipse. When $\epsilon < 1$, the equation of the general conic given above becomes

$$\left(\frac{x - c}{1 - \epsilon^2}\right)^2 + \frac{y^2}{1 - \epsilon^2} = \frac{\epsilon^2 c^2}{(1 - \epsilon^2)^2} \quad (\epsilon < 1)$$

This is an equation of the ellipse with focus $(c, 0)$ and directrix the y axis. Putting $X = x - c/(1 - \epsilon^2)$, $Y = y$, the standard form $X^2/A^2 + Y^2/B^2 = 1$ of the equation is obtained, where $A = \epsilon c/(1 - \epsilon^2) > 0$, $B = \epsilon c/(1 - \epsilon^2)^{1/2} > 0$. This is equivalent to referring the ellipse to a new set of coordinate axes obtained by translating the origin to the point $[c/(1 - \epsilon^2), 0]$. The coordinates of the focus F in the XY coordinate system are $[-c\epsilon^2/(1 - \epsilon^2), 0]$ and the equation of the directrix is $X = -c/(1 - \epsilon^2)$. Since the standard form shows the curve to be symmetric with respect to each of the new axes and the origin, it is clear that $F[c\epsilon^2/(1 - \epsilon^2), 0]$ and $X = c/(1 - \epsilon^2)$ may be taken as focus and directrix. Putting $C = c\epsilon^2/(1 - \epsilon^2)$, then $C^2 = A^2 - B^2$, and $X^2/A^2 + Y^2/B^2 = 1$ is seen to be the locus of points $P(X, Y)$, the sum of whose distances from $(C, 0)$ and $(-C, 0)$ is $2A$. This relation is frequently used to define an ellipse. The chord containing the foci is the major axis; its length is $2A$. The chord of length $2B$ through the midpoint of the foci, perpendicular to the major axis, is the minor axis. The tangent to $X^2/A^2 + Y^2/B^2 = 1$ at $P(X_1, Y_1)$ on the ellipse is $X_1X/A^2 + Y_1Y/B^2 = 1$. Lines PF, PF' make equal angles with the tangent, so sound or light that emanates from one focus is reflected to the other. An irreducible quadratic $Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$ is an equation of an ellipse provided that $AC - B^2 > 0$. *See* ELLIPSE.

Hyperbola. When $\epsilon > 1$, the equation for the conic is written $[x + c/(\epsilon^2 - 1)]^2 - y^2/(\epsilon^2 - 1) = \epsilon^2c^2/(1 - \epsilon^2)^2$, $(\epsilon > 1)$. Putting $X = x + c/(\epsilon^2 - 1)$, $Y = y$ gives the standard form of a hyperbola $X^2/A^2 - Y^2/B^2 = 1$, where $A = \epsilon c/(\epsilon^2 - 1) > 0$, $B = \epsilon c/(\epsilon^2 - 1)^{1/2} > 0$. It is clear that the curve consists of two branches and is symmetric to the axes and the origin. Putting $C = c\epsilon^2(\epsilon^2 - 1)$, then both $F(C, 0)$, $F'(-C, 0)$ are foci of the hyperbola, with respective directrices $X = c/(\epsilon^2 - 1)$, $X = -c/(\epsilon^2 - 1)$, and $A^2 + B^2 = C^2$. Then $X^2/A^2 - Y^2/B^2 = 1$ is seen to be an equation of the locus of points $P(X, Y)$ such that $PF - PF' = 2A$. The lines $X/A \pm Y/B = 0$ are asymptotes of the hyperbola. As a point P traverses either branch of the hyperbola, its distances from an asymptote approach zero. An irreducible quadratic $Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$ is an equation of an hyperbola provided $AC - B^2 < 0$. *See* HYPERBOLA.

Three-dimensional space. Let cartesian coordinate systems be established on each of three pairwise mutually perpendicular lines of three-space that intersect in O, the common origin of the systems. Suppose equal scales and call the lines the x axis, y axis, and z axis. To each point P of space an ordered triple (x, y, z) of real numbers is attached as rectangular cartesian coordinates, where x is the coordinate of the foot of the perpendicular from P to the x axis, and y and z are similarly defined. Thus every point of space has unique coordinates, and each ordered triple of real numbers is the coordinates of a point of space. If d denotes the distance of two points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, then $d = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]^{1/2}$. Let g denote any directed line and g' the line through O parallel to g and directed in the same sense. If α, β, γ denote the angles that g' makes with the x, y, and z axes, respectively (the direction angles of g), then $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are the direction cosines of g. They satisfy the relation $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$, and any three numbers λ, μ, ν such that $\lambda^2 + \mu^2 + \nu^2 = 1$ are the direction cosines of a (directed) line. Three numbers a, b, c proportional to the direction cosines λ, μ, ν of a directed line g are direction numbers of g. Clearly direction numbers of parallel lines are proportional. If $P_1(x_1, y_1, x_1)$, $P_2(x_2, y_2, z_2)$ are distinct points of g, then $x_2 - x_1, y_2 - y_1, z_2 - z_1$ are direction numbers of g. It follows that $P(x, y, z)$ is on g if and only if

$x - x_1 = t(x_2 - x_1)$, $y - y_1 = t(y_2 - y_1)$, $z - z_1 = t(z_2 - z_1)$, $-\infty > t > \infty$. These are parametric equations of g (t is the parameter), from which the symmetric equations $(x-x_1)/(x_2-x_1) = (y-y_1)/(y_2 - y_1) = (z-z_1)/(z_2-z_1)$ follow at once. The direction cosines of g, directed from P_1 to P_2, are $\cos\alpha = (x_2 - x_1)/d$, $\cos\beta = (y_2 - y_1)/d$, $\cos\gamma = (z_2 - z_1)/d$, where $d = [(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]^{1/2}$.

If α_1, β_1, γ_1, and α_2, β_2, γ_2 are direction angles of directed lines g_1, g_2, respectively, and θ denotes the angle between them, then $\cos\theta = \cos\alpha_1\cos\alpha_2 + \cos\beta_1\cos\beta_2 + \cos\gamma_1\cos\gamma_2$. Hence g_1, g_2 are mutually perpendicular if and only if $a_1a_2 + b_1b_2 + c_1c_2 = 0$, where a_1, b_1, c_1 and a_2, b_2, c_2 are direction numbers of g_1, g_2, respectively. Let the plane π go through $P_0(x_0, y_0, z_0)$ and be perpendicular to a line g with direction numbers a, b, c. Then $P(x, y, z)$ is in π if and only if $P = P_0$ or the line joining it to P_0 is at right angles to a line through P_0 that is parallel to g; that is, if and only if $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$. Hence to each plane corresponds a linear equation. Conversely, if P_0 (x_0, y_0, z_0) satisfies the linear equation $Ax + By + Cz + D = 0$, with A, B, C not all zero, then $A(x - x_0) + B(y - y_0) + C(z - z_0) = 0$ is an equation of a plane through P_0, perpendicular to a line with direction numbers A, B, C, and so $Ax + By + Cz + D = 0$, with $A, B, C \neq 0, 0, 0$, is an equation of a plane. Clearly $x = 0$ is an equation for the plane determined by the y and z axes; equations of the other two coordinate planes are $y = 0$ and $z = 0$. If a directed perpendicular g from O to a plane meets the plane at P (g is directed from O to P in case the plane is not through O) the plane has equation $x\cos\alpha + y\cos\beta + z\cos\gamma = p$, where α, β, γ are the direction angles of g and $p = OP$. This is the normal form of equation of a plane. The general form is reduced to it upon dividing by $\pm[A^2 + B^2 + C^2]^{1/2}$. The distance from $Ax + By + Cz + D = 0$ to $P(x_0, y_0)$ is $(Ax_0 + By_0 + Cz_0 + D)/\pm(A^2 + B^2 + C^2)^{1/2}$, where the sign is selected opposite that of D. (In case $D = 0$, other conventions are used.) Two planes $A_ix + B_iy + C_iz = 0$, with $i = 1, 2$, are parallel in case the number triples A_1, B_1, C_1 and A_2, B_2, C_2 are proportional and are mutually perpendicular provided $A_1A_2 + B_1B_2 + C_1C_2 = 0$ (since A_1, B_1, C_1 and A_2, B_2, C_2 are direction numbers of lines that are perpendicular to the respective planes). If $P_i(x_i, y_i, z_i)$, with $i = 1, 2, 3$, is not collinear, the plane determined has the equation

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0$$

Three planes $A_ix + B_iy + C_iz + D_i = 0$, with $i = 1, 2, 3$, intersect in one point if and only if

$$\begin{vmatrix} A_1 & B_2 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} = 0$$

The locus of points whose coordinates satisfy an equation $f(x, y, z) = 0$ is a surface. Curves may be thought of as intersections of two surfaces, and as such the coordinates of their points satisfy two equations $f(x, y, z) = 0$, $g(x, y, z) = 0$. Thus a line, considered as the intersection of two planes, is given by the two (simultaneous) equations $A_ix + B_iy + C_iz = D_i = 0$, with $i = 1, 2$. That line has direction numbers

$$\begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix} \quad \begin{vmatrix} C_1 & A_1 \\ C_2 & A_2 \end{vmatrix} \quad \begin{vmatrix} A_1 & B_2 \\ A_2 & B_2 \end{vmatrix}$$

Curves are also represented parametrically by equations $x = f(t)$, $y = g(t)$, $z = h(t)$, the parameter t varying in an interval (a, b), finite or infinite. Parametric equations for the line have already been given. Additional examples are $x = r\cos t, y = r\sin t, z = 0$, with $0 \leq t \leq 2\pi$, parametric equations of the circle in the xy plane, with center at O and radius r; $x = a \cdot \cos t, y = a \cdot \sin t, z = kt$, with $k \neq 0, -\infty < t < \infty$, parametric equations of a circular helix—the curve of the thread of a machine (untapered) screw.

Special surfaces. It follows from the definition of a sphere and the formula for distance of two points that $(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2$ is an equation for the sphere with center (a, b, c) and radius r, and by completing the squares of the x, y, and z terms in the equation $x^2 + y^2 + z^2 + 2Dx + 2Ey + 2Fz + G = 0$, it is seen that the locus of such an equation is a sphere with positive or zero radius, or there is no (real) locus.

Any equation in just two of the three coordinates is an equation of a cylinder whose elements are parallel to the axis of the missing variable. Thus the locus in 3-space of $x^2 + y^2 = r^2$ is a (right) circular cylinder whose elements are parallel to the z axis and which intersects the xy plane in the circle $x^2 + y^2 = r^2$, $z = 0$.

Any equation $f(x, y, z) = 0$, with $f(x, y, z)$ homogeneous in x, y, z (for example, $4xy - xz + yz = 0$, $x^3 - xy^2 + z^3 = 0$) has a cone with vertex O as locus.

A surface of revolution is obtained by rotating a plane curve C about a line g of its plane. If $f(x, y) = 0$, $z = 0$ are equations of C, and g is the x axis, the resulting surface of revolution has the equation $f(x, \sqrt{y^2 + z^2}) = 0$. Thus the surface generated by revolving the circle $x_2 + (y - b)^2 = a^2$, $z = 0$, about the x axis (the torus or anchor ring, if $b > a$) has the equation $x^2 + (\sqrt{y^2 + z^2} - b)^2 = a^2$.

A quadric surface is the locus of points whose coordinates satisfy an equation of the form $Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz + Gx + Hy + Jz + K = 0$, where at least one coefficient of a second-degree term is not zero. Some surfaces obtained by rotating conics about a line belong to this class, for example, spheres, prolate and oblate spheroids (given by rotating an ellipse about its major and minor axes, respectively), hyperboloids and paraboloids resulting from rotations of hyperbolas and parabolas about their axes of symmetry, and right circular cones and cylinders. Cylinders with conics for directrix curves are also members. Apart from degenerate cases, such as two planes, the remaining quadric surfaces and the standard forms of their equations are (1) ellipsoid,

$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$; (2) hyperboloid of one sheet, $x^2/a^2 + y^2/b^2 - z^2/c^2 = 1$; (3) hyperboloid of two sheets, $x^2a^2 - y^2/b^2 - z^2/c^2 = 1$; (4) elliptic paraboloid, $x^2/a^2 + y^2/b^2 = 2z$; and (5) hyperbolic paraboloid, $x^2/a^2 - y^2/b^2 = 2z$. Hyperboloids of one sheet and hyperbolic paraboloids are ruled surfaces; that is, each contains an infinity of straight lines, called generators. In fact, each of those surfaces contains two sets of generators. *See* QUADRIC SURFACE.

n-Dimensions. Let cartesian coordinate systems with equal scales be established on each of n pairwise mutually perpendicular lines intersecting in the common origin O, and label the lines $OX_1, OX_2, \ldots, OX_n$. To each point P of n space an ordered n-tuple $(x_1, x_2, \ldots, x_n)$ of numbers is attached as coordinates, where x_i is the coordinate of the foot of the perpendicular from P to OX_i, with $i = 1, 2, \ldots, n$. Two points $P(x_1, x_2, \ldots, x_n)$, $Q(y_1, y_2, \ldots, y_n)$ have distance $d = [(x_1 - y_2)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2]^{1/2}$. Direction angles $x_1, x_2, \ldots, x_n$ of a directed line are defined as in three-dimensional analytic geometry, and the direction cosines satisfy the relation $\cos^2 \alpha_1 + \cos^2 \alpha_2 + \cdots + \cos^2 \alpha_n = 1$. Direction cosines of the line $g(P, Q)$, directed from P to Q, are $(y_1 - x_1)/d, (y_2 - x_2)/d, \ldots, (y_n - x_n)/d$, and numbers proportional to them (for example, the numerators) are direction numbers of $g(P, Q)$. Hence $(X_1, X_2, \ldots, X_n)$ are coordinates of a point on $g(P, Q)$ if and only if $(X_1 - x_1)/(y_1 - x_1) = (X_2 - x_2)/(y_2 - x_2) = \cdots = (X_n - x_n)/(y_n - x_n)$. These are symmetric equations for the line determined by P, Q. Denoting the common value of the quotients by t gives the n parametric equations $X_i = x_i + t(y_i - x_i)$, with $i = 1, 2, \ldots, n, -\infty < t < \infty$. Let g be a line through $C(c_1, c_2, \ldots, c_n)$ with direction numbers $a_1, a_2, \ldots, a_n$. Point C, together with all points X such that the line $g(C, X)$ is perpendicular to g, defines an $(n - 1)$-dimensional subspace (hyperplane). Its equation is $a_1(X_1 - c_1) + a_2(X_2 - c_2) + \cdots + a_n(X_n - c_n) = 0$, since $X_i - c_i$, with $i = 1, 2, \ldots, n$, are direction numbers of $g(C, X)$ and the equation is the condition that g and $g(C, X)$ be mutually perpendicular. It is readily seen that the locus of every linear equation $A_1X_1 + A_2X_2 + \cdots + A_nX_n + K = 0$ is a hyperplane perpendicular to a line with direction numbers $A_1, A_2, \ldots, A_n$. It may be put in normal form by dividing by $\pm(A_1^2 + A_2^2 + \cdots + A_n^2)^{1/2}$, and the distance from it to a point $C(c_1, c_2, \ldots, c_n)$ if found by substituting the coordinates of C for $X_1, X_2, \ldots, X_n$. Subspaces of dimension k $(1 \leq k < n)$ are given by systems of $n - k$ linear equations. An equation of the hyperplane determined by n points is readily expressed in determinant form, as well as the condition that $n + 1$ points be on a hyperplane [substitute the coordinates of the $(n + 1)$-st point in the first row of the determinant equation of the hyperplane]. Discussion of loci of higher order lies outside of the scope of this article.

This brief sketch of analytic geometry has dealt only with that coordinate system most frequently used. For other coordinate systems (polar), as well as a discussion of transformation of coordinates, *See* ALGEBRA; COORDINATE SYSTEMS; DIFFERENTIAL GEOMETRY; EUCLIDEAN GEOMETRY; PROJECTIVE GEOMETRY; TRIGONOMETRY.

Leonard M. Blumenthal

Bibliography. G. Fuller, *Analytic Geometry*, 7th ed., 1993; D. F. Riddle, *Analytic Geometry*, 6th ed., 1996.

Analytic hierarchy

A framework for solving a problem. The analytic hierarchy process is a systematic procedure for representing the elements of any problem. It organizes the basic rationality by breaking down a problem into its smaller constituents and then calls for only simple pairwise comparison judgments, to develop priorities in each level.

The analytic hierarchy process provides a comprehensive framework to cope with intuitive, rational, and irrational factors in making judgments at the same time. It is a method of integrating perceptions and purposes into an overall synthesis. The analytic hierarchy process does not require that judgments be consistent or even transitive. The degree of consistency (or inconsistency) of the judgment is revealed at the end of the analytic hierarchy process.

Human reasoning. People generally provide subjective judgments that are based on feelings and intuition rather than on well-worked-out logical reasoning. Also, when they reason together, people tend to influence each other's thinking. Individual judgments are altered slightly to accommodate the group's logic and the group's interests. However, people have very short memories, and if asked afterward to support the group judgments, they instinctively go back to their individual judgments. Repetition is needed to effect deep-rooted changes in people's judgment.

People also find it difficult to justify their judgments logically and to explicate how strong these judgments are. As a result, people make great compromises in their thinking to accommodate ideas and judgments. In groups, there is a willingness to compromise. If truth is to be an objective reality, reality must be very fuzzy because the search for truth often ends in compromise. What is regarded as truth may often be essentially a social product obtained through interaction rather than by pure deduction.

Logical understanding does not seem to permeate judgment instantaneously. It apparently needs time to be assimilated. However, even when logical understanding has been assimilated, people still offer judgment in a spontaneous emotional way without elaborate explanation. Even when there is time, explanations tend to be fragmentary, disconnected, and mostly without an underlying clear logical foundation.

Outline of the process. People making comparisons use their feelings and judgment. Both vary in intensity. To distinguish among different intensities, the scale of absolute numbers in **Table 1** is useful.

The analytic hierarchy process can be decomposed into the following steps. Particular steps may be emphasized more in some situations than in

TABLE 1. Scale of relative importance

Intensity of relative importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Slight importance of one over another	Experience and judgment slightly favor one activity over another
5	Essential or strong importance	Experience and judgment strongly favor one activity over another
7	Demonstrated importance	An activity is strongly favored and its dominance is demonstrated in practice
9	Absolute importance	The evidence favoring one activity over another is of the highest possible order of affirmation
2, 4, 6, 8	Intermediate values between the two adjacent judgments	When compromise is needed
Reciprocals of above nonzero numbers	If an activity has one of the above numbers assigned to it when compared with second activity, the second activity has the reciprocal value when compared to the first	

others. Also as noted, interaction is generally useful for stimulation and for representing different points of view.

1. Define the problem and determine what knowledge is sought.

2. Structure the hierarchy from the top (the objectives from a broad perspective) through the intermediate levels (criteria on which subsequent levels depend) to the lowest level (which usually is a list of the alternatives).

3. Construct a set of pairwise comparison matrices for each of the lower levels, one matrix for each element in the level immediately above. An element in the higher level is said to be a governing element for those in the lower level since it contributes to it or affects it. In a complete simple hierarchy, every element in the lower level affects every element in the upper level. The elements in the lower level are then compared to each other, based on their effect on the governing element above. This yields a square matrix of judgments. The pairwise comparisons are done in terms of which element dominates the other. These judgments are then expressed as integers according to the judgment values in Table 1. If element A dominates element B, then the whole number integer is entered in row A, column B, and the reciprocal (fraction) is entered in row B, column A. Of course, if element B dominates element A, the reverse occurs. The whole number is then placed in the B,A position with the reciprocal automatically being assigned to the A,B position. If the elements being compared are equal, a one is assigned to both positions. The numbers used express an absolute rather than an ordinal relation.

4. There are $n(n - 1)/2$ judgments required to develop the set of matrices in step 3 (taking into account the fact that reciprocals are automatically assigned in each pairwise comparison), where n is the number of elements in the lower level.

5. Having collected all the pairwise comparison data and entered the reciprocals together with n unit entries down the main diagonal (an element is equal to itself, so a "one" is assigned to the diagonal positions), the eigenvalue problem $Aw = \lambda_{max}w$ is solved and consistency is tested, using the departure of λ_{max} from n (see below).

6. Steps 3, 4, and 5 are performed for all levels and clusters in the hierarchy.

7. Hierarchial composition is now used to weigh the eigenvectors by the weights of the criteria, and the sum is taken over all weighted eigenvector entries corresponding to those in the lower level of the hierarchy.

8. The consistency ratio of the entire hierarchy is found by multiplying each consistency index by the priority of the corresponding criterion and adding them together. The result is then divided by the same type of expression, using the random consistency index corresponding to the dimensions of each matrix weighted by the priorities as before. The consistency ratio should be about 10% or less to be acceptable. If not, the quality of the judgments should be improved, perhaps by revising the manner in which questions are asked in making the pairwise comparisons. If this should fail to improve consistency, it is likely that the problem should be more accurately structured; that is, similar elements should be grouped under more meaningful criteria. A return to step 2 would be required, although only the problematic parts of the hierarchy may need revision.

If the exact answer in the form of hard numbers was actually available, it would be possible to normalize these numbers, form their ratios as described above, and solve the problem. This would result in getting the same numbers back, as should be expected. On the other hand, if firm numbers were not available, their ratios could be estimated to solve the problem.

Example of the process. In the following example the analytic hierarchy process is used to assist a young family (a father, a mother, and a child) of specified income to buy a new car, say, either model A, B, or C. The choice will be determined through four important criteria.

TABLE 2. Comparison matrix comparing criteria for buying a car

Decision to buy a car	Price	Running cost	Comfort	Status	Priorities
Price	1	3	7	8	.582
Running cost	1/3	1	5	5	.279
Comfort	1/7	1/5	1	3	.090
Status	1/8	1/5	1/3	1	.050

$\lambda_{max} = 4.198$
C.I. = .066
C.R. = .073

The hierarchy of such a decision often takes the form shown in the **illustration**. In this hierarchy, level 1 is the single overall objective: Best New Car to Buy. On level 2 are the criteria which are perceived to compose what is meant by Best New Car, such as Price and Running Cost (operating and maintenance). On level 3 are the various alternative cars from which the family will choose.

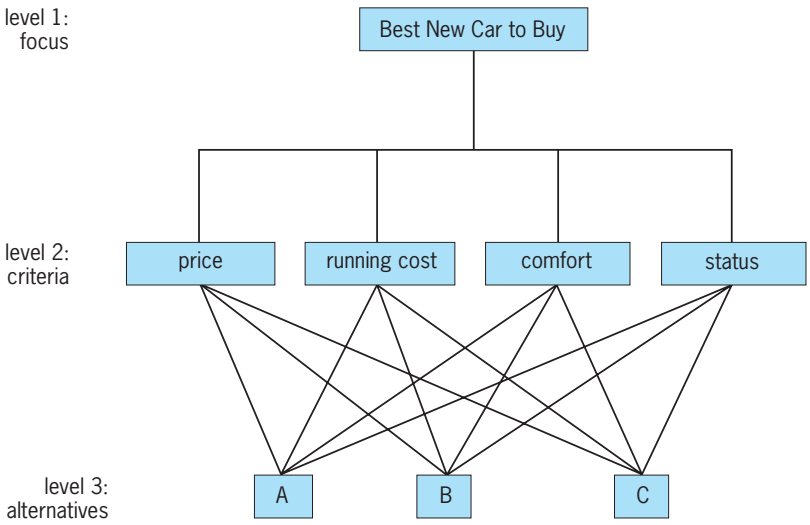
This downward decomposition format can easily be used on a wide class of problems. In addition, a slight further modification to incorporate feedback loops will cover an even wider range.

The questions asked when comparing each criterion are of the following kind: Of the two alternatives being compared, which is considered more important by the family buying a car and how much more important is it? The comparison matrix of **Table 2** is then formed. Since Price and Running Cost have the highest priorities, the other factors are discarded and only these two are used in continuing the process. Care must be taken in doing this. The original priorities are then normalized by dividing each of their sum to obtain the new relative priorities in Eqs. (1).

$$.67 = \frac{.58}{.58 + .28}$$
$$.33 = \frac{.28}{.58 + .28} \qquad (1)$$

The process is then repeated for the third level, where each car is compared with respect to the two high-priority factors from the second level, as shown in **Table 3**.

The two priority columns are then recorded as in **Table 4**. All entries of the first column are then multiplied by .67, the priority of Price, and those of the second column by .33, the priority of Running Cost, and added. This gives the third column. Thus Car B was selected for its efficient operation even



Analytic hierarchy used to assist a family in buying a new car.

though its initial price is considerably higher than Car A. A car dealer in the income neighborhood of this family may find it profitable to stock up on the cars in the respective proportions.

One of the most powerful contributions that the analytic hierarchy process makes is to test out the degree of inconsistency or incompatibility of new ideas or new policies adopted with older, more familiar, better tried successful methods. For example, in doing the above problem the participants were not sure whether the judgments for Price over Running Cost should be 7, 5, or 3. Each one was tried separately, and it was found that 3 yielded the highest consistency. Those who voted for 3 won that argument.

Priorities and consistency. An easy way to get a good approximation of the priorities is to use the geometric mean. This is done by multiplying the elements in each row and taking their nth root, where n is the number of elements. Then, normalize the column of numbers thus obtained by dividing each entry by the sum of all entries. Alternatively, normalize the elements in each column of the matrix and then average each row.

The consistency index can also be determined by hand calculations. Add the numbers in each column of the judgment matrix, multiply the first sum by the first priority, the second by the second, and so on, and add. For the first matrix the column sums (1.60, 4.40, 13.33, 17) are obtained, and multiplying by (.582, .279, .090, .050) gives 4.20. This number is

TABLE 3. Comparison matrices comparing alternative cars

Price	Car A	Car B	Car C	Priorities
Car A	1	2	3	.540
Car B	1/2	1	2	.297
Car C	1/3	1/2	1	.163
				$\lambda_{max} = 3.009$ C.I. = .005 C.R. = .008

Running cost	Car A	Car B	Car C	Priorities
Car A	1	1/5	1/2	.106
Car B	5	1	7	.745
Car C	2	1/7	1	.150
				$\lambda_{max} = 3.119$ C.I. = .059 C.R. = .103

TABLE 4. Composition of priorities

	Price (priority .67)	Running cost (priority .33)	Composite priority of cars
Car A	.540	.106	.396
Car B	.297	.745	.445
Car C	.163	.150	.159

denoted by λ_{max}. The consistency index is given by Eq. (2).

$$\text{C.I.} = \frac{\lambda_{max} - n}{n - 1} \tag{2}$$

The consistency is now checked by taking the ratio (C.R.) of C.I. with the appropriate one of the set of numbers in **Table 5** to see if it is about 10% or less (20% may be tolerated in some cases but not more). Otherwise the problem must be studied again and judgments revised. The consistency of the hierarchy in the above example, as given by Eq. (3), is .06, which is good.

$$\frac{.066 \times 1 + .005 \times .67 + .059 \times .33}{.900 \times 1 + .580 \times .67 + .580 \times .33} = \frac{.089}{1.48} = .06 \tag{3}$$

Judgment formation. When several people participate, judgments are often debated. Sometimes the group accepts a geometric average of their combined judgments. If there is strong disagreement, the different opinions can each be taken and used to obtain answers. Those which subsequently display the highest consistency within the group are the ones usually retained.

The analytic hierarchy process incorporates equally both tangible factors, which require hard measurements, and such intangible factors as comfort, which require judgment. Eventually one finds that so-called hard numbers have no meaning in themselves apart from their utilitarian interpretations. In the above example, buying a $10,000 car is more than twice as "painful" as buying a $5000 car.

The interdependence of criteria, such as Comfort and Price, has to be considered carefully since there may be some perceived overlap. For example, higher price buys more comfort, but it also buys other desirable attributes. Judging the relative importance of such things as price and comfort, therefore, must be done as independently as possible with avoidance of overlaps.

Validation by physical laws. Using the scale 1–9 has been justified and demonstrated by many examples. However, the following simple optics illustration, carried out with small children, shows that perceptions, judgments, and these numbers lead to results which can be validated by laws of physics. In this example, four identical chairs were placed at distances of 9, 15, 21, and 28 yards (1 yd = 0.9144 m) from a floodlight. The children stood by the light, looked at the line of chairs, and compared the first with the second, the first with the third and then with the fourth, and so on for the second, third, and fourth chairs. Each time the children said how much brighter one chair was, compared to the other.

Their judgments were entered in the matrix of **Table 6** to record the relative brightness of the chairs. The reciprocals were used in the transpose position.

TABLE 6. Comparison matrix comparing perceived brightnesses of chairs

	Chair 1	Chair 2	Chair 3	Chair 4	Brightness ratios
Chair 1	1	5	6	7	0.61
Chair 2	1/5	1	4	6	0.24
Chair 3	1/6	1/4	1	4	0.10
Chair 4	1/7	1/6	1/4	1	0.05

λ_{max} = 4.39
C.I. = 0.13
C.R. = 0.14

The inverse-square law of optics is now used to test these judgments. Since the distances are 9, 15, 21, and 28 yd, these numbers are squared and their reciprocals calculated. This gives .0123, .0044, .0023, and .0013 respectively. Normalization of these values gives .61, .22, .11, and .06, which are very close to the brightness ratios obtained in the test using the analytic hierarchy process.

Structuring a hierarchy. There are no rules for structuring a hierarchy. However, a typical analytic hierarchy for allocating resources—either by measuring costs or by measuring benefits—will often be stratified roughly as follows. The top level will include the overall objectives of the organization or system. Benefit-cost criteria may appear in the next level. A subordinate level may further clarify these criteria in the context of the particular problem by itemizing specific tasks which are to be accomplished at some level of performance. This is followed by the alternatives being evaluated.

Capabilities. Designing an analytic hierarchy—like the structuring of a problem by any other method—is more art than science. It necessitates substantial knowledge of the system in question. A very strong aspect of the analytic hierarchy process is that the knowledgable individuals who supply judgments for the pairwise comparisons usually also play a prominent role in specifying the hierarchy.

Although a hierarchy to be used in resource allocation will tend to have the vertical stratification

TABLE 5. Random consistencies

n:	1	2	3	4	5	6	7	8	9	10
Random consistency:	0	0	.58	.90	1.12	1.24	1.32	1.41	1.45	1.49

indicated above, it can also be much more general. The only restriction is that an element on a higher level must serve as a governing element for at least one element (which can be the element itself) on the immediately lower level. The hierarchy need not be complete; that is, an element at an upper level need not function as a criterion for all the elements in the lower level. It can be partitioned into nearly disjoint subhierarchies sharing only a common topmost element. Thus, for instance, the activities of separate divisions of an organization can be structured separately. As suggested above, the analyst can insert and delete levels and elements as necessary to clarify the task or to sharpen a focus on one or more areas of the system.

The analytic hierarchy process has already been successfully applied in a variety of fields, including planning the allocation of energy to industries; designing a transport system for the Sudan; planning the future of a corporation and measuring the impact of environmental factors on its development; designing future scenarios for higher education in the United States; selecting candidates and winners in elections; setting priorities for the top scientific institute in a developing country; solving a faculty promotion and tenure problem; and predicting oil prices. *See* DECISION THEORY; SYSTEMS ENGINEERING.

Thomas L. Saaty

Bibliography. B. I. Golden (ed.), *The Analytic Hierarchy Process*, 1989; T. L. Saaty, *The Analytic Hierarchy Process*, 1980; T. L. Saaty, *Decision Making for Leaders*, 1982, reprint 1990; T. L. Saaty and L. Vargas, *The Logic of Priorities*, 1981, reprint 1991.

Analytical chemistry

The science of chemical characterization and measurement. Qualitative analysis is concerned with the description of chemical composition in terms of elements, compounds, or structural units, whereas quantitative analysis is concerned with the measurement of amount.

Scope. Originally, chemical analysis conformed closely to its literal meaning and consisted of the separation of a sample into its components, which were weighed. Later, such gravimetric techniques were supplemented by volumetric or, more appropriately, titrimetric analysis, which consists of measurement of the amount of a standardized solution of a reagent that reacts with a measured portion of the sample to reach an end point—a process known as titration. The end point is usually detected by means of an indicator, that is, a substance that changes color when a given amount of reagent has been added. These forms of chemical analysis are called classical or wet methods to distinguish them from the instrumental methods, which have greatly enlarged the scope of analytical chemistry. *See* GRAVIMETRIC ANALYSIS; TITRATION.

Instrumental methods, which might better be termed physicochemical methods, have been developed to the degree that they make up the vast bulk of both qualitative and quantitative analysis. They include measurements of purely physical characteristics, as well as the use of physical measurements, to follow the course of chemical reactions.

Over the years, the scope of analytical chemistry has been enlarged from its original meaning, which was limited to the determination of chemical composition in terms of the relative amounts of elements or compounds in a sample. The discipline has been expanded to involve, for example, the spatial distribution of elements or compounds in a sample, the distinction between different crystalline forms of a given element or compound, the distinction between different chemical forms (such as the oxidation state of an element), the distinction between a component on the surface or in the interior of a particle, and the detection of single atoms or molecules of a substance. To permit these more detailed questions to be answered, as well as to improve the speed, accuracy, sensitivity, and selectivity of traditional analysis, a large variety of physical measurements are used. These methods are based on spectroscopic, electrochemical, chromatographic, chemical, and nuclear principles.

Modern analysis has also placed significant demands on sampling techniques. It has become necessary, for example, to handle very small liquid samples [in the nanoliter (10^{-9} liter) range or less] as part of the analysis of complex mixtures such as biological fluids, and to simultaneously determine many different components. The sample may be a solid that must be converted through vaporization into a form suitable for analysis.

Spectroscopy. One important area is spectroscopy, which includes the measurement of emission, absorption, reflection, and scattering phenomena resulting from interaction of a sample with gamma rays and x-rays at the high-energy end of the spectrum and with the less energetic ultraviolet, visible, infrared, and microwave radiation. *See* SPECTROSCOPY.

Molecular spectroscopy. Lower-energy forms of excitation such as ultraviolet, visible, or infrared radiation are used in molecular spectroscopy. Ultraviolet radiation and visible radiation, which are reflective of the electronic structure of molecules, are used extensively for quantitative analysis. The radiation absorbed by the sample is measured. It is also possible to measure the radiation emitted (emission, fluorescence). The absorption of infrared radiation is controlled by the properties of bonds between atoms, and accordingly it is most widely used for structure identification and determination. It is less widely used for quantitative analysis except for gases such as carbon monoxide (CO) and hydrocarbons. X-rays are used through emission of characteristic radiation, absorption, or diffraction. In the last case, characteristic diffraction patterns reveal information about specific structural entities, such as a particular crystalline form. Extended x-ray absorption fine structure (EXAFS) is based on the use of x-rays from a synchrotron source to reveal structural details such as interatomic distances. *See* EMISSION SPECTROCHEMICAL ANALYSIS; EXTENDED X-RAY ABSORPTION

FINE STRUCTURE (EXAFS); INFRARED SPECTROSCOPY; X-RAY FLUORESCENCE ANALYSIS.

Mass spectrometry. Mass spectrometry is an important and increasingly applied method of analysis, especially for organic and biological samples. Among the applications are the analysis of more than 70 elements (spark-source mass spectrometry), surface analysis (secondary ion mass spectrometry and ion-probe mass spectrometry), and the determination of the structure of organic molecules and of proteins and peptides (high-resolution mass spectrometry, Fourier-transform mass spectrometry). *See* MASS SPECTROMETRY; SECONDARY ION MASS SPECTROMETRY (SIMS).

Nuclear magnetic resonance. This technique measures the magnetic environment around individual atoms and provides one of the most important means for deducing the structure of a molecule. Atoms possessing nuclear spin are probed by monitoring the interaction between their nuclear spin and an applied external magnetic field. For large molecules, these interactions are complex, and a variety of nuclear excitation techniques have been developed that permit establishment of the connectivity between the various atoms in a molecule. Since the technique is nondestructive, it can be used to monitor living systems. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

Electron spectroscopy. Several forms of spectroscopy are especially useful for surface analysis. The scanning electron microscope (SEM) involves a finely collimated electron beam that sweeps across the surface of a sample to produce an image. At the same time, the surface atoms are excited to emit characteristic x-rays, making it possible to obtain an image of the surface along with its spatially resolved elemental composition. The resolution of this technique (electron microprobe) is in the micrometer (10^{-4} cm) range. Images with a resolution of nanometers (10^{-9} cm) have been obtained by using the techniques of atomic force microscopy (AFM) and scanning tunneling microscopy (STM), and this resolution corresponds to the dimensions of individual atoms. A significant advantage of the latter two techniques is that a high vacuum is not required, so samples can be analyzed at atmospheric pressure. *See* ELECTRON-PROBE MICROANALYSIS; ELECTRON SPECTROSCOPY; SCANNING ELECTRON MICROSCOPE; SCANNING TUNNELING MICROSCOPE.

There are a number of other important surface analysis techniques involving the interaction of radiation with surface atoms. Low-energy electron diffraction (LEED) involves the observation of the diffraction pattern of a reflected electron beam to give structural information about the crystallographic orientation of the atoms on the surface. Several techniques are used to determine composition and structural information by measuring the energies of electrons emitted from surface atoms. These include ultraviolet photoelectron spectroscopy (UPS) and x-ray photoelectron spectroscopy (also called electron spectroscopy for chemical analysis, or ESCA), which differ in the energy of the photon beam striking the surface. Other techniques include Auger electron spectroscopy (AES) and electron energy loss spectroscopy (EELS). *See* AUGER EFFECT; ELECTRON DIFFRACTION; SURFACE AND INTERFACIAL CHEMISTRY.

Radiochemistry. The radiation emitted by a radionuclide can be determined accurately, forming the basis of a number of important analytical techniques. In some cases, these nuclides are incorporated directly into a molecule, rendering it radioactive [carbon-14 (^{14}C), hydrogen-3 (^{3}H), iodine-125 (^{125}I), phosphorus-32 (^{32}P)]. In other cases, the radioactivity is induced in the sample by bombardment with neutrons or photons (neutron or photon activation analysis). The radionuclides are identified by their characteristic radiation. For example, gamma-ray spectrometry is used to measure the characteristic radiation from nuclides emitting gamma rays. Many nuclides can be analyzed simultaneously by this method. *See* ACTIVATION ANALYSIS; GAMMA-RAY DETECTORS; RADIOACTIVITY.

Electrochemical analysis. There are a number of widely applied techniques that are based mainly on four electrochemical phenomena: (1) potentiometry, involving the relationship between the measurement of an electrical potential between two electrodes, one of which responds selectively to the concentration (activity) of a species in solution; (2) amperometry, involving the measurement of a current proportional to the concentration of a species in solution at a given applied potential between two electrodes; (3) coulometry, based on the relationship between the charge passed in an electrochemical cell and the amount of a species generated (Faraday's law); and (4) conductimetry, involving the relationship between the impedance (resistance) of a solution and the concentration of ionic species present. *See* ELECTROCHEMISTRY.

Potentiometry. This is undoubtedly the most widely applied electrochemical technique, since it includes a variety of ion-selective electrodes, the most important of which is the glass electrode used to measure pH. Other important ion-selective electrodes measure ions of sodium, potassium, calcium, sulfide, chloride, and fluoride. When the electrodes are used in conjunction with gas-permeable membranes, gases such as ammonia, carbon dioxide, and hydrogen sulfide can be measured. *See* ION-SELECTIVE MEMBRANES AND ELECTRODES; pH; POLAROGRAPHIC ANALYSIS.

Amperometry. This method depends on the detection of species that can undergo electrolysis; many substances such as oxygen are detected in this way. If the detection electrode is covered with an enzyme layer, a variety of substances in complex mixtures can be determined by taking advantage of the selectivity of the enzymatic reaction. An electroactive product of the enzymatic reaction is detected, and this configuration is called an enzyme electrode. These determinations are typically performed at a constant applied potential. A number of important techniques are based on measuring the current resulting from time-dependent variation of the potential. These techniques include cyclic

voltammetry, chronocoulometry, pulse techniques (such as pulsed amperometric detection), and chronoamperometry. They are used to study electrochemical reactions at electrodes, to study properties of electrode surfaces, and to analyze (for example, to detect an effluent from a chromatographic column or to determine a trace metal in solution).

Other phenomena. Coulometry is used in quantitatively generating analytical reagents and in connection with a system capable of detecting an end point. Conductimetry is a nonselective technique that is used to measure the overall ionic content of an electrically conducting solution. *See* COULOMETER.

Herbert A. Laitinen; George S. Wilson

Trace analysis. Trace analysis involves the determination of those elemental constituents of a sample that make up approximately 0.01% by weight of the sample or less. There is no sharp boundary between nontrace and trace constituents. The lower limit of detected concentration is set by the sensitivity of the available analytical methods and, in general, is pushed downward with progress in analytical techniques. A large number of different physical and chemical techniques have been developed for the measurement of the elemental composition at the microgram-per-gram and nanogram-per-gram level, thereby constituting the field of trace analysis.

Methods. All trace analytical methods can be divided into three component steps: sampling, chemical or physical pretreatment, and measurement. Depending upon the type of information desired in an analysis and the requirements of sensitivity, precision, and other performance figures of merit, an appropriate measurement technique is selected. Most frequently, this will be an instrumental approach. Therefore, it is essential to know the capabilities and limitations of the various methods of instrumental measurement available for trace determinations. Once they are known, appropriate steps can be taken in the sampling and pretreatment steps to provide sufficient amounts of the microconstituents that are free of interferences and in the appropriate form for the final measurement. In a number of methods the pretreatment step may be omitted, and in others the sampling and measurement occur simultaneously. In spite of possible deviations, these steps are interrelated and require different degrees of emphasis, depending upon the individual analytical situation. In many cases, the analyst uses a particular physical or physicochemical method in which manifestations of energy provide the basis of measurement. These methods are indirect in the sense that the emission or absorption of radiation or transformation of energy must be related in some way to the mass or concentration of the species that are being determined. The establishment of these relations almost invariably requires calibration, with the use of standards of known content of the constituent in question. As such, many of the available techniques do not provide absolute results.

Trace analysis ranges from the more classical chemical methods of colorimetric and absorption spectrophotometric analysis to modern instrumental approaches. The wide diversity of methods is apparent from the number of approaches and the attendant classes and subclasses in the **table**. Within any one of these categories, there are techniques that are more specialized.

Of the various criteria used in the selection of an

Some trace analysis methods and limits of detection

Method	Absolute, g	Limits of detection: concentration, ppm
Chromatography		
Thin-layer	10^{-5} to 10^{-3}	
Gas-liquid		10 to 10^{6}
Liquid-liquid		10^{-3} to 10^{0}
Electrochemical		
Coulometry	10^{-9} to 10^{0}	
Ion selective electrode		10^{-2} to 10^{2}
Polarography		
Conventional		10^{0} to 10^{3}
Modern		10^{-3} to 10^{3}
Laser probe microanalysis		10^{2} to 10^{4}
Nuclear		
Neutron activation		10^{-3} to 10^{-1}
Electron probe		10^{2} to 10^{3}
Ion probe		10^{-1} to 10^{1}
Mass spectrometry		
Isotope dilution		10^{-5} to 10^{6}
Spark source		10^{-3} to 10^{1}
Electrical plasma*		10^{-5} to 10^{0}
Organic microanalysis	$>10^{-5}$	
Optical-absorption		
Atomic		
Flame		10^{-3} to 10^{1}
Nonflame	10^{-15} to 10^{-9}	
Molecular		
UV-visible		10^{-3} to 10^{2}
Infrared		10^{3} to 10^{6}
Microwave		10^{0} to 10^{3}
Optical-emission		
AC spark		10^{1} to 10^{3}
DC arc		10^{-2} to 10^{2}
Electrical plasma†		10^{-4} to 10^{2}
Optical-fluorescence		
Atomic		
Flame		10^{-3} to 10^{2}
Nonflame	10^{-15} to 10^{-9}	
Molecular		10^{-3} to 10^{-1}
Optical-phosphorescence		10^{-3} to 10^{2}
Optical-Raman		10^{0} to 10^{5}
Spectrometric-resonance		
Nuclear magnetic		10^{1} to 10^{5}
Electron spin	10^{-9} to 10^{-6}	
Thermal analysis	10^{-5} to 10^{-4}	
Wet chemistry		
Gravimetry	10^{-3} to 10^{-2}	
Titrimetry		10^{-2} to 10^{4}
X-ray spectrometry		
Auger		10^{3} to 10^{5}
ESCA		10^{3} to 10^{5}
Fluorescence		10^{-1} to 10^{2}
Mössbauer		10^{0} to 10^{3}
Photoelectron		10^{0} to 10^{3}
Immunoassay	10^{-2} to 10^{-9}	10^{-3} to 10^{0}

*Inductively coupled plasma or microwave-induced plasma.
†Inductively coupled plasma or direct-current plasma.

appropriate trace analytical method, sensitivity, accuracy, precision, and selectivity are of prime importance. Other important considerations, such as scope, sampling and standards requirements, cost of equipment, and time of analyses, are of great practical significance.

Sensitivity and detection limits. For all analysis regimes, the measure x of some physical parameter is related to the concentration c of the analyte in a certain sample. The sensitivity S is the slope of the analytical calibration curve, which is the plot of the measure x versus the concentration of the analyte in a series of standards having known analyte concentrations. The term sensitivity must not be used to indicate either the limit of detection or the concentration required to give a certain signal.

The standard deviation σ is given by Eq. (1), where

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \tag{1}$$

x_i is the individual measure, and $\bar{x}$ is the mean of n measurements. If n is less than about 10, σ is replaced by the term s. The relative standard deviation (RSD) is given by Eqs. (2).

$$\text{RSD} = \frac{\sigma}{\bar{x}} \quad \text{or} \quad \text{RSD} = \frac{s}{\bar{x}} \tag{2}$$

Precision is defined as the random uncertainty for the measure x of the corresponding uncertainty in the estimate of the concentration. It is often expressed as the relative standard deviation.

Accuracy is a measure of the agreement between the estimated concentration and the true value. Accuracy can never be better than the precision of the analytical procedure. Bias (systematic errors) in the calibration procedure always causes the estimated uncertainties to disagree with the true value by an amount equal to the bias.

The limit of detection c_L is the smallest concentration that can be detected with a reasonable certainty. This concentration is given by Eq. (3), where

$$c_L = \frac{x_L - \bar{x}_b}{S} \tag{3}$$

x_L is the limiting detectable measure and x_b is the average blank measure. The value for x_L is given by Eq. (4), where k is a factor determining a margin of

$$x_L = x_b + ks_b \tag{4}$$

confidence and s_b is the standard deviation of a limited number of measures of the blank. Generally, k is taken as 3; thus, the value of c_L can be determined by Eqs. (5). This means that any concentration re-

$$c_L = \frac{3s_b}{S} \quad \text{or} \quad c_L = \frac{3N_b}{S} \tag{5}$$

sulting in a signal three times the background standard deviation is considered just detectable. To evaluate x_b and s_b, at least 20 measurements should be used. If the major sources of variation are electrical noises, s_b can be replaced with N_b, the background noise level. If $3s_b$ is chosen, the confidence level is 99.86% for a purely Gaussian distribution of errors. At low concentrations, broader and asymmetric distributions are likely, so $3s_b$ corresponds to a practical confidence level of about 90%. It is noted that at the limit of detection the relative standard deviation is about 0.5; that is, there is equal confidence in deciding whether the analyte is detected or not.

As a result of widely varying pathways of development of many analytical techniques, there is a lack of consistency in the definition or specification of detection limits in the analytical literature. Consequently, it is difficult to make critical comparisons between methods for a particular element. However, since a particular element may be determined by a number of different techniques, depending upon the matrix in which it is being sought, a summary of the experimental values that have been published is pertinent. *See* ACTIVATION ANALYSIS; ANALYTICAL CHEMISTRY; CHROMATOGRAPHY; ELECTROCHEMICAL TECHNIQUES; GAS CHROMATOGRAPHY; SPECTROSCOPY.

Andrew T. Zander

Separation techniques. This collection of techniques includes various forms of chromatography and electrophoresis. They are based on the separation of a mixture of species in a sample due to differential migration. Two forces act in opposition: a stationary phase acts to retard a migrating species, while the mobile phase tends to promote migration. The mobile phase may be liquid (liquid chromatography) or gaseous (gas chromatography), while the stationary phase may be a solid or a solid covered with a thin film of liquid. The stationary phase is typically packed in a column through which the mobile phase is pumped. High-performance liquid chromatography (HPLC) has become especially important for the separation of complex mixtures of nonvolatile materials, because the use of high pressure emphasizes the differences in the rate of diffusion among the species to be separated. Separations may often be accomplished in a matter of several minutes. The stationary phase can preferentially interact with the migrating species according to charge, size, and hydrophobicity, or in some cases because of the special affinity that a species has for the stationary phase (affinity chromatography). The stationary phase can also be a thin layer of solid support deposited on a plate (thin-layer chromatography). *See* CHROMATOGRAPHY; GAS CHROMATOGRAPHY; LIQUID CHROMATOGRAPHY.

Alternatively, the driving force for separation will be the migration of charged species in an electric field (electrophoresis). The stationary phase may be a gel on a plate or in a tube, or a solution maintained in a capillary through which the analytes move. The important techniques in this area are capillary electrophoresis, isotachophoresis, and isoelectric focusing. *See* ELECTROPHORESIS.

In order to detect the various species after they are separated, most of the techniques described above have been employed. In some cases, it is desir-

able to have detection methods that are specific for the species of interest. In other cases, such specificity is not necessary since the species have already been separated and a detector with a general response can be used.

Chemical methods. Chemical and biological reactions form the basis of many analytical methods, besides the classical forms of gravimetric and volumetric analysis. From biology have come a number of widely used techniques: immunoassays, enzyme-catalyzed reactions, gene probes, and bioluminescence. These methods are often based on the rate of a reaction, which is proportional to the concentration of the analyte of interest. *See* BIOLUMINESCENCE; IMMUNOASSAY; KINETIC METHODS OF ANALYSIS.

Thermal methods. Thermal methods are based on the heating of a sample over a range of temperatures. This approach may result in absorption of heat by the sample or in evolution of heat due to physical or chemical changes. Thermogravimetry involves the measurement of mass; differential thermal analysis involves a detection of chemical or physical processes through a measurement of the difference in temperature between a sample and a stable reference material; differential thermal calorimetry evaluates the heat evolved in such processes. A variety of calorimetric techniques are used to measure the extent of reactions that are otherwise difficult to evaluate. *See* CALORIMETRY.

Sample handling and processing methods. As the demands increase for more rapid and reliable analysis at lower cost, automation of sampling and sample handling, measurement, and data analysis have become extremely important. Applications may involve monitoring a manufacturing process stream, air quality, or critical diagnostic analyses of a patient undergoing surgery. Fortunately, the advent of low-cost but powerful computers has greatly aided development in this area.

Computers. The digital computer has had a major impact on modern analytical chemistry. The microprocessor has been incorporated as an integral part of a variety of analytical instruments to permit detailed programming of their operation. More complex programming and automation are possible through the personal computer, which is being used not only to operate a single instrument but also to tie various operations together. These systems are increasingly linked to networks, which provide for storage of and access to data. *See* DIGITAL COMPUTER; MICROCOMPUTER; MICROPROCESSOR.

Chemometrics. Chemometrics refers to the use of statistical and other mathematical approaches to data handling. An important application is the optimization of an instrumental procedure by processing the initial instrument output rapidly and feeding the information back in order to control the settings of the instrument, thereby achieving optimal accuracy and sensitivity from the instrument. Still more complex computer operations, such as pattern recognition, are designed to draw conclusions through comparisons of arrays of data. *See* CHEMOMETRICS.

Automation. Analytical methods can be automated by several approaches. Clinical methods are often carried out by means of an automated analyzer. The sample is introduced into a flowing stream (unsegmented flow) or separated from the next sample by an air bubble (segmented flow), and a sequence of analytical operations is performed on the sample, culminating in a final measurement step based on an instrumental reading. The instrument is calibrated by running a standard sample through the system. A computer readout compares the analytical result with the normal range expected for the analysis. Other physical principles are also used in automated analytical systems. In robotics, a computer-controlled robot carries out the routine steps of chemical methods, including sample preparation and cleanup, weighing, dissolution, adjustment of conditions, reagent manipulation, and instrument reading. This capability increases sample throughput and reduces errors caused in many cases by operator boredom and inattention due to the repetitive nature of the manipulations. Such analyses can be performed very rapidly with small volumes (nanoliters) of solution. *See* COMBINATORIAL CHEMISTRY; ROBOTICS.

Instrumentation. As there is a constant need for improved instrumentation, the development of new instrumentation, rather than simple application of old techniques, is an area of active interest in analytical chemistry research. *See* CHEMISTRY; INSTRUMENT SCIENCE.

George S. Wilson

Bibliography. D. C. Harris, *Quantitative Chemical Analysis*, 6th ed., 2002; A. G. Howard and P. J. Statham, *Inorganic Trace Analysis: Philosophy and Practice*, 1994; J. R. Lawrence (ed.), *Trace Analysis*, vol. 1, 1981, vol. 2, 1982; J. F. Rubinson and K. A. Rubinson, *Contemporary Chemical Analysis*, 1998; D. A. Skoog et al., *Principles of Instrumental Analysis*, 5th ed., 1998.

Anamnia

Those vertebrate animals, sometimes called Anamniota, which lack an amnion in development. The amnion is a protective embryonic envelope that encloses the embryo and its surrounding liquid, the amniotic fluid, during fetal life. An amnion is present in mammals, birds, and reptiles, but is absent in fishes and amphibians. In early classifications the vertebrates were commonly separated on this basis, and the expressions Amniota and Anamniota are still useful in grouping higher and lower vertebrates. It should be recognized that these terms represent grades of development, however, and do not carry the connotation of established classificatory ranks. Anamnia, then, is a group name that includes the Recent classes Agnatha, Chondrichthyes, Osteichthyes, and Amphibia and, by presumption, the class Placodermi, which is known only from fossils. *See* AMNION; AMNIOTA; AMPHIBIA; PISCES (ZOOLOGY).

Reeve M. Bailey

Anaphylaxis

A generalized or localized tissue reaction occurring within minutes of an antigen-antibody reaction. Similar reactions elicited by nonimmunologic mechanisms are termed anaphylactoid reactions. The term anaphylaxis was introduced by C. R. Richet to describe acute adverse reactions which followed reinjection of an eel toxin into dogs. Instead of becoming immune (nonreacting) to the toxin, the animals developed severe and often fatal congestion and hemorrhage of the gastrointestinal tract, lungs, pleura, and endocardium.

It was initially thought that anaphylaxis in humans was associated with precipitating antibody (known today to be of the IgG class of immunoglobulins), and that it differed from the common allergic manifestations of hay fever due to "reaginic" antibody. The reaginic antibody has now been identified as IgE. *See* ALLERGY; IMMUNOGLOBULIN.

Signs and symptoms. In animals undergoing anaphylaxis, one organ system is usually affected more than others (thus the term shock-organ). In the guinea pig, the manifestations of anaphylaxis are respiratory, with cough, wheezing, and gasping respirations caused by severe bronchoconstriction. In the dog and the rat, the manifestations are gastrointestinal with hemorrhage. Severe pulmonary vasoconstriction leading to heart failure and death prevail in the rabbit. In humans, the clinical manifestations of anaphylaxis include reactions of the skin with itching, erythema, and urticaria; the upper respiratory tract with edema of the larynx; the lower respiratory tract with dyspnea, wheezing, and cough; the gastrointestinal tract with abdominal cramps, nausea, vomiting, and diarrhea; and the cardiovascular system with hypotension and shock. Individuals undergoing anaphylactic reactions may develop any one, a combination, or all of the signs and symptoms.

Anaphylaxis may be fatal within minutes, or may occur days or weeks after the reaction, if the organs sustained considerable damage during the hypotensive phase. Tissue changes in acute anaphylaxis are minimal and include edema and a mild eosinophilic infiltration. Myocardial ischemia and ischemic changes of the renal tubules and the brain may be seen after prolonged hypotension. Macroscopic changes are variable, and in fatal cases usually involve the respiratory system, where pulmonary edema, pulmonary infiltrates, and increased bronchial secretions are noted.

Mechanism. Anaphylaxis in humans is most often the result of the interaction of specific IgE antibody fixed to mast cells and antigen. Two molecules of IgE are bridged by the antigen, which may be a complex protein or chemical (hapten) bound to protein. The antigen-antibody interaction leads to increased cell-membrane permeability, with influx of calcium and release of either preformed or newly formed pharmacologic mediators from the granules. Preformed mediators include histamine and eosinophilic or neutrophilic chemotactic factors. Newly formed molecules include leukotrienes or slow-reacting substance of anaphylaxis and prostaglandins. The mediator action induces bronchoconstriction, vasodilation, cellular infiltration, and increased mucus production. Degranulation of mast cells and basophils is augmented when intracellular cyclic guanine monophosphate (cGMP) is increased; the reaction is inhibited in the presence of increasing intracellular concentrations of cyclic adenine monophosphate (cAMP). *See* EICOSANOIDS.

Another mechanism for induction of anaphylaxis in humans occurs when antigen binds to preformed IgG antibody and complement components interact with the antigen-antibody complex. The early components of the complement system bind to the antibody molecule, leading to activation of other complement components. During the activation, components known as anaphylatoxins (C3a and C5a) are released which may directly cause bronchoconstriction with respiratory impairment, and vasodilation with hypotension or shock. Activation of the complement system by agents like iodinated radiocontrast media, or degranulation of mast cells by drugs like opiates, will elicit similar manifestations without an antibody-antigen interaction. *See* COMPLEMENT.

Anaphylaxis due to IgE mechanisms has been associated with foreign proteins such as horse antitoxins, insulin, adrenocorticotropic hormone (ACTH), protamine, and chymopapain injected into herniated discs; drugs such as penicillin and its derivatives; foods such as shellfish, nuts, and eggs; and venom of stinging insects. Anaphylaxis mediated by IgG is seen in blood-transfusion reactions and following the use of cryoprecipitate, plasma, or immunoglobulin therapy. Direct mast-cell degranulation with release of mediators occurs with opiates, radiocontrast media, and curare.

In certain individuals, mast cell degranulation occurs after exercise and following exposure to cold. Individuals with systemic mastocytosis spontaneously release mediators from the increased number of tissue mast cells, which may cause an anaphylactic reaction. When no cause for the anaphylaxis can be found, the reaction is termed idiopathic.

Prevention and treatment. After the identification of the inciting agent for the anaphylactic reaction, prevention is the best mode of therapy. Individuals should not be given antibiotics to which they are sensitive, and foods which cause reactions should be avoided. Individuals sensitive to insect-sting venom should take preventive measures to avoid stings and should use insect repellents. Individuals who react to exercise should avoid such activities; when cold is the cause of anaphylaxis, warm clothes should be worn, and submerging in the cold water of swimming pools should be avoided. Immunotherapy with insect venom and desensitization with certain drugs are effective prophylactic measures.

Individuals with recurrent episodes of anaphylaxis, when the etiological cause is unknown and preventive measures are impractical, should be provided with epinephrine in a form that can be self-administered whenever symptoms occur.

Administration of corticosteroids may also be helpful. Such prophylaxis may be life-saving. *See* EPINEPHRINE.

The treatment of anaphylaxis is aimed at reducing the effect of the chemical mediators on the end organs and preventing further mediator release. The drug of choice for this is epinephrine given subcutaneously in repeated doses. Additionally, a clear airway and appropriate oxygenation must be maintained; hypotension should be treated, as should any cardiac arrhythmia. Aminophyllin, a rapid-onset bronchodilator, and corticosteroids, which alter prostaglandin and leukotriene metabolism but have an onset of action in several hours, are second-line indications to be used in patients who do not respond adequately to epinephrine. Cromolyn sodium, known to prevent mast-cell degranulation, has been effective in preventing bronchoconstriction, but has no place in the treatment of the acute anaphylactic reaction. Such effective therapeutic measures must be carried out as soon as possible because of the potentially fatal nature of anaphylaxis. *See* HYPERSENSITIVITY; SHOCK SYNDROME.

Shlomo Bar-Sela; Jordan N. Fink

Bibliography. E. Bacal, R. Patterson, and C. R. Zeiss, Evaluation of severe (anaphylactic) reactions, *Clin. Allergy*, 8:295, 1978; E. Middleton, C. E. Reed, and E. F. Ellis (eds.), *Allergy: Principles and Practices*, 4th ed., 1993; R. Patterson (ed.), *Allergic Diseases: Diagnosis and Management*, 5th ed., 1997; P. L. Smith et al., Physiologic manifestations of human anaphylaxis, *J. Clin. Invest.*, 66:1072–1080, 1980.

Anaplasmosis

A disease of mammals caused by a specialized group of gram-negative bacteria in the order Rickettsiales, family Anaplasmataceae, genus *Anaplasma*. *Anaplasma* is an obligate intracellular parasite that infects erythrocytes (red blood cells) of cattle, sheep, goats, and wild ruminants in much of the tropical and subtropical world. A newly reclassified species (*Anaplasma phagocytophilum*) that infects the white blood cells of mammals was recently added to this genus. However, the most important species and the one traditionally associated with the disease anaplasmosis is *A. marginale*. Anaplasmosis is one of the most important diseases of cattle, causing anemia and sometimes death, resulting in significant economic losses worldwide.

Occurrence and transmission. *Anaplasma* resides in erythrocytes within a vacuole formed from the host cell membrane. The rickettsiae remain within the vacuole and multiply by binary fission, producing numerous rickettsiae. The vacuole—which can be identified in stained blood films as a round, densely staining body approximately 1 micrometer in diameter—is referred to as an inclusion body. *Anaplasma marginale* contains both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), with a total genome size of 1200–1260 kilobase pairs (kbp). Molecular studies based on the analysis of ribosomal RNA relate *Anaplasma* closely to *Ehrlichia* spp.

Transmission of *Anaplasma* occurs biologically by multiple species of ticks. In the United States *Dermacentor* spp ticks are the primary vectors, but some 20 species act as vectors worldwide. Rickettsiae are ingested with a blood meal and undergo complex development beginning in the tick gut cells. *Anaplasma* is transmitted from the salivary glands while ticks feed on the host. Infections acquired by one tick stage may be transmitted after molting to the next stage (transtadial). Ticks may remain infected indefinitely (without feeding) and serve as a reservoir for infection of susceptible cattle. Mechanical transmission occurs when the mouthparts of biting flies (such as horse flies and stable flies) become contaminated while feeding on infected cattle and then quickly move to uninfected animals. There is no development of rickettsiae in these mechanical vectors. Contaminated needles and instruments for dehorning, castration, and tagging may also transmit the organism throughout a herd. Isolates of *Anaplasma* from different geographic areas differ from each other in biology, morphology, and antigenic characteristics.

Epidemiology. Once cattle are infected with *Anaplasma*, they remain persistently infected for life and may serve as reservoirs of infection for other cattle or tick vectors. During this time the parasitemia (presence of parasites in the blood) fluctuates and at times may be undetectable. The spread of anaplasmosis may occur when these carrier cattle are moved to nonendemic areas. Ticks have been shown to transmit new infections after feeding on animals with undetectable parasitemias. Carrier animals may also have relapse infections, producing new infections in susceptible herds.

Five major surface proteins of *Anaplasma* have been identified, and some have been found to be encoded by polymorphic multigene families. This variability results in distinct geographic strains of *Anaplasma* which are antigenically different and complicate vaccine strategies.

Clinical disease. Once an animal is infected it takes approximately 21 days before organisms can be identified in blood smears; but animals may not exhibit clinical disease for as long as 60 days postinfection. In acute anaplasmosis, cattle exhibit depression, loss of appetite, increased temperature, labored breathing, dehydration, jaundice, and a decrease in milk production. Erythrocyte count, packed cell volume, and hemoglobin values decrease significantly, and death may result from severe anemia. Abortions may also occur following infection. These clinical manifestations of anemia are associated with the rapid removal of infected and uninfected erythrocytes by macrophages (phagocytic cells of the immune system). Recovered animals gradually regain condition but remain chronically infected and are subject to periodic relapses. They may also serve as a source of infection for susceptible animals. The disease is most severe in older animals not previously exposed; calves up to 1 year old become infected but usually

have a mild or subclinical (showing no symptoms) reaction.

Diagnosis. Parasites can be identified as a round inclusion body at the periphery of red blood cells in a Giemsa-stained blood film. Infection can be confirmed by identifying specific antibodies in serological tests such as the competitive enzyme-linked immunosorbent assay (C-ELISA). Molecular amplification techniques such as the polymerase chain reaction (PCR) are also used, particularly when parasite density is very low or undetectable.

Immunity. Cattle that recover from clinical anaplasmosis are protected from subsequent exposure to the same geographic isolate of *Anaplasma*. Calves have a natural immunity to clinical disease if exposed within their first year. Protection is attributed to a complex combination of antibody and cell-mediated responses, with initial responses directed against the outer-membrane proteins of *Anaplasma*. Animals develop high levels of antibody during acute and recovery phases of anaplasmosis, but passive transfer of the antiserum (the serum component of blood that contains antibodies specific to one or more antigens) does not convey protection to naive animals (which have not been previously exposed to the disease). Splenectomy of recovered animals results in recrudescence (return of symptoms) of the rickettsemia (presence of rickettsiae in the blood), suggesting a role for cellular immunity in protection. Cell-mediated mechanisms of protection may involve macrophage activation and specific T-lymphocyte responsiveness. Specific antibodies may coat infected erythrocytes, enhancing their susceptibility to phagocytosis by activated macrophages. The ability of the rickettsiae to persist within the host at low levels despite a strong immune response is thought to be the result of antigenic variants created during cyclical rickettsemia. *See* ANTIBODY; ANTIGEN; CELLULAR IMMUNOLOGY; IMMUNITY.

Control and treatment. The clinical manifestations of anaplasmosis can be halted or prevented if tetracycline antibiotics are administered early. The drug inhibits rickettsial protein synthesis but does not kill the organism. Tetracyclines have been added to feeds to prevent clinical symptoms, but this does not prevent infection. Controlling fly and tick populations by chemical spraying or dipping have also been used to reduce the spread of disease.

Vaccination with killed *A. marginale* from erythrocytes has been shown to provide some protection against the acute disease but it does not prevent infection. Blood-derived vaccines are no longer used in most of the United States. The related and less virulent species, *A. centrale*, is used as a live-blood vaccine in some countries and provides some protection against *A. marginale*, but it can revert to a virulent form. Recombinant surface proteins of *A. marginale* have been evaluated as candidate vaccines and may be included in next-generation vaccines. In areas where anaplasmosis is endemic and control measures are not applied, most calves become infected, resulting in endemic stability with minimal clinical disease but high infection rates. *See* ANTIBIOTIC; VACCINATION.

Economic impact. Anaplasmosis continues to be a major disease of cattle in the tropical and subtropical world. Economic losses result from antibiotic control and vaccine costs, vector control, and production losses (such as decreased weight gain) from infected and recovered cattle and deaths. It is the most important tick-borne disease of cattle in the United States. *See* RICKETTSIOSES.

Edmour F. Blouin

Bibliography. A. F. Barbet, Recent developments in the molecular biology of anaplasmosis, *Vet. Parasitol.*, 57:43–49, 1995; E. F. Blouin et al., Applications of a cell culture system for studying the interaction of *Anaplasma marginale* with tick cells, *Animal Health Res. Rev.*, 3(2):57–68, 2002; J. de la Fuente et al., Molecular phylogeny and biogeography of North American isolates of *Anaplasma marginale* (Rickettsiaceae: Ehrlichiae), *Vet. Parasitol.*, 97:65–76, 2001; D. M. French et al., Expression of *Anaplasma marginale* major surface protein 2 variants during persistent cyclic rickettsemia, *Inf. Immun.*, 66(3):1200–1207, 1998; K. M. Kocan, Adaptations of the tick-borne pathogen, *Anaplasma marginale*, for survival in ticks and cattle, *Exp. Appl. Acarol.*, 28:9–25, 2002; G. H. Palmer and T. F. McElwain, Molecular basis for vaccine development against anaplasmosis and babesiosis, *Vet. Parasitol.*, 57:233–253, 1995.

Anapsida

Formerly, a group of reptiles that was recognized in rank-based classifications. Amniotes were once divided into subclasses, based on the patterns of temporal fenestrae (openings in the skull roof posterior to the openings for the eyes). Subclass Synapsida included taxa that exhibit a single pair of temporal fenestrae (synapsid condition). Subclass Diapsida included taxa that feature two pairs of temporal fenestrae (diapsid condition) and those that are descended from ancestors that had them. The absence of temporal fenestrae is known as the anapsid condition. Mesosaurids, pareiasaurids, and captorhinids are examples of fossil reptiles that lack temporal fenestrae, whereas turtles are the only living reptiles that lack them; these reptiles were placed together in the subclass Anapsida. Certain Paleozoic reptiles (for example, millerettids) that exhibited synapsid-like temporal fenestrae but clearly did not belong in Synapsida (or Diapsida) were placed in Anapsida. *See* AMNIOTA; CAPTORHINIDA; CHELONIA; DIAPSIDA; MESOSAURIA; REPTILIA; SYNAPSIDA.

Studies of early amniotes that use cladistic methodology, in which derived features are used to group taxa, revealed that the absence of temporal fenestrae is a primitive feature. Moreover, amniote taxa that had been placed in the subclass Anapsida in rank-based classifications were found to form a paraphyletic (artificial) group at the base of Reptilia in cladograms. Phylogenetic systematists refuse to recognize artificial groups with formal taxon names;

but with the advent of phylogenetic nomenclature, attempts were made to conserve traditional group names by linking them with extant representatives, in order to adapt them as "clade names." Thus, Anapsida was linked to a clade that included turtles. In the earliest cladistic studies of amniotes, turtles and captorhinids formed a clade, which was called Anapsida. Subsequent studies placed turtles in a clade that contained pareiasaurs, mesosaurs, and their close relatives, and the name Anapsida was transferred to that clade. That act generated confusion because captorhinids, which are regarded by paleontologists as exemplary fossil reptiles with anapsid skulls, were no longer to be regarded as anapsid reptiles. *See* ANIMAL SYSTEMATICS; PHYLOGENY; TAXONOMY.

Recent discoveries have led taxonomists to advocate the abandonment of Anapsida as a taxonomic entity. Paleontological research indicates that synapsid-like temporal fenestrae arose at least four times in the clade of pareiasaurs, mesosaurs, and their close relatives. Morphological and molecular studies now place turtles in Diapsida, a corollary of which is that the anapsid condition of turtles can be interpreted as a derived feature of turtles. If turtles are truly nested within Diapsida, the name Anapsida can no longer be used as a clade name (because the phylogenetic definition for Anapsida used Diapsida as a reference taxon).

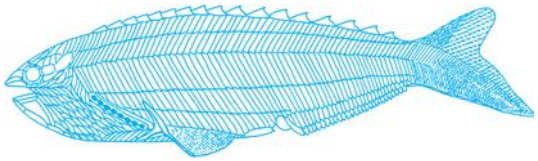
Whereas it is no longer appropriate to use the term anapsid in a taxonomic context, it can still be used in a morphological sense to describe the skull of an amniote that lacks temporal fenestrae. Thus, the anapsid condition was common among Paleozoic amniotes, where it represented the retention of a primitive feature in those taxa. Conversely, the absence of temporal fenestrae in turtles is a derived feature of these reptiles (if the diapsid origin of turtles is correct). Sean Modesto

Bibliography. M. J. Benton, *Vertebrate Palaeontology*, 3d ed., Blackwell, Oxford, 2004; R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, New York, 1988; A. S. Romer, *Vertebrate Paleontology*, 3d ed., University of Chicago Press, 1966.

Anaspida

A group of fossil jawless fishes, roughly similar in appearance to extant lampreys, known from Early Silurian–Early Devonian deposits (about 410–430 million years ago) that represent marine, coastal environments. Anaspid fossils are known predominantly from the Northern Hemisphere.

Most anaspids, such as *Rhyncholepis*, are small, rarely exceeding 15 cm (6 in.), with a rather slender and laterally compressed body. A slanting row of 7–15 branchial (gill) openings is present between the head region and the trunk. Triradiate postbranchial spines at the ventral end of the branchial row characterize the group. The head of most anaspids is covered by a series of smaller plates, in contrast to the massive headshield seen in most other contemporary jawless fishes (such as Osteostraci, Heterostraci), and the trunk bears rod-shaped scales arranged in a chevronlike pattern in five variously inclined overlapping rows. The relatively fragile nature of this dermal skeleton means that well-preserved articulated fossils are rare but isolated scales and plates are fairly common. The anaspids are also characterized by a median dorsal row of peculiar, often hook-shaped spines/scales. The tail is strongly hypocercal (the muscular lobe containing the notochord turns downward). Some articulated specimens have traces of paired ventrolateral fins, which vary in length from genus to genus (see **illustration**).



Reconstruction of *Rhyncholepis parvula* from the Silurian of Norway. (*Modified from A. Ritchie, 1980; reproduction by permission of Royal Society of Edinburgh, Trans. Roy. Soc. Edinburgh: Earth Sci., 92:263–323, 2002*)

Due to their close external similarities to living lampreys, anaspids have traditionally been regarded as close relatives or ancestors of the former. However, current views on the relationships of early vertebrates suggest that anaspids, together with other fossil jawless vertebrates, are more closely related to the jawed vertebrates (gnathostomes) than to lampreys. Their exact position, though, on the family tree is still a matter of considerable debate, mainly due to a lack of information regarding their internal anatomy. *See* ACANTHODII; HETEROSTRACI; JAWLESS VERTEBRATES; OSTEOSTRACI. Henning Blom

Bibliography. H. Blom, T. Märss, and G. C. Miller, Silurian and earliest Devonian Birkeniid anaspids from the Northern Hemisphere, *Trans. Roy. Soc. Edinburgh: Earth Sci.*, 92:263–323, 2002; P. Janvier, *Early Vertebrates*, Oxford Monogr. Geol. Geophys. 33, Oxford University Press, 1996.

Anaspidacea

An order of the crustacean superorder Syncarida, the most primitive of the Eumalacostraca. The families Anaspididae and Koonungidae have living representatives, while the Clarkecaridae are true fossils. The families Gampsonychidae and Palaeocaridae, formerly included in this order, now are assigned to a new order, Palaeocaridacea. The Anaspidacea arose during the Paleozoic Era, and the recent species are now restricted as living fossils to freshwater habitats in Tasmania and South Australia (not New Zealand). They exhibit a great adaptability to special habitats. The first thoracic somite is incorporated with cephalic tagmata.

Anaspididae. *Anaspides tasmaniae* has a body which is similar to that of the amphipods. It is about 2 in. (5 cm) long and is found in mountain lakes and cold rivers at elevations of 2000–4000 ft

(600–1200 m) on Mount Wellington. *Paranaspides lacustris*, from the same region, is 1.2 in. (3 cm) long and resembles *Mysis*. *Micrasides calmani* is the smallest species and is found associated with *Sphagnum*. Anaspidites antiquus is from the Triassic of Australia. *See* AMPHIPODA.

Koonungidae. *Koonunga cursor* is 0.8 in. (2 cm) long and is found in the mud bottom of temporary springs near Melbourne. The eyes are sessile and the first thoracic limb is modified for digging. *See* SYNCARIDA.

Hans Jakobi

Bibliography. R. D. Barnes, *Invertebrate Zoology*, 1968; H. K. Brooks, On the fossil Anaspidacea, with a revision of the classification of the Syncarida, *Crustaceana*, 4:239–242, 1962; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; R. Siewing, Uber die Verwandtschaftsbeziehungen der Anaspidaceen, *Zool. Anz. Suppl.*, 18:242-252, 1955; G. Smith, On the Anaspidacea, living and fossil, *Quart. J. Microsc. Sci.*, 53:489-578, 1909.

Anatomy, regional

The detailed study of the anatomy of a part or region of the body of an animal, most commonly applied to regional human anatomy. This is in contrast, but supplementary, to the study of organ systems, such as the cardiovascular, where all the structures pertaining to the system are studied in their continuity. To obtain the maximum information about the anatomy of an animal, both methods are used, as well as histological, embryological, and functional studies.

The various regions of the body have particular interest for certain medical and other scientific specialists. For example, the obstetrician has a detailed knowedge of the perineal region, and the ophthalmologist of the eye. Regional anatomy is of great importance to the surgeon, and certain texts concerning it are labeled surgical anatomy.

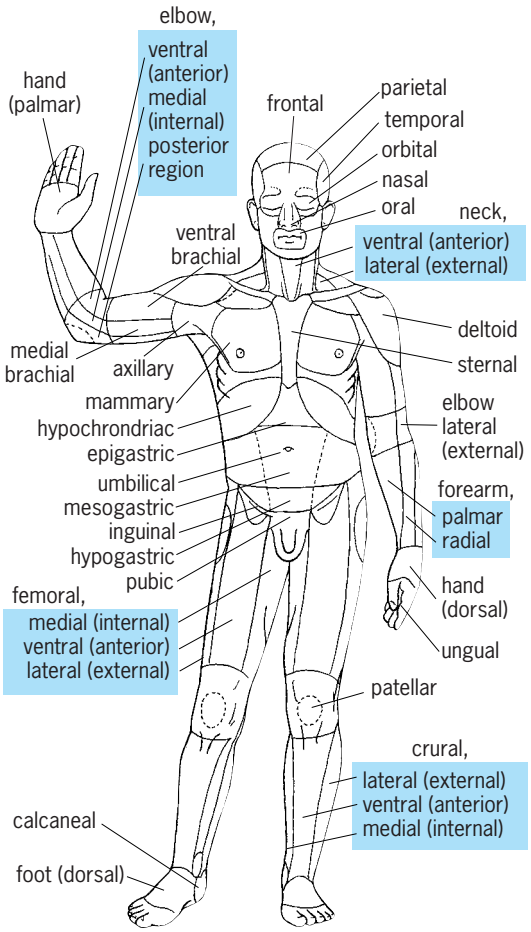
Closely related to regional anatomy is topographic anatomy, in which bony and soft tissue landmarks on the surface of the body are used to indicate the known location of deeper structures. An example is the point halfway between the umbilicus and the ventral prominence of the right hipbone; this marks the approximate location of the vermiform appendix and is known as McBurney's point.

There are many methods of dividing the body into regions for study, and one such means of classification is shown in the **illustration**. This system includes only externally visible areas; other systems would include special internal regions as well.

The head, trunk, and extremities are the principal regions. Subdivision of these can be carried out, as illustrated, so that many distinct areas or especially vital regions are indicated. There is no end to the dividing and subdividing a specialist may do to make his task eventually less difficult.

Regional anatomy is the natural and historical approach to the study of human and animal forms. The ancient scientists did not know much about the systems of the body except through conjecture. They approached the dissection of animals and cadavers in a manner intended to display the structures of some part, such as limb. The integration of such information led to the development of systematic anatomy.

Thomas S. Parsons



Some of the major regions of the human body seen in ventral view. (*After B. Anson, ed., Morris' Human Anatomy, 12th ed., McGraw-Hill, 1966*)

Andalusite

A nesosilicate mineral, composition Al_2SiO_5, crystallizing in the orthorhombic system. It occurs commonly in large, nearly square prismatic crystals. The variety chiastolite has inclusions of dark-colored carbonaceous material arranged in a regular manner. When these crystals are cut at right angles to the *c* axis, the inclusions form a cruciform pattern (**Fig. 1**). There is poor prismatic cleavage; the luster is vitreous and the color red, reddish-brown, olive-green, or bluish. Transparent crystals may show strong dichroism, appearing red in one direction and green in another in transmitted light. The specific gravity is 3.1–3.2; hardness is 7.5 on Mohs scale, but may be less on the surface because of alteration.

Occurrence and use. Andalusite was first described in Andalusia, Spain, and was named after this locality. It is found abundantly in the White Mountains near

Fig. 1. Andalusite, variety chiastolite. Prismatic crystal specimens from Worchester County, Massachusetts. (*American Museum of Natural History specimens*)

Laws, California, where for many years it was mined for manufacture of spark plugs and other highly refractive porcelain. Chiastolite, in crystals largely altered to mica, is found in Lancaster and Sterling, Massachusetts. Water-worn pebbles of gem quality are found at Minas Gerais, Brazil. *See* SILICATE MINERALS.

Cornelius S. Hurlbut, Jr.

Aluminum silicate phase relations. The three polymorphs of Al_2SiO_5 are andalusite (Fig. 1), kyanite, and sillimanite. These minerals occur in medium- and high-grade metamorphic rocks, and andalusite and sillimanite occur rarely in granites. Laboratory determination of the pressure-temperature stability fields of these minerals has been difficult and remains a subject of controversy, but the phase diagram shown in **Fig. 2** is generally accepted. It is consistent with thermochemical data and natural occurrence of the minerals. The univariant boundaries of the three stability fields meet at an invariant point located at 3.8 kilobars (380 megapascals) and 930°F (500°C).

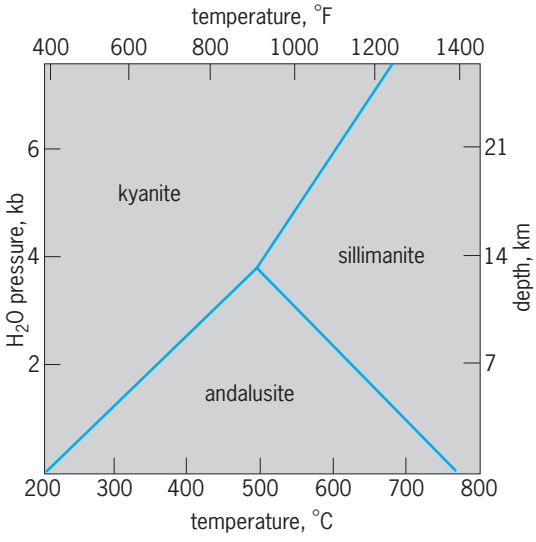


Fig. 2. Phase diagram for the polymorphs of aluminum silicate (Al_2SiO_5). 1 km = 0.6 mi; 1 kb = 10^2 MPa. (*After M. J. Holdaway, Stability of andalusite and the aluminum silicate phase diagram, Amer. J. Sci., 271:97–131, 1971*)

Grade (degree) of metamorphism is commonly a function of increasing temperature at some nearly constant or slowly changing pressure. Temperature variation is a reflection of changing temperature gradients (or heat sources) in the Earth, and pressure is an indication of thickness of overlying rock at the time of metamorphism. In pelitic (former shale) schists the grade or degree of metamorphism may be expressed by index minerals, such as the aluminum silicates. In regional metamorphic terranes the sequence of kyanite followed by sillimanite with increasing grade or temperature corresponds to pressures above 3.8 kbar (380 MPa) or depths greater than 8 mi (13 km) in the Earth's crust. In other regional metamorphic terranes the sequence andalusite to sillimanite indicates shallower levels in the crust (Fig. 2). For the New England region of the United States, the sequence is kyanite to sillimanite in Vermont, western New Hampshire,western Massachusetts, and Connecticut, while in Maine, eastern New Hampshire, eastern Massachusetts, and Rhode Island the sequence is andalusite to sillimanite, all of which formed during Early Devonian metamorphism. This indicates that central and western New England were metamorphosed at greater pressure (deeper in the crust) than northern and eastern New England. Thus, natural occurrence of aluminum silicates combined with knowledge of their phase relationships is a useful aid in deducing extent of vertical movement and erosion occurring in the Earth's crust after metamorphism. *See* KYANITE; METAMORPHISM; SILLIMANITE.

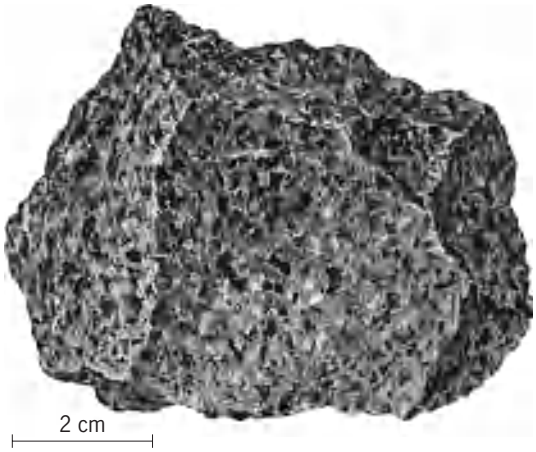
M. J. Holdaway

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 1a: *Orthosilicates* 2d ed., 1997; M. J. Holdaway, Stability of andalusite and the aluminum silicate phase diagram, *Amer. J. Sci.*, 271:97–131, 1971; D. M. Kerrick, The Al_2SiO_5 Polymorphs, *Reviews in Mineralogy*, vol. 22, 1990; C. Klein, *Manual of Mineralogy*, 21st ed., 1993; E-an Zen, W. S. White, and J. B. Hadley (eds.), *Studies of Appalachian Geology: Northern and Maritime*, 1968.

Andesine

A plagioclase feldspar with composition $Ab_{70}An_{30}$ to $Ab_{50}An_{50}$ (Ab = $NaAlSi_3O_8$; An = $CaAl_2Si_2O_8$). Andesine occurs primarily in igneous rocks, often in a glassy matrix as small, chemically zoned, lathlike crystals known as microlites. The rock types may be called andesinites (if dominantly feldspar), andesites (see **illus.**), andesitic basalts (or olivine-bearing andesites, as in Hawaiian lava flows), or pyroxene-, hornblende- or biotite-andesites (depending on the second most dominant mineral—all are volcanic). *See* ANDESITE.

The symmetry of andesine is triclinic, hardness on the Mohs scale 6, specific gravity 2.69, melting point ~1210°C (2210°F). If quenched at very high



Andesine grains with biotite in a specimen from Zoutpansberg, Transvaal, South Africa. Andesine is rarely found except as grains in igneous rocks.

temperatures, andesine has an albitelike structure, with aluminum (Al) and Silicon (Si) essentially disordered in the structural framework of the crystals. But, in the course of cooling, most natural andesines develop an Al-Si ordered structure with unusual, strain-induced, periodic features detectable only by single-crystal x-ray diffraction or by transmission electron microscopy. These are called e-plagioclases, based on the presence of non-Bragg diffraction maxima (e-reflections) in x-ray or electron patterns. *See* ALBITE; CRYSTAL STRUCTURE.

Calcic andesines and labradorites ($Ab_{55}An_{45}$–$Ab_{40}An_{60}$) may exsolve into two distinctly intergrown lamellar phases whose regularity of stacking produces beautiful interference colors like those in the feathers of a peacock. Polished specimens of this material are called spectrolite in the gem trade, and at some localities (notably eastern Finland) crystals up to 10 in. (25 cm) are mined by hand. Smaller crystals are made into cabochons for jewelry. They may be abundant enough in the host rock to be valued as a decorative stone. Hematite (Fe_2O_3) is present in rare oligoclases, and some andesines as micrometer-scale thin flakes, oriented parallel to certain structurally defined planes, producing brilliant reddish-gold reflections. These semiprecious stones are called aventurine or sunstone. *See* FELDSPAR; GEM; HEMATITE; IGNEOUS ROCKS; LABRADORITE.

Paul H. Ribbe

Andesite

A typical volcanic rock erupted from a volcano associated with convergent plate boundaries. The process of subduction, which defines convergent plate boundaries, pushes oceanic lithosphere beneath either oceanic lithosphere or continental lithosphere. Andesites are the principal rocks forming the volcanoes of the "ring of fire," the arcuate chains of volcanoes which rim the Pacific Ocean basin. The Marianas and Izu-Bonin islands, the islands of Japan, the Aleutian Islands, the Cascades Range of the northwest United States, the Andes mountain chain of South America, and the Taupo Volcanic Zone of New Zealand (**Fig. 1**) are andesitic. *See* LITHOSPHERE; PLATE TECTONICS; VOLCANO.

Andesite volcanoes have been responsible for some of the most destructive, globally significant eruptions of historic and modern times. Examples include the 1815 eruption of Tambora and the 1883 eruption of Krakatau in Indonesia, the 1902 eruption of Mount Pelee in the West Indies, the 1980 eruption of Mount St. Helens in the United States, the 1991 eruption of Pinatubo in the Philippines, and the 1995–1998 eruption of the Soufriere Hills in the West Indies. The activity of most active andesite volcanoes is monitored continuously, but loss of life can still occur. For example, mud flows (lahars) from Nevado del Ruiz volcano in Colombia caused around 25,000 deaths in November 1985.

Andesites are mostly dark-colored vesicular volcanic rocks which are typically porphyritic (containing larger crystals set in a fine groundmass). Phenocrysts (the larger crystals) comprise plagioclase; calcium-rich, calcium-poor pyroxene; and iron-titanium oxides set in a fine-grained, frequently glassy, groundmass. Some andesites contain phenocrysts of olivine, and some contain amphibole and biotite; these latter rocks generally contain more potassium. The porphyritic nature of andesites is derived from a complicated history of magmatic crystallization and evolution as the melts rise toward the surface from deep in the Earth. Phenocryst minerals commonly are strongly zoned and show evidence for disequilibrium during growth, consistent with an origin involving crystal fractionation and mixing processes. Andesites are readily classified in terms of their silicon dioxide (SiO_2) content, between 53 and 63 wt %, and potassium oxide (K_2O) content at a given SiO_2 content (**Fig. 2***a*; **table**). They can also be readily discriminated on a total alkali versus SiO_2 diagram (the TAS diagram; Fig. 2*b*). Most andesite volcanoes erupt lavas and tephras (volcanic ash) which range in composition from basaltic andesite to dacite. Eruptions are often explosive, reflecting the relatively high water and gas content of the magmas. During an eruption, a column may

Fig. 1. Mount Ruapehu (2797 m; 9140 ft) in the North Island, New Zealand, erupting July 1996. This active andesite volcano is at the southern end of the Taupo Volcanic Zone continental margin arc system. Tephra (volcanic ash) is shown falling from the eruptive column to blanket the landscape to the north (left) of the volcano.

Chemical analyses (expressed in percent) of typical andesites from a continental margin arc (Mount Ruapehu, New Zealand) and an oceanic island arc (Raoul Island, Kermadec Islands, Southwest Pacific)

Component	Ruapehu	Raoul	Average continental crust
	Percent		
SiO_2	59.44	57.27	59.1
TiO_2	0.67	1.23	0.70
Al_2O_3	16.92	14.10	15.8
Fe_2O_3	6.37	13.31	6.6†
MnO	0.10	0.23	0.11
MgO	4.03	2.93	4.4
CaO	6.79	7.52	6.4
Na_2O	3.38	2.69	3.2
K_2O	1.70	0.62	1.9
P_2O_5	0.12	0.19	0.2
LOI*	0.27	−0.47	
Total	99.79	99.62	98.41
	Trace elements, ppm		
Sc	16.7	41	22
V	152	294	131
Cr	51	4	119
Ni	25	4	51
Cu	49	102	24
Zn	68	122	73
Ga	17	13	16
Rb	59	8.5	58
Sr	274	188	325
Y	19	37	20
Zr	130	75	123
Nb	4.54	0.82	12
Cs	5.13	0.35	2.6
Ba	339	202	390
La	13.54	4.4	18
Ce	29.24	13.0	42
Pr	3.8		5.0
Nd	14.92	11.7	20
Sm	3.49	3.84	3.9
Eu	1.04	1.35	1.2
Gd	3.87	4.61	3.6
Tb	0.57	0.93	0.56
Dy	3.29	5.41	3.5
Ho	0.69		0.76
Er	1.87	3.15	2.2
Tm	0.30		
Yb	2.03	3.95	2.0
Lu	0.31	0.62	0.33
Hf	4.26	2.37	3.7
Ta	0.81	0.066	1.1
Pb	14.53	2.7	12.6
Th	5.47	0.563	5.6
U	1.63	0.233	1.4

* Loss on ignition (≈volatile content).
† All Fe as Fe_2O_3.

rise tens of kilometers above the volcano, leading to stratospheric dispersion of tephra and volcanic gases and generation of pyroclastic flows (for example, Pinatubo). Pyroclastic flows are a particular feature of andesite-type volcanism and are among the most dangerous of volcanic hazards. Indeed, it was pyroclastic flows from Vesuvius which overwhelmed Pompeii in A.D. 79; and at Mount Pelee on the island of Martinique in 1902 the pyroclastic flow engulfed the town of St. Pierre, killing around 28,000 people. *See* BASALT; LAVA; PYROCLASTIC ROCKS.

Less commonly, rocks of andesite and basaltic-andesite composition are associated with sites of intraplate volcanism, such as Iceland, Galápagos, and Hawaii, and are unrelated to subduction or orogenic processes. These rocks have basaltic-andesite to andesite composition (that is, in the composition range of 53–63% SiO_2) and have been called icelandite and hawaiite. They have formed in response to fractional crystallization of basaltic magma.

It is generally recognized that most andesites cannot be direct partial melts of peridotitic mantle. As such, andesites are evolved magmas and the end product of a plethora of processes, including crystal

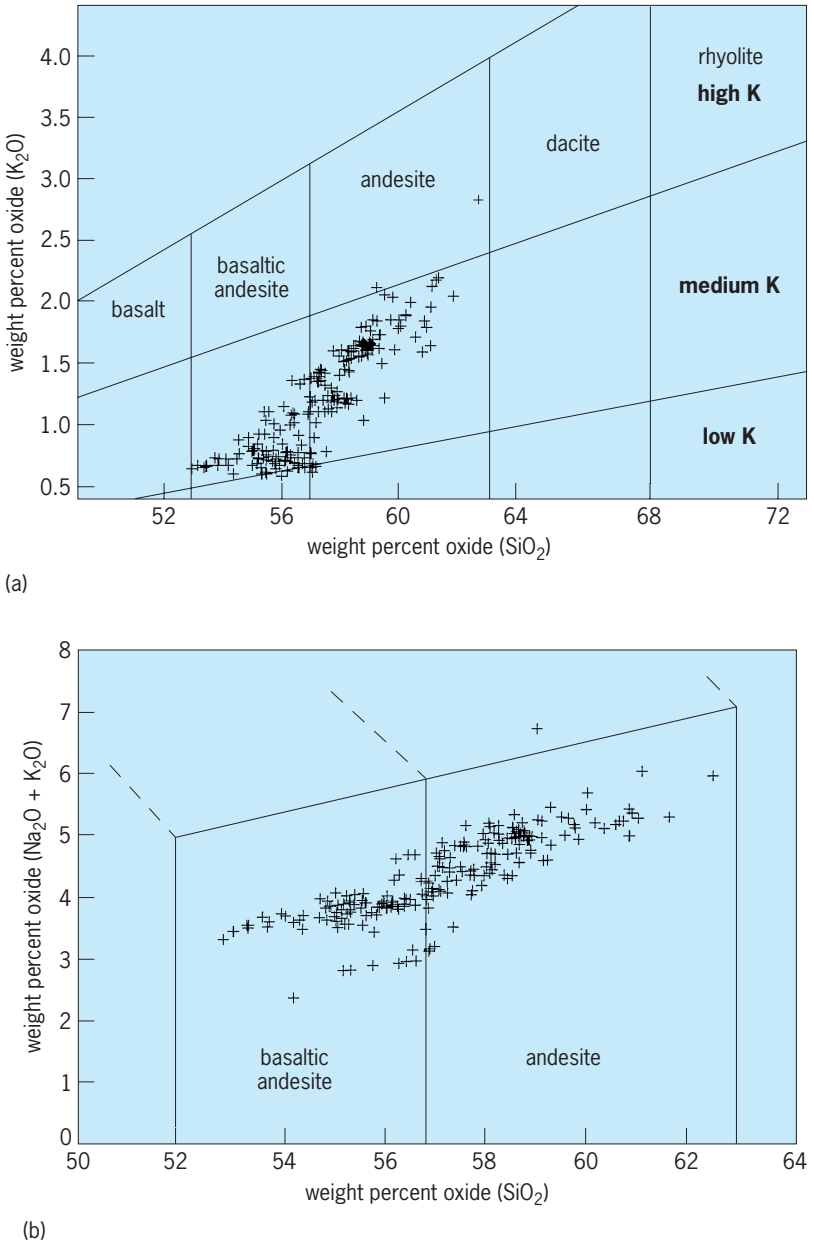


Fig. 2. Igneous rock and andesite classification. (*a*) Igneous rocks, in terms of SiO_2 versus K_2O. Note that a continuum exists between basalt, basaltic andesite, andesite, dacite, and rhyolite, and that it is possible to subdivide igneous rocks into low-, medium-, and high-K varieties. The data shown are andesites from Mount Ruapehu, which classify in this diagram as medium-K basaltic andesites and andesites. (*b*) Andesitic rocks, in terms of SiO_2 versus ($Na_2O + K_2O$), showing a part of the total alkali versus silica (TAS) diagram. (After *M. J. Le Bas et al., A chemical classification of volcanic rocks based on the total alkali-silica diagram, J. Petrol., 27:745–750, 1986*)

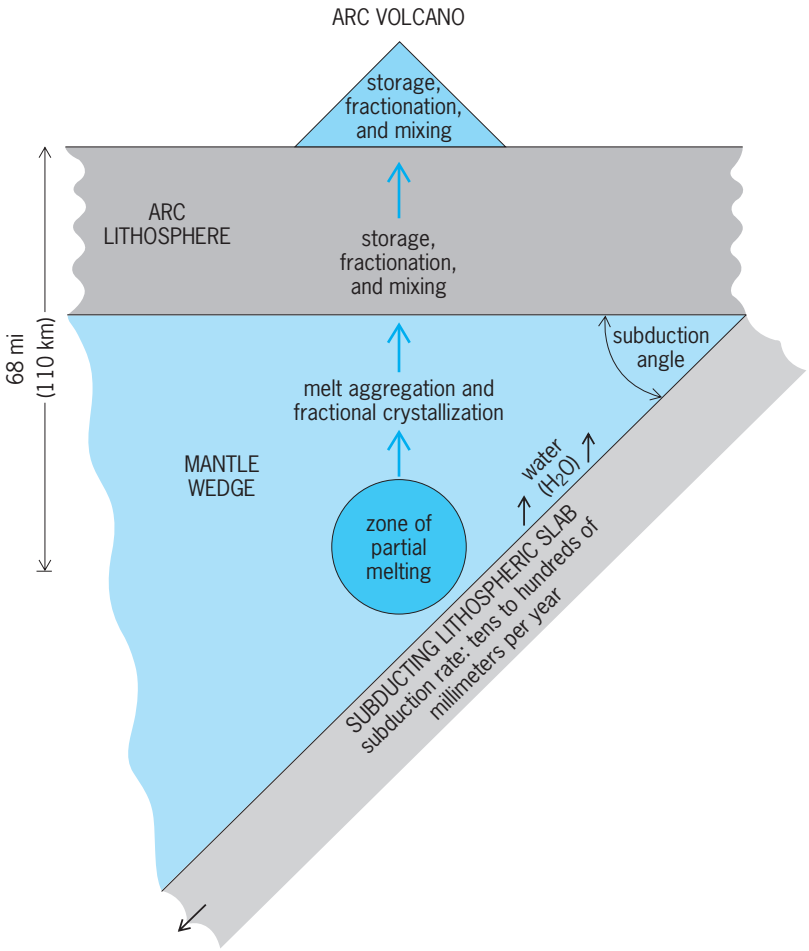


Fig. 3. Schematic cross section through a typical subduction system. The major components of a convergent plate boundary are the arc lithosphere, the subducting lithospheric slab, and the mantle wedge. Note that arc volcanoes overlie the subducting slab by around 110 km (68 mi), that the volume of the mantle wedge is determined by arc lithosphere thickness and slab dip, and that the mantle wedge is the principal site of magma production. After magmas form by partial melting, they can evolve further by fractional crystallization, mixing, and storage in reservoirs over a range of depths.

fractionation, magma mixing, mingling, assimilation, and storage, which have acted on primary basaltic magmas produced by partial melting in the mantle wedge. *See* MAGMA.

In more detail, the process of subduction recycles oceanic lithosphere back into the deep Earth; and as such, it complements the process of sea-floor spreading which generates new basalt crust and oceanic lithosphere at the mid-ocean ridges. The zone along which subduction takes place is demarcated at the Earth's surface by a deep oceanic trench and within the Earth by an increase of earthquake intensity along an inclined plane, the Wadati-Benioff Zone. This zone derives from the fact that cold, brittle material (the subducting oceanic lithosphere) is pushing into warmer upper mantle, and the earthquakes reflect the contrasting physical properties between the two mediums. The subduction process takes place at rates varying from a few tens to a few hundreds of millimeters per year and at angles of around 20° (shallow subduction) to nearly 90° (steep subduction). Typically, andesite volcanoes are located about 110 km (68 mi) above the surface of the subducting plate (**Fig. 3**). Between the surface of the subducting plate, which geologists refer to as the subducting slab, and the overlying plate is the mantle wedge, where basaltic melts form by the process of partial melting. Partial melting involves the generation of melt by dissolution of minerals, leaving a residual, more refractory, mineral assemblage. The melts aggregate and then move toward the surface by virtue of their lower density and viscosity. The melting occurs because the subducting slab, which may have spent many millions of years on the ocean floor as oceanic crust, contains minerals with seawater locked into their structure. When the minerals are subjected to higher pressure and temperature as subduction proceeds, chemical reactions in the minerals in the slab lead to loss of water and other volatile constituents which diffuse upward into the mantle wedge, together with an array of readily soluble elements such as potassium, rubidium, cesium, barium, and uranium. The mantle wedge becomes relatively enriched in these elements. The introduction of water is especially significant as it depresses the melting temperature and facilitates partial melting among constituent mineral grains. Importantly, the mantle wedge is recognized as the source of most arc-related magmas. On rare occasions, where subducting oceanic lithosphere is young, is still hot, and is subducting rapidly, the slab itself may experience partial melting, generating a unique suite of high-silica, low-potassium rocks of andesite composition, called adakites (after Adak Island in the Aleutian island arc chain). Still another unique suite of andesitic rocks, called boninites, which have high-magnesium and typically low-titanium contents, lacking plagioclase phenocrysts but with magnesium-rich, calcium-poor pyroxene phenocrysts, are associated with some suites of island arc rocks (for example, the Bonin island arc, which is the type area), where water-saturated melting of depleted mantle peridotite produced partial melts of high-silica, highmagnesium content. *See* SUBDUCTION ZONES.

The bulk composition of typical andesites is similar to that of continental crust (see table), contrasting with oceanic crust which is basaltic. Over geological time, subduction and the resulting andesite volcanism has proved a means of generating new continental crust, possibly at continental margin arc systems.

John Gamble

Bibliography. P. W. Francis, *Volcanoes: A Planetary Perspective*, Oxford University Press, 1993; J. A. Gamble et al., A fifty year perspective of magmatic evolution on Ruapehu Volcano, New Zealand: Verification of open system behaviour in an arc volcano, *Earth Planet. Sci. Lett.*, 170:301–314, 1999; J. B. Gill, *Orogenic Andesites and Plate Tectonics*, Springer-Verlag, Berlin, 1981; M. J. Le Bas et al., A chemical classification of volcanic rocks based on the total alkali-silica diagram, *J. Petrol.*, 27:745–750, 1986; A. R. McBirney, *Igneous Petrology*, Jones and Bartlett, Boston, 1993; R. L. Rudnick, Making continental crust, *Nature*, 378:571–578, 1995; R. S. Thorpe (ed.), *Andesites: Orogenic Andesites and Related Rocks*, Wiley, Chichester, 1982;

M. Wilson, *Igneous Petrogenesis: A Global Tectonic Approach*, Unwin Hyman, London, 1989.

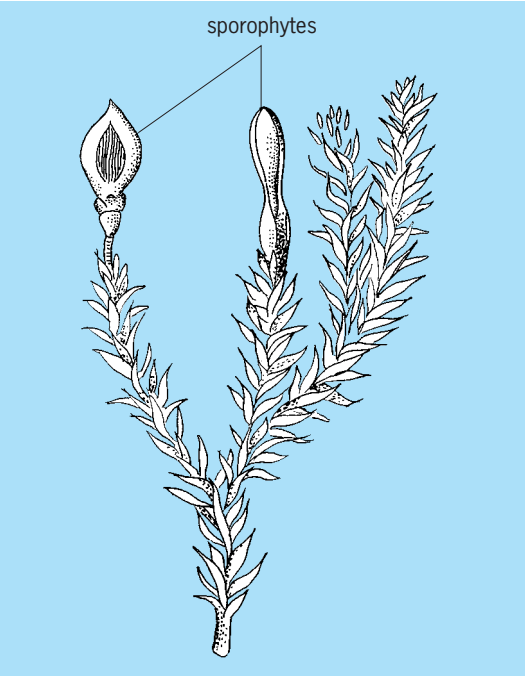
Andreaeopsida

A class of the plant division Bryophyta, containing plants commonly called granite mosses. The class consists of one order and two families, the Andreaeobryaceae and the Andreaeaceae. The large tapered foot and short, massive seta of the Andreaeobryaceae show possible linkage to the Bryopsida, but the differences seem more fundamental and more significant than the similarities.

The small, brittle plants grow as perennials on rocks in montane regions in dark red, red-brown, or blackish tufts. The stems are erect, simple or forked, and grow from a single apical cell with three cutting faces; they consist of a homogeneous tissue of thick-walled cells. The rhizoids are multicellular and filamentous or platelike. The leaves are spirally arranged, of various shapes and with or without a midrib. The leaf cells are thick-walled and often papillose. Paraphyses are present in both male and female inflorescences.

The sex organs are superficial in origin. The antheridia are long-ellipsoid and stalked. The archegonia are flask-shaped. The terminal sporophytes consist of a small foot and an ellipsoid capsule elevated on an extension of the gametophytic axis (see **illus**.), the pseudopodium (or, in *Andreaeobryum*, they consist of a well-developed foot, a short seta and no pseudopodium). The capsule dehisces by 4–8 vertical slits. The slender columella is surrounded (and overarched) by sporogenous tissue of endothecial origin. The wall consists of solid tissue. Stomata, peristomes, and elaters are lacking. The spores begin division within the spore wall before dehiscence of the capsule. The protonema is a branched filament producing numerous buds; erect filaments may produce at their tips leaflike plates. The calyptra is small and mitrate or, in *Andreaeobryum*, larger and campanulate-mitrate. *See* ANTHOCEROTOPSIDA; BRYOPHYTA; BRYOPSIDA; HEPATICOPSIDA; SPHAGNOPSIDA.

Howard Crum



Granite moss (*Andreaea ruperstris*), showing gametophyte with sporophytes. (*After G. M. Smith, Cryptogamic Botany, vol. 2, 2d ed., McGraw-Hill, 1955*)

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; W. Schultze-Motel, Monographie der Laubmoosgattung *Andreaea*, I: Die costaten Arten, *Willdenowia*, 6:25–110, 1970; W. C. Steere and B. M. Murray, *Andreaeobryum macrosporum*, a new genus and species of Musci from northern Alaska and Canada, *Phytologia*, 33:407–410, 1976.

Androgens

A class of steroid hormones produced in the testes and adrenal cortex which act to regulate masculine secondary sexual characteristics. The major naturally occurring androgens, such as testosterone, are C-19 steroids (that is, they all contain 19 carbon atoms) with a hydroxyl or ketone group at the C-3 and C-17 positions (see testosterone in **Fig. 1**).

Reproductive functions. Androgens play several important roles related to male reproductive functions. These steroids are responsible for the development of the entire reproductive tract, including the epididymis, seminal vesicles, and vas deferens, in the developing male embryo as well as the external genitalia, including the penis. At the onset of puberty, androgen secretion increases dramatically; once testosterone levels reach a plateau, they remain fairly constant until a man reaches the age of 70–80, after which they slowly decline. During puberty, these androgenic hormones are essential for development of the masculine secondary sex characteristics, which include a deepening of the voice, growth of a beard and coarse body hair, growth of the penis, and increased development of muscles. They are also required for development of the male sexual drive, or libido; initiation and maintenance of spermatogenesis, or sperm production; and the growth and function of accessory sex glands, including the prostate gland. *See* PROSTATE GLAND; SPERMATOGENESIS.

Although the interstitial cells (Leydig cells) in the male testes secrete testosterone, other less potent androgens, including dehydroepiandrosterone (DHEA) and androstenedione, are secreted by the cortex of the adrenal glands in both males and females. In females these adrenal androgens modulate the physiological function of reproductive and sexual organs including the ovaries, vagina, clitoris, and mammary glands. They are also responsible for growth of hair in the armpits and appear to be required for sexual arousal of the female. Overproduction of these adrenal androgens in a female can cause

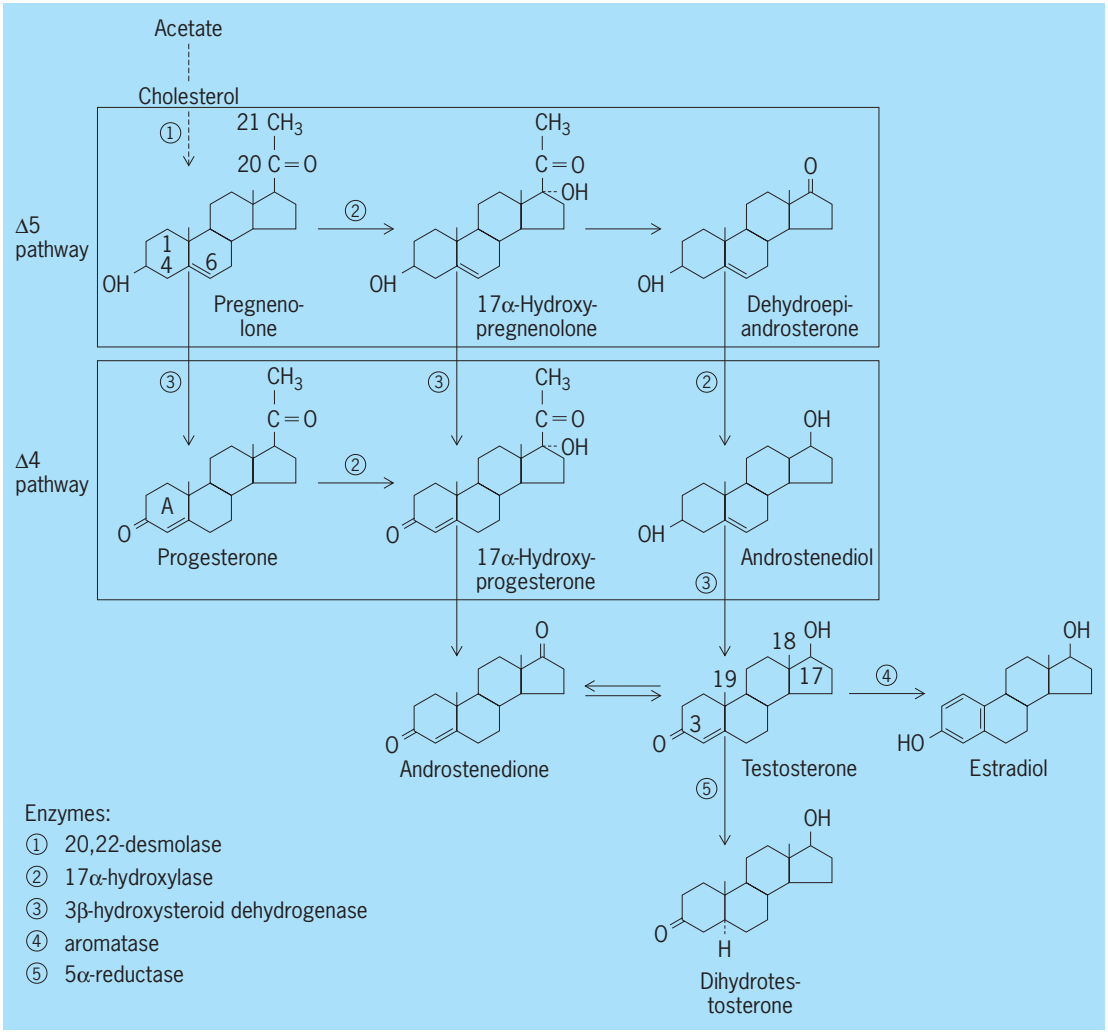


Fig. 1. Major pathways of testicular androgen and estrogen synthesis.

masculinization, which includes the growth of coarse body hair. Although the synthesis and secretion of adrenal androgens in both males and females is stimulated by adrenocorticotropic hormone, testosterone secretion by the Leydig cells of the male testes is specifically stimulated by luteinizing hormone. Both adrenocorticotropic hormone and luteinizing hormone are tropic hormones secreted by the anterior pituitary gland. *See* ADENOHYPOPHYSIS HORMONE; ADRENAL CORTEX; OVARY; REPRODUCTIVE SYSTEM; TESTIS.

Nonreproductive functions. Androgens also produce nonreproductive effects that include increased secretory activity of the sebaceous glands in a boy's skin at the time of puberty and fusion of the epiphyseal growth plate in the long bones, preventing an adult male from continuing to grow taller. DHEA and its sulfated ester, DHEA(S), have been reported to exert a number of potential health benefits that range from antiaging and anticancer effects to enhanced weight loss and stimulation of the immune system. However, the use of synthetic androgens, often referred to as anabolic steroids, to enhance muscle mass and athletic ability can have deleterious consequences. In a teenage boy, these undesirable side effects include not only acne and breast development but also a shutting down of bone growth. In adult males, overuse of these anabolic steroids can result in: infertility (decreased sperm production); liver damage; acceleration of the growth of prostate tumors; and the development of liver and brain tumors.

Biosynthesis. Like all steroid hormones, testosterone is synthesized from cholesterol, which in turn is either synthesized from acetate (Fig. 1) or derived from serum lipoproteins. The rate-limiting step for testosterone synthesis is the side-chain cleavage of cholesterol to form pregnenolone. The desmolase enzyme, located within the mitochondria of Leydig cells, catalyzes this key reaction. Once pregnenolone is generated, it is converted to testosterone via two different pathways. The $\Delta 5$ pathway is the major pathway in humans; in this pathway side-chain cleavage (carbons 20 and 21) and reduction of the 17-keto group occur before oxidation of the first ring (A ring) of the steroid nucleus. In the $\Delta 4$ pathway this sequence is required. The enzymes that catalyze these steroid modifications are identical for both pathways

and are located in the smooth endoplasmic reticulum of testosterone-secreting cells. Once the synthesized testosterone has been released into the bloodstream, it circulates throughout the body bound to proteins that are synthesized in the liver. Most of the secreted testosterone binds with high affinity to a serum transport protein known as sex steroid-binding globulin. Roughly 40% of the circulating testosterone is bound with low affinity to serum albumin. Although only 1–3% of the circulating testosterone is unbound, it is this free form of testosterone that is biologically active. *See* CHOLESTEROL; STEROID.

Further reduction of testosterone to generate the more potent androgen dihydrotestosterone (DHT) occurs in some androgen-responsive tissues, including cells in the prostate gland, that express the 5α-reductase enzyme (Fig. 1). This potent androgen binds to intracellular androgen receptors more tightly, or with higher affinity, than testosterone itself. Once either testosterone or dihydrotestosterone has bound to these receptor proteins located within target cells, these hormone-receptor complexes interact with other molecules (coactivators and corepressors) and bind directly to specific nuclear deoxyribonucleic acid (DNA) sequences. This nuclear binding subsequently regulates the expression or transcription of androgen-responsive genes. *See* GENE.

Conversion to estradiol. Testosterone can also be converted to estradiol via an enzyme that catalyzes the aromatization (introduction of double bonds) of the A ring of the steroid (Fig. 1). This reaction occurs primarily within the testicular Sertoli cells when they are stimulated by another anterior pituitary tropic hormone, follicle stimulating hormone (FSH). Similar amounts of estradiol can also be generated from testosterone within adipose tissue and the central nervous system. Although estradiol is a well-characterized female reproductive hormone, normal concentrations are also required for spermatogenesis, as well as regulation of bone density in males. Overproduction of estradiol in obese men can lead to feminization, including enlargement of the breasts. *See* ESTROGEN.

Catabolism. The metabolic breakdown, or catabolism, of testosterone takes place fairly rapidly in the liver. The major metabolites that are produced are 17-ketosteroids, generated from either androstenedione or dihydrotestosterone (**Fig. 2**). The major end product of this metabolic pathway is androsterone, which is excreted in the urine in the free form of conjugated with glucuronate or sulfate. DHEA, the predominant androgen secreted by the adrenal glands, is the major source of these excreted 17-ketosteroids. Other metabolites of testosterone that are found in small amounts in urine include

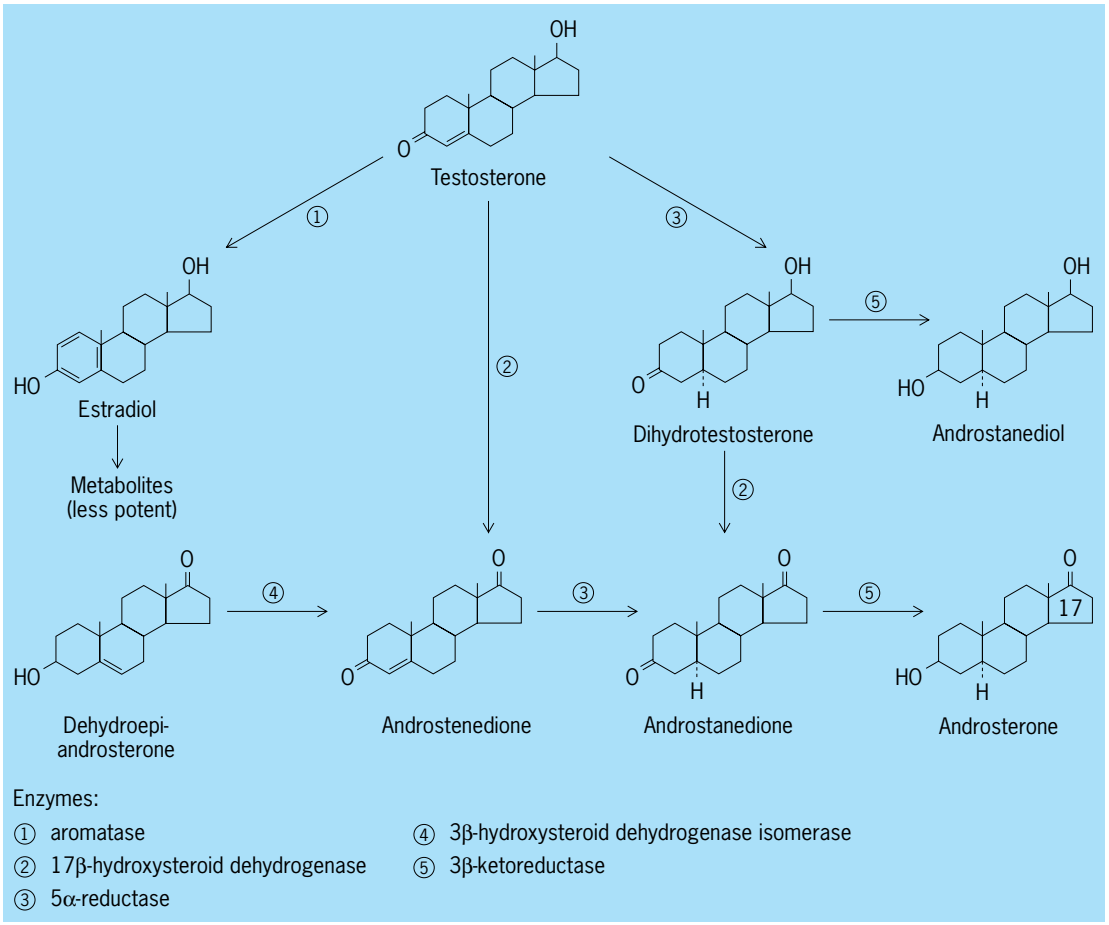


Fig. 2. Catabolism of androgens to generate active metabolites or inactive 17-ketosteroids.

androstanediol, which is formed by the reduction of the 3-keto group of dihydrotestosterone, and estrogen metabolites (estrone and estriol) which are formed after testosterone has been converted to estradiol. Thomas J. Schmidt

Bibliography. C. J. Bagatelli and W. J. Bremner, *Androgens in Health and Disease*, Humana Press, Totowa, NJ, 2003; R. M. Berne et al., *Physiology*, 5th ed., Mosby, St. Louis, 2003; C. Chang, *Androgens and Androgen Receptor: Mechanisms, Functions, and Clinical Applications*, Kluwer Academic Publishers, Boston, 2002.

Andromeda

A prominently located constellation in the northern sky (see **illustration**), named for the daughter of Cassiopeia in Greek mythology: When Cassiopeia bragged that her daughter Andromeda was more beautiful than Poseidon's daughters, the Nereids, Poseidon created Cetus, the sea monster. (In some versions of the myth, Cassiopeia boasted of her own beauty.) The situation required Andromeda's sacrifice. However, Perseus saved Andromeda by showing Cetus the head of Medusa, turning Cetus to stone. *See* CASSIOPEIA; PERSEUS.

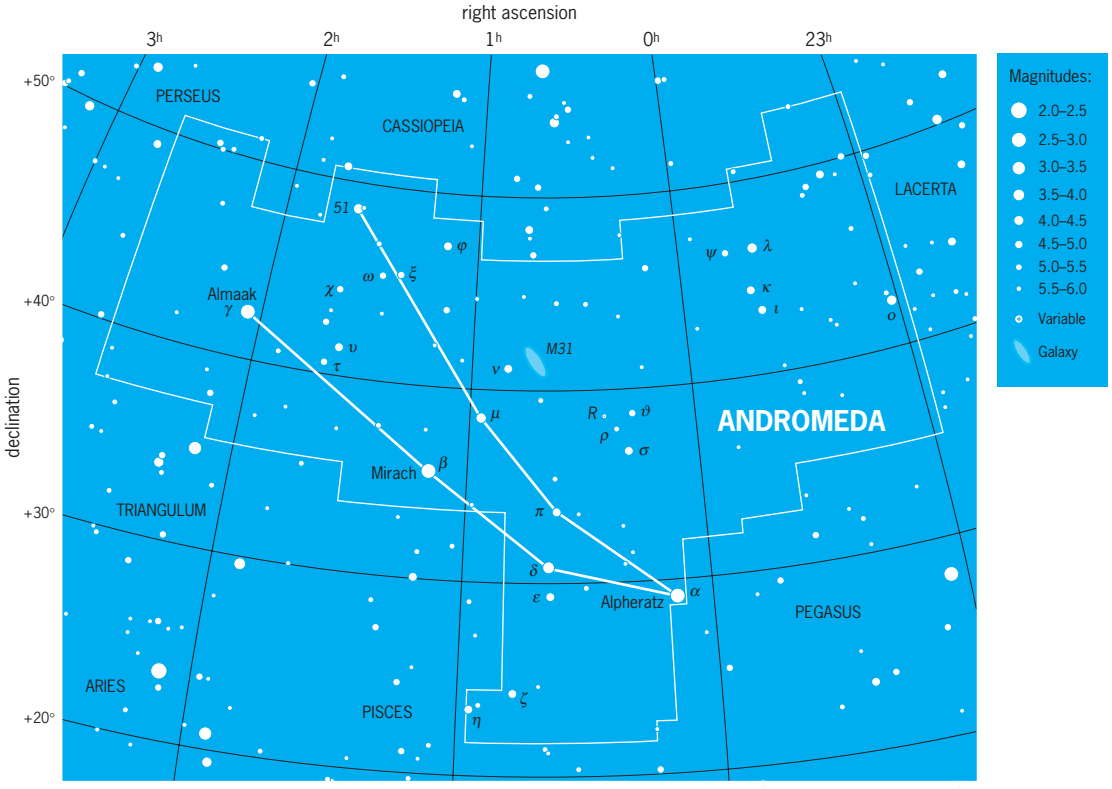
The Great Galaxy in Andromeda (M31) is the nearest spiral galaxy to the Earth, and along with the galaxy M33 in Triangulum is the farthest object that can be seen with the unaided eye, since the glow that comes from its core can be seen even though it is 2.4 million light-years away. Only photographs show the galaxy's extent. *See* ANDROMEDA GALAXY.

The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION. Jay M. Pasachoff

Andromeda Galaxy

The giant spiral galaxy nearest to Earth. The Andromeda Galaxy is seen in the constellation of Andromeda, which is visible from Earth in the autumn skies. The galaxy is also known as M31 and NGC 224, based on its entries in two widely used catalogs. *See* ASTRONOMICAL CATALOG; CONSTELLATION; MESSIER CATALOG.

The Andromeda Galaxy can be seen without a telescope, one of a very few external galaxies that are sufficiently close and bright. Visually through a telescope, it appears as a large, faint, nearly featureless object. But photography, especially with modern electronic detectors, shows that it is an immense system of stars and other material, arranged in a nearly edge-on disk and large, spherical, faint halo (**Fig. 1**). It has been mapped in great detail at many wavelengths, including radio, infrared, visual, ultraviolet, and x-ray.



Modern boundaries of the constellation Andromeda, daughter of Cassiopeia. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

Fig. 1. Photographic image of the Andromeda Galaxy obtained with the 1.2-m Schmidt telescope of the Palomar Observatory. (*Courtesy of Paul Hodge*)

The Andromeda Galaxy is one of 36 galaxies that are clustered together in the Local Group. It and the Milky Way Galaxy are the largest members and are two of only three spiral galaxy members. Its distance can be measured in a number of ways, most accurately by the use of the Cepheid variable period-luminosity law. The distance is found to be 2,500,000 light-years (1.5×10^{19} mi or 2.4×10^{19} km). *See* CEPHEIDS; LOCAL GROUP.

Size. The main body of the Andromeda Galaxy is about 3° across in the sky, corresponding to about 135,000 light-years (8×10^{17} mi or 1.3×10^{18} km). However, the hydrogen gas in the galaxy's disk can be detected to almost twice that size, and the dark-matter halo, the extent of which is only roughly measured, is even larger. (Dark matter is material of an unknown nature that can be detected only by its gravitational influence on visible matter. It appears to be the major constituent of galaxies and galaxy clusters.)

Contents. There are a few hundred billion stars (a few times 10^{11}) in the Andromeda Galaxy, most of them apparently quite similar in their properties to the stars in the Milky Way Galaxy. The Andromeda Galaxy also is similar in containing hundreds of star clusters, both globular clusters (**Fig. 2**) and smaller, young clusters, many of which have been studied in detail by the *Hubble Space Telescope*. *See* HUBBLE SPACE TELESCOPE; STAR; STAR CLUSTERS.

Between the stars and clusters there is an interstellar medium of gas, mostly atomic hydrogen, detected by radio telescopes. The gas is very tenuous except in certain areas where it is dense enough to condense into new stars. Star-forming regions are made up of both gas and dust, the gas in these places being largely molecular hydrogen, as well as other molecules. The inner parts of the Andromeda Galaxy are relatively free of atomic hydrogen, but contain much cool molecular gas. Currently star formation is concentrated at intermediate distances from the center, where the spiral arms are most conspicuous. The total mass of gas is a few billion times the mass of the Sun. *See* INTERSTELLAR MATTER; RADIO ASTRONOMY.

There is also dust located throughout the disk of the galaxy, visible both at optical and at infrared wavelengths. The northwest side of Andromeda is heavily obscured by this dust, as seen in telescopic images. The total mass of dust in the galaxy is about 1/100 the mass of gas.

Early results from space-borne x-ray telescopes searching for x-ray sources in the Andromeda Galaxy turned up a few dozen sources, mostly located in the inner parts. Searches with the *Chandra X-ray Observatory* have pinpointed hundreds of x-ray sources, many of which correspond in position with globular clusters. Most of them are neutron stars like those found in Milky Way clusters. *See* CHANDRA X-RAY OBSERVATORY; NEUTRON STAR; X-RAY ASTRONOMY.

Structure. The nucleus of the Andromeda Galaxy is unusual. As discovered by the *Hubble Space Telescope*, it is a double nucleus, the two components being too close together to have been resolved by ground-based telescopes. The two revolve around each other, and one contains a black hole of several million solar masses. *See* BLACK HOLE.

Surrounding the nucleus is an ellipsoidal bulge, made up of old and intermediate-age stars.

Fig. 2. Globular star cluster G1 in the Andromeda Galaxy, imaged by the *Hubble Space Telescope*. (*NASA*)

Enveloping the whole disk and extending even farther out than the bulk of the disk stars is a nearly spherical halo, primarily containing thinly spaced old stars. However, unlike the halo of the Milky Way Galaxy, that of the Andromeda Galaxy is not mainly made up of ancient stars with low heavy-element abundances. The origin and evolution of the halo differs in some unexplained way from those of the Milky Way Galaxy. Paul Hodge

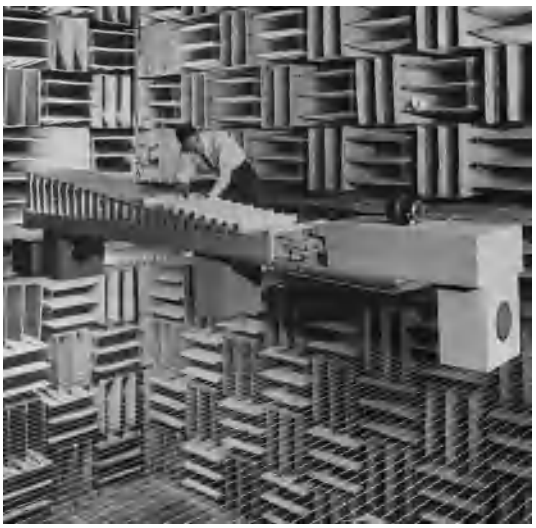
Bibliography. P. Hodge, *The Andromeda Galaxy*, Kluwer Academic, 1992; P. Hodge, *Atlas of the Andromeda Galaxy*, University of Washington Press, 1981; T. R. Lauer et al., Planetary camera observations of the double nucleus of M31, *Astron. J.*, 106: 1436–1447, 1993; B. F. Williams, Recent star formation history of the M31 disk, *Astron. J.*, 126:1312–1325, 2003.

Anechoic chamber

A room whose boundaries absorb effectively all the waves incident on them, thereby providing free-field conditions. A free field is a field whose boundaries exert negligible effect on the incident waves. In practice, it is a field in which the effects of the boundaries are negligible over the frequency range of interest. Acoustic chambers and radio-frequency and microwave chambers will be discussed.

Acoustic chambers. An acoustic anechoic chamber is a room in which essentially an acoustic free field exists. It is sometimes referred to as a free-field or dead room. The word anechoic is derived from the Greek, meaning "without echo." However, a room that is free from echoes (in the usual sense of the word) is not necessarily an anechoic room: in addition, there must be no significant reflections of sound waves from the boundaries of the room.

Free-field conditions can be approximated when the absorption by the boundaries of the room approaches 100%. To reduce sound reflected by the boundaries to a minimum, the absorption coefficient must be very high and the surface areas of the boundaries should be large.



Interior of an acoustic anechoic chamber. The room's special construction eliminates 99.9% of reflected sound.

The sound absorptive material usually installed in such rooms consists of glass fibers held together with a suitable binder. In order to achieve large surface area, a wall construction (see **illus.**) is used that includes wedges of sound absorptive material, the base of which is usually between 8 × 8 in. (20 × 20 cm) and 8 × 24 in. (20 × 60 cm), and the length of which is usually 3 to 5 ft (0.9 to 1.5 m). These wedges resemble stalagmites and stalactites and absorb about 99% of incident sound energy over most of the audio-frequency ranges. A horizontal net of thin steel cables just above floor wedges permits walking in the room. *See* REFLECTION OF SOUND; SOUND ABSORPTION. Cyril M. Harris

Radio-frequency and microwave chambers. The radio-frequency or microwave anechoic chamber is a shielded (screened) enclosure in which internal reflections of electromagnetic waves are reduced to an absolute minimum, thus providing a zone that is free from radio noise, broadcast transmissions, and other extraneous signals, and that simulates free-space conditions. Its main use is to provide a controlled and well-defined environment for the testing of electronic equipment.

Shielded or screened enclosures may vary in size from small metallic boxes that are intended for the protection of circuit elements to large rooms capable of housing vehicles or aircraft. The purpose of such enclosures is to isolate sources of electromagnetic energy from electronic, radio, or other sensitive equipment which might suffer degradation of performance as a result of interaction with transmitters, radio-frequency heating equipment, and so forth. The provision of a radio-frequency quiet zone is an essential feature of many applications in electromagnetic tests and investigations, for example, the determination of antenna characteristics and the measurement of low-level emissions during electromagnetic compatibility assessments. *See* ELECTRICAL SHIELDING; ELECTROMAGNETIC COMPATIBILITY; MICROWAVE FREE-FIELD STANDARDS.

The mechanism by which shielded enclosures function is a combination of absorption as the electromagnetic wave progresses through the material of the shield and reflection of the wave at the interface. The principle of reciprocity applies, and reflection occurs at the internal as well as at the external interface. In most practical shielded enclosures any radio-frequency signals generated internally are reflected and the bulk of the energy remains within the enclosure. At certain frequencies where the dimensions of the enclosure are comparable with half a wavelength, reinforcement of reflected waves take place and the enclosure becomes a resonant cavity. The internal conditions are thus somewhat different from free space, and it is the attempt to reproduce free-space or open-field conditions within a shielded enclosure which has led to the development of resistive material to line the enclosure walls to absorb rather than reflect the

radio-frequency energy and hence create an anechoic chamber. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; CAVITY RESONATOR; RECIPROCITY PRINCIPLE; REFLECTION OF ELECTROMAGNETIC RADIATION.

Theoretical investigations indicate that reflection-free conditions can be established at discrete frequencies by provision of sheets of resistive material that have a surface impedance of 120π (377) ohms per square, that is, the free-space or plane-wave impedance, and are spaced at one-quarter wavelength from the reflecting surfaces. Extension of this principle to achieve broadband coverage and hence an anechoic enclosure capable of use over a wide frequency range led to the development of various types of material having the appropriate resistive, lossy characteristics. Such material must have good basic dielectric properties, with a relative permittivity near unity to minimize reflection at the interface, and to incorporate increasing resistive loss through the material. In practice, foam plastics are ideal for the purpose, and these can be readily loaded with carbon, ferrite, or other conducting particles on a graded or layer basis to provide the increasing loss. Furthermore, the use of cones or pyramids of the material assists considerably in the transition from minimal reflection at the apex to full absorption at the base where the resistive loading is at a maximum. The lining of the walls and ceiling of a shielded enclosure with such materials provides for absorption of any internal signals and therefore anechoic or minimal reflection conditions simulating propagation in free space. *See* DIELECTRIC MATERIALS; ELECTROMAGNETIC RADIATION; PERMITTIVITY.

In the microwave band, that is, at frequencies greater than 1 GHz, the wavelength is less than 1 ft (30 cm), and hence a lining thickness of only 1 or 2 in. (2.5 or 5 cm) provides for a reflectivity better than −40 dB or less than 1%. Enclosures that are used for test and measurement purposes have walls and ceilings with conical or pyramidal linings up to 7 ft (2 m) thick and are capable of simulating free-space conditions with reflections less than 10% over a wide frequency range extending down to 50 MHz. Special precautions are taken with the installation of lossy material on the floor, because the latter generally needs to be mechanically strong in addition to having desirable radio-frequency characteristics. *See* RADAR-ABSORBING MATERIALS. G. A. Jackson

Bibliography. P. A. Chatterton and M. A. Houlden, *EMC: Electromagnetic Theory in Practical Design*, 1992; C. M. Harris, *Handbook of Acoustical Measurements and Noise Control*, 3d ed., 1991.

Anemia

A reduction in the total quantity of hemoglobin or of red blood cells (erythrocytes) in the circulation. Because it generally is impractical to measure the total quantity, measures of concentration are used instead. Hemoglobin is contained in red blood cells, which are suspended in plasma, the liquid component of blood. Therefore, concentration is affected not only by quantities of hemoglobin and red blood cells but also by plasma volume. Thus, the apparent anemia found in many women in the third trimester of pregnancy is not really anemia at all: the red cell mass is actually increased, but the plasma volume is expanded even more. In other words, hemodilution is present. Conversely, in dehydration and other circumstances of hemoconcentration, the plasma volume is reduced, thereby tending to mask anemia. *See* BLOOD; HEMOGLOBIN.

The three measures of concentration most often employed are the hemoglobin, the red blood cell count, and the volume of packed red cells. Hemoglobin is determined photocolorimetrically. The red blood cell count is determined by automated optical-electronic devices; until recently it was determined by microscopic observation of a counting chamber—a less accurate method. The volume of packed red cells, often called the hematocrit after the instrument originally used in its measurement, is a measure of the proportion of volume that red cells occupy in a specimen of whole blood, and is determined by centrifugal methods, or by computation from light scatter or displacement volume of individual cells during automated enumeration procedures. When total red cell mass must be known, it is measured by removing some of the patient's red cells and tagging them with the radioisotope chromium-51, reinjecting them, and then computing the red cell mass from the resulting dilution of the tagged red cells by the unlabeled red cells in the individual's bloodstream. For diagnostic and therapeutic purposes, anemia is categorized according to the average volume of the individual red cells (mean corpuscular volume) and the concentration of hemoglobin within the red cells (mean corpuscular hemoglobin concentration).

General considerations. In a group of healthy individuals, the values for hemoglobin, red cell count, and hematocrit approximate a "normal" distribution. Values that are less than 2.5 standard deviations below the mean are indicative of anemia if other clinical factors do not indicate a condition of hemodilution. The mean values are greater for adult males than for adult females, and greater for adults than children. Lower atmospheric oxygen tension at higher altitudes results in higher mean values for hemoglobin, red cells, and hematocrit in healthy individuals living under these conditions; hence, anemia would be defined at a higher value.

Red blood cells have a finite life-span, averaging 120 days in the circulation in healthy persons, but the span may be shortened in certain diseases. Every day, about 45 billion aged red cells are removed from each liter of blood, or a total of 2×10^{11} red cells per day in the adult of average size. In the state of health, the rate of production of new red blood cells (erythropoiesis) equals the rate of removal of senescent red blood cells. Anemia occurs if the rate of erythropoiesis is reduced below normal. It also occurs if hemorrhage or destruction (hemolysis) of red blood cells within the body increases the rate of loss of

erythrocytes and the rate of erythropoiesis does not increase enough to compensate.

Mechanisms. Red blood cell production may be affected by several different mechanisms. Erythropoietin, a growth factor produced by healthy kidneys, stimulates the bone marrow to produce more erythroblasts and accelerates their maturation. An intact bone marrow can respond by increasing production of red blood cells by about sevenfold after a few days or weeks. The proliferative response of the bone marrow to anemia may be defective or absent if the production of erythropoietin is diminished or absent, as occurs in chronic renal disease and some endocrine disorders. Erythropoietin produced synthetically has become available on prescription to overcome that deficiency and thereby resolve that specific type of anemia. *See* KIDNEY.

Ineffective erythropoiesis results when, in the intact, stimulated bone marrow, red blood cell precursors either fail to mature, or die in the bone marrow prior to their delivery to the circulation as erythrocytes.

Aplastic anemia occurs when the bone marrow stem cells, that give rise to precursors of erythroblasts, are markedly diminished in number or may respond inadequately to erythropoietin. This condition occurs when the bone marrow is adversely affected by certain chemicals or autoantibodies; is injured by irradiation; atrophies, or is replaced by fat; is replaced by fibrous (scar) tissue; or is infiltrated by cancer cells. Drugs and other chemicals that have been associated with aplastic anemia include those whose effects are dose-related and are an extension of their intended purpose, for example, some of the drugs used in cancer chemotherapy. Other drugs, notably chloramphenicol and benzol, seem to operate in an idiosyncratic fashion; many other agents have been implicated, though each in fewer cases.

Defective synthesis of deoxyribonucleic acid (DNA) and abnormal nuclear maturation result from malabsorption of vitamin B_{12} (as in pernicious anemia) or dietary deficiency of folic acid or its malabsorption (sprue). Methotrexate, fluorouracil, hydroxyurea, mercaptopurine, and other drugs used in chemotherapy affect purine and pyrimidine metabolism and can produce the same lesion. All of these anemias are generally characterized by large red blood cells, which are called macrocytes, and a specific morphological abnormality of the nuclear chromatin of erythroblasts which characterizes them as megaloblasts, and the condition as megaloblastic anemia. *See* VITAMIN B_{12}.

Defective synthesis of hemoglobin impairs cytoplasmic maturation. The majority of cases are due to deficiency of body stores of iron and to abnormal release of iron from reticuloendothelial stores. The former occurs in iron-deficiency anemia, and the latter in the anemia of chronic inflammatory diseases. Qualitative abnormalities of globin synthesis occur in the hereditary hemoglobinopathies (as in sickle-cell anemia, hemoglobin-C disease), whereas reductions in the rate of synthesis of either the α- or β-globin chains, in the absence of qualitative abnormalities, occurs in the thalassemia syndromes. The hemoglobinopathies and thalassemia may be inherited independently of one another or simultaneously. Heterozygotes have little if any anemia, whereas homozygotes and double heterozygotes have moderate or marked anemia. Rarely, disorders affect the porphyrin moiety of the heme prosthetic group in globin (as in lead intoxication and hereditary erythropoietic porphyria).

Acute blood loss reduces the total blood volume and produces symptoms of weakness, dizziness, thirst, faintness, and shock, in that order, according to increasing magnitude of blood loss. The anemia which results is not detectable by measures of concentration until hemodilution occurs over subsequent days, or more rapidly if replacement fluids are given intravenously. The proliferative response of the healthy marrow will correct the anemia in 2–6 weeks, depending upon the size of the deficit and provided there are sufficient body stores of iron, folic acid, and vitamin B_{12}, required for hemoglobin synthesis. Chronic blood loss results in iron-deficiency anemia.

Hemolysis, the accelerated destruction of red blood cells, also induces a proliferative response from the marrow. However, it differs from hemorrhage because red blood cells are lost without plasma, and it thus diminishes the measures of concentration at the outset. Disorders external to the red blood cell (such as autoantibodies, chemicals including drugs, splenomegaly, irradiation, and thermal injury) are almost always acquired, and may destroy otherwise healthy cells of the host, as well as transfused, normal red blood cells. Even trauma to the red blood cells in the circulation (such as from artificial heart valves, abnormal capillaries, sometimes marching or running on hard surfaces) has caused hemolytic anemia. Disorders intrinsic to the red blood cell are almost always hereditary, and may affect the membrane, or envelope itself (as in hereditary spherocytosis and elliptocytosis), or the glycolytic enzymes of the cytosol (as in hereditary deficiency of pyruvate kinase or glucose-6-phosphate dehydrogenase), or be the consequence of anomalies of globin synthesis. When an external factor interacts with the intrinsic defect (such as the spleen: hereditary spherocytosis; a redox drug or chemical: abnormality of glucose-6-phosphate dehydrogenase), the anemia is abated when the external factor is removed, even though the hereditary factor remains. In hemolytic anemia the iron is recovered and used again for hemoglobin synthesis, except when there is intravascular hemolysis, in which case it may be lost in the urine.

Diagnosis and treatment. Pallor, weakness, and fatigue are common to all anemias. They may not be noticed until anemia is advanced, if it is of gradual onset and there has been time for cardiovascular and biochemical adaptation. Faint jaundice in the sclerae is a feature of hemolytic anemia, whereas in anemia due to lack of vitamin B_{12}, glossitis and neuropathy occur and may be noted. The morphological study of the blood is often the most simple

guide. Macrocytic anemias are often found to be megaloblastic and due to deficiency of vitamin B_{12} or folic acid. The latter contention can be proved by measuring the quantity of those substances in the plasma. Administration of one of those vitamins will only cure individuals in whom its specific deficiency is established. Microcytic anemias are often also hypochromic. Hypochromia is a reduction in the concentration of hemoglobin within individual red blood cells. It is detected by microscopic examination of the thin blood film after Wright's stain, and by finding the ratio of hemoglobin concentration/hematocrit to be less than 0.31.

Microcytic-hypochromic anemias are most often due to iron deficiency. The administration of iron will resolve iron-deficiency anemia. Since the body normally conserves iron and has no excretory route for it, iron deficiency implies a pathological blood loss which must be explained. The administration of iron or iron-containing over-the-counter medications without investigation of the underlying cause may result in the failure to detect an asymptomatic neoplasm of the gastrointentinal tract which is causing occult blood loss until it has progressed to the incurable state. In the absence of proved deficiency, the administration of iron, vitamin B_{12}, and folic acid are not of value to the anemic individual. Further, the administration of iron, especially by injection, may be harmful. Prednisone and other adrenal corticosteroids are helpful in hemolytic anemias associated with autoantibodies.

Hereditary disorders are generally not amenable to therapy, except for those hemolytic diseases which may benefit after splenectomy. In all other cases, the treatment of the anemia is achieved by treating the underlying disease, such as hypothyroidism, rheumatoid arthritis, or leukemia. Blood transfusions are reserved for acute blood loss when symptoms of hypovolemia and shock are present, or in chronic anemia if there are signs of inadequate cardiovascular or pulmonary compensation and an underlying cause cannot be found or treated. Since blood transfusions entail a risk of immune reactions, hepatitis, acquired immune deficiency syndrome, and iron overload, they should be avoided when other alternatives for treatment are possible. *See* CLINICAL PATHOLOGY; HEMATOLOGIC DISORDERS. Arthur Haut

Bibliography. G. R. Lee et al., *Clinical Hematology*, 8th ed., 1981; W. J. Williams et al. (eds.), *Hematology*, 3d ed., 1983.

Anesthesia

Loss of sensation with or without loss of consciousness. There are several ways of producing anesthesia, with the choice dependent on the type of surgery and the medical condition and preference of the patient. Each person responds differently to a given anesthetic, and anesthetic techniques and drugs often have marked effects on bodily functions, especially those of the cardiovascular and respiratory systems. Therefore, these systems are monitored closely during anesthetic administration, with measurements such as heart sounds, blood pressure, heart action (electrocardiogram), temperature, and oxygenation taken using a variety of sophisticated devices.

General anesthesia. A state of complete insensitivity or unconsciousness is produced when anesthetic gases are inhaled; adjuvant drugs are often given intravenously. Anesthetics commonly used by inhalation are halogenated-hydrocarbon liquids (isoflurane, enflurane, and halothane), which are vaporized and administered in oxygen, usually with nitrous oxide. Adjuvant drugs include thiopental (to produce rapid loss of consciousness for induction of anesthesia), narcotic opioids (for supplemental analgesia, that is, insensitivity to pain), muscle relaxants, and sedatives. Although the mechanism of general anesthesia is unknown, the anesthetics act on the upper reticular formation of neurons in the thalamus and midbrain (neuronal structures necessary for activating the cerebral cortex and maintaining an active, attentive state).

Regional anesthesia. Analgesia, without loss of consciousness, results from injecting a solution of local anesthetic drug either into the cerebrospinal fluid surrounding the spinal cord (spinal anesthesia) or into the epidural space surrounding the cerebrospinal fluid (epidural anesthesia), usually in the lower back (lumbar) area. Epidural injection is also performed at the base of the spine (caudal anesthesia). The duration and extent of analgesia depend on the nature and amount of local anesthetic injected, among other factors. The local anesthetic acts by blocking the conduction of nerve impulses. Narcotic opioids are injected postoperatively into either the epidural space or cerebrospinal fluid for pain relief.

Local anesthesia. Analgesia can be localized to a small area, for example, the forearm, by injecting a local anesthetic solution around nerves supplying the area (the bracheal plexus in the upper arm and chest supplies the forearm). Large peripheral nerves may also be blocked individually by using this method. The duration of analgesia is determined by the choice and amount of local anesthetic.

Injection of a dilute solution of local anesthetic into the skin and superficial tissues provides very localized analgesia for minor surgery.

Acupuncture. Acupuncture is an ancient procedure, once used only in China but now practiced in the United States and elsewhere. It involves inserting needles into specific points around the body, as determined from historical charts, and manipulation of the needles; sometimes electric current is applied. Weak analgesia results through alteration of pain perception. Although sometimes helpful for chronic pain, acupuncture generally has not been found satisfactory for surgical anesthesia. *See* CENTRAL NERVOUS SYSTEM; PAIN. Fredrick K. Orkin

Bibliography. M. J. Cousins and P. O. Bridenbaugh (eds.), *Neural Blockade in Clinical Anesthesia and Management of Pain*, 3d ed., 1998; R. D. Dripps, J. E. Eckenhoff, and L. D. Vandam (eds.), *Introduction to Anesthesia: The Principles of Safe Practice*,

7th ed., 1988; R. D. Miller (ed.), *Anesthesia*, 4th ed., 1994; R. K. Stoelting, *Pharmacology and Physiology in Anesthetic Practice*, 3d ed., 1998.

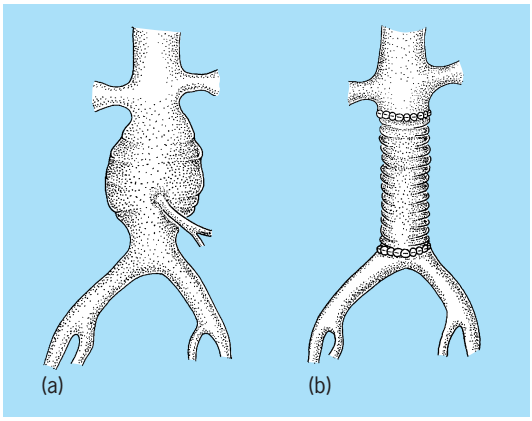
Aneurysm

A localized, abnormal arterial dilation usually caused by a weakening of the vessel wall. Aneurysms are commonly found in the abdominal aorta, intracranial arteries, and thoracic aorta; however, they may also involve the femoral, popliteal, splenic, carotid, or renal arteries. Aneurysms vary in size from less than 1 in. (2–3 cm) to more than 4 in. (10 cm). They are usually classified as true or false: true aneurysms involve all layers of the artery (inner endothelium or intima, middle muscular layer or media, and outer connective tissue or adventitia); false aneurysms do not involve all layers. With the exception of intracranial aneurysms, atherosclerotic vessel disease is generally the cause of aneurysms; other causes include syphilis, trauma, cystic medial necrosis, bacterial infections, and arteritis.

Complications. The major clinical sign of all aneurysms (except intracranial) is a large pulsatile mass; other manifestations depend on the location of the aneurysm. Aneurysms of the thoracic aorta can produce symptoms of compression or erosion as well as dissection. In dissection, the blood creates a channel between the intima and media of the aorta, which may result in blockage of major branches of the aorta, a condition that is often fatal. Debris in aneurysms may also embolize; that is, it may break free and block distal arteries. The major complication of an aneurysm is rupture, the possibility of which is directly related to its size. The mortality rate for a ruptured abdominal aortic aneurysm is 60–70%.

Intracranial aneurysms, often called berry aneurysms, are congenital weaknesses in the intracranial arteries. These aneurysms are saclike outpocketings of the vessel and vary in size from 0.2 in. (0.4 cm) to greater than 0.6 in. (1.5 cm). The major manifestations are usually related to bleeding and can vary from a severe headache, to neurologic impairment, to death. In many instances, a small initial hemorrhage (a "herald bleed") is often followed by a second bleeding episode, which is much more severe and often fatal. *See* HEMORRHAGE.

Diagnosis and treatment. The best method of diagnosis is ultrasound or a computer x-ray analysis (computed tomography scan). The recommended treatment for all aneurysms is surgery. In patients whose aneurysms are due to atherosclerosis, associated diseases can increase the risk of surgery. Abdominal and thoracic aneurysms are generally treated with resection of the aneurysm and replacement with a graft made of synthetic material (see **illus.**). Blood flow must be established to the branches of the aorta. Intracranial aneurysms are also treated surgically; however, surgery involves excluding the aneurysm from the cerebral circulation by placing a metallic clip across the neck of the aneurysm sac. A ruptured aneurysm or dissecting aneurysm is a true emergency and requires immediate surgical intervention. *See* ARTERIOSCLEROSIS; CARDIOVASCULAR SYSTEM; CIRCULATION DISORDERS; HYPERTENSION; MEDICAL IMAGING.



Surgical repair of an abdominal aortic aneurysm. (*a*) Before resection. (*b*) Tubular-type graft anastomosis. *(After M. D. Kerstein, P. V. Moulder, and W. R. Webb, eds., Aneurysms, Williams and Wilkins, 1983*)

Donald L. Akers; Morris D. Kerstein

Bibliography. C. Arbost, M. Allary, and N. Economou, Resection of an aneurysm of the abdominal aorta: Reestablishment of the continuity by a preserved human arterial graft which results after five months, *Arch. Surg.,* 64:405, 1952; M. D. Kerstein, P. V. Moulder, and W. R. Webb (eds.), *Aneurysms*, 1983; S. I. Schwartz and G. T. Shires (eds.), *Principles of Surgery*, 7th ed., 1998.

Angiogenesis

The origin and development of blood vessels. Blood vessels are composed of two basic cell types, vascular endothelial cells and peri-endothelial cells (including vascular smooth muscle cells and elongated contractile cells called pericytes, both of which support the underlying endothelial cells). The inner epithelial lining of all blood vessels, adjacent to the lumen, is a single layer of endothelial cells (**Fig. 1**). In larger blood vessels, such as arteries and veins, the inner endothelial lining, called the tunica intima, is surrounded by a medial layer, the tunica media, composed of multiple layers of vascular smooth muscle cells embedded in elastin-rich extracellular matrix. The tunica media layer is surrounded by an extracellular matrix-rich layer called the tunica adventitia. In contrast, capillary walls consist of only a single layer of endothelial cells, sometimes surrounded by pericytes.

Arteries and veins are the two fundamental types of blood vessels, carrying blood away from and toward the heart, respectively. Although both have the basic structure noted above, higher-pressure arteries generally have thicker medial smooth muscle–containing layers, whereas larger veins have thinner, more elastic walls and valves to prevent backflow. Recent studies of the process of blood vessel formation, or angiogenesis, have shown that the endothelial cells contributing to arteries and veins have

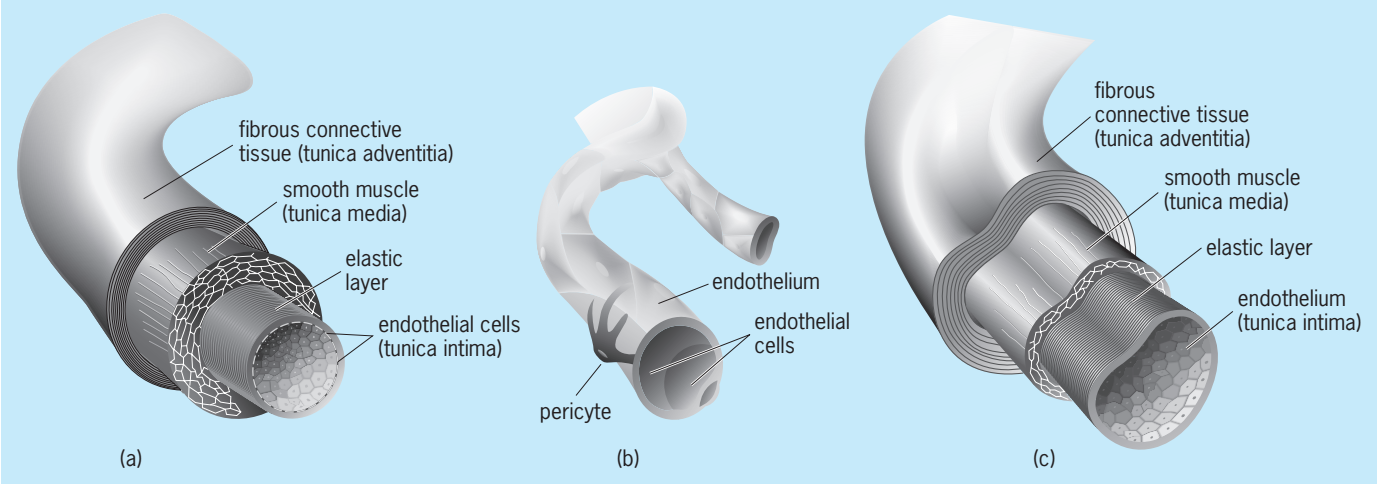


Fig. 1. Structure of blood vessels. (*a*) Arteries and (*c*) veins have the same tissue layers. (*b*) Capillaries are composed of only a single layer of endothelial cells, sometimes surrounded by elongated contractile cells known as pericytes (not to scale). (*Adapted from P. H. Raven and G. B. Johnson, Biology, 6th ed., McGraw-Hill, New York, 2002*)

distinct molecular and functional identities that are established early during development.

Blood vessel formation. Blood vessels are found only within the chordates, but within the vertebrates the anatomical architecture of the major blood vessels of the circulatory system is surprisingly well conserved, as highlighted by recent detailed characterization of the anatomy of the developing zebrafish vasculature. The first major vessels to emerge during embryonic development form by coalescence of individual mesodermal endothelial progenitor cells, or angioblasts, into cords of attached cells which then form open vascular cells, or lumenize. This process, called vasculogenesis, was thought to be restricted to early embryonic development; however, recent evidence has shown that vasculogenesis also occurs later in development and even postnatally. Most later developmental and postnatal blood vessel formation, however, occurs by sprouting and elongation of new vessels from preexisting vessels or remodeling of preexisting vessels, collectively known as angiogenesis.

Vessels generally form first as unlined endothelial tubes but generally become rapidly associated with supporting pericyte or vascular smooth muscle cells. A variety of studies have shown that the acquisition of these supporting cells is critical for proper morphogenesis, stability, and survival of nascent blood vessels. Although most larger blood vessels form during development and are relatively stable thereafter, significant growth and remodeling of blood vessels continues throughout adult life, for example during uterine cycling or in conjunction with gain or loss of fat or muscle mass.

Postnatal vessel growth also occurs in pathologic contexts, notably cancer. Tumors recruit new blood vessels to provide themselves with access to the host circulation and obtain oxygen and nutrients. Acquiring a vascular blood supply is essential for the progression and survival of a tumor. New tumor vessels form primarily through endothelial proliferation and growth from preexisting, adjacent host blood vessels. These vessels are usually morphologically abnormal, poorly structured, and deficient in vascular smooth muscle cells. Some tumor vascular spaces appear to lack even an endothelial lining and are surrounded only by tumor-derived cells.

Molecular regulators. Recently, interest in factors that might promote or inhibit blood vessel growth and the application of molecular methods to vascular biology and the study of vascular development have led to the discovery of many new genes encoding proteins involved in blood vessel growth and assembly. The vascular endothelial growth factor (VEGF) family of ligands and their tyrosine kinase vascular endothelial growth factor receptors (VEGFRs) expressed on the surface of endothelial cells play critical roles in the differentiation of blood vessels. VEGF-A plays a particularly central role—it is essential for the survival, proliferation, migration, and arterial differentiation of endothelial cells. Targeted disruption or "knockout" of just one of the two copies of the gene encoding this protein in mice causes (heterozygous) embryos to die early in development with a dramatic reduction in endothelial cell number. VEGFR2 is a key endothelial receptor transducing VEGF-A signals, and loss of the gene encoding this protein is also lethal (although only in homozygotes with both functional copies of the gene knocked out).

Other related VEGF ligands and VEGFR receptors have also been uncovered, and these also play important roles in modulating the development of blood vessels. VEGF-C signaling through the VEGFR3 receptor plays an essential role in the formation of lymphatic vessels. Another set of ligands, the angiopoietins, play a critical role in vascular endothelial–vascular smooth muscle cell interaction and blood vessel remodeling. Angiopoietin-1 binds to the endothelially expressed Tie-2 receptor to promote the acquisition of vascular smooth muscle cells and stabilization and maturation of nascent blood vessels. Loss of either angiopoietin-1 or Tie-2 function is lethal

during embryogenesis; mice with knockouts of the genes encoding these proteins die early in development with highly enlarged, poorly remodeled blood vessels deficient in vascular smooth muscle cells. Many additional genes have been identified that also play important roles in the specification, differentiation, remodeling, and maturation of blood vessels. It has become clear from the myriad of molecular studies of blood vessel formation conducted to date that regulation of vessel assembly is a highly complex and exquisitely regulated process that we have only just begun to understand.

Clinical importance. Much of the recent explosion of scientific interest in the mechanisms regulating growth and assembly of blood vessels has been driven by potential antiangiogenic (inhibiting vessel development) or proangiogenic (enhancing vessel development) therapeutic applications.

Antiangiogenic therapy. The idea that growth of tumors can be inhibited by targeting the blood vessels that supply them, rather than the tumors themselves (**Fig. 2**), grew out of observations that tumors acquire blood vessels, and that their ability to promote local angiogenesis correlates with progression from relative dormancy to rapid growth and aggressiveness (the term "angiogenic switch" has been coined to describe this transition). A variety of tumor-secreted factors mediate the angiogenic switch, including the factors described above that regulate normal vessel growth and development, notably the critical vascular regulator VEGF. A worldwide search is now on for antiangiogenic compounds that can halt or reverse the growth of tumor blood vessels and thereby stop tumor growth or even cause tumor regression. One focus of attention has been on endogenously produced antiangiogenic factors, many of them normal products of the in-vivo processing of extracellular matrix or plasma proteins. More recently, some of the prominent molecular players described above (particularly VEGF) have been directly targeted for antiangiogenic cancer therapy using various means, including chemical inhibitors, interfering fragments of receptors or ligands, or interfering antibodies. Although human trials of antiangiogenic therapies are mostly in early stages, animal studies and preliminary human trials have yielded encouraging findings, including a lack of general toxicity (unlike chemotherapeutic agents traditionally used for attacking tumor cells directly, for example) and efficacy against many tumors in animal models. Fortuitously, it appears that tumor vessels are differentially sensitive to antiangiogenic therapies compared with normal host blood vessels, probably because the morphologically abnormal tumor vessels deficient in pericyte/smooth muscle cells are less stable than normal host vessels.

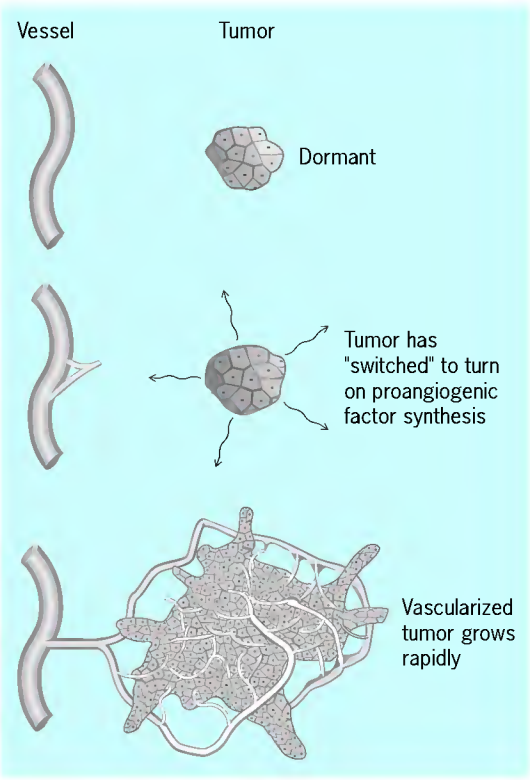


Fig. 2. Tumor angiogenesis and antiangiogenic therapy.

Proangiogenic therapy. Treatment of limb or cardiovascular ischemia (insufficient blood supply) using proangiogenic therapy is also an exciting new possibility, promoting regrowth of new vessels into tissues that because of disease and/or injury have become deficient in blood supply. To date, attempts at proangiogenic gene therapy using the genes encoding VEGF and FGFs (fibroblast growth factors) have been mixed.

Animal developmental models. One difficulty in developing effective therapies for either antiangiogenic treatment of cancer or proangiogenic treatment of ischemia is that although we have now identified many important molecular regulators of blood vessel growth and maintenance, we still have a relatively rudimentary understanding of how all of these regulators function together in the complex orchestration of blood vessel morphogenesis. A great deal of the insight into the functional roles of the key vascular signaling molecules as well as the receptors and intracellular pathways that transduce these signals has come from studies of blood vessel formation during development in animal model organisms. Targeted disruption and overexpression experiments in mice have provided definitive assays for the in-vivo functions of vascular genes during development. In most cases these genes have later been shown to play analogous functional roles in pathologic or nonpathologic angiogenesis in adults (neoangiogenesis). Studies in developing avians, *Xenopus* (frogs), and zebrafish have also begun to provide important new insights into vessel specification, differentiation, and assembly during early development.

Zebrafish. The zebrafish shows particular promise as a new model organism for understanding how blood vessels are assembled in vivo. Zebrafish are tropical fish native to Southeast Asia that have a number of advantageous attributes for developmental studies, particularly studies of vascular development.

Zebrafish adults are small; fully grown animals are only about an inch long, so large numbers of fish can be housed in a small space. Zebrafish have a relatively short generation time, and adult females lay hundreds of eggs every few weeks, so large numbers of progeny can be obtained for genetic or experimental studies. The eggs are fertilized and develop externally to the mother; thus zebrafish embryos and larvae are readily accessible for noninvasive observation or experimental manipulation at all stages of their development. Early development is very rapid: Blood circulation begins within 24 hours after fertilization, larvae hatch by approximately 2.5 days, and larvae are swimming and feeding by 5–6 days.

Two additional features of zebrafish make them particularly useful for studying vascular development. First, developing zebrafish are very small—a 2-day postfertilization (dpf) embryo is just 2 mm long. The embryos are so small, in fact, that the cells and tissues of the zebrafish receive enough oxygen by passive diffusion to survive and develop in a reasonably normal fashion for the first 3 to 4 days, even in the complete absence of blood circulation. This makes it fairly straightforward to assess the cardiovascular specificity of genetic or experimental defects that affect the circulation. Second, zebrafish embryos and early larvae are virtually transparent. The embryos of zebrafish (and many other teleosts) are telolecithal; that is, yolk is contained in a single large cell separate from the embryo proper. The absence of obscuring yolk proteins gives embryos and larvae a high degree of optical clarity. Genetic variants deficient in pigment cells or pigment formation are even more transparent. This remarkable transparency is probably the most valuable feature of the fish for studying blood vessels in vivo, and has been exploited in a number of different ways.

Confocal microangiography, a technique that permits high-resolution three-dimensional visualization of fluorescent structures, has been used to study functioning blood vessels throughout the zebrafish at every stage of embryonic or early larval development. Transgenic zebrafish expressing green fluorescent proteins in blood vessels allow visualization of vascular endothelial cells and their angioblast progenitors prior to initiation of circulation through a vessel, and even before vessel assembly. A complete atlas of the anatomy of the developing vasculature throughout early development has been prepared using confocal microangiography, the first such atlas to be compiled for any vertebrate. Together these and other tools make it possible to perform in-vivo genetic and experimental dissection of the mechanisms of blood vessel formation in the zebrafish. This facility has already been used to dissect the nature of molecular signals guiding arterial-venous differentiation of endothelial cells during development, and promises to make the fish an important resource for understanding the cues and signals that guide the patterning of growing blood vessels. *See* ARTERY; BLOOD VESSELS; CANCER (MEDICINE); CAPILLARY (ANATOMY); CIRCULATORY SYSTEM; DEVELOPMENTAL BIOLOGY; TUMOR; VEIN. Brant M. Weinstein

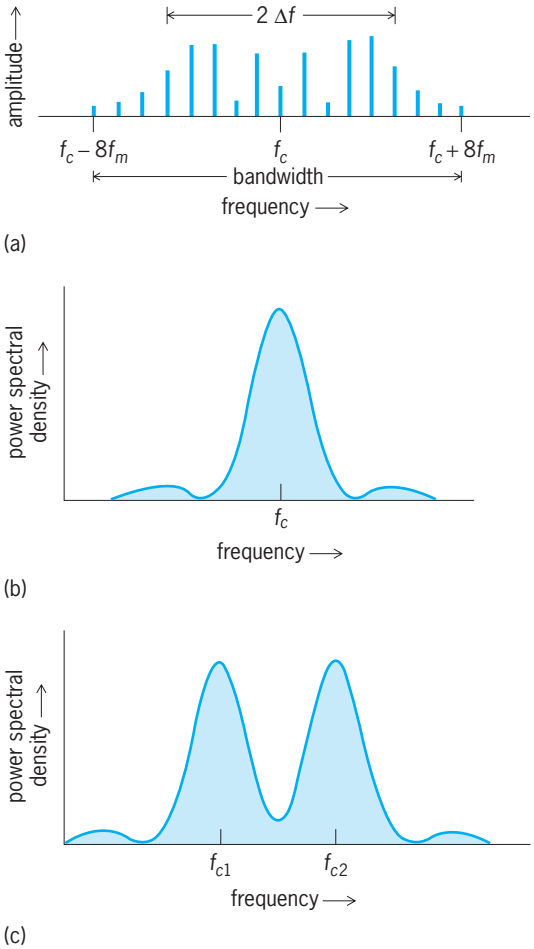
Bibliography. E. M. Conway, D. Collen, and P. Carmeliet, Molecular mechanisms of blood vessel growth, *Cardiovasc. Res.*, 49:507–521, 2001; N. Ferrara, VEGF and the quest for tumour angiogenesis factors, *Nat. Rev. Cancer*, 2:795–803, 2002; J. M. Isner, Myocardial gene therapy, *Nature*, 415:234–239, 2002; S. Isogai, M. Horiguchi, and B. M. Weinstein, The vascular anatomy of the developing zebrafish: An atlas of embryonic and early larval development, *Dev. Biol.*, 230:278–301, 2001; R. Kerbel and J. Folkman, Clinical translation of angiogenesis inhibitors, *Nat. Rev. Cancer*, 2:727–739, 2002; N. D. Lawson and B. M. Weinstein, Arteries and veins: Making a difference with zebrafish, *Nat. Rev. Genet.*, 3:674–682, 2002; B. Weinstein, Vascular cell biology in vivo: A new piscine paradigm?, *Trends Cell Biol.*, 12:439–445, 2002.

Angle modulation

Modulation in which the instantaneous angle varies in proportion to the modulating waveform. When used in an analog manner, it is referred to as either frequency modulation (FM) or phase modulation (PM), depending upon whether the instantaneous frequency or instantaneous phase varies with the modulation. When used in a digital modulation format, it is referred to as either frequency-shift keying (FSK) or phase-shift keying (PSK). Both FSK and PSK can be used with either a binary or an M-ary alphabet, where M is an integer greater than 2. Indeed, the most common forms of PSK correspond to an input alphabet size of either two or four; while for FSK alphabets of two, four, or eight, symbols are typically employed.

The spectrum of an angle-modulated signal for an arbitrary modulating waveform (that is, an arbitrary message) is difficult to determine. However, if that modulating waveform is a sine wave of frequency f_m, modulating a carrier of frequency f_c, then it is straightforward to show that the modulated signal has a discrete frequency spectrum with components at f_c, $f_c \pm f_m$, $f_c \pm 2f_m$, $f_c \pm 3f_m$, ... (**illus.** *a*). The amplitude of the nth harmonic of the spectrum, at frequency $f_c + nf_m$, is given by the magnitude of $J_n(\beta)$, a Bessel function of order n and argument β. This quantity β is known as the modulation index. For phase modulation, β is the maximum deviation of the signal phase from the carrier phase, in radians; while for frequency modulation, it is the ratio $\Delta f/f_m$, where Δf is the maximum deviation of the signal frequency from the carrier frequency. *See* MODULATION.

For a PSK waveform modulated by a message that is modeled as a sequence of random binary data, the power spectral density is a continuous function of frequency, f, which has a peak centered at the carrier frequency, f_c (illus. *b*). It has the form $A \sin^2 x/x^2$, where A is a constant and x is proportional to $f-f_c$. If, instead of phase-shift-modulating a carrier, the message frequency-shift modulates the signal between the carrier frequencies f_{c1} and f_{c2}, the power spectral



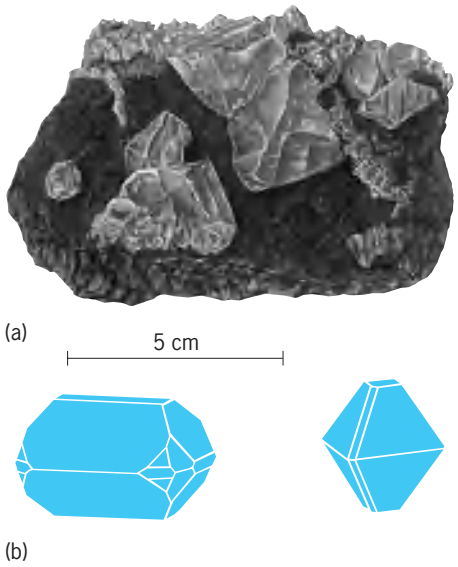
Spectral diagrams of angle-modulated signals. (*a*) Amplitude-frequency spectrum of a signal that is frequency-modulated or phase-modulated by a sine wave, with modulation index $\beta = 5$ (*after M. Schwartz, Information Transmission, Modulation, and Noise, 4th ed., McGraw-Hill, 1990*). (*b*) Power spectral densities of PSK signal and (*c*) FSK signal (*after B. P. Lathi, Modern Digital Communications Systems, 3d ed., Oxford University Press, 1998*).

density has peaks at both these frequencies (illus. *c*). *See* FREQUENCY MODULATION; MODULATION; PHASE MODULATION. Laurence B. Milstein

Bibliography. M. Schwartz, *Information Transmission, Modulation and Noise*, 4th ed., 1990; H. Taub and D. L. Schilling, *Principles of Communication Systems*, 2d ed., 1986; R. E. Ziemer and W. H. Tranter, *Principles of Communications*, 4d ed., 1994.

Anglesite

A mineral with the chemical composition $PbSO_4$. Anglesite occurs in white or gray, orthorhombic, tabular or prismatic crystals or compact masses (see **illus.**). It is a common secondary mineral, usually formed by the oxidation of galena. Fracture is conchoidal and luster is adamantine. Hardness is 2.5–3 on Mohs scale and specific gravity is 6.38. Anglesite fuses readily in a candle flame. It is soluble with difficulty in nitric acid. The mineral does not occur in large enough quantity to be mined as an ore of lead, and is therefore of no particular commercial value. Fine exceptional crystals of anglesite have been found throughout the world. In the United States good crystals of anglesite have been found at the Wheatley Mine, Phoenixville, Chester County, Pennsylvania, and in the Coeur d'Alene district of Shoshone County, Idaho. Edward C. T. Chao



Anglesite. (*a*) Crystals on galena from Phoenixville, Pennsylvania. (*Bryn Mawr College specimen*). (*b*) Crystal habits (*after L. G. Berry and B. Mason, Mineralogy, W. H. Freeman, 1959*)

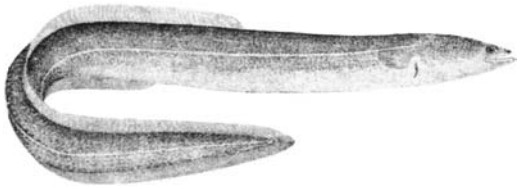
Anguilliformes

The true eels, a large order of actinopterygian fishes, also known as the Apodes. The Anguilliformes are related to Saccopharyngiformes (sack pharynx fishes), Elopiformes (tarpons), and Notacanthiformes (spiny eels and halosaurs); all have a leptocephalous larval stage in development. The chief characteristics of the Anguilliformes include an elongate body with numerous vertebrae; a pectoral girdle which, when present, is free from the head; absence of a pelvic girdle and pelvic fins in Recent adults; dorsal and anal fins confluent with the caudal fin (caudal fin absent in some ophichthids); loss of skeletal parts, especially about the head (for example, the orbitosphenoid, posttemporal, symplectic, posttemporal bones, and gular plate are absent); mesocoracoid and postcleithra absent; swim bladder present, usually physostomous (that is, with a connection to the esophagus); restricted gill openings; no fin spines; scales usually absent or, if present, cycloid and embedded in the skin; and 6 to 49 branchiostegal rays. *See* ACTINOPTERYGII; EEL; ELOPIFORMES; NOTACANTHOIDEI; SACCOPHARYNGIFORMES; SWIM BLADDER.

The Anguilliformes comprise three suborders, 15 families, 141 genera, and about 791 species. The largest family of each suborder is treated below.

Suborder Anguilloidei. Frontal bones are divided (sutured), and scales present or absent. There are three families, five genera, and about 29 species.

The largest family is the Anguillidae (freshwater eels). Unlike most eels of the world, anguillids have small scales embedded in the skin; however, scales do not develop until an individual is 2 or 3 years old. They are further distinguished by having pectoral fins, and a caudal fin continuous with the dorsal and anal fins; branchiostegals do not overlap; the anterior nostril is tubular, near the tip of the snout; the posterior nostril is round and in front of the eye. One genus and about 16 species occur worldwide, except in mainland Asia, the eastern Pacific, and most of South America. Adult anguillids live in freshwater streams or in brackish-water estuaries and, upon reaching maturity, migrate to sea to spawn.



American eel (*Anguilla rostrata*). (*After G. B. Goode, Fishery Industries of the U.S., 1884*)

In the case of the common American eel (*Anguilla rostrata*) [see **illustration**], the life cycle encompasses several morphological and physiological stages: egg > leptocephalus > glass eel > elver > yellow eel > silver eel > death after spawning. The act of spawning has never been observed. The leptocephalus is a planktonic larva; the glass eel is the first stage of the metamorphosed leptocephalus and lacks pigment; the elver is completely metamorphosed and pigmented; the yellow eel is the nonmigratory juvenile; and the mature silver eel is ready for seaward migration. The leptocephali drift toward the coast of the Americas, a journey requiring about one year, at which time they attain 60–80 mm (2.4–3.2 in.) in length and begin to metamorphose. During metamorphosis, larvae shrink, leaving the glass eels shorter than their precursors. Female eels ascend Atlantic and Gulf coastal rivers, often far into the heartland of North America. There they remain 5–40 years (depending on latitude) before returning to sea. In the meantime, males remain on or near the coast. Males are generally smaller than females and develop larger eyes just before migration.

It remains a mystery as to the exact place of spawning. Adult American eels have never been captured beyond the continental shelf, and neither spawning adults nor their eggs have been observed. The only empirical evidence of a possible spawning site is in 2000 m (6560 ft) of water of a geologically old, deep basin of the Bahamas where two adults were photographed.

Suborder Muraenoidei. Frontals are divided (sutured), and scales absent. There are three families, 24 genera, and about 207 species.

The largest family is the Muraenidae (moray eels). Members of the family, consisting of 15 genera and about 185 species, lack pectoral fins; the gill opening is restricted to a small oval aperture; the teeth are long and fanglike; posterior nostrils are above and in front of the eyes; and some species of *Gymnothorax* are ciguatoxic (ciguatoxin is a poisonous substance accumulated up the food chain in the flesh and viscera of some fish). Maximum length is about 3 m (9.8 ft). Moray eels occur in the tropical and temperate seas of the world.

Suborder Congroidei. Frontals are fused scales rarely present. There are nine families, 112 genera, and about 555 species.

The largest family is the Ophichthidae (snake eels and worm eels). Ophichthidae is a family of marine eels, none of which regularly occurs in freshwater. The family differs from other eels in the unusual form of the branchiostegal rays, which are long, slender, and numerous, do not attach basally to the hyoid bone, and broadly overlap those of the other side, leaving the branchiostegal region inflated or slightly bulbous. It is the most diverse and speciose family of eels, consisting of about 250 known species in 55 genera. Ophichthids occur in coastal marine habitats of tropical and warm temperate waters of the world.

Herbert Boschung

Bibliography. E. B. Böhlke (ed.), *Fishes of the Western North Atlantic*, Part 9, vol. 1: *Orders Anguilliformes and Saccopharyngiformes* (pp. 1–655), vol. 2: *Leptocephali* (pp. 657–1055), Sears Foundation for Marine Research Memoir (Yale University), New Haven, 1989; J. E. McCosker and R. H. Rosenblatt, A revision of the Eastern Pacific snake-eel genus *Ophichthus* (Anguilliformes: Ophichthidae) with the description of six new species, *Proc. Calif. Acad. Sci.*, 50(19):397–432, 1998; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Angular correlations

An experimental technique that involves measuring the manner in which the likelihood of occurrence (or intensity or cross section) of a particular decay or collision process depends on the directions of two or more radiations associated with the process. Traditionally, these radiations are emissions from the decay or collision process. However, a variant on this technique in which the angular correlations are between an incident and emitted beam of radiation has been widely used; this variant is known as angular distributions.

The fundamental reason for performing such measurements, rather than just scrutinizing a single radiation in a particular direction or measuring the total intensity for a process, is that the angular correlation or angular distribution measurement provides much more information on both the decay or collision process and on the structure and properties of the emitter of the radiation. The technique is used to study a variety of decay and collision processes in atomic and molecular physics, condensed-matter

(solid-state) and surface physics, and nuclear and particle physics.

Detection of multiple emissions. The principal use of this technique in nuclear physics has been to determine the angular momentum, or spin, and parity of excited nuclear states which are radioactive, that is, decay spontaneously, by measuring in coincidence the radiation in specific directions from two successive transitions in the radioactive cascade. The measurements are generally of coincidences between gamma rays, but coincidences between gamma rays and electrons (beta particles) are also used. The form of the angular correlation, the measured intensity as a function of the angle between the two radiations, gives the information about the intermediate excited state in the cascade. *See* GAMMA RAYS; NUCLEAR SPECTRA; NUCLEAR STRUCTURE; RADIOACTIVITY.

In atomic and molecular collisions as well as in nuclear and particle collisions, this technique is employed as a means of completely specifying the dynamics of the collision, with the added proviso that the energies of the emitted radiations are also to be measured. Wide use has been made of angular correlations in the impact ionization of atoms by electrons where the directions of both the scattered electron and the ejected electron are measured. This application is known as $(e,2e)$ spectroscopy. These studies, particularly just above the ionization threshold energy, have revealed the details of the law governing threshold ionization by charged particles, known as the Wannier law. Furthermore, these measurements reveal considerable information about the quantum-mechanical wave function of the system after the ionization process has taken place, that is, the final state. This type of final state, two electrons and one positive ion, is an example of a three-body continuum Coulomb state. *See* ATOMIC STRUCTURE AND SPECTRA; SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Angular distributions. This technique is much simpler experimentally than is the general angular correlations technique, since coincidence measurements are not required. Consequently, it can be employed in situations where the probability (or cross section) for the process is so small that a coincidence experiment would not lead to enough counts to make the result meaningful. This technique allows the extraction of the multipolarity of the incident radiation, provided the target orientation is known. The widest application of this technique has been in atomic, molecular, and condensed-matter investigations, and both emitted electrons and electromagnetic radiation (x-rays, ultraviolet, and visible) have been studied. The most important incident particles are electromagnetic radiation, generated principally by synchrotron radiation. Other incident particles include electrons, protons, and heavy ions and atoms. *See* MULTIPOLE RADIATION; SYNCHROTRON RADIATION.

Photoelectron spectroscopy. In photoionization, the study of the angular distribution of photoelectrons emitted from a target as a result of ionization by electromagnetic radiation (photons), the technique is used in several different ways. The atomic and molecular targets are generally randomly oriented. If unpolarized photons are used, the photoelectron direction with respect to the photon beam direction is measured; for linearly polarized photons, the photoelectron direction with respect to the polarization is measured. This technique is used primarily at photon energies low enough (below about 500 electronvolts, corresponding to photon wavelengths longer than about 2.5 nanometers) that they constitute electric dipole radiation, to an excellent approximation. The general form of the angular distribution is determined by a parameter β, known as the asymmetry parameter, that contains dynamical information about both the target and the photoionization process. The dependence of β on photon energy is extremely helpful in understanding the details of the photoionization cross section, particularly in energy regions where resonances or minima occur. In addition, since β depends on the quantum-mechanical phases of each of the possible final states, phase information, which is unavailable from total cross section measurements, can be obtained. *See* ELECTRON SPECTROSCOPY; PHOTOIONIZATION; POLARIZED LIGHT.

For oriented targets, that is, target atoms or molecules whose magnetic moments, spins, or angular momenta are all aligned, the general form of the angular distribution is more complex and depends upon both the degree of alignment and the state of polarization of the incident radiation. This is often the case for studies of surfaces and crystalline solids where the crystal structure aligns the constituent atoms and molecules; then the technique is generally known as angle-resolved photoelectron spectroscopy. In surface studies, this technique can often aid in the understanding of the details of how the atoms and molecules are attached to the surface along with the geometry of their orientation with respect to the surface. *See* ANGULAR MOMENTUM; SPIN (QUANTUM MECHANICS); SURFACE PHYSICS.

Decay of excited states. When an excited state decays, energy is given off in the form of some kind of radiation. Measuring the angular distribution of the radiation provides certain information about the excited state. In particular, if the emission is isotropic, that is, independent of direction, the excited state was also isotropic. An angular distribution that does depend upon direction (anisotropic) implies that the excited state too had a preferred direction. Thus, the technique provides information about the process that created that excited state. The technique is used in nuclear physics with observation of emitted gamma rays, particularly in connection with radioactive decay. It is also used extensively in atomic and molecular physics. When electromagnetic radiation emitted from atoms or molecules is observed, the technique is called fluorescence spectroscopy, and with emitted electrons it is known as Auger spectroscopy. *See* AUGER EFFECT; X-RAY FLUORESCENCE ANALYSIS.

Steven T. Manson

Bibliography. J. Boyle and M. Pindola (eds.), *Many-Body Interactions in Atomic Physics,* 1995; P. C. Deshmukh, V. Radojevič, and S. T. Manson, Photoionization of the outer shells of radon and radium, *Phys. Rev. A,* 45:6339–6348, 1992.

Angular momentum

In classical physics, the moment of linear momentum about an axis. A point particle with mass m and velocity $\mathbf{v}$ has linear momentum $\mathbf{p} = m\mathbf{v}$. Let $\mathbf{r}$ be an instantaneous position vector that locates the particle from an origin on a specified axis. The angular momentum $\mathbf{L}$ can be written as the vector cross-product in Eq. (1) The **illustration** shows the geometrical meaning of this equation.

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} \tag{1}$$

See CALCULUS OF VECTORS; MOMENTUM.

The time rate of change of the angular momentum is equal to the torque $\mathbf{N}$. A rigid body satisfies two independent equations of motion (the dynamical equations) given by Eqs. (2) and (3), where d/dt

$$\frac{d}{dt}\mathbf{p} = \mathbf{F} \tag{2}$$

$$\frac{d}{dt}\mathbf{L} = \mathbf{N} \tag{3}$$

denotes the rate of change, the derivative with respect to time t. Only Eq. (1) is required for a point particle. Equation (2) indicates that a rigid body acts as a point particle located at its center of mass. The motion of the center of mass depends upon the net force $\mathbf{F}$, which is the vector sum of all applied forces. Equation (3) gives the angular motion about the center of mass. The case of statics occurs when the net force and net torque both vanish. *See* KINETICS (CLASSICAL MECHANICS); STATICS; TORQUE.

Vectors and pseudovectors. Given the definition of angular momentum in Eq. (1), although it has three components, it is not a vector. The parity transformation π transforms vectors such as $\mathbf{r}$ and $\mathbf{p}$ according to Eqs. (4). Thus, $\pi \cdot \mathbf{L} \rightarrow +\mathbf{L}$, so that

$$\mathbf{r}' = \pi \cdot \mathbf{r} = -\mathbf{r} \tag{4a}$$

$$\mathbf{p}' = \pi \cdot \mathbf{p} = -\mathbf{p} \tag{4b}$$

$\mathbf{L}$ is called a pseudovector or axial vector. Mathematically, these quantities are called tensors of odd relative weight. Examples of pseudovectors include angular momentum, torque, and magnetic fields. Two or more pseudovectors are combined by adding their components, so that the total angular momentum is the sum of the angular momenta of the individual particles.

Among the fundamental particles, the photon field is a massless vector field and the rho meson is a pseudovector field. *See* ELEMENTARY PARTICLE; PARITY (QUANTUM MECHANICS).

Rigid-body motion. To study the nonrelativistic mechanics of a rigid body made up of N particles, it is convenient to write the instantaneous velocity $\mathbf{v}_\alpha$ of the α-th particle in terms of angular velocity ω as $\mathbf{v}_\alpha = \omega \times \mathbf{r}_\alpha$. For this system, Eq. (1) becomes Eq. (5),

$$\mathbf{L} = \sum_{\alpha=1}^{N} m_\alpha \mathbf{r}_\alpha \times (\omega \times \mathbf{r}_\alpha) = \mathbf{I} \cdot \omega \tag{5}$$

where $\mathbf{I}$ is the moment of inertia tensor. The moment of inertia tensor can be written in matrix form as Eq. (6). Representative components of this tensor are given by Eqs. (7). Since $\mathbf{I}$ is symmetric, it can always be diagonalized and expressed in principal axis (or symmetry axis) form, $\mathbf{I}_{PA} = \text{diag}\,(I_1, I_2, I_3)$, and in this form it is constant in time: $\dot{\mathbf{I}}_{PA} = 0$. The Euler angle coordinates (α, β, γ) are useful because they can be used to rotate a rigid body into its principal axis frame. *See* EULER ANGLES; MOMENT OF INERTIA; ROTATIONAL MOTION.

$$I_{ij} = \sum_{\alpha=1}^{N} m_\alpha [r_\alpha^2 \delta_{ij} - (r_\alpha)_i (r_\alpha)_j] \tag{6}$$

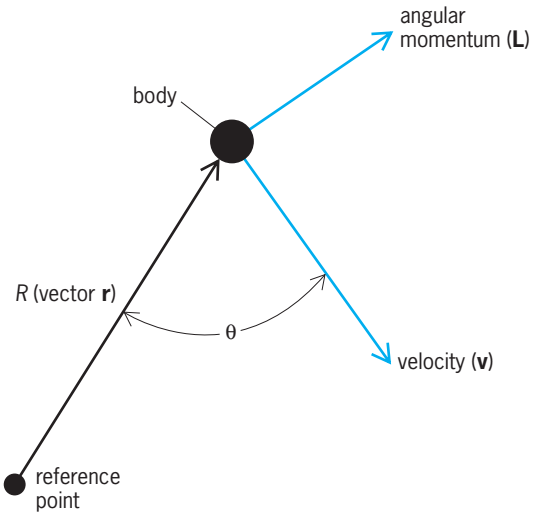
$$I_{11} = \sum_{\alpha=1}^{N} m_\alpha (y_\alpha^2 + z_\alpha^2) \tag{7a}$$

$$I_{12} = -\sum_{\alpha}^{N} m_\alpha x_\alpha y_\alpha \tag{7b}$$

When all of the forces that act on a system are conservative (that is, $\nabla \times \mathbf{F} \equiv \mathbf{O}$), a potential energy function $V(\mathbf{r})$ exists for which $\mathbf{F} = -\nabla V(\mathbf{r})$. Consider a rigid body of mass M and inertia tensor $\mathbf{I}$ subject to two conservative forces $\mathbf{F}_i = -\nabla_R V_1(\mathbf{R})$ and $\mathbf{F}_2 = -\nabla_{\alpha 0} V_2(\alpha_0)$, where $\mathbf{R}$ is the position vector of the center of mass, and $\alpha_0 = (\alpha, \beta, \gamma)$ are moment-arm-weighted Euler angles. The quantities ∇_R and $\nabla_{\alpha 0}$ are gradient operators. The equations of motion (2) and (3) become Eqs. (8) and (9). Using $\mathbf{I}_{PA}$ and writing out Eq. (9) in component form in the rotating principal-axis frame yields Eqs. (10).

$$M\mathbf{R} = \mathbf{F}_1 \tag{8}$$

$$\mathbf{I} \cdot \dot{\omega} = \mathbf{N}_2 = \mathbf{r} \times \mathbf{F}_2 \tag{9}$$



Geometrical definition of angular momentum. If we choose a reference point and draw a line *R* from it to a moving body, then the line *R* and the velocity v of the body of mass *m* define a plane. The angular momentum is a vector that lies perpendicular to this plane. If the momentum of the body is p = *m*v and the line *R* is a vector r, then the magnitude of the angular momentum L is *pr* sin θ.

$$I_1\dot{\omega}_x - (I_2 - I_3)\omega_y\omega_z = N_x \qquad (10a)$$

$$I_2\dot{\omega}_y - (I_3 - I_1)\omega_x\omega_z = N_y \qquad (10b)$$

$$I_3\dot{\omega}_z + (I_1 - I_2)\omega_x\omega_y = N_z \qquad (10c)$$

See ENERGY; RIGID-BODY DYNAMICS.

Conservation. A symmetry is a transformation that leaves a physical system unchanged. A physical quantity is called invariant under a transformation if it remains the same after being transformed. For example, the solutions to Eqs. (2) and (3), or equivalently Eqs. (8) and (9), are invariant under change of the coordinate origin or orientation of the **i**, **j**, and **k** axes. The freedom to choose any orientation of coordinate axis is called rotational invariance, because one choice of axes can be rotated into another.

In physics, the rotational invariance follows from the isotropy and homogeneity of space that has been experimentally established to high accuracy. The diagonalization transformation $\mathbf{I} \rightarrow \mathbf{I}_{PA}$ is carried out by rotating the axes, and thus is a consequence of rotational invariance. The principal axes of a rigid body are called the symmetry axes because their mechanical description is simplest there. (Rotational invariance says that the mechanics can be solved in any orientation of the axes but not that an arbitrary orientation is simplest.)

The study of symmetry shows that one of the deepest relations in physics is that between dynamics and conservation. A physical quantity B is conserved if Eq. (11) is satisfied; that is, B is constant in time

$$\frac{dB}{dt} = 0 \qquad (11)$$

although it may vary in space. Noether's theorem states that if a physical system is invariant under a continuous symmetry, a conservation law exists, provided that the observable in question decreases rapidly enough at infinity. Thus, when the force is zero everywhere (the system is invariant under translation in space), the linear momentum is conserved. If the torque is zero everywhere (the system is invariant under rotation), the angular momentum is conserved. If the system is invariant under translations in time, the total energy is conserved. The converse to Noether's theorem is false, as conserved quantities do not imply continuous symmetries. For example, the parity transformation π of Eqs. (4) can be conserved but is not continuous.

The set of all three-dimensional rotations form a group called SO(3): S for special (det = +1), O for orthogonal (length-preserving), and 3 for the space dimension. The set of 2×2 complex, unitary (length-preserving) special transformations is called SU(2). There is a 2:1 homorphism of SU(2) $\rightarrow$ SO(3), and SU(2) is called the covering group of SO(3).

Angular momentum is important in the evolution of celestial objects. The shapes of these bodies and collections of these bodies, spiral galaxies for instance, follow from the space-time variation of their torques. There is a puzzling problem along these lines: If the universe can rotate, it is not clear why it rotates so slowly; that is, why the anisotropics are so small. A proposed solution is based on the hypothesis of an inflationary epoch in the very early universe, about 10^{-35} s after the big bang. *See* BIG BANG THEORY; CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM; COSMOLOGY; SYMMETRY LAWS (PHYSICS); UNIVERSE.

Quantum angular momentum. Quantum mechanics has a richer and more complicated structure than classical physics. Because of this, the relationship between symmetry and conservation is even more useful. Whereas in classical physics the observables and states coincide, in quantum mechanics the states of the system correspond to probability amplitudes ψ and the observables **A** to self-adjoint operators (hermitian operators with suitable boundary conditions). According to the Heisenberg uncertainty principle, all observables cannot be simultaneously known, rather only those that commute. Two commuting self-adjoint operators can be diagonalized by a single unitary transformation. The complete set of commuting observables specifies the maximum amount of sharp information possible about a quantal system. The time dependence of a state, represented in **x**-space, where **x** is the position coordinate, as $\psi(\mathbf{x}, t)$, is given by the Schrödinger equation (12), where $\hat{H}$ is the system hamiltonian

$$\hat{H}\psi(\mathbf{x}, t) = i\hbar\left(\frac{\partial\psi}{\partial t}\right)(\mathbf{x}, t) \qquad (12)$$

(energy) operator and $\hbar = 1.05 \times 10^{-34}$ joule-second and is Planck's constant divided by 2π. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Rotational invariance under SO(3) transformations in quantum mechanics implies the angular momentum commutation relations for a single particle given in vector form by Eq. (13). The z component of this equation can be written as Eq. (14), and the x

$$\mathbf{J} \times \mathbf{J} = i\hbar\,\mathbf{J} \qquad (13)$$

$$[J_x, J_y] = i\hbar\,J_z \qquad (14)$$

and y components are expressed in similar fashion, where $[A, B]$ is the commutator of A and B, defined by Eq. (15). The quantum-mechanical momentum

$$[A, B] = AB - BA \qquad (15)$$

J is the most general observable satisfying Eq. (13). The orbital angular momentum $\mathbf{L} = i\hbar\ \mathbf{r} \times \nabla$, where ∇ is the gradient operator, is a special case of a quantum-angular momentum. The total angular momentum $\mathbf{J} = J_x\mathbf{i} + J_y\mathbf{j} + J_z\mathbf{k}$ has squared length $J^2 = {J_x}^2 + {J_y}^2 + {J_z}^2$. It is straightforward to use the operator identity $[A^2,B] = A[A,B] + [A,B]A$ together with Eq. (13) to show that Eq. (16) is satisfied for

$$[J^2, J_i] = 0 \qquad (16)$$

$i = x, y, z$. An inspection of Eqs. (13) and (16) shows that the angular momentum complete set of commuting observables is $\{J^2, J_i\}$, where any single component J_i can be simultaneously diagonal with J^2. Conventionally J_z is chosen as the diagonal component, and J^2 is called the Casimir operator of SO(3).

The two nonself-adjoint ladder operators $J_\pm = J_x \pm J_y$ are useful for determining the properties of angular momentum states.

Let $\psi(J, M, r)$ be a simultaneous eigenfunction of the operators J^2 and J_z with eigenvalues λ and μ, so that Eqs. (17) and (18) are satisfied. The variable r in

$$J^2\psi(J, M, r) = \lambda\psi(J, M, r) \quad (17)$$

$$J_z\psi(J, M, r) = \mu\psi(J, M, r) \quad (18)$$

ψ stands for all other observables which commute with J^2 and J_z. Repeated use of the angular momentum commutation relations of the operators J_+ and J^2 proves that $\lambda = J(J + 1)\hbar^2$ and that $\mu = M\hbar$, where $M = -J, -J + 1, \ldots, J - 1, J$ for each value of J. Further, $J = n/2$, where n is a positive integer, so that $0, {}^1/_2, {}^3/_2, 1, {}^5/_2, \ldots$ are the allowed values of total angular momentum. The eigenfunctions for half-odd-integer values of total angular momentum are called spinors. A value S of spin, together with S_z, labels intrinsic spin states. Spinors are new features of rotational invariance. They have the surprising property of changing sign under an SO(3) rotation through 2π radians, $\mathbf{R}(2\pi)$; that is, Eq. (19) is satisfied. The states

$$\mathbf{R}(2\pi)\psi(J, M, r) = -\psi(J, M, r) \quad (19)$$

$\psi(J, M, r)$ can be spinor-valued since measured probability densities depend only upon $|\psi(J, M, r)|^2 dV$ (where dV is a volume element), whereas the observables must be tensors to have sensible classical limits. The doublet Zeeman splitting of an unpaired electron in an external magnetic field is an indication that the electron has this strange internal structure. Thus, $\mathbf{J} = \mathbf{L} + \mathbf{S}$ is the total angular momentum. *See* SPIN (QUANTUM MECHANICS).

Quantum addition of angular momenta. The discussion above gives the properties of a single angular momentum. Two angular momenta, $\mathbf{J}_1$ and $\mathbf{J}_2$, are called independent if all components of one commute with all components of the other. From the preceding discussion, the sets of eigenfunctions $\psi(J_1, M_1, r)$ and $\psi(J_2, M_2, r)$ can be determined. Let $\mathbf{J} = \mathbf{J}_1 + \mathbf{J}_2$. Then the eigenfunction $\psi(J, M, r)$ can be expressed as a linear combination of the $\psi(J_1, M_1, R)$ and $\psi(J_2, M_2, r)$ by Eq. (20), where

$$\psi(J, M, r) = \sum \mathrm{C}(J_1 J_2 J, M_1 M_2)\psi(J_1, M_1, r) \cdot \psi(J_2, M_2, r) \quad (20)$$

$\mathrm{C}(J_1 J_2 J, M_1 M_2)$ is the Wigner coefficient (often erroneously called the Clebsch-Gordon coefficient). The summation is over values of M_1 and M_2 such that $M = M_1 + M_2$, and J must have one of the set of values $J = J_1 + J_2, J_1 + J_2 - 1$.

The Wigner coefficients play another important role in the Wigner-Eckart theorem. Let $\mathbf{T}(JM)$ be a family of tensor operators with angular momentum JM, and consider the quantum matrix elements $\langle j'm'|\mathbf{T}(JM)|jm\rangle$. These matrix elements represent transition probability amplitudes from quantum state jm to state jm' caused by $\mathbf{T}(JM)$. The Wigner-Eckart theorem states that these matrix elements can be expressed by Eq. (21), where $\langle j'|\mathbf{T}(J)|j\rangle$ is the

$$\langle j'm'|\mathbf{T}(JM)|jm\rangle = \mathrm{C}(jJj', mM)\langle j'|\mathbf{T}(J)|j\rangle \quad (21)$$

reduced matrix element and is independent of mMm'. The angular momentum conservation is contained in the Wigner coefficient; the details of the particular operator $\mathbf{T}$ are contained only in the reduced matrix element.

The operator $\mathbf{T}$ in atomic spectroscopy is a perturbation hamiltonian from an incident electromagnetic wave that drives atomic transitions. Thus, the Wigner-Eckart theorem, through the Wigner coefficients, determines which transitions vanish, that is, the selection rules governing the process. Much of atomic spectroscopy follows from the rotational symmetry, or the quantum theory of angular momentum. *See* ATOMIC STRUCTURE AND SPECTRA; SELECTION RULES (PHYSICS). Brian De Facio; John L. Safko

Bibliography. L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics*, 1981; E. U. Condon and H. Odabasi, *Atomic Structure*, 1980; E. P. Wigner, *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra*, 1959.

Anharmonic oscillator

An oscillator that does not obey Hooke's law. This law is an idealized expression that assumes that a system displaced from equilibrium responds with a restoring force whose magnitude is proportional to the displacement. The use of Hooke's law results in a linear equation of motion that fails to describe many properties of the real world. Nature demonstrates two fundamentally different forms of nonlinearity, which may be called elastic anharmonicity and damping anharmonicity. To understand their difference, and the nature of a harmonic oscillator, it is necessary to understand potential functions. *See* HARMONIC MOTION; HARMONIC OSCILLATOR; HOOKE'S LAW; NONLINEAR PHYSICS.

Atomic interaction. In the classical formulation, where the simplest case is one-dimensional, the x-variable in the potential function $V(x)$ represents displacement away from the stable position of equilibrium that was established between competing forces of attraction and repulsion. For example, in the case of two atoms constrained against rotation, the potential describing their interaction can be approximated by the asymmetric potential curve in **Fig. 1***a*. It is not symmetric because of the Pauli exclusion principle, involving the overlap of electronic orbitals. Thus, it is more difficult to push the atoms together than it is to pull them apart. Therefore, the potential rises more quickly for compression than it does for extension. *See* EXCLUSION PRINCIPLE; INTERMOLECULAR FORCES; MOLECULAR STRUCTURE AND SPECTRA.

Although the potential of Fig. 1 was generated by means of an empirical formula, based on classical physics, $V(x)$ is foundational to the Hamiltonian with which the Schrödinger equation is derived. The Schrödinger equation, in turn, is used to calculate the probability of observing various states of position or

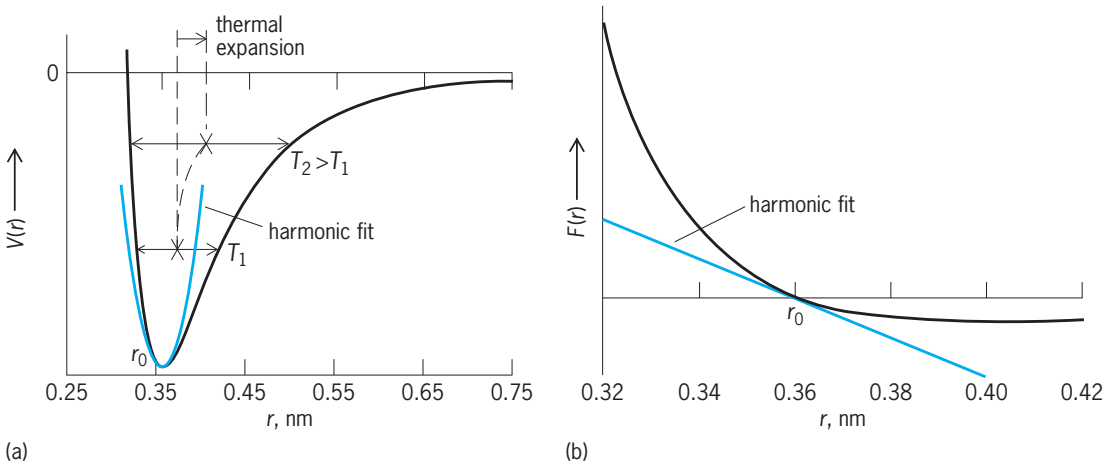


Fig. 1. Interaction between atoms. (*a*) Potential function of the interaction, *V*, plotted against *r*, the separation distance of the atomic nuclei. *V*(*r*) is plotted on a linear scale in units of negative joules. (*b*) Force, *F*(*r*), that is associated with the potential and results from its negative derivative [$F(r) = -dV/dr$] when the atoms are displaced from the equilibrium position r_0. *F*(*r*) is plotted on a linear scale in units of newtons.

momentum. Shown superposed on the anharmonic potential of Fig. 1 is a parabola that represents the harmonic approximation. An appreciation for its limitations can be realized by considering two different oscillatory motions, whose turning points are illustrated by the pair of horizontal lines with arrows on each end. These correspond to two different temperatures, T_1 and T_2, and the amplitude of oscillation is greater at the higher temperature, T_2. The shift toward larger time-averaged separation of the atoms as the temperature increases is the basis for thermal expansion. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUATION; THERMAL EXPANSION.

There are numerous other cases in which the harmonic potential is known to be inadequate. To cite one example, its use precludes the proper description of the heat capacity of solids. *See* SPECIFIC HEAT OF SOLIDS.

Simple pendulum. Elastic nonlinearity at large amplitudes of the motion is exhibited also by many macroscopic-scale oscillators, such as the pendulum. **Figure 2** shows the potential for a rigid, planar simple pendulum, in which the bob mass is able to go past vertical, that is, past the unstable equilibrium at the angle $\theta = \pi$ radians. This is not possible for a mass on a string, since the string can support only tension and not compression. *See* PENDULUM.

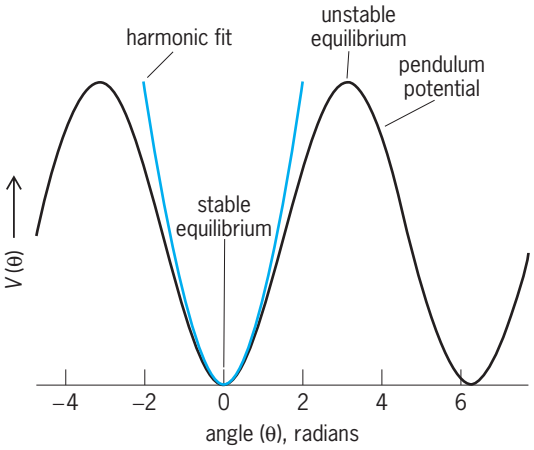


Fig. 2. Potential function for the rigid, planar simple pendulum.

The pendulum has become a macroscale archetype of chaos. For chaotic motion, the energy with which the damped oscillator is driven must be large enough to allow winding modes, in which the bob is allowed to move "over the top." Otherwise, it is a system demonstrating that nonlinearity is a necessary but not sufficient condition for chaos. At low levels, but not so small as to introduce complications due to damping nonlinearities, the motion of the pendulum is nearly harmonic and thus isochronous, meaning that the period is independent of amplitude. At high levels, the motion can be periodic and yet highly nonlinear. Highly nonlinear cases involve subharmonic response, in which it takes longer for the pendulum to repeat its motion than is required for the drive. Such modes cannot be explained by a linear equation of motion. *See* CHAOS.

Highly nonlinear but nonchaotic pendulum motion is a realm of amplitude jumps, as drive of the instrument is "tuned" back and forth in frequency across "resonance." Unlike the harmonic oscillator, when the pendulum is driven at large levels there is not a well-defined resonance frequency. If the drive frequency is below the pendulum's small-motion resonance frequency, small phase perturbations introduced by an external disturbance can seriously disrupt the steady-state motion. The pendulum can transition from a high-level, long-period mode to a low-level, short-period mode, even though the driving force was not altered. Such a change is suggestive of an analogy in physiology, ventricular fibrillation. *See* HEART DISORDERS.

Damping anharmonicity and internal friction. Anharmonicity manifests itself in different ways at different scales. At both the atomic scale and the macroscopic scale, the assumption of a smooth potential function is oftentimes reasonable. At the mesoscale,

this is no longer true, as defect organization takes place via the exceedingly complex interactions responsible for friction. When nature fills the vacuum it abhors, it rarely does so with perfection. At finite temperatures, for example, solids contain lattice sites devoid of an atom. These vacancies are called Schottky defects. Materials are never as strong as theoretical estimates of strength, based on the assumption of perfect lattices, because of defect structures such as dislocations. *See* CRYSTAL DEFECTS.

When material is strained in a cyclic manner, stress/strain hysteresis is observed, and it has been thoroughly studied by engineers. The anharmonic nature of this hysteresis has been described in terms of its influence on oscillator damping as a mechanism of internal friction. It differs in type from elastic anharmonicity, discussed above, and is illustrated by the graph of **Fig. 3**. *See* HYSTERESIS.

This form of anharmonicity has been the source of challenges to those who seek to study gravitational waves. Metastabilities, illustrated by the metastable equilibria in the enlargement in Fig. 3, are present in the springs used to isolate the mirrors of the LIGO (Laser Interferometer Gravitational-Wave Observatory) instrument from earth noise. Their presence disallows the occurrence of an absolute equilibrium position. A primary problem with trying to explain friction from first principles is that the fine structure responsible for these metastabilities is not static. Rather, because they are both temperature- and stress-dependent, defect structures evolve in time. *See* GRAVITATIONAL RADIATION; LIGO (LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY).

Damping anharmonicity is of a totally different character than viscous damping, and only the latter case is well known to physics. Because damping anharmonicity is incompatible with the Hooke's-law description of a spring, the description of energy loss must be modified (**Fig. 4**). *See* DAMPING.

In the internal friction case (Fig. 4*a*), it is seen that the minimum of the potential shifts back and forth as the spring is cycled in strain during free decay. This is due to material creep responsible for hysteresis, and it involves alterations in the configuration of defect structures. The amount of a given shift is a function of the amplitude of motion in the half cycle that just preceded the shift. For common exponential damping, the shift is proportional to amplitude. An energy-based description of exponential damping is provided by the equation below for the friction force shown in **Fig. 5**, which involves the function

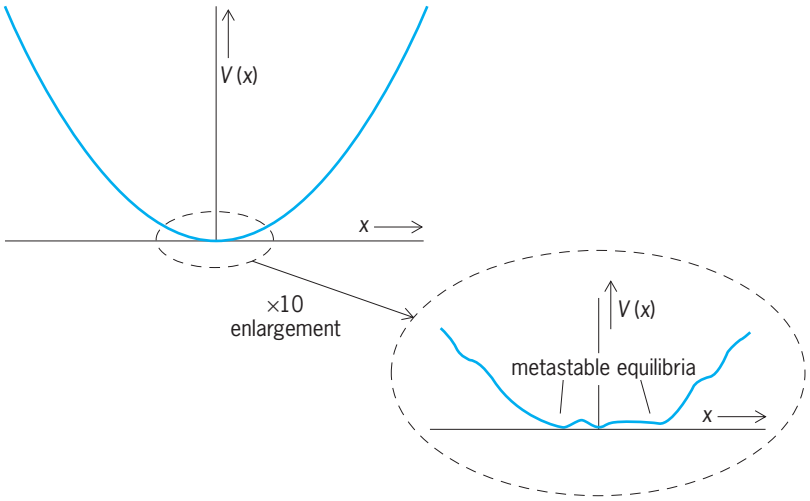


Fig. 3. Anharmonic potential responsible for internal friction damping. Only at low levels (as noted by the ×10 enlargement) does the nonlinear nature of the potential become observable.

$$\text{Friction force} = \text{(per unit mass)} \begin{cases} \dfrac{\omega}{Q} v & \text{for viscous damping} \\ \dfrac{\pi\omega}{4Q}\sqrt{\omega^2 x^2 + v^2}\,\text{sgn}(v) & \text{for hysteretic damping} \end{cases}$$

sgn(v). [When $v > 0$, sgn(v) = 1, and when $v < 0$, sgn(v) = −1.] Here, v is the velocity, ω is the angular frequency, and Q is the quality factor. *See* CREEP (MATERIALS); Q (ELECTRICITY).

Unlike viscous damping which is linear, hysteretic damping is nonlinear. Since the term under the

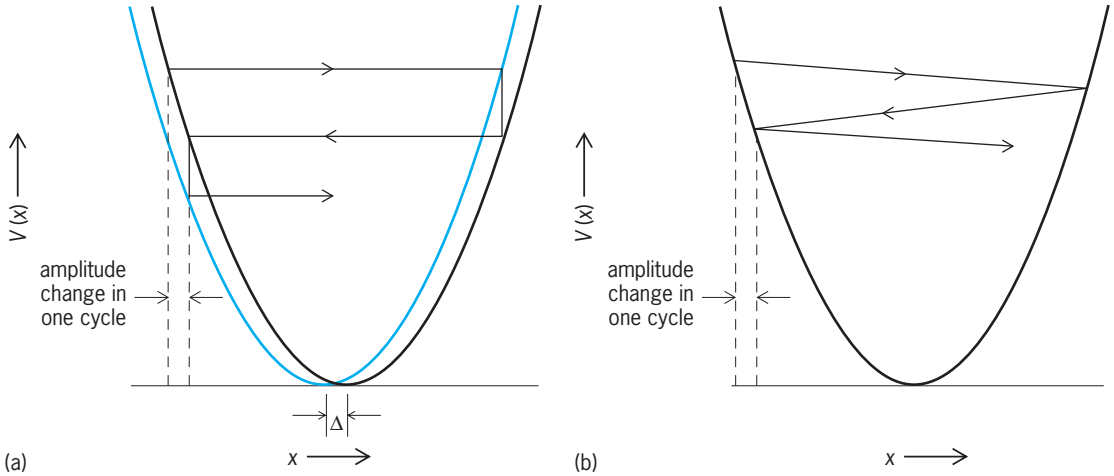


Fig. 4. Comparison of nonlinear and linear oscillator damping. (*a*) Nonlinear damping due to internal friction. (*b*) Linear damping due to an external force.

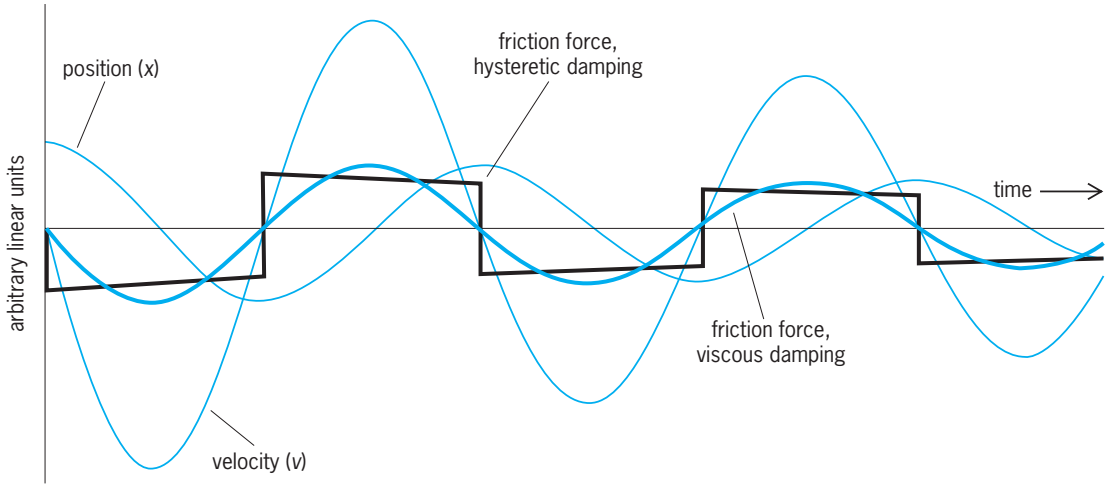


Fig. 5. Comparison of (1) viscous (linear) damping that derives from an external-to-the-oscillator friction force and (2) hysteretic (nonlinear) damping of the internal friction type. Both oscillators are in free decay with quality factor $Q = 10$ and angular frequency $\omega = 3.07$ Hz.

square root in the equation above is twice the energy of oscillation per unit mass, the velocity is seen to be important only by way of its algebraic sign. The two forms are similar only in that they yield the same quality factor Q for the exponential free decay. If one considers a Fourier series to describe the nonlinear friction force, it can be shown that the fundamental component of the series is the only term responsible for the energy loss per cycle that determines the Q. For this reason, many have failed to recognize the ubiquitous nature of nonlinear damping. The two forms differ greatly with regard to frequency dependence of Q. Because $\omega/Q = 2\beta$, where β is the coefficient commonly referred to as the damping constant, it is seen that Q must be proportional to frequency in the case of viscous damping. When monitored at low frequency and low energy, the vast majority of mechanical oscillators exhibit a Q dependence that is quadratic in the frequency, which is consistent with the nonlinear damping model presented. *See* FOURIER SERIES AND TRANSFORMS.

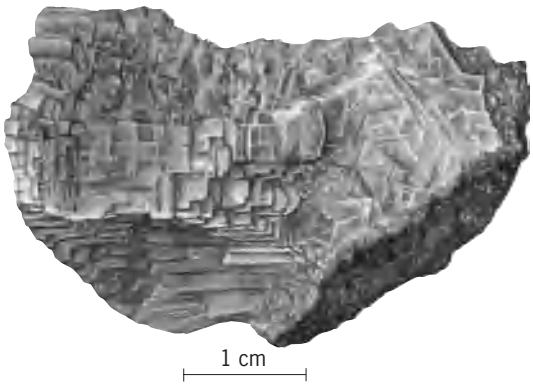
The properties of internal friction responsible for $Q \propto \omega^2$ were first discovered early in the twentieth century, and the investigators claimed that internal friction in solids appeared to be universal. Experiments since the 1990s support this belief, which implies a closer connection than previously recognized to the well-known case of surface friction described by C. A. Coulomb. In an additional early study by L. Portevin and F. Le Chatelier, which has been also mostly overlooked by physics, it was found that strain of alloys under stress is not smooth but jerky. The discontinuous changes observed are reminiscent of the Barkhausen effect, also discovered early in the twentieth century. *See* BARKHAUSEN EFFECT; FRICTION.

This jerky behavior, related to stick-slip features of friction, implies that atomic-scale quantum mechanics is not best suited to the description of internal friction. The electronvolt is infinitesimal compared to the quantum changes observed. There is evidence for an energy of mesoscale quantization, given by mc^2/α, where $m = 9.1 \times 10^{-31}$ kg is the electron mass, $\alpha = 1/137$ is the fine-structure constant, and $c = 3 \times 10^8$ m/s is the speed of light. This Compton scale of energy, having the mesoscale value of 1.1×10^{-11} J, is thought to regulate, through self-organizing metastabilities, the processes "still unknown from first principles" responsible for internal friction. It is possible that one of the last frontiers of physics actually lies in one of the most accessible parts of our world. Randall D. Peters

Bibliography. G. L. Baker and J. P. Gollub, *Chaotic Dynamics, an Introduction*, 2d ed., Cambridge University Press, 1996; R. Desalvo et al., Study of quality factor and hysteresis with the state-of-the-art passive seismic isolation system for Gravitational Wave Interferometric Detectors, *Nucl. Instrum. Meth. Phys. Res., Sec. A*, 538:526–537, 2005; C. Kittel, *Introduction to Solid State Physics*, 8th ed., Wiley, 2005; M. R. Matthews, C. F. Gauld, and A. Stinner (eds.), *The Pendulum: Scientific, Historical and Educational Perspectives*, Springer, 2005; R. D. Peters, Friction at the mesoscale, *Contemp. Phys.*, 45:475–490, 2004; R. D. Peters, Resonance response of a moderately driven rigid planar pendulum, *Amer. J. Phys.*, 64:170–173, 1996.

Anhydrite

A mineral with the chemical composition $CaSO_4$. Anhydrite occurs commonly in white and grayish granular masses, rarely in large, orthorhombic crystals (see **illus.**). Fracture is uneven and luster is pearly to vitreous. Hardness is 3–3.5 on Mohs scale and specific gravity is 2.98. It fuses readily to a white enamel. It is soluble in acids and slightly soluble in water. Anhydrite is an important rock-forming mineral and occurs in association with gypsum, limestone, dolomite, and salt beds. It is deposited directly by evaporation of seawater of high salinity at or above 108°F (42°C). Anhydrite can be produced artificially by dehydration of gypsum at about 390°F (200°C). Under natural conditions anhydrite



Anhydrite from Montanzas, Cuba. (*Specimen from Department of Geology, Bryn Mawr College*)

hydrates slowly, but readily, to gypsum. It is not used as widely as gypsum. Anhydrite is of worldwide distribution. Large deposits occur in the Carlsbad district, Eddy County, New Mexico, and in salt-dome areas in Texas and Louisiana. *See* GYPSUM; SALINE EVAPORITES. Edward C. T. Chao

Animal

Any living organism which possesses certain characteristics that distinguish it from plants is a member of the animal kingdom. There is no single criterion that can be used to distinguish all animals from all plants. Animals usually lack chlorophyll and the ability to manufacture foods from raw materials available in the soil, water, and atmosphere. Animal cells are usually delimited by a flexible plasma or cell membrane rather than a cell wall composed either of cellulose or chitin, as are the cells of most plants. Animals generally are limited in their growth and most have the ability to move in their environment at some stage in their life history, whereas plants are usually not restricted in their growth and the majority are stationary.

The presence or lack of chlorophyll in an organism does not determine its affinity to the plant or animal kingdom. Among the protozoa, the class Phytamastigophora includes animals, such as the euglenids, which have chromatophores containing chlorophyll. These organisms are considered to be animals by zoologists and plants by phycologists. The vestige of a feeding apparatus in these protozoa indicates that they have descended from forms that ingested food particles, that is, animals. Higher parasitic plants and the large plant group Fungi also lack chlorophyll. Another borderline group is the slime molds: the Mycetozoa of zoologists and the Myxomycophyta of the botanists. These organisms exhibit both plant and animal characteristics during their life history. Movement is not a characteristic restricted to the animal kingdom; many of the thallophytes such as *Oscillatoria*, numerous bacteria, and colonial chlorophytes are motile.

The grouping of living organisms into kingdoms is still unsettled. Previously biologists used only two groups, the Animalia and the Plantae, based on the large, familiar organisms then known to them. The discovery of microorganisms, such as bacteria and viruses, introduced difficulties for this simple scheme. Today biologists recognize up to five kingdoms, but with considerable differences of opinion. Most, however, place the one-celled animals and plants, sometimes along with algae and certain other groups, into the Protista. Other kingdoms are the Monera for the bacteria and blue-green algae, and the Fungi for the slime molds and true fungi. These schemes for recognizing additional kingdoms have the practical advantage of eliminating the difficulties of delimiting and describing the kingdoms of multicellular animals and plants. *See* ANIMAL KINGDOM; PLANT; PLANT KINGDOM. Walter J. Bock

Animal communication

A discipline within the field of animal behavior that focuses upon the reception and use of signals. Animal communication could well include all of animal behavior, since a liberal definition of the term signal could include all stimuli perceived by an animal. However, most research in animal communication deals only with those cases in which a signal, defined as a structured stimulus generated by one member of a species, is subsequently used by and influences the behavior of another member of the same species in a predictable way (intraspecific communication). In this context, communication occurs in virtually all animal species, if only as a means by which a member of one sex finds its partner.

The field of animal communication includes an analysis of the physical characteristics of those signals believed to be responsible in any given case of information transfer. A large part of this interest is due to technological improvements in signal detection (for example, the use of tape and video recorders and gas chromatographs), coupled with analysis of the signals obtained with such devices.

Communication modes. Information transmission between two individuals can pass in four channels: acoustic, visual, chemical, and electrical. An individual animal may require information from two or more channels simultaneously before responding appropriately to reception of a signal. Furthermore, a stimulus may evoke a response under one circumstance but be ignored in a different context.

Acoustic. Sound signals have characteristics that make them particularly suitable for communication, and virtually all animal groups have some forms which communicate by means of sound. Sound can travel relatively long distances in air or water, and obstacles between the source and the recipient interfere little with an animal's ability to locate the source.

Sounds are essentially instantaneous and can be altered in important ways. Both amplitude and frequency modulation can be found in sounds emitted by animals; in some species sound signals have discrete patterns due to frequency and timing of utterances. Since a wide variety of sound signals are

possible, each species can have a unique set of signals in its repertoire. Animals can travel directly toward a sound source, despite variable wind speed and direction, the onset of darkness, or simultaneous signaling by thousands of other animals of various species.

Sound signals are produced and received primarily during sexual attraction, including mating and competition. They may also be important in adult–young interactions, in the coordination of movements of a group, in alarm and distress calls, and in intraspecific signaling during foraging behavior. *See* REPRODUCTIVE BEHAVIOR.

The human hearing range (for example, of frequency and intensity) is somewhat limited, but with new technology sounds outside the range of human perception can be tape recorded. Such research reveals that many species of animals, particularly insects, apparently can communicate at frequencies much higher or lower than the human hearing range; a few species, perhaps elephants, communicate at frequencies below the human range. *See* HEARING (HUMAN); PHONORECEPTION.

Visual. Visual signaling between animals can be an obvious component of communication, as is the case for fireflies. Travel time is essentially instantaneous; however, visual signals have very different characteristics from those of sound. Sound requires a medium for transmission (such as air, water, or a substrate), but light travels unimpeded unless blocked by obstacles. Besides the normal range of human vision (visible light), visual signals, that is, the entire electromagnetic spectrum, include additional frequencies in the infrared (heat) and ultraviolet ranges.

The quality of light that is often considered is color, but other characteristics are important in visual communication. Alterations of brightness, pattern, and timing also provide versatility in signal composition. Two other features of vision deserve mention. First, an animal equipped with binocular vision can easily locate the precise source of a signal. Second, color patterns in animals (for example, butterflies) can serve as relatively permanent signals.

Even with all its favorable characteristics, the visual channel suffers from the important limitation that all visual signals must be line of sight. Information transfer is therefore largely restricted to the daytime (except for animals such as fireflies) and to rather close-range situations because of obstacles normally present in the environment.

Intraspecific visual signaling, as with acoustical and chemical signaling, appears to occur primarily during mate attraction. The color dimorphism of birds, the interesting patterns of butterfly wings, the posturing of some fish, and firefly flashing are examples. Among fireflies, for example, flying males flash at regular intervals in specific patterns as they cruise along. Stationary females of the same species, if receptive, flash at a set interval of time after the male, enabling the male to travel directly to them and not to a member of another species.

Some parent–young interactions involve visual signaling. A young bird in the nest may open its mouth when it sees the underside of its parent's beak. Other examples are the synchronized behavior observed in schooling fish and flocking birds.

As with sounds, the human range of perception of visual signals is somewhat limited. Some other species, such as honeybees, see well into the ultraviolet spectrum, and a few forms, such as some beetle species, can perceive objects in the infrared range.

Chemical. Chemical signals, like visual and sound signals, can travel long distances, but with an important distinction. Distant transmission of chemical signals requires a movement of air or water. Therefore, an animal cannot perceive an odor from a distance; it can only perceive molecules brought to it by a current of air or water. Animals do not hunt for an odor source by moving other than upwind or upcurrent in water, because chemical signals do not travel in still air or water since diffusion is far too slow.

The fact that chemical signals comprise molecules means that they have special attributes important in information transmission, attributes not shared by the other channels. Although acoustical or visual signals are essentially instantaneous, chemical signals have a time lag. A single molecule might stimulate a flying insect to head upwind, but if no other similar molecules are perceived, the insect might well abandon the search or drop downwind until it perceives other molecules of the same type.

Chemical signals at times have to be of an appropriate concentration if they are to be effective. A chemical normally considered to be an attractant can serve as a repellent if it is too strong. Chemical signals may persist for a while, and time must pass before the concentration drops below the threshold level for reception by a searching animal. Since molecules of different sizes and shapes have varying degrees of persistence in the environment, the chemical channel is often involved in territorial marking, odor trail formation, and mate attraction to a fixed location. This channel is particularly suitable where acoustical or visual signals might betray the location of a signaler to a potential predator.

The array of molecular structure is essentially limitless, permitting a species-specific nature for chemical signals. Unfortunately, that specificity can make interception and analysis of chemical signals a difficult matter for research.

Pheromones are chemical signals that are produced by an animal and are exuded through glandular openings (or otherwise) and influence the behavior of other members of the same species. If pheromones are incorporated into a recipient's body (by ingestion or absorption), they may chemically alter the behavior of such an individual for a considerable period of time. *See* PHEROMONE.

In social insects, long-term regulation of the behavior of individuals within a colony appears to be mainly, if not exclusively, due to chemical inhibition or enhancement of the endocrine system of individuals by pheromones received from other members of the colony. Queen honeybees or queen termites produce one or more chemical substances that inhibit

other females (workers) from becoming functional queens. *See* SOCIAL INSECTS.

All animal species seem to have a somewhat limited range of types of molecules that they can perceive. A substance with a particular odor or taste for humans may not even be perceived by a member of another species. The converse is also true: another species may well be able to smell or taste a compound that humans do not notice. *See* CHEMICAL ECOLOGY; CHEMORECEPTION.

Electrical. Some electric fish and electric eels live in murky water and have electric generating organs that are really modified muscle bundles. Communication by electric signaling is rapid; signals can travel throughout the medium (even murky water), and rather complex signals can be generated, permitting species-specific communication during sexual attraction. However, the electrical mode is apparently restricted to those species that have electric generating organs. *See* ELECTRIC ORGAN (BIOLOGY).

Context. Any of the above types of signals can be effective under the right circumstance but ineffective under other circumstances. For example, when recruited honeybees are searching for a potential hive site found by experienced foragers (scout bees), they are attracted by the chemical components of an exudate from a gland at the upper side and tip end of the abdomen. Similarly, when a honeybee swarm is moving from the location of its parent colony to a new site, an odor from the gland orients bees during transit and stimulates settling once the new site has been reached. However, recruited bees searching for flowers exploited by foragers from their colony apparently ignore those same chemicals.

Similarly, male moths of some species are first attracted to the odor of the host plant of their future offspring (plant-feeding caterpillars) and are then attracted to the pheromones exuded by females perched on that host plant. In some cases, if males are not first exposed to the odor of the larval host plant, they may not respond to pheromones exuded by females of their species. *See* PLANT.

Methodology. Animal communication is one of the most difficult areas of study in science for several reasons. First, experiments must be designed and executed in such a manner that extraneous cues (artifacts) are eliminated as potential causes of the observed results. Second, once supportive evidence has been obtained, each hypothesis must be tested. In animal communication studies, adequate tests often rely upon direct evidence—that is, evidence obtained by artificially generating the signal presumed responsible for a given behavioral act, providing that signal to a receptive animal, and actually evoking a specific behavioral act in a predictable manner.

A classic example of such a test entails holding a wooden dowel with a red dot on its underside over a gull chick, eliciting the same behavior in the chick as when a parent gull holds its beak over the nest. Whereas this example is quite simple and can be repeated with little equipment, acoustic, visual, chemical, and electric signals can be quite complex, requiring information derived from physics, chemistry, and mathematics and employing electronics and computers.

Nonhuman language. The fact that human beings have a language indicates that language is possible in an animal species, and this possibility has been investigated repeatedly in several nonhuman species. Dog, cat, and horse owners, for example, may have a strong bias on this point; those who work with personable animals, such as chimpanzees and dolphins, serve as another example.

Experiments conducted by Karl von Frisch in the 1940s yielded results suggestive of a language among honeybees in the form of movements by scout bees through which the location of food sources is communicated to other bees. Although the hypothesis drew wide interest and support because of its emotional appeal, experiments in the 1960s, coupled with a careful analysis of the design and controls of the original series of experiments, indicated that Frisch's own earlier (1937) odor-search hypothesis more than adequately accounted for the observation.

Continued attempts to demonstrate language use in various species (bees, pigeons, chimpanzees, dolphins), and experiments using a computer-simulated bee, have not been able to satisfy critics of the design or interpretation of the experiments. It is possible that human beings may well be the only animal species capable of using symbolic language. *See* ETHOLOGY; PSYCHOLINGUISTICS. Adrian M. Wenner

Bibliography. T. Lewis (ed.), *Insect Communication*, 1984; T. A. Sebeok (ed.), *How Animals Communicate*, 1977; A. M. Wenner, *Concept-centered vs. Organism-centered Research*, 1989; A. M. Wenner, D. Meade, and L. J. Friesen, Recruitment, search behavior and flight ranges of honeybees, *Amer. Zool.*, 31:768–782, 1991; A. M. Wenner and P. H. Wells, *Anatomy of a Controversy: The Question of a "Language" among Bees*, 1990.

Animal evolution

The theory that modern animals are the modified descendants of animals that formerly existed and that these earlier forms descended from still earlier and different organisms.

Animals are multicellular organisms that feed by ingestion of other organisms or their products, being unable to derive energy through photosynthesis or chemosynthesis. Animals are currently classed into about 30 to 35 phyla, each of which has evolved a distinctive body plan or architecture; representatives of the major phyla are depicted in **Fig. 1**.

Invertebrates

All phyla began as invertebrates, but lineages of the phylum Chordata developed the internal skeletal armature, with spinal column, which was exploited in numerous fish groups and which eventually gave rise to terrestrial vertebrates. The number of phyla is uncertain partly because most of the branching patterns and the ancestral body plans from which

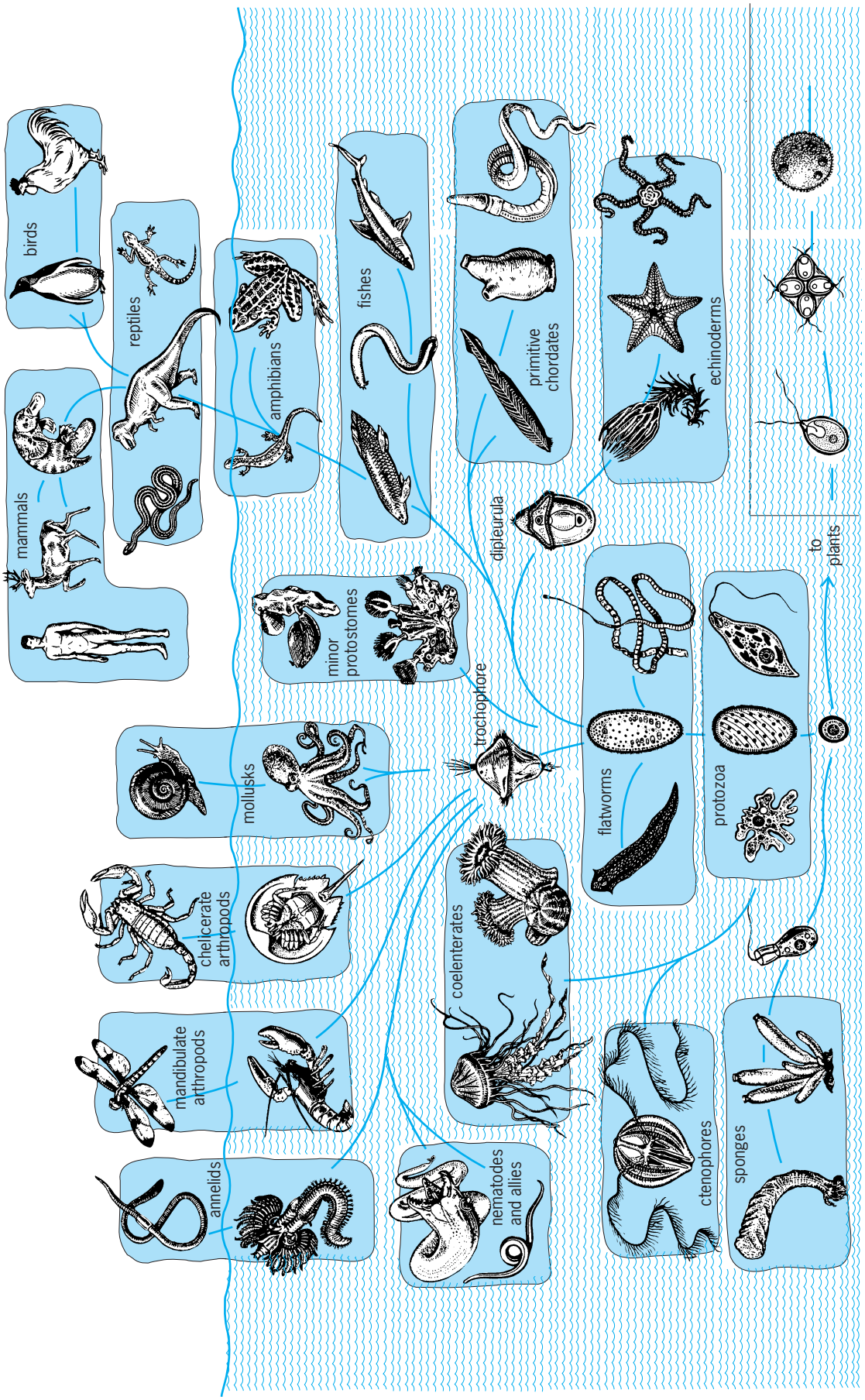


Fig. 1. Evolution of animal groups, showing hypothetical relationships. (*After G. B. Moment, General Zoology, 2d ed., Houghton Mifflin, 1967*)

putative phyla have arisen are not yet known. For example, arthropods (including crustaceans and insects) may have all diversified from a common ancestor that was a primitive arthropod, in which case they may be grouped into a single phylum; or several arthropod groups may have evolved independently from nonarthropod ancestors, in which case each such group must be considered a separate phylum. So far as known, all animal phyla began in the sea. *See* ANIMAL; ANIMAL KINGDOM.

Several lines of evidence bear significantly upon the ancestry of these animals and of their major subdivisions. Classically, animals were grouped according to the similarities of their adult body plans, and while the body plans have provided a basis for the recognition of most of the phyla and other distinctive animal groups, they have not provided definitive evidence of the interrelations of the phyla. A second approach is to compare patterns of early development, which should be more conservative than adult features; such comparisons have permitted the grouping of some phyla into likely alliances. Hypotheses as to the body plans of phylum ancestors have been erected from these morphologic data. It has not proven possible to corroborate any hypotheses, however, for the earliest members of the phyla (and of invertebrate classes) appear abruptly in the fossil record, and their ancestors cannot be traced. Furthermore, numbers of extinct phyla or other major animal groups also appear in the fossil record, adding branches to the tree of life which must be reckoned with, but with no definitive indication of either their origins or branching patterns. Finally, genetic changes that have accumulated after branching events separated the major groups should have left a pattern of variation in the structure of genes that may be reconstructed to yield a phylogeny. If genetic changes have occurred in a reasonably regular manner, then the amount of divergence in the structure of genes and gene products, such as protein and ribonucleic acid (RNA) molecules, should be proportional to the time since branching occurred. Some genes change so fast that their pattern of divergence is useful only for recent branching events, while others are so conservative that they are used to study divergences that occurred billions of years ago. These methods are promising when used in conjunction with other evidence.

Major animal groups. The simplest living phyla may have no direct relationships to other animals (**Fig. 2**), but indicate the versatility of multicellular design. Porifera (sponges) contain several differentiated cell types that form body walls, but their architecture is limited to folding these walls into chambers that are sometimes highly convoluted. Coelenterata (=Cnidaria; sea anemones, corals, and jellyfish) have two cellular body layers, separated by a gelatinous layer; the inner tissue forms a digestive cavity. *See* CNIDARIA; PORIFERA.

Flatworms (Platyhelminthes) and their allies have solid bodies, with both external (ectoderm) and internal (endoderm) tissue layers separated by a third cellular layer (mesoderm); there is a gut lined with endoderm. All phylogenetic evidence suggests that a flatwormlike form gave rise to the remaining phyla, which possess body cavities (other than a gut) in which reproductive and other organs can be sequestered. In about 15 of these phyla this cavity (the coelom) is lined by mesoderm and usually functions as a hydrostatic skeleton. *See* PLATYHELMINTHES.

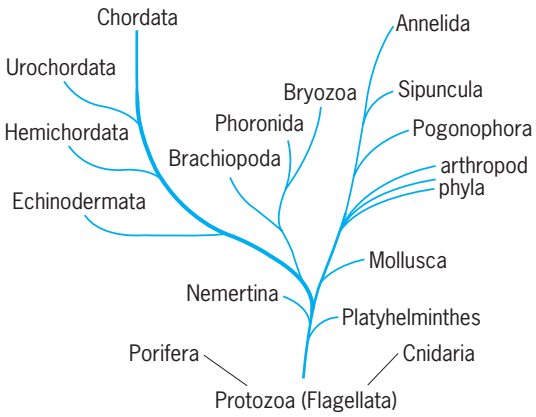


Fig. 2. Phylogenetic tree depicting the possible relations among the phyla according to evidence from all sources. The branching pattern is constrained by models of development and of body-plan evolution and by molecular data. Branch length is not to scale.

The coelomates constitute all of the familiar, larger-bodied animals and are commonly grouped into two main divisions (Fig. 2). In one group, the body was originally segmented, as in annelid worms or arthropods, though the segmentation may be reduced or lost in some branches. In the other group, the coelom is divided into two or three regions; these include the Enchinodermata (starfish and sea urchins), Tunicata (sea squirts), and Chordata (including humans). Additionally, there are coelomate phyla that cannot yet be assigned to either of these groups; some, such as the Brachiopoda, Phoronida, and Bryozoa, display an enigmatic mixture of traits that otherwise characterize one or the other of the main groups. In about 10 phyla (not shown in Fig. 2) the body cavity is not a true coelom but lies between endoderm and mesoderm and is termed a pseudocoel. These pseudocoelomates are soft-bodied, small to minute, and commonly parasitic; some of them have never been found as fossils. *See* ANNELIDA; BRACHIOPODA; BRYOZOA; CHORDATA; ECHINODERMATA; PHORONIDA; TUNICATA.

Origins of animals. Some features of the cells of primitive animals resemble those of the single-celled Protozoa, especially the flagellates, which have long been believed to be animal ancestors. Molecular phylogenies have supported this idea and also suggest that the phylum Coelenterata arose separately from all other phyla that have been studied by this technique. Thus animals may have evolved at least twice from organisms that are not themselves animals, and represent a grade of evolution and not a single branch (clade) of the tree of life. Sponges have also been suspected of an independent origin, and it is possible

that some of the extinct fossil phyla arose independently or branched from sponges or cnidarians. *See* PROTOZOA.

The earliest undoubted animal fossils (the Ediacaran fauna) are soft-bodied, and first appear in marine sediments nearly 650 million years (m.y.) old. This fauna lasted about 50 m.y. and consisted chiefly of cnidarians or cnidarian-grade forms, though it contains a few enigmatic fossils that may represent groups that gave rise to more advanced phyla. Then, nearly 570 m.y. ago, just before and during earliest Cambrian time, a diversification of body architecture began that produced most of the living phyla as well as many extinct groups. The body plans of some of these groups involved mineralized skeletons which, as these are more easily preserved than soft tissues, created for the first time an extensive fossil record. The earliest mineralized skeletons are chiefly minute, the small shelly fauna of earliest Cambrian age; and while some of these fossils are primitive members of living phyla, many—perhaps 10 to 20 kinds of them—are distinctive and quite possibly represent phyla or at least major branches of phyla that soon became extinct, geologically speaking. The soft-bodied groups were also markedly diversified, though their record is so spotty that their history cannot be traced in detail. A single, exceptionally preserved soft-bodied fauna from the Burgess Shale of British Columbia that is about 530 m.y. old contains not only living soft-bodied worm phyla, but extinct groups (perhaps a dozen) that cannot be placed in living phyla and do not seem to be ancestral to them. Among these is *Wiwaxia,* a novel form which bears spinelike elements similar to some of the small shelly fossils of Early Cambrian age, thus providing evidence that the small shelly forms did indeed include novel types. Clearly, an early radiation of animals produced a vast array of novel body plans in an evolutionary episode that has never been surpassed. *See* BURGESS SHALE; CAMBRIAN; EDIACARAN BIOTA; FOSSIL.

History of major groups. Following the early phase of rampant diversification and of some concurrent extinction of phyla and their major branches, the subsequent history of the durably skeletonized groups can be followed in a general way in the marine fossil record (**Fig. 3**). The composition of the fauna changed continually, but three major associations can be seen: one dominated by the arthropodlike trilobites during the early Paleozoic (Fig. 3*a*), one dominated by articulate brachiopods and crinoids (Echinodermata) in the remaining Paleozoic (Fig. 3*b*), and one dominated by gastropod (snail) and bivalve (clam) mollusks during the

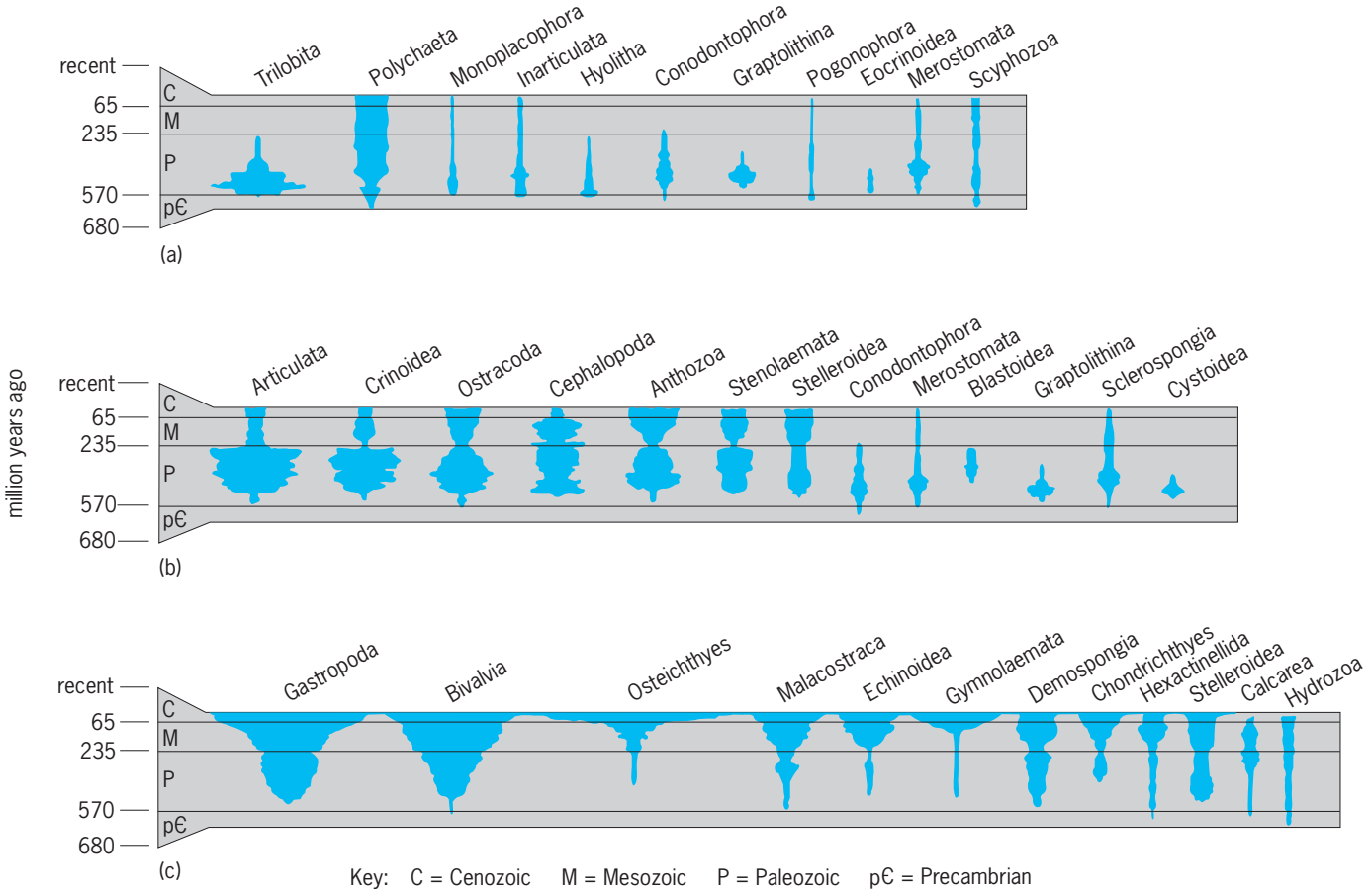


Fig. 3. Marine fossil record of major groups during the last 680 million years. The width of the spindles is proportional to the number of families present. The shifting of dominance from (*a*) trilobites and other forms during the early Paleozoic to (*b*) articulate brachiopods, crinoids, and others in the late Paleozoic and then to (c) gastropod and bivalve mollusks can be seen. (*After J. J. Sepkoski, Jr., in J. W. Valentine, ed., Phanerozoic Diversity Patterns: Profiles in Macroevolution, Princeton University Press, 1985*)

Mesozoic and Cenozoic (Fig. 3*c*). The mass extinction at the close of the Paleozoic that caused the contractions in so many groups (Fig. 3) may have extirpated over 90% of marine species and led to a reorganization of marine community structure and composition into a modern mode. Resistance to this and other extinctions seems to have been a major factor in the rise of successive groups to dominance. On land, the fossil record is too spotty to trace in detail the invasions of the terrestrial environment, which must have been under way by the mid-Paleozoic. Annelids, arthropods, and mollusks are the more important invertebrate groups that made the transition to land. The outstanding feature of terrestrial fauna is the importance of the insects, which appeared in the late Paleozoic and later radiated to produce the several million living species, surpassing all other life forms combined in this respect. *See* ANNELIDA; ARTHROPODA; CENOZOIC; INSECTA; MESOZOIC; MOLLUSCA; PALEOZOIC.

James W. Valentine

Chordate Origins

The phylum Chordata consists largely of animals with a backbone, the Vertebrata, including humans (**Fig. 4**). The group, however, includes some primitive nonvertebrates, the protochordates: lancelets, tunicates, acorn worms, pterobranchs, and possibly the extinct graptolites and conodonts. The interrelationships of these forms are not well understood. With the exception of the colonial graptolites, they are soft-bodied and have only a very limited fossil record. They suggest possible links to the Echinodermata in developmental, biochemical, and morphological features. In addition, some early Paleozoic fossils, the carpoids, have been classified alternatively as chordates and as echinoderms by various students, again suggesting a link. In spite of these various leads, the origin of the chordates remains basically unclear. *See* VERTEBRATA.

Chordates are characterized by a hollow, dorsal, axial nerve chord, a ventral heart, a system of slits in the pharynx that serves variously the functions of feeding and respiration, a postanal swimming tail, and a notochord that is an elongate supporting structure lying immediately below the nerve chord. The protochordates were segmented, although sessile forms such as the tunicates show this only in the swimming, larval phase. Even the free-swimming forms were not truly active, and they used the gill apparatus primarily for feeding on small particles

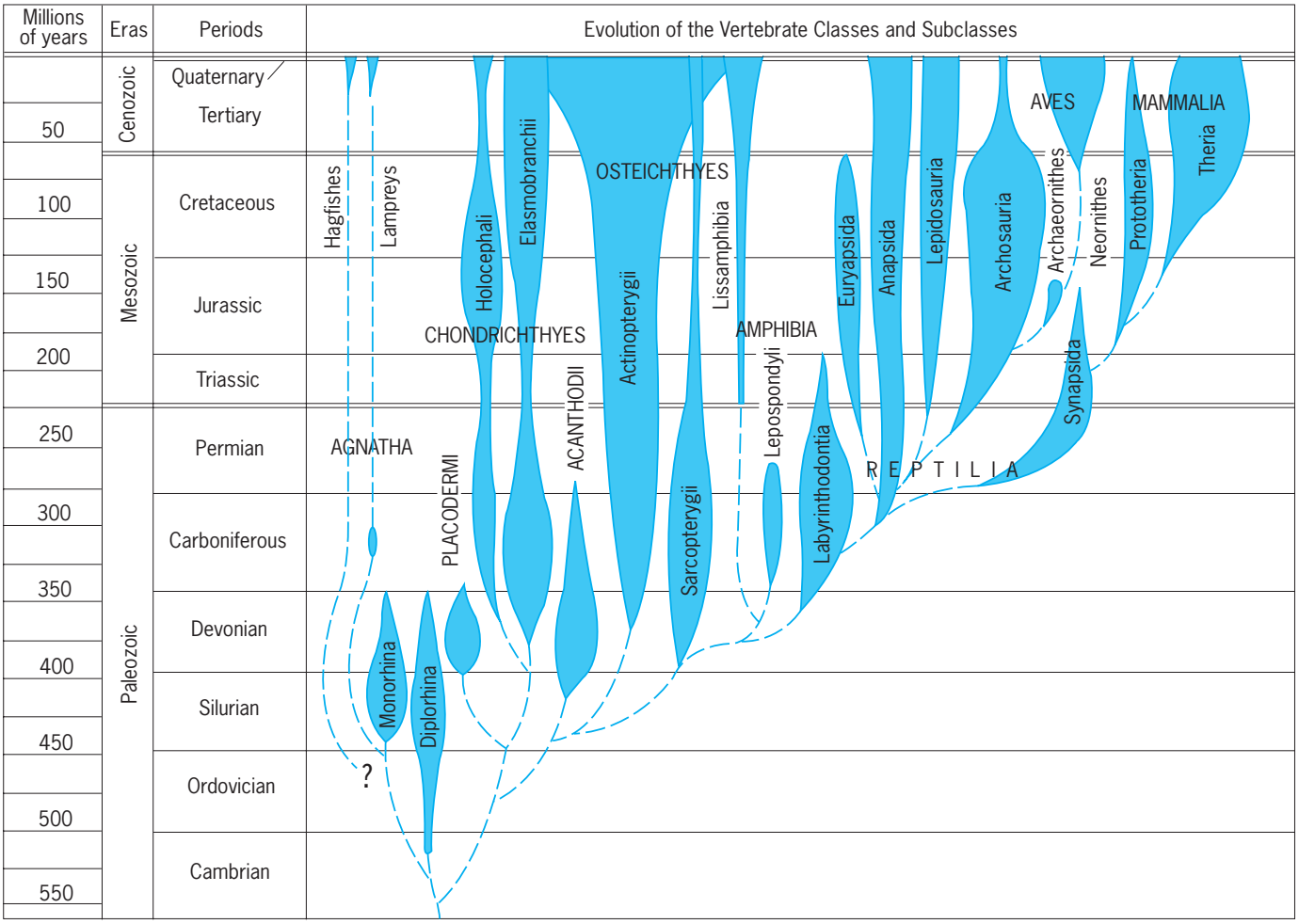


Fig. 4. Vertebrate evolution in relation to geologic time scale. Classes are shown in capital letters; subclasses are shown in lowercase letters, except that "Hagfishes" and "Lampreys" are common names for orders. (*After M. Hildebrand, Analysis of Vertebrate Structure, Wiley, 1982*)

trapped on mucous streams generated in the pharynx. There are two hypotheses concerning the nature of the protochordates. One holds that all protochordates were originally sessile, with, however, active larvae, and that the vertebrates arose from the larval phase through increased development of the segmented muscular trunk and associated nervous system. The alternative hypothesis proposes that the sessile condition seen in many protochordates is secondary.

The first vertebrates were fishlike animals in which the pharyngeal slits formed a series of pouches that functioned as respiratory gills. An anterior specialized mouth permitted ingestion of food items large in comparison with those of the filter-feeding protochordates. Vertebrates are first known from bone fragments found in rocks of Cambrian age, but more complete remains have come from the Middle Ordovician. Innovations, related to greater musculoskeletal activity, included the origin of a supporting skeleton of cartilage and bone, a larger brain, and three pairs of cranial sense organs (nose, eyes, and ears). At first the osseous skeleton served as protective scales in the skin, as a supplement to the notochord, and as a casing around the brain. In later vertebrates the adult notochord is largely or wholly replaced by bone, which encloses the nerve chord to form a true backbone. All vertebrates have a heart which pumps blood through capillaries, where exchanges of gases with the external media take place. The blood contains hemoglobin in special cells which carry oxygen and carbon dioxide. During the exchanges the oxygen content of the cells is increased and that of carbon dioxide decreased. In most fishes the blood passes from the heart to the gills and thence to the brain and other parts of the body. In most tetrapods, and in some fishes, blood passes to the lungs, is returned to the heart after oxygenation, and is then pumped to the various parts of the body.

Fish evolution. The jawless fish, known as Agnatha, had a sucking-rasping mouth apparatus rather than true jaws. They enjoyed great success from the Late Cambrian until the end of the Devonian. Most were heavily armored, although a few naked forms are known. They were weak swimmers and lived mostly on the bottom. The modern parasitic lampreys and deep-sea scavenging hagfish are the only surviving descendants of these early fish radiations. *See* DEVONIAN; JAWLESS VERTEBRATES.

In the Middle to Late Silurian arose a new type of vertebrate, the Gnathostomata, characterized by true jaws and teeth. They constitute the great majority of fishes and all tetrapod vertebrates. The jaws are modified elements of the front parts of the gill apparatus, and the teeth are modified bony scales from the skin of the mouth. With the development of jaws, a whole new set of ecological opportunities was open to the vertebrates. Along with this, new swimming patterns appeared, made possible by the origin of paired fins, forerunners of which occur in some agnathans. *See* GNATHOSTOMATA; SILURIAN.

Four groups of fishes quickly diversified (Fig. 2). Of these, the Placodermi and Acanthodii are extinct. The Placodermi were heavily armored fishes, the dominant marine carnivores of the Silurian and Devonian, rivaled only by the large eurypterid arthropods. The Acanthodii were filter-feeders mostly of small size. They are possibly related to the dominant groups of modern fishes, the largely cartilaginous Chondrichthyes (including sharks, rays, and chimaeras) and the Osteichthyes (the higher bony fishes). These also arose in the Late Silurian but diversified later. *See* ACANTHODII; CHONDRICHTHYES; OSTEICHTHYES; PLACODERMI.

Conquest of land. The first land vertebrates, the Amphibia, appeared in the Late Devonian and were derived from an early group of osteichthyans called lobe-finned fishes, of which two kinds survive today, the Dipnoi or lungfishes, and the crossopterygian coelacanth *Latimeria*. They were lung-breathing fishes that lived in shallow marine waters and in swamps and marshes. The first amphibians fed and reproduced in or near the water. True land vertebrates, Reptilia, with a modified (amniote) egg that could survive on land, probably arose in the Mississippian. *See* AMNIOTA; AMPHIBIA; DIPNOI; MISSISSIPPIAN.

Reptile radiations. By the Middle Pennsylvanian a massive radiation of reptiles was in process. In it can be traced the line leading to mammals as well as the lineages of the great reptiles that dominated the Mesozoic Era. The most prominent reptiles belong in the Diapsida: dinosaurs, lizards and snakes, and pterosaurs (flying reptiles). The birds, Aves, which diverged from the dinosaur radiation in the Late Triassic or Early Jurassic, are considered to be feathered dinosaurs, and thus members of the Diapsida, whereas older authorities prefer to treat them as a separate class. In addition, there were several Mesozoic radiations of marine reptiles such as ichthyosaurs and plesiosaurs. Turtles (Chelonia) first appeared in the Triassic and have been highly successful ever since. *See* AVES; DINOSAURIA; JURASSIC; PENNSYLVANIAN; REPTILIA.

Mammalian origins. The line leading to mammals can be traced to primitive Pennsylvanian reptiles, Synapsida, which diversified and spread worldwide during the Permian and Triassic. The first true mammals, based on characteristics of jaw, tooth, and ear structure, arose in the Late Triassic. Derived mammals, marsupials (Metatheria) and placentals (Eutheria), are known from the Late Cretaceous, but mammalian radiations began only in the early Cenozoic. By the end of the Eocene, all the major lines of modern mammals had become established. Mammals are easily separated into distinct groups (orders), but their relationships are not easy to discover from fossil records because of the explosion of mammalian evolution in the early Cenozoic. Molecular analyses (blood proteins, deoxyribonucleic acid, ribonucleic acid) of living mammals show that the most primitive group of placentals is the edentates (sloths, armadillos, and anteaters). An early large radiation included the rodents, primates (including monkeys, apes, and

humans), and bats, possibly all closely related to the insectivores and carnivores. The newest radiations of mammals are of elephants and sea cows, while the whales are related to the artiodactyls (cattle, camels).

Because of biogeographic isolation, marsupials came to flourish in Australia and South America, whereas the placentals diversified widely in Eurasia and North America. *See* CENOZOIC; CRETACEOUS; EOCENE; MAMMALIA; PERMIAN; SYNAPSIDA; TRIASSIC.

Primates and humans. Among the early and primitive lines of placental mammals were the Paleocene and Eocene plesiadapoids from which the primates arose. They were mostly nocturnal, insect- and fruit-eating animals with forwardly directed eyes and a locomotor system well developed for an arboreal life. The living primates consist of prosimians; tarsiers; lemurs; monkeys and anthropoids; apes; and humans and near relatives, the hominids. During the Oligocene, monkeys diverged into Old and New World lines. From the former, the apes arose in the late Oligocene. Known as the Pongidae, they are represented today by gibbons, orangutans, chimpanzees, and gorillas. *See* APES; OLIGOCENE.

The Hominidae, which include the human subspecies *Homo sapiens sapiens*, diverged from the apes about 5 million years ago. The closest living relatives among the apes are chimpanzees and gorillas. The earliest known hominids consist of fossils of the genus *Australopithecus*, excavated from rocks formed about 3.75 million years ago in eastern Africa. Several distinct lines of *Australopithecus*, under various generic names, coexisted for a long period of time; and also, during the later part of their existence, with *H. habilis*, which lived from about 2 million to 1.75 million years ago. This early species of *Homo* was notable for the relatively large brain, and probably included the first tool user. It was superseded by *H. erectus*, a larger animal with a greatly increased relative brain size. About 1 million years ago, *H. erectus* spread from Africa to eastern and southern Asia. It too has been given several generic names, of which *Pithecanthropus* and *Sinanthropus* from eastern Asia are the most familiar. *See* AUSTRALOPITHECINE; FOSSIL HUMANS.

Skeletal remains anatomically similar to those of *H. sapiens* have been found in northern Africa in beds formed about 300,000 years ago. Truly modern forms appeared some 100,000 years ago. They penetrated Europe some 30,000 to 40,000 years ago and spread throughout the Earth, occupying the Old World first, penetrating to Australia, and about 12,000 to 15,000 years ago entering North America and spreading rapidly to South America. *See* PALEONTOLOGY.

Keith S. Thomson

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988; K. G. Field et al., Molecular phylogeny of the animal kingdom, *Science*, 239: 748–752, 1988; M. F. Glaessner, *The Dawn of Animal Life*, 1984; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966; J. W. Valentine (ed.), *Phanerozoic Diversity Patterns: Profiles in Macroevolution*, 1985; H. B. Whittington, *The Burgess Shale*, 1985; J. Z. Young, *Life of the Vertebrates*, 3d ed., 1991.

Animal feeds

Substances suitable for the nutrition of animals. This article discusses the main category—livestock and poultry feed— then briefly describes the secondary area—pet foods.

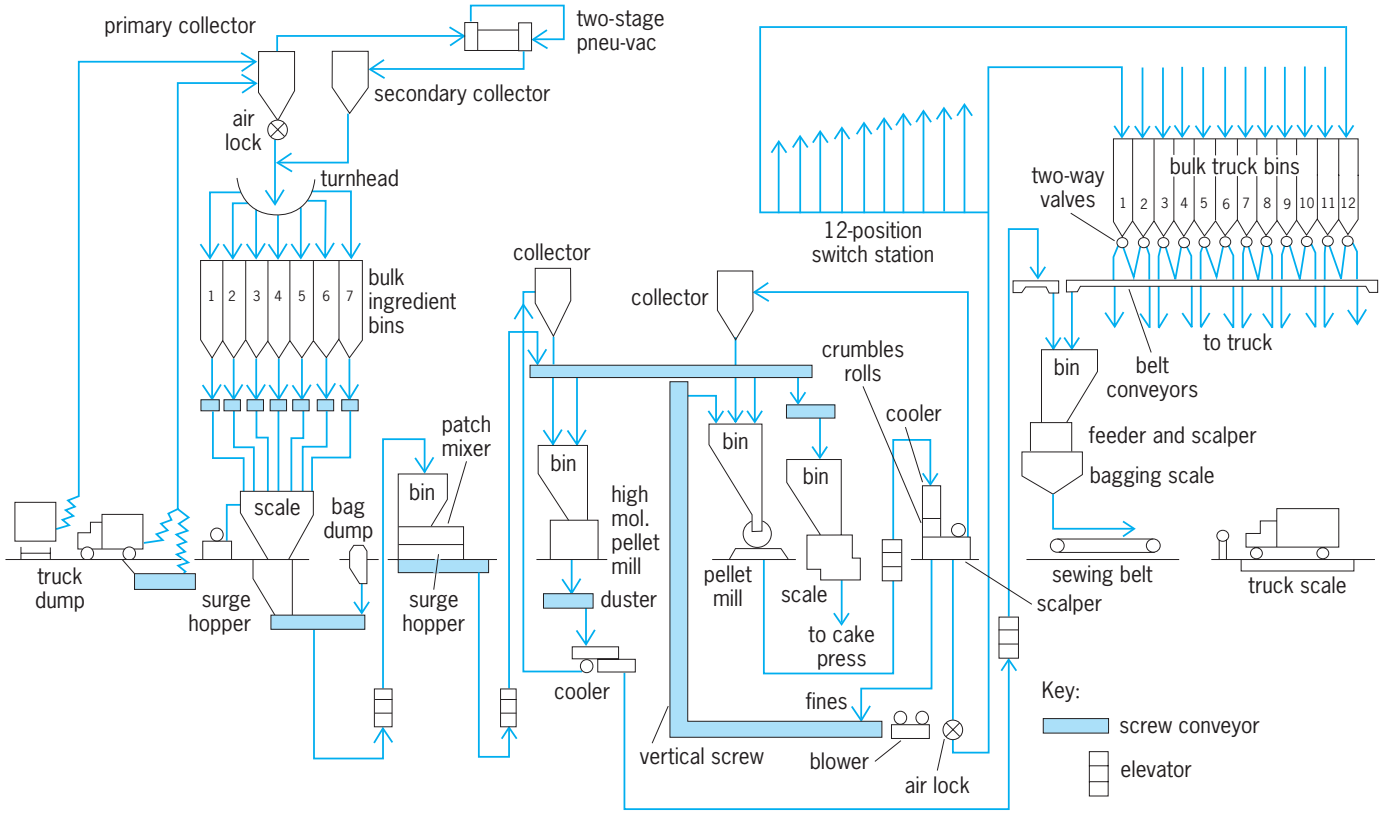
Livestock and Poultry Feed

In formulating livestock and poultry diets, animal nutritionists select from a variety of bulk feed ingredients including forages, feed grains, protein concentrates, by-product feeds, and vitamin-mineral supplements. These are formulated into complete rations in modern feed manufacturing plants (see **illus.**) to provide a balance of energy, protein, vitamins, and minerals to meet the nutrient requirements of various classes of livestock and poultry. Drugs and other additives may also be needed to meet specific needs. Today most feed manufacturers use high-speed electronic computers and various linear programming techniques to formulate diets. The overall goal is to prepare the lowest-cost ration that will provide economical livestock and poultry products for the consumer while maximizing returns to the producer.

Forages. In a broad sense, forages consist of grasses and legumes fed to ruminants (cattle, sheep, goats) and horses. Because of symbiotic microbes (bacteria and protozoa) in the digestive tract of these animals, they are uniquely equipped to utilize nutrients in forages; thus they do not compete with humans for the same food source. *See* FORAGE CROPS.

Grazing. Pasture and rangeland areas are covered with forage that is harvested by the grazing animal. Pasture provides about one-third of the nutrients consumed by dairy cattle and about half of those used by beef cattle and sheep. Cattle and sheep on rangeland obtain more than half their feed from forage. Grazing quality forage has the advantages of providing more digestible protein per acre than most other feeds, eliminating harvesting and feeding costs, and reducing soil erosion that results from the action of water and wind. While a portion of the forage is returned to the land as manure, appreciable forage may be lost due to trampling and selection by the grazing animals. To preclude such loss, forages are often mechanically harvested and fed as hay, silage, or dehydrated material. Although this practice minimizes loss and permits forage to be harvested at the optimum stage of maturity and highest quality, additional labor and energy are required to operate the harvesting, processing, and feeding equipment. *See* GRASS CROPS.

Hay and silage. In the United States, cool-season grasses, such as tall fescue and orchard grass, grow most rapidly during the spring and fall when rainfall is adequate and temperatures are moderate. Conversely, warm-season species, such as bahia grass, bermuda grass, and digit grass, grow best during the summer. Furthermore, because of variable climatic conditions, forage species do not grow at a uniform rate throughout the growing season. Because of this uneven distribution in forage production, most farms produce more forage than can be consumed by the



Flow diagram of formula feed plant. (*Sprout Waldron and Co., Inc.*)

livestock. This surplus forage is conserved as hay or silage and either fed or sold at later periods.

Hay is dried forage which has been cut and sun-cured in the field to about 10–20% moisture and then baled or otherwise packaged. It is commonly fed to ruminants and horses when grazing is inadequate or when the animals are confined. In the United States, alfalfa is the major hay crop, but other legumes, as well as annual and perennial grasses, are also extensively grown. The type of hay produced in a region depends primarily on those species which are best adapted to that area. For example, in the Northeast, timothy and alfalfa are important hay crops, while in the humid Southeast, warm-season perennial grasses such as Coastal bermuda grass and Pensacola bahia grass predominate.

Silage is fermented forage which is produced by the action of acetic and lactic acid–producing anaerobic bacteria in closed structures called silos. Although almost any forage crop can be preserved as silage, corn ranks first in the United States because good-quality, chopped, whole corn plant ensiles well and produces more digestible nutrients per acre than any other forage. Other silage consists of sorghum, grass, legumes, small grains, and miscellaneous crops. *See* CLOVER; LEGUME FORAGES.

Dehydration. Preserving forage by dehydration is relatively new compared to hay-making or ensiling. Although this industry is based largely on alfalfa, other crops, including Coastal bermuda grass, Johnson grass, and whole corn plant are also dehydrated.

In this process, freshly harvested or field-wilted forage is chopped and fed directly into the combustion gases (1600–2000°F or 870–1090°C) inside rotating horizontal drum dryers. These may be single- or triple-pass and range 8–12 ft (2.4–3.7 m) in diameter and 24–65 ft (7–20 m) in length. The evaporative capacity ranges 3–30 tons (2.7–27 metric tons) of water per hour. During drying, moisture diffuses to the surface of the forage particles as fast as it is evaporated. This rapid evaporation of surface moisture keeps the plant material cool enough to prevent burning. After remaining in the drum for about 2–10 min, the dehydrated forage is separated from the moisture-laden exit gases (250–350°F or 120–180°C) by means of a cyclone separator or other type of separator. Next, the forage is ground in a hammer-mill and pelleted for ease of storage, transportation, and feeding. To control dust, about 1.0% animal fat or vegetable oil may be added prior to grinding.

Mechanically dehydrated forages such as alfalfa and Coastal bermuda grass usually do not compete with hay or silage as roughages for ruminants. These products are high in protein, minerals, carotene (provitamin A), tocopherol (vitamin E), vitamin K, xanthophylls (substances that provide yellow color to poultry skin and egg yolks), and unidentified growth and reproduction factors. They are used primarily as supplements in swine and poultry rations and pet foods. The importance of carotene, xanthophylls, and vitamin E is such that producers of pelleted, dehydrated alfalfa or Coastal bermuda grass usually store the product under inert gas or add an

antioxidant (ethoxyquin) to minimize oxidation and ensure a high-quality product.

Feed grains. Feed grains are any of several grains commonly used as livestock and poultry feeds; they include corn, barley, oats, and grain sorghum (milo). Although wheat is second only to corn as a cereal grain in the United States, wheat is not usually fed to livestock. Most of the wheat crop is used for the manufacture of flour and other human foods. However, when properly used, wheat is a satisfactory feed for all classes of livestock. Accordingly, when the price of feed grains is high and surplus wheat is available at low price, considerable amounts may be fed to animals. *See* BARLEY; CORN; OATS; SORGHUM; WHEAT.

The purpose of processing feed grains is to improve their palatability, digestibility, and overall feed efficiency. For many years, stockmen sought to do this by grinding, crushing, rolling, and soaking feed grains, but since the late 1950s several new processing techniques have been used. One such technique is to ensile high-moisture (32%) ground ear corn. This has been shown to result in 10–15% greater utilization by cattle compared to regular ground ear corn. Additional techniques for improving grains by processing include reconstitution (moistening), flaking, roasting, micronizing, popping, and exploding.

Reconstitution. The process in which water is added to grain to bring its moisture content up to 20–30% is known as reconstitution. Usually grain so treated is held in limited-oxygen storage for a minimum of 3 weeks. Generally, cattle fed reconstituted corn or milo gained 8–15% more weight than those fed regular grain; and the feed efficiency of the reconstituted grain is increased by 7–11%.

Flaking. In a modification of steam rolling, the grain is cooked or steamed at approximately 200°F (93°C) at atmospheric pressure for 15 to 30 min, or at higher temperature and pressure, approximately 300°F (150°C) and 50 lb/in.^{2} (345 kilopascals) for 1 to 2 min, and then is rolled into flakes $^1/_{\sqrt{32}}$ in. (0.8 mm) thick. Because this process increases the moisture content to about 15–20%, the rolled flakes are immediately dried to 15% moisture. Flaking was the first modern process whereby the rate of gain and feed efficiency of milo were markedly increased. For beef cattle, gain and feed efficiency increased up to about 11% and 15%, respectively.

Roasting. In roasting corn, the grain is heated to 300°F (150°C). This results in expansion of the kernel, as evidenced by the fact that the roasted material weighs 39 lb/ft^3 (625 kg/m^3) compared to 45 lb/ft^3 (720 kg/m^3) for the raw corn material. Moisture content also is decreased about 2–4%. For fattening cattle, roasted corn improves weight gain by about 11% and has a feed efficiency about 14% greater than that of unroasted corn.

Micronizing. Another method used in the feed industry involves heating grain to 300°F (150°C) with microwaves. This micronized grain contains about 7% moisture. It is then rolled to form a uniform, free-flowing product. Prior to feeding, water is usually added to adjust the moisture content to 10%. Cattle fed micronized grain sorghum (milo) compare favorably with those fed steam-flaked grain sorghum in rate of gain and feed efficiency. However, the cost of processing favors the micronizing technique over the steam-flaking method.

Popping. This method involves rapidly heating grain containing about 15–20% moisture to 300–310°F (150–155°C). At this temperature, the internal moisture of the grain is volatilized and explodes the kernel. The exploded product is then dry-rolled to form a uniform product and reduce storage space. All grains can be processed by this method; however, it appears to be particularly effective in processing milo. Feeding trials conducted in Texas revealed that cattle consumed less of the popped milo than the regular grain. Although this resulted in lower average daily gains, feed efficiency was increased 17%.

Explosion puffing. In another method, explosion puffing, live steam is injected into high-tensile-strength steel bottles which hold about 200 lb (90 kg) of grain. About 20 s after the pressure reaches 250 lb/in.^{2} (1.7 megapascals) a valve is opened. Rapid release of the pressurized moisture within the grain explosively puffs the kernels. In practice, tandem vessels are alternately filled and pressurized to achieve a continuous operation. The resulting product resembles puffed breakfast cereal. Cattle fed puffed milo have exhibited feed intake, gain, and feed efficiency similar to that of cattle fed flaked grain.

Protein supplements. Protein supplements are feeds which contain more than 20% protein or protein equivalent. They are primarily added to rations containing feed grains or forages which are low in protein. Addition of protein supplements depends upon the animal's requirement for protein and the amount and quality of grain and forage in the ration.

Plant sources. High-protein meals can be made from soybean, cottonseed, sunflower seed, safflower seed, rapeseed, peanut, linseed, and coconut (copra). These meals are by-products of the oilseed-crushing industry; they vary in feeding value (that is, protein quantity and quality) depending upon the amount of hull or seedcoat remaining in the meal, conditions of time and temperature during extraction, and constituent amino acids of the protein. For example, although soybean meal and sunflower meal contain about the same amount of protein (45.8% versus 46.8%, respectively), they differ in amino acid content. Compared to sunflower meal, soybean meal is high in lysine (6.6 versus 3.8 g/16 g nitrogen) and low in methionine (1.1 versus 2.1 g/16 g nitrogen). In the United States, soybean meal is the most widely used protein supplement.

Quality forage and other leafy plants contain considerable protein (12–30%). The amino acid content of this protein compares favorably to that of soybean meal, fishmeal, and other high-quality proteins. Considerable research has been conducted in the United States, England, and Pakistan to investigate the preparation of leaf protein concentrates from forages. In the United States, the Pro-Xan process was developed for concentrating protein from alfalfa. The final product constitutes 4.5% of the starting material and

contains 40–50% protein and 600–900 mg/kg xanthophyll. Feeding trials with chicks have indicated that the product is a good source of protein and xanthophyll. If the cost of processing can be lowered, leaf protein concentrate may contribute significantly as an alternate protein source in the production of animal protein—meat, milk, and eggs—for humans.

Animal sources. Protein supplements of animal origin primarily consist of inedible products derived from meat packing or rendering plants, marine sources, poultry processing plants, and dairies. In general, such products are high in protein and of excellent quality, being well balanced in amino acids and providing vitamins and minerals. Feather meal, a by-product of the poultry processing industry, has also been used as a protein source. Although this product contains about 85% protein, it is poorly digested and must be hydrolyzed for good utilization. Animal-derived protein supplements are subject to autoxidation and rancidity because of their high fat content, and are often a source of bacterial (particularly *Salmonella*) contamination of poultry and meat products used for human food. It is therefore important that proper sanitation be practiced during processing and handling of these protein supplements to prevent bacterial contamination of human food.

Single-cell sources. There has been considerable interest in the production of protein from single-cell organisms such as yeasts, bacteria, and algae. A wide variety of carbon sources can be used as substrate for the growth of these organisms, including petroleum, methane, alcohol, starch, molasses, cellulose, and paper pulp waste liquor. Potential yields are quite high. One thousand pounds (450 kg) of single-cell organisms might produce 50 tons (45 metric tons) of protein per day. Although production of single-cell protein as a feedstuff for livestock appears promising, problems relating to palatability, digestibility, nucleic acid content, toxins, protein quality, and economics must be solved before this becomes a reality.

Microbial synthesis. Cattle and other ruminants are able to derive a portion of their dietary protein from nonprotein nitrogen via microbial protein synthesis in the rumen. Amino acids, amides, ammonium salts, and other nonprotein nitrogen sources provide nitrogen from which symbiotic rumen bacteria synthesize microbial protein. This microbial protein is then digested by the ruminant animal in the abomasum and gastrointestinal tract. Providing that adequate energy and other nutrients are available, about one-fourth to one-third of the protein requirements of ruminants can be met by the feeding of nonprotein nitrogen. If this amount is exceeded and the animals are not managed properly, toxicity and death may result. Commercially available products include urea and the ammoniated products of molasses, citrus pulp, beet pulp, cottonseed meal, and rice hulls. Both solid (range cubes and pellets) and liquid (lick tanks) products containing urea, molasses, trace minerals, vitamin A, and so forth, are available. Slow-release urea products also are marketed. With concern about world food supply and competition between humans and animals for available protein, the feeding of nonprotein nitrogen to ruminants is increasingly important. *See* PROTEIN.

Vitamins. Vitamins are complex organic compounds which are required in exceedingly small amounts by one or more animal species. Most function as cofactors in enzyme systems involving the regulation of metabolism and transformation of energy. Certain vitamins (such as vitamin K) are apparently needed by only a few species; others (such as vitamin A and thiamine) are required by all animals. Some vitamins (such as vitamin A) must be present in the diet; others are synthesized in the tissues of the animals (such as vitamin C) or by bacteria in the digestive tract (such as B-complex vitamins). The absence in the diet of one or more of the required vitamins may prevent the animal from growing or reproducing, or may cause deficiency diseases. If the deficiency is severe, the animals may die. Farm animals fed diets containing high-quality green forage usually receive enough of most required vitamins. Because monogastric animals (poultry, swine) thrive when small amounts of leaf meal are added to their diets, alfalfa and other forages are usually described as containing unidentified growth factors.

High temperatures and oxidation destroy certain vitamins; therefore, care must be taken in the processing and storage of feedstuffs to protect the vitamins they contain. Certain feedstuffs also may contain antivitamins, which prevent the proper assimilation and functioning of vitamins. To ensure that rations contain sufficient vitamins to meet dietary requirements, manufacturers often fortify feed formulations with one or more vitamins. *See* VITAMIN.

Minerals. Animals, like plants, require minerals to grow, and if deficiencies occur, livestock become unthrifty, lose weight, or exhibit other deficiency symptoms. Minerals serve many vital functions. The skeletons of vertebrates are composed chiefly of calcium and phosphorus. Not only must livestock receive adequate calcium and phosphorus, but they must also receive these minerals in the proper proportion, because a great excess of one or the other may be detrimental. Also, if the proper calcium-phosphorus ratio is provided, less vitamin D is needed. Minerals also are necessary constituents of soft tissues and body fluids, and they function in maintaining the proper acid-base balance of body tissues. In addition to macroelements (for instance, calcium, phosphorus, sodium, chlorine, and iron), trace elements (zinc, cobalt, manganese, and selenium, among others) are required for proper nutrition. Trace elements primarily function as cofactors in metabolic reactions. Feedstuffs produced in particular areas may be deficient in or have an excess of certain minerals. For example, in the Great Lakes and Northwest regions of the United States, soils are deficient in iodine. Also, soils containing more than 0.5 ppm (part per million) of selenium are potentially dangerous for production of livestock feed. The form and availability of certain minerals also are an important consideration in formulating livestock diets. Phytin phosphorus,

for example, is not a good source of phosphorus for poultry or swine; however, it appears to be satisfactory for ruminants. To ensure an adequate supply of minerals, feed manufacturers often add mineral supplements. Farm animals may also be permitted free access to a mineral mixture. If this is allowed, care should be taken, particularly with trace minerals, to see that the animals do not consume too much, or toxicity may result. Like vitamins, use of mineral supplements adds to production costs. Therefore, the need for such supplements should be given careful consideration.

By-products. This group of feeds includes a number of by-products formed in the processing of livestock, plants, and other materials. To be acceptable as livestock feeds, such products should be palatable, provide digestible nutrients or roughage, and present no health hazard. Use of by-products as feed not only lowers the cost of animal production but helps to reduce the competition between livestock and humans for available grain and protein supplies. Another advantage is that industry derives income from products that otherwise might be wasted and disposed of at a cost. By-product feeds include inedible products and residues from animal, poultry, and marine processing plants, fruit and vegetable processing plants (citrus pulp, tomato pomace), the brewing and distilling industry (brewers' and distillers' grains), dairy processing plants (whey, cheesemeal), wood and paper mills (wood molasses, treated wood scraps), as well as miscellaneous fibrous materials such as cottonseed gin trash, cottonseed, soybean, and peanut hulls. Research has been conducted on the feeding of animal wastes (manure and litter) to livestock. There is also interest in the treatment of crop residues (cornstalks, straw, and so forth) with alkali (sodium hydroxide) to improve their digestibility for ruminants. Although certain by-products (such as peanut hulls) are available only in a limited area, others (such as brewers' grains) are more widely available.

Feed additives and implants. These may be defined as nonnutritive substances which enhance the utilization of a feed or productive performance of the animal. It is estimated that in the United States 75% of the finishing cattle, finishing lambs, and growing-finishing pigs receive feed additives or implants. Use of these products and guidelines for administration and withdrawal of approved products are regulated and prescribed by the U.S. Food and Drug Administration (FDA). Also, feed-control officials in individual states may regulate use of these substances for livestock marketed intrastate. Prior to receiving approval for marketing a product, the manufacturer must prove its safety and efficacy through a rigorous testing program. Since the early 1950s, antibiotics have been used to stimulate growth and promote feed efficiency in ruminants, swine, and poultry. However, there is concern that continuous feeding of antibiotics to livestock may increase the resistance of certain strains of bacteria and that such resistance may be transferred to bacteria associated with human diseases. It is therefore expected that the use of such antibiotics as penicillin and the tetracyclines in feeds may come under increased scrutiny.

In 1954 the hormone diethylstilbesterol (DES) was approved for use in cattle finishing rations. Two years later, implants for steers were approved. In 1973 the use of DES was banned when it was proved to be carcinogenic and present in trace amounts in the meat of animals. However, in 1974 a Federal Court of Appeals overruled this ban. On November 1, 1979, the 1974 decision was reversed, and DES may no longer be used as a feed additive or implant for cattle and sheep.

Another hormone, melengestrol acetate (MGA), is marketed for feeding to heifers being finished for slaughter. This compound, a synthetic progestin, affects metabolic rate and suppresses estrus, thereby increasing feed efficiency of feedlot heifers.

Other hormones or hormonelike compounds which are marketed for implantation in beef cattle to improve growth and feed efficiency are Synovex and Ralgro. Synovex-S contains progesterone and estradiol benzoate, and is used for steers; Synovex-H contains testosterone and estradiol benzoate, and is used for heifers. Although both hormones are FDA-approved, they are being evaluated to determine the advisability of withdrawing approval of these products in consideration of human safety. Ralgro is the trade name for zeranol, a substance produced by the mold *Giberella zeae*. Although not a hormone itself, this compound is thought to influence the production and release of certain hormones in the body. It is approved for implantation in beef cattle.

In fermenting feeds, rumen microorganisms produce carbon dioxide and methane in addition to fatty acids. Methane accounts for about 10% of the total energy of feedstuffs; hence its production reduces the energy efficiency of feeds. Therefore, compounds which may inhibit methane production in the rumen are being investigated. Dichlorovinyldimethyl phosphate, hemiacetal derivatives of chloral and starch, and halogenated analogs of methane have shown promise.

A compound which has proved effective in altering rumen metabolism is monensin, a fermentation product of *Streptomyces cinnamonensis*. The proprietary product, Rumensin, has been shown to significantly improve feed efficiency in feedlot cattle. This compound acts by altering the proportion of fatty acids produced in the rumen; more propionic acid (and less methane) is produced than acetic and butyric acids. Because conversion of feed to propionic acid is energetically more efficient than its conversion to the other acids, overall production efficiency increases. *See* NUTRITION.

International feed nomenclature. Because feed names have evolved through usage, the name of a particular feedstuff has not always been descriptive enough to fully identify and characterize it. For example, the term Linseed meal does not convey whether the product is obtained by mechanical or solvent extraction or whether it contains 33%, 35%, or 37% crude protein. To overcome this deficiency, a

system was developed that makes it possible to identify the contents and other characteristics of a feed from its name. This system, known as the National Research Council (NRC) Feed Nomenclature System, or International Feed Nomenclature System, gives to each feed a specific name that may consist of up to nine terms: scientific name, origin (or parent material), species (variety or kind), part actually eaten, process(es) and treatment(s) to which the parent material or the part eaten was subject before being fed to the animal, stage of maturity (applicable only to forages), cutting or crop, grade (quality designations and guarantees), and classification (according to nutritional characteristics). In addition, each feed is grouped into one of eight feed classes, each of which is designated by a number in parenthesis: (1) dry forage or dry roughage; (2) pasture, range plant, and forages fed green; (3) silages; (4) energy feeds; (5) protein supplements; (6) minerals; (7) vitamins; (8) additives; then each feed is given an international reference number. The classification number, for example, (1), is the last term in the international feed name and the first digit of the feed reference number. An example of an international feed name for dehydrated alfalfa pellets is: alfalfa, aerial part, dehydrated ground pelleted, early bloom, cut 2, (1). The international reference number for this feed would be 1-07-733. This six-digit reference number is the numerical name of the particular feedstuff and may be used when formulating rations using linear programming and computers. Only the first digit is related to a specific feed characteristic. The remaining digits are used only to number the feed. This international system of naming feeds is being widely adopted for use in scientific writing and journal articles. International reference numbers particularly are being used to identify feeds in tables and computer formulations.

Donald Burdick

Pet Foods

Pet foods are usually nutritionally complete products which adequately support growth and life, even if fed exclusively.

Pet foods are made from proper mixtures of meat, fish, fowl, bone, cereal, milk, minerals, and vitamins. The minerals and vitamins are added in quantities necessary to augment the natural foodstuffs. They may be prepared in various finished forms, canned, frozen, or dried.

The canned products are sterilized at high temperatures to permit storage without refrigeration. The frozen variety often is limited to raw meat or seafood. Dry items are stable at room temperature, though usually not as long as canned items.

Those pet foods which include meat contain a fairly large proportion of highly nutritious organ meats as well as other meat products.

The manufacture of the canned variety is fairly standardized in that the grinding, blending, filling, and sterilizing procedures are well known. The dry variety, particularly dry dog food, can be manufactured in a number of different ways, ranging from pelletizing of dry ingredients to hot-air and vacuum drying of moistened, expanded pastes and extrusions. Little or no preparation is needed for the frozen products since these are generally raw.

A significant innovation is intermediate-moisture pet food that keeps without canning or refrigeration. Careful control of amount of "active water" in the product avoids spoilage.

Aseptic canning of certain pet foods appears entirely possible. In this procedure the product is sterilized outside the can and then filled into a sterilized container, whereupon it requires no further cooking and is stable at room temperature for indefinite periods.

Clarence K. Wiesman

Bibliography. P. R. Cheeke, *Applied Animal Nutrition*: *Feeds and Feeding*, 2d ed., 1998; D. C. Church, *Livestock Feeds and Feeding*, 4th ed., 1997; A. Cullison and R. S. Lowrey, *Feeds and Feeding*, 4th ed., 1986; R. M. Ensminger, *The Stockman's Handbook*, 7th ed., 1992; M. E. Ensminger, J. E. Oldfield, and W. W. Heinemann, *Feeds and Nutrition*, 2d ed., 1998; D. Natz (ed.), *Feed Additive Compendium*, 1977; W. G. Pond et al., *Basic Animal Nutrition and Feeding*, 1995.

Animal growth

The increase in size or weight of an animal. A number of processes are involved in growth. These include hyperplasia, or increase in cell number due to cell proliferation (cell division or mitosis) and recruitment from stem cell populations; hypertrophy, or increase in the size of cells; and differentiation (precursor cells achieving mature functioning—often this process cannot be reversed). Schematic diagrams showing muscle and adipose development are included below. *See* ANIMAL MORPHOGENESIS; CELL DIFFERENTIATION; CELL DIVISION; CELL SENESCENCE AND DEATH; EMBRYONIC DIFFERENTIATION; MITOSIS.

Measurement of growth. The indices of growth include height (as for humans and horses), body weight (agricultural animals), dry weight, length (rodents, lizards, fish), total body DNA or nitrogen retention (used in agricultural animals as an indicator of muscle growth), or rates of whole-body protein synthesis and/or accretion:

$$\text{Protein accretion (net protein synthesis)} = \text{Protein synthesis} - \text{Protein degradation}$$

Mathematical models for growth. Growth appears to follow an S or sigmoidal curve. There are multiple equations that mathematically describe growth. These include the logistic model and the Gompertz equation

$$W = \int A \cdot e^{-b \cdot e - kt}$$

$$dW/dt = k \cdot W \cdot (A/W)$$

The Gompertz equation describes growth of an animal within a specific species very well from embryonic stage to mature adult size. Here W is the weight at any given age (time t following the beginning of

development), A is the maximal weight (mature or adult weight), k is the growth rate or constant, and b is a second constant.

At significant periods of growth, a linear function can easily be applied for growth.

Telomeres. Telomeres are multiple TTAGGG tandem repeats at the end of chromosomes in at least higher vertebrates. The telomeres stablize the chromosomal ends and protect the genes in the subtelomeric regions. Telomeres shorten with progressive cell cycles (either in the animal or in tissue culture), leading to the view that shortened telomeres are related to senescence or aging. Addition of telomeric repeats can be accomplished by the enzyme telomerase, an RNA-dependent DNA polymerase. High levels of telomerase activity are observed in tissues with a high rate of cell proliferation, for example, malignant tissues and fetal tissues. Given the number of cell divisions, it is not surprising that cloned animals resulting from nuclear transfer from somatic cells have reduced telemetric length, with corresponding effects on growth rate. *See* CHROMOSOME.

Embryonic and fetal development. Development begins following fertilization of the ovum (egg) by the spermatozoon. There is cell division to form first a two-cell stage, then four-cell, then eight-cell stage. Cell divisions continue but become desynchronized.

Prior to implantation, in mammals, the part of the conceptus that will develop into the animal (as opposed to the membranes and placenta) is called the embryo. At the end of embryonic development, the major organs are normally formed and "true" bone formation begins (replacing cartilage). The fetus develops in the uterus until birth. Considerable growth and development occurs with organs achieving the capability of full functioning.

In mammals, the placenta plays an important role influencing fetal growth. The availability of critical nutrients and oxygen to the growing fetus is related to placental size and functioning. In addition, the placenta can produce hormones that influence fetal growth. Other critically important hormones affecting fetal growth are insulin and insulinlike growth factor-II (IGF-II). In birds, the yolky egg provides the nutrients for the developing embryo. In amphibians, there is a free-living larval (for example, "tadpole") stage that metamorphoses into the adult form as stimulated by the hormone thyroxine. *See* EGG (FOWL); HORMONE; PLACENTATION.

Birth and postnatal growth and development. At birth, all animals have their organs formed. However, animals can be born (mammals) or hatched (birds) either with their systems well functioning (precocial development) or with the systems not completely functional (altricial development). For instance, horses can walk very soon after birth, where as cats and rats have offspring that are blind at birth and rats have hairless offspring.

Following birth, the vast majority of growth (increase in weight) occurs. There is not complete synchrony of growth with different organ systems. First, the rate of skeletal growth is maximal, then comes maximal rate of muscular growth, followed by maximal rate of adipose growth. Growth in most mammals ceases shortly after puberty (as discussed below). At puberty, there is an increase in sex steroid hormones in the circulation. The hormones induce a "growth spurt" in many species, including humans. Another result is the closure of the growth plate and the cessation of growth (discussed below).

Muscle growth. Muscle development and growth involves the massive cell division (proliferation) of the myogenic precursor cells. These cells migrate to specific sites in the body, for instance, the developing limbs. They are attracted to the sites by gradients of growth factors and are undergoing some differentiation. The precursor cells then form myoblast cells, which are the last stage of cells that can undergo division. The myoblasts differentiate into myocytes. Multiple myocytes then aggregate to form multinucleated myotubes, which start to produce the contractile proteins and then fibers of muscle. Postnatally, muscle growth is accomplished by increasing the size of the muscle fibers. More nuclei are added by the recruitment of satellite cells. Muscle growth and development may be summarized schematically:

Mesodermal somites
↓
Myogenic precursor cells (→ Satellite cells)
↓
Proliferating, differentiating, and migrating
(embryonic and fetal life)
↓
Myoblast cells
↓
Proliferating and differentiating myocyte cells
(capable of synthesizing many muscle proteins)
↓ Differentiating
Myotubes
(multinucleated cells that are assembling
muscle contractile fibers)
↓
Myofibers (multinucleated)
↓
Muscles growing in size due to hypertrophy
as well as recruitment of satellite cells

See MUSCLE.

Bone and cartilage growth. Longitudinal growth or lengthening of the long bones of the skeleton occurs at the epiphyseal or growth plate. Growth of this cartilage (including proliferation of cartilage cells or chondrocytes) occurs under the influence of the hormone insulinlike growth factor (IGF-I) (discussed below). At puberty or sexual maturation in most higher vertebrates, there is closing or ossification of epiphyseal plates, and longitudinal growth of bones ceases. This occurs under the influence of sex steroid hormones. *See* BONE; CARTILAGE.

Adipose tissue growth. Adipose tissue provides energy storage as triglyceride. This is critically important to animals at a time of nutritional deprivation or during hibernation or when energy consumption is not meeting their needs (for example, during pregnancy and lactation). Unlike other organs, adipose

tissue can continue to increase in size after sexual maturation. *See* ADIPOSE TISSUE; LIPID; TRIGLYCERIDE.

Adipose tissue grows by proliferation and migration of adipose precursor cells. These differentiate into preadipocytes, which are the last cells in the adiocyte lineage capable of division. They may divide multiple times to form adipocytes. Adipocytes do not divide but produce the key enzymes necessary for triglyceride or fat formation and they fill with triglyceride. Adipose tissue growth and development may be summarized schematically:

Adipose precursor cells
↓
Proliferating, differentiating, and migrating
(embryonic and fetal life)
↓
Preadipocyte cells
↓
Proliferating and differentiating adipose cells
(capable of synthesizing adipose proteins
and triglyceride/fat)
↓
Mature fat-filled adipose cells
(growing by increasing in size)

Growth hormone. Multiple hormones influence growth. A prime example is growth hormone (GH), produced by specific cells, the somatotropes, found in the anterior pituitary gland. The development of the somatotropes is induced by pit-1. Release of GH is controlled by peptides released from specialized nerve terminals in the hypothalamus. These peptides include GH releasing hormone (GHRH) which stimulates GH release and somatostatin (SRIF) which inhibits GH release. Growth hormone acts by binding to GH receptors (GHR) on the cell membrane of cells in the liver and some other tissues. In turn, GH increases the release of IGF-I. The IGF-I binds reversibly to IGF-binding proteins (IGF-BPs) in the circulation. Unbound IGF-I stimulates skeletal growth:

Hypothalamus (in brain)
↓ ↓
$GHRH^{+}SRIF^{-}$
↓ ↓
Anterior pituitary gland (somatotropes)
↓
Growth hormone (GH)
↓
Liver and other tissues (GHR)
↓ ↓
IGF-I IGF-BPs
↓ ↓
Growing tissues, such as cartilage

Lack of sufficient production of GH and/or IGF-I is associated with dwarfism in, for example, humans, cattle, chickens, dogs, and mice. Excess GH production causes giantism and, in adult animals, acromegaly.

Other hormones influencing growth. Normal levels of thyroid hormones are required for growth. Glucocorticoid hormones, produced by the adrenal cortex [cortisol in humans and domestic mammals, and corticosterone in rodents and poultry and other birds], inhibit growth.

Manipulation of growth rate. Some chemical agents (related to hormones) are used to stimulate growth in animals. These include the use of GH in children with abnormally low growth rates (and likely low adult heights) due to hypopituitary dwarfism or low height for age. Muscle growth is stimulated in many but not all higher vertebrates by androgens. In commercial cattle, growth can be enhanced by the administration of estrogens, alone or with androgens. This treatment is approved for commercial use in some countries. There are beta-adrenergic agonists (for instance, Ractopamine) that enhance growth, particularly muscle growth. This treatment is used commercially in pigs in some countries.

Factors influencing growth rate. Growth rates within a species are influenced considerably by the environment and by the animal's genetics.

Growth may be reduced by lack of availability of nutrients, the presence of toxicants in the environment, environmental stressors (for example, high or low temperatures), and disease. Diets deficient in critical nutrients, for example, energy components, proteins (and even specific amino acids such as lysine, methionine, or tryptophan), vitamins, and minerals, will result in reduced growth rates. When full nutrition is restored, compensatory or catch-up growth occurs. Not only is the growth rate restored but it rebounds to a higher level, above that normally seen. One mechanism by which stressors reduce growth is via increased release of the glucocorticoid hormones, cortisol and corticosterone. Other mechanisms of growth reduction include suppression of IGF-I levels, for instance, by disease; and toxicants that influence the thyroid gland (goitrogens). *See* NUTRITION; PROTEIN.

Genetics. Genetics profoundly influences growth rate. There is marked sexual dimorphism in growth rates, with males of most species growing faster and larger than females. Different breeds of animals have been produced with markedly different growth rates and mature body sizes. Perhaps the best example is seen in the widely different breeds of dogs and even the disparate-sized poodles such as miniature and toy poodles. Selective breeding can greatly influence growth rate. An example is seen in chickens (meat or broiler type), where since 1945 there has been increased growth rates by more than fivefold with consequent improvements in efficiency and reduced cost to the consumer. More recently, geneticists working for the livestock industry have employed quantitative trait loci (QTL) following genome scans or specific point mutations (for instance, single-nucleotide polymorphorisms, or SNPs) in candidate genes to identify superior parents or grandparents, and then the geneticists have used these in select breeding programs. *See* BREEDING (ANIMAL); GENETICS.

Gene imprinting. Fetal growth rates are influenced by imprinting of genes. This is the result of methylation of DNA. The pattern of imprinting varies with the parent of origin of the genes. An example

is the imprinting of the IGF-II gene in eutherian mammals. There are distinct differences in imprinting sites, depending on maternal or paternal origin of the gene. There is also imprinting of the IGF-II receptor gene (responsible for degradation of IGF-II); this is found in most mammals but not in the primate lineage.

Colin G. Scanes

Bibliography. S. Brody, *Bioenergetics and Growth*, Waverley Press, Baltimore, 1945; M. F. Rothschild, Porcine genomics delivers new tools and results: This little piggy did more than just go to market, *Genet. Res.*, 83:1–6, 2004; C. G. Scanes, *Biology of Growth of Domestic Animals*, Iowa State Press/Blackwell, 2003.

Animal kingdom

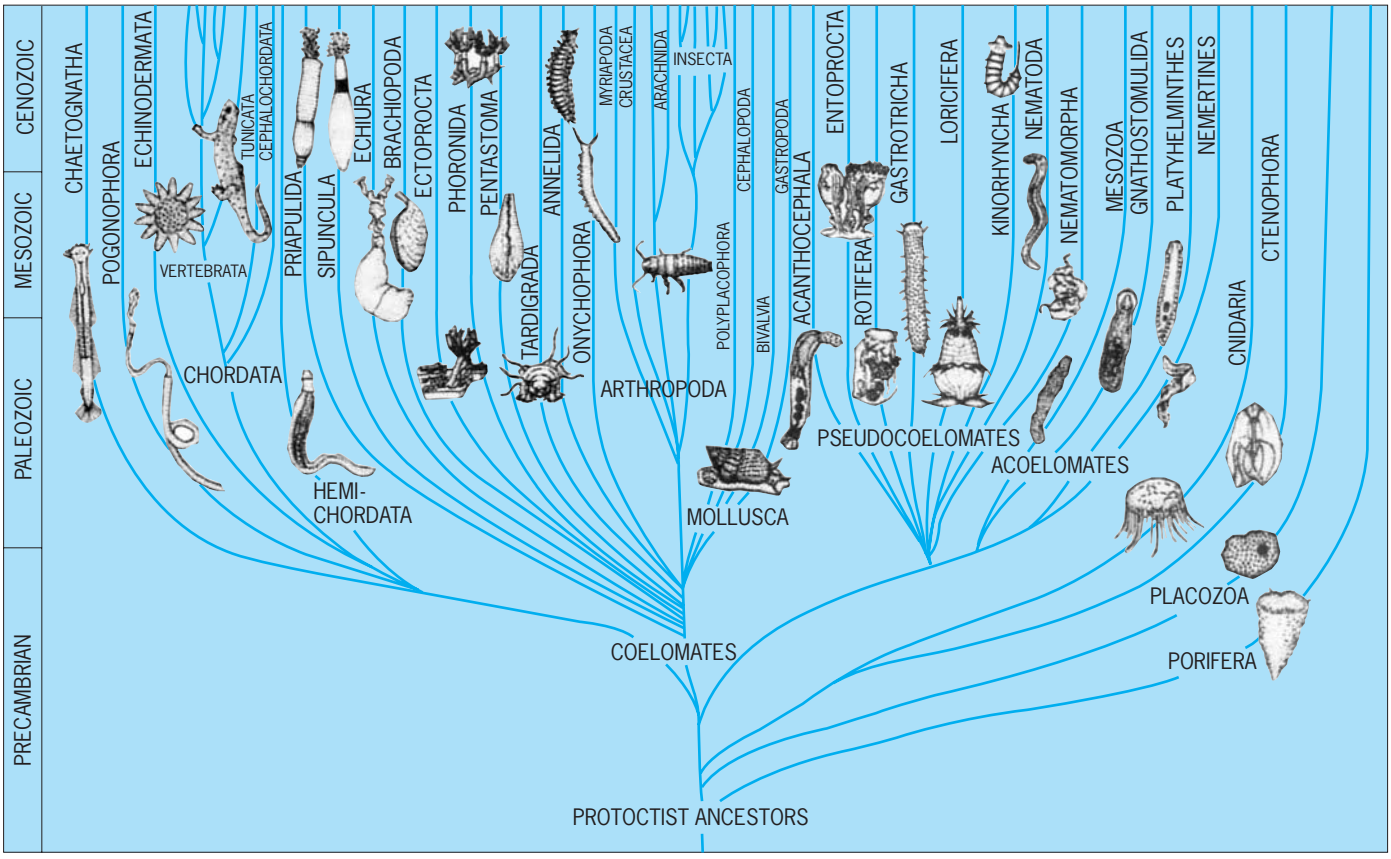
One of five kingdoms of organisms: Animalia, Plantae, Fungi, Protoctista (or, in part, in some classifications Protozoa), and Procaryotae (bacteria). Animals are eukaryotic multicellular organisms that take food into their bodies and that develop from blastula embryos. More than 30 million species of organisms have been described in all five kingdoms; the animal kingdom includes over 1 million described species. Animal species are organized into phyla (see **illus.**) that are defined according to comparative patterns of development, body structures, behavior, biochemical pathways, modes of nutrition, and ancestry. *See* ANIMAL SYSTEMATICS.

Form and habitat. Animals take diverse forms. The largest animals are great blue whales, which grow up to 100 ft (30 m) in length; the smallest are marine loriciferans, measuring as little as 300 micrometers. Thus, the largest animals are about 100,000 times as large as the smallest.

Some one-celled organisms are classified as members of the Protoctista kingdom. Most protoctists are microorganisms, but some are larger than animals. For example, rotifers, nematodes, and loriciferans are smaller than protoctists such as ciliates.

Most animal phyla are aquatic; that is, they are found in shallow seas, ponds, puddles, and streams. Some members of two phyla are called terrestrial because they live on the land surface. These phyla are Chordata, including humans and other mammals, reptiles, amphibians, fish, and birds; and Arthropoda, including moths, beetles, and millipedes. Animals that fly are referred to as aerial; insects (phylum Arthropoda), birds, bats, and flying dinosaurs are included in this category. Some animals inhabit a moist, underground environment but are not strictly terrestrial; annelid worms, nematodes, and other soil-dwelling animals are adapted to life in the thin film of water surrounding soil particles.

Diversity. There are approximately 1.2 million animal species, but new species and phyla are



Animal phylogeny or family tree, diagramming the relationships among phyla of living animals. The ancestors of all present-day animals probably were soft-bodied marine forms, such as the placozoa. (*After L. Margulis and K. V. Schwartz, "The Animal Kingdom" Teachers' Guide, Wards Natural Science, Rochester, New York, 1987*)

constantly being identified. Of the known animal species, some 75% are arthropods. One phylum, Vestimentifera, includes giant tube worms that measure 5 ft (1.5 m) long. These animals of the abyss live in hot, submarine gardens some 1.5 mi (2.5 km) below the ocean surface. Another phylum, Loricifera, includes spiny marine animals that live in offshore sediments. Much remains to be learned about loriciferan food and development because these animals stick tenaciously to shelly gravel. Unfortunately, animals along with their habitats are being obliterated thousands of times faster than any of the earlier mass extinctions recorded in the geological record. Biologists estimate that about 99.9% of all animal species that ever lived are now extinct. *See* EXTINCTION (BIOLOGY).

Nutrition and development. Unlike plants, seaweeds, and some bacteria which photosynthesize (autotrophs), animals are fed by other sources (that is, they are heterotrophs). Animals obtain food by ingestive or absorptive nutrition. Most animals take food into their bodies (ingestive), but some aquatic animals absorb dissolved nutrients through their cell membranes either from symbiotic partners or from the surrounding water. The energy source is organic compounds.

Animals may reproduce sexually or asexually or by both methods. Sexual reproduction involves two different reproductive cells, a female egg cell and a smaller male sperm cell. At fertilization these two cells join and produce a zygote that is destined to become a new organism. The animal zygote divides to form a blastula embryo that is a hollow, liquid-filled ball of cells. Blastulas are found only in the animal kingdom. Animal species also vary broadly in their mode of embryonic development. Embryos of some species (human, for example) develop directly into smaller versions of the adult, but in other species the embryos develop first into a larva, an immature stage preceding adulthood. For example, frog embryos develop into a tadpole, which is a larval stage, and then the tadpole dramatically reorganizes through the process of metamorphosis into the adult form. Adult animals produce sex cells that carry on the continuous life cycle of the species. Different patterns of development are one line of evidence utilized to deduce ancestry of a phylum.

In contrast to sexual reproduction, asexual reproduction is by a single individual without fertilization of egg by sperm. New individuals are produced asexually by various processes, such as budding by sponges, detaching of body parts by annelids, parthenogenesis by rotifers, and pseudogamy by ticks. Asexual reproduction occurs in about half of the animal phyla. *See* BLASTULATION; EMBRYOGENESIS; REPRODUCTION (ANIMAL).

Multicellularity. Animals are multicellular, but since all five kingdoms include certain multicellular species multicellularity alone does not define an animal. Animals, like plants, fungi, and protoctists, but unlike bacteria, are also eukaryotic. Eukaryotes have membrane-bound nuclei; more than one chromosome; and organelles (that is, "little organs" within a cell, such as mitochondria and a nucleus). In contrast to plant cells, animal cells lack cell walls (which contain cellulose). Moreover, animal cells intercommunicate and support one another by means of elaborate junctions. In addition, the cells of animals form tissues which specialize in a wide array of functions; examples of such tissues are muscle, skin, and blood. The only exception is animals of phyla Porifera (sponges) and Placozoa, which have poorly defined tissues. *See* CELL (BIOLOGY); EUKARYOTAE; PLANT CELL; TISSUE.

Invertebrates and vertebrates. Traditionally, animals have been grouped into invertebrates (without backbones) and vertebrates (with backbones). Vertebrates include mammals, amphibians, reptiles, birds, fish, all members of the phylum Chordata. However, a few chordates, such as tunicates, are invertebrates. Members of all other animal phyla, more than 98% of all animal species, are invertebrates. Although invertebrates lack backbones, they achieve physical support by structures ranging from delicate glass spicules, to tough rings and rods, to hydrostatic pressure. The number of animal species within a single invertebrate phylum varies considerably: Placozoa comprises a single species, but the great phylum Arthropoda comprises more than 800,000 species. If tropical species were better described, the arthropods might include as many as 10 million living insect species. *See* AMPHIBIA; AVES; CHONDRICHTHYES; CHORDATA; MAMMALIA; OSTEICHTHYES; REPTILIA.

Animal ancestry. Microscopic evidence, fossils, geographic distribution, function (physiology), behavior, embryonic patterns, and form (morphology) are lines of evidence from which contemporary classifications are drawn, both within the animal kingdom and in linking it to other kingdoms. The common ancestors of living animals probably evolved over 700 million years ago. Most biologists agree that animals evolved from choanomastigotes, which are microscopic protozoa ancestral to the sponges. Examination of ribosomal ribonucleic acid sequences indicate that animals and fungi share an evolutionary history: their common ancestor may well be a single-celled organism similar to choanomastigotes. *See* ANIMAL EVOLUTION; PHYLOGENY; PLANT KINGDOM.

Karlene V. Schwartz

Bibliography. R. M. Alexander, *The Chordates*, 1975; M. F. Glaessner, *The Dawn of Animal Life*, 1984; E. D. Hanson, *Origin and Early Evolution of Animals*, 1977; R. Lewin, *Thread of Life*, 1991; L. Margulis and K. V. Schwartz, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*, 3d ed., 1998; S. P. Parker, *Synopsis and Classification of Living Organisms*, 2 vols., 1982; E. E. Ruppert and R. D. Barnes, *Invertebrate Zoology*, 6th ed., 1994.

Animal reproduction

The formation of new individuals, which may occur by asexual or sexual methods. In the asexual methods, which occur mainly among the lower animals, the offspring are derived from a single individual.

Sexual methods are general throughout the animal kingdom, with offspring ordinarily derived from the paired union of special cells, the gametes, from two individuals. Basic to all processes of reproduction is the origin of the new individual from one or more living cells of the parent or parents. There has been no acceptable demonstration of the origin of a new living organism from other material than preexisting living cells. *See* PREBIOTIC ORGANIC SYNTHESIS.

Asexual reproduction. Asexual processes of reproduction include binary fission, multiple fission, fragmentation, budding, and polyembryony. Among the protozoans and lower metazoans, these are common methods of reproduction. However, the last-mentioned process can occur in mammals, including humans.

Binary fission involves an equal, or nearly equal, longitudinal or transverse splitting of the body of the parent into two parts, each of which grows to parental size and form. This method of reproduction occurs regularly among protozoans, in which it is essentially the process of cell division, with complete separation of the daughter cells. To a limited extent, binary fission may be observed among metazoans, such as sea anemones, as longitudinal fission and among planarians as transverse fission.

Multiple fission, schizogony, or sporulation produces several new individuals from a single parent. It is common among the Sporozoa, such as the malarial parasite, which form cystlike structures containing many cells, each of which gives rise to a new individual. The cyst cells arise from a series of divisions of the nucleus, followed by cytoplasmic divisions of the original cell. *See* SPOROZOA.

Fragmentation is a form of fission occurring in some metazoans, especially the Platyhelminthes, or flatworms; the Nemertinea, or ribbon worms; and the Annelida, or segmented worms. The parent worm breaks up into a number of parts, each of which regenerates missing structures to form a whole organism. It occurs also in certain starfish, as *Linckia*, in which single arms may pinch off and regenerate a complete animal.

Budding is a form of asexual reproduction in which the new individual arises from a relatively small mass of cells that initially forms a growth or bud on the parental body. The bud may assume parental form either before separation from the body of the parent as in external budding, or afterward, as in internal budding. External budding is common among sponges, coelenterates (**Fig. 1**), bryozoans, flatworms, and tunicates. Among certain of the coelenterates, such as the colonial hydroid *Obelia*, buds give rise to medusae, or jellyfish, rather than to the parental-type polyp. The medusae represent the sexual generation. They are free-swimming and of separate sexes, producing eggs and sperm, respectively. Upon fertilization of the eggs the asexual, polyp-type individual develops. Another example of asexual individuals budding sexual individuals is found in the cestodes, or tapeworms. Here, the head or scolex, by which the animal is attached to the host tissue, produces a series of segments, termed proglottids, each of which is a sexual individual. This phenomenon is also known as strobilization. In some species of sponges, coelenterates, bryozoans, and tunicates, budding may occur without separation of the buds and thus lead to the formation of an organic colony, accompanied in some cases by specialization of parts for particular functions, such as in the Portuguese man-of-war, *Physalia*.

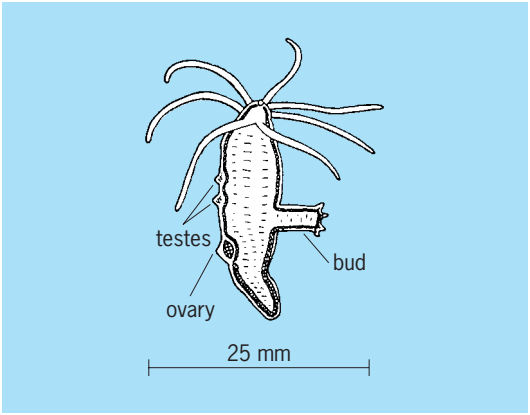


Fig. 1. Budding in *Hydra*.

Internal budding occurs among fresh-water sponges and bryozoans. In the sponges the internal buds, termed gemmules, consist of groups of primitive cells surrounded by a dense capsule formed by the body wall. If the parent animal dies as a result of desiccation or low temperature, the cells of the gemmules can later be released and form new sponges. In the bryozoans the similarly functioning buds are known as statoblasts.

Polyembryony is a form of asexual reproduction, occurring at an early developmental stage of a sexually produced embryo, in which two or more offspring are derived from a single egg. Examples are found scattered throughout the animal kingdom, including humans; in humans it is represented by identical twins, triplets, or quadruplets. In some flatworms, polyembryony is illustrated by the rediae, each of which in turn produces many young tadpole-like cercariae. A striking example of polyembryony is found among the insects in the hymenopteran *Litomastix*, which is a parasite on the egg of the moth *Plusia*. The embryo of this wasp subdivides so extensively that about 1500 individuals are formed.

In mammals, polyembryony is a regular feature of the development of armadillos. The four offspring produced at a single birth in the nine-banded armadillo (*Dasypus novemcinctus*) are identical quadruplets. The quadrupling process occurs during the late blastocyst stage of development. In humans, twins occur in about 1.1% of births, triplets in 0.012%, and quadruplets in 0.00014%. About one-third of the twins are identical and have thus arisen by polyembryony, probably occurring in a late blastocyst stage at the onset of gastrulation.

Sexual reproduction. Sexual reproduction in animals assumes various forms which may be classified under conjugation, autogamy, fertilization

(syngamy), and parthenogenesis. Basically, the various processes all involve the occurrence of certain special nuclear changes, termed meiotic divisions, preliminary to the production of the new individual. *See* GAMETOGENESIS; MEIOSIS.

Conjugation occurs principally among the ciliate protozoans, such as *Paramecium*, and involves a temporary union of two individuals during which each is "fertilized" by a micronucleus from the other. In this process the macronucleus of each conjugant breaks down, and the micronucleus undergoes two meiotic divisions to form four nuclei, of which three degenerate. The fourth nucleus divides again, and each conjugant transfers one of these micronuclei to its partner, where it fuses with the stationary micronucleus. The conjugants then separate and the fusion-micronucleus in each divides three times. Of the eight nuclei, four form macronuclei and three degenerate. The exconjugants then divide twice, along with mitotic division of the micronuclei, so that each of the four cells obtains one macro- and one micronucleus. The parental state is thus restored, and further reproduction occurs by simple fission. The ability to conjugate again is not attained until after a large number of divisions. Conjugation does not ordinarily occur within clones, which are organisms derived by mitotic division from a single individual. Strains that are capable of conjugating with one another are designated mating types, rather than males or females, because there are no correlated morphological distinctions between them, and moreover a particular species may have several mating types that are interfertile and thus represent more than two sexes.

In autogamy the nuclear changes described for conjugation take place, but since there is no mating, there is no transfer of micronuclei. Instead, the prospective migratory micronucleus reunites with the stationary one. The process may be considered related to parthenogenesis.

Fertilization, or syngamy, comprises a series of events in which two cells, the gametes, fuse and their nuclei, which had previously undergone meiotic divisions, fuse. In metazoans, the gametes are of two morphologically distinct types: spermatozoa, or microgametes, and eggs, also called ova or macrogametes. These types are produced by male and female animals, respectively, but in some cases both may be produced by a single, hermaphroditic individual. The nucleus of the spermatozoon has half the number of chromosomes characteristic of the ordinary (somatic) cells of the animal. The nucleus of the ripe egg in some animals, for instance, coelenterates and echinoderms, also has attained this haploid condition, but in most species of animals it is at an early stage of the meiotic divisions when ready for fertilization. In the latter situation, the meiotic divisions of the egg, characterized by formation of small, nonfunctional cells termed polar bodies, are completed after the sperm enters, whereupon the haploid egg nucleus fuses with the haploid sperm nucleus. Fertilization thus produces a zygote with the diploid chromosome number typical of the somatic cells of the species (23 pairs in humans) and this is maintained during the ensuing cell divisions (**Fig. 2**). *See* FERTILIZATION.

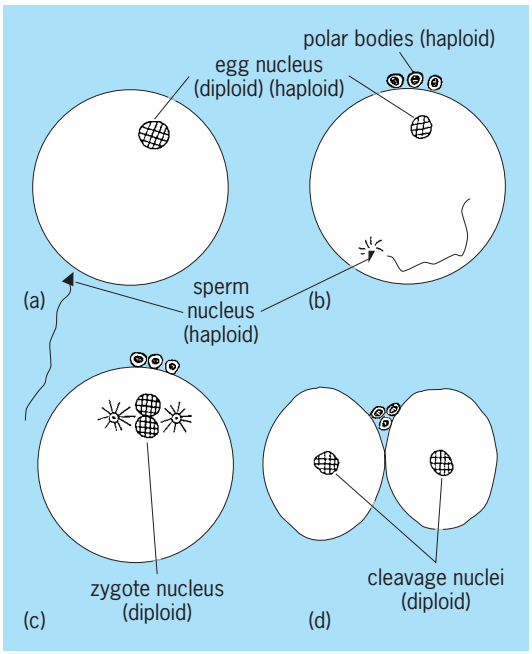


Fig. 2. Stages (*a–d*) of fertilization in animals. In different species the sperm enters at one or another of the stages between *a* and *b*, during which the two meiotic divisions of the egg occur.

In many protozoans, such as the flagellate *Polytoma*, the fusing gametes are not visibly different from one another nor from the ordinary organism. They are termed isogametes. In others they may differ (anisogametes), usually in size as in *Chlamydomonas braunii*; in some cases (*C. coccifera*), one gamete is motile and the other not.

Some metazoans are able to produce functional gametes while still in a larval condition. Reproduction of this type, termed neoteny, is exhibited in amphibians by the axolotl, and also in various insects. The Larvaceae among the tunicates are considered to represent persistent larvae reproducing neotenically.

Hermaphroditism is rare among the vertebrates but is found among most groups of invertebrates. Despite the production of eggs and sperm by the same individual, most hermaphroditic animals reproduce by cross-fertilization. In some cases, this is due to differences in time of ripening of the eggs and sperm. In others, with internal fertilization, crossing is assured by differences in location of sperm and egg ducts. In ascidians there is a physiological block to the union of the spermatozoon with an egg from the same individual. The latter situation illustrates that the interaction of egg and sperm in fertilization is not only generally species-specific but may also be individual-specific. *See* HERMAPHRODITISM.

Parthenogenesis is the development of the egg without fertilization by a spermatozoon. It is listed as a form of sexual reproduction because it involves development from a gamete. Rotifers, crustaceans,

and insects are the principal groups in which it occurs naturally. It has also been induced (artificial parthenogenesis) in species from all the major phyla by various kinds of chemical or physical treatment of the unfertilized egg. Even in mammals, several adult rabbits have reportedly been thus produced. *See* DEVELOPMENTAL BIOLOGY.

The honeybee provides a classic example of natural parthenogenesis. The males, or drones, all develop from unfertilized eggs and are haploid. The females, or workers and queen, arise from fertilized eggs and are diploid. Since the queen is inseminated only once, during the nuptial flight, she stores the sperm during her egg-laying life, which is 5 years or more, and can evidently permit or prevent them from fertilizing the eggs that are laid.

In certain animals periods of parthenogenesis may alternate with fertilization. Thus aphids produce parthenogenetic females during part of the year, but as winter approaches males appear, whereupon fertilization ensues. The fertilized eggs hatch out into females the following spring.

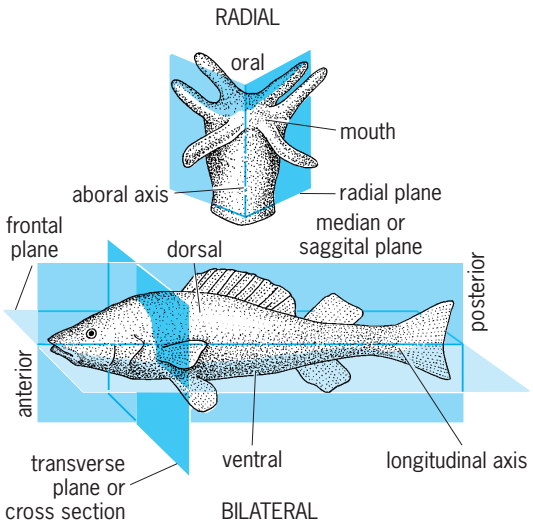
Among certain of the Lepidoptera, like the gypsy moth, *Lymantria*, parthenogenesis occurs at times when there appears to be a scarcity of males, and the unfertilized eggs can form males or females. There is also sporadic occurrence of parthenogenesis in chickens and turkeys. Mostly the embryos are abnormal and die at an early stage, but in the turkeys many develop quite far, and a few have been reared to adults.

Parthenogenetic development may also occur in eggs that are produced during a larval stage, as in the gallfly, *Miastor*. This condition is termed pedogenesis and is the parallel of neoteny. *See* ESTRUS; OOGENESIS; OVUM; SPERM CELL; SPERMATOGENESIS.

Albert Tyler; Howard L. Hamilton

Animal symmetry

Animal symmetry relates the organization of parts in animal bodies to the geometrical design that each type suggests. The term asymmetrical applies to most sponges and some protozoa because the body lacks definite form or geometry, as it cannot be subdivided into like portions by one or more planes. Spherical symmetry is exhibited by some protozoa, such as the Heliozoia and Radiolaria. The body is spherical with its parts concentrically around, or radiating from, a central point. Radial symmetry is exemplified by the echinoderms and most coelenterates. The body is structurally a cylinder, tall or short, having a central axis named the longitudinal, anteroposterior, or oral-aboral axis (see **illus.**). Any plane through this axis divides the animal into like halves. Often several planes, from the axis outward, can divide the body into a number of like portions, or antimeres, five in most echinoderms. Ctenophores and many sea anemones and corals possess biradial symmetry, basically radial but with some parts arranged on one plane through the central axis. Most animals have bilateral, or two-sided, symmetry, in which a



Types of symmetry and the axes, planes, and regions in animal bodies. (*After T. I. Storer et al., eds., General Zoology, 6th ed., McGraw-Hill, 1979*)

median or sagittal plane divides the body into equivalent right and left halves, each a mirror image of the other.

Tracy I. Storer

Animal virus

A small infectious agent that is unable to replicate outside a living animal cell. Unlike other intracellular obligatory parasites (for example, chlamydiae and rickettsiae), they contain only one kind of nucleic acid, either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), but not both. They do not replicate by binary fission. Instead, they divert the host cell's metabolism into synthesizing viral building blocks, which then self-assemble into new virus particles that are released into the environment. During the process of this synthesis, viruses utilize

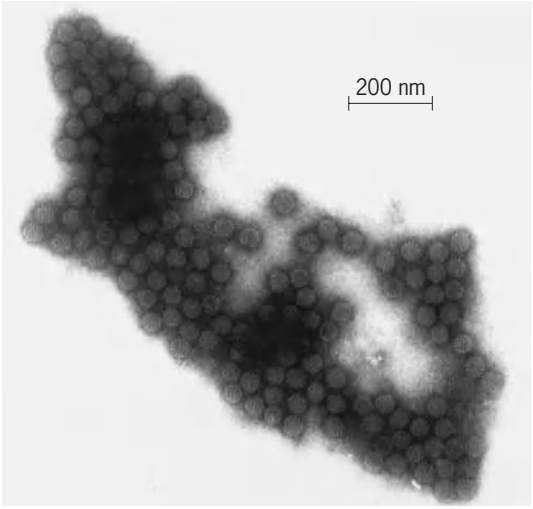


Fig. 1. Electron micrograph of negatively stained, purified bovine papilloma virions with icosahedral symmetry. (*Courtesy of David Hill and John P. Sundberg*)

cellular metabolic energy, many cellular enzymes, and organelles which they themselves are unable to produce. For this reason they are incapable of sustaining an independent synthesis of their own components. Animal viruses are not susceptible to the action of antibiotics. The extracellular virus particle is called a virion, while the name virus is reserved for various phases of the intracellular development. *See* DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

Morphology. Virions are small, 20–300 nanometers in diameter, and pass through filters which retain most bacteria. Due to this property, viruses were called filterable agents of disease. However, large virions (for example, vaccinia, which is 300 nm in diameter) exceed in size some of the smaller bacteria.

The major structural components of the virion are proteins and nucleic acid, but some virions also possess a lipid-containing membranous envelope. The protein molecules are arranged in a symmetrical shell, the capsid, around the DNA or RNA. The shell and the nucleic acid constitute the nucleocapsid.

In electron micrographs of low resolution, virions appear to possess two basic shapes: spherical and cylindrical. High-resolution electron microscopy and x-ray diffraction studies of crystallized virions reveal that the "spherical" viruses are in fact polyhedral in their morphology, while the "cylindrical" virions display helical symmetry. The polyhedron most commonly encountered in virion structures is the icosahedron, in which the protein molecules are arranged on the surface of 20 equilateral triangles. Based on these morphological features, viruses are classified as helical or icosahedral (**Fig. 1**). Certain groups of viruses do not exhibit any discernible features of symmetry and are classified as complex virions (**Fig. 2**). Further distinction is made between virions containing RNA or DNA as their genomes and between those with naked or enveloped nucleocapsids. These assignments are shown in the **table**, in which the major groups of animal viruses are listed.

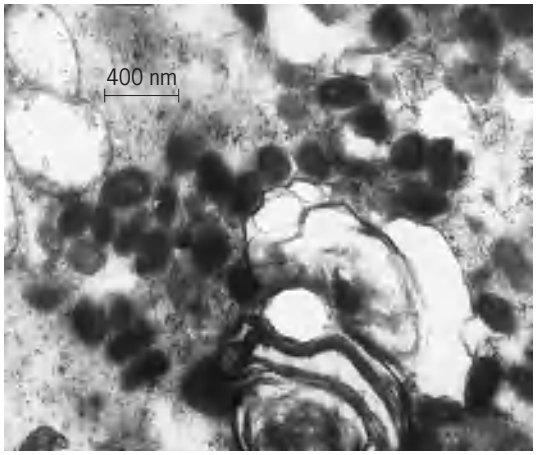


Fig. 2. Electron micrograph of a thin section of squirrel cells afflicted with squirrel fibroma. The brick-shaped pox virions contain a characteristic dumbbell-shaped nucleocapsid. (*Courtesy of David Hill and John P. Sundberg*)

Viral nucleic acid. The outer protein shell of the virion furnishes protection to the most important component, the viral genome, shielding it from destructive enzymes (ribonucleases or

Classification of animal viruses

Family*	Prototype	Nucleic acid	Structure
Poxviridae	Vaccinia virus	Double-stranded DNA	Complex, enveloped
Parvoviridae	Adeno-associated	Single-stranded DNA	Icosahedron, naked
Reoviridae	Reovirus	Double-stranded RNA, fragmented	Icosahedron, naked
Rhabdoviridae	Vesicular stomatitis virus rabies	Single-stranded RNA negative strand	Bullet-shaped, helical nucleocapsid, enveloped
Herpesviridae	Herpes simplex	Double-stranded DNA	Icosahedron, enveloped
Adenoviridae	Adenovirus	Double-stranded DNA	Icosahedron, naked
Papovaviridae	Simian virus 40	Double-stranded DNA, circular	Icosahedron, naked
Retroviridae	Rous sarcoma	Single-stranded RNA	Complex, enveloped
Paramyxoviridae	Newcastle disease	Single-stranded RNA negative strand	Helical nucleocapsid, enveloped
Orthomyxoviridae	Influenza	Single-stranded RNA negative strand, fragmented	Helical nucleocapsid, enveloped
Togaviridae	Alpha: Sindbis Flavi: Yellow fever	Single-stranded RNA	Icosahedron, enveloped
Coronaviridae	Avian infectious bronchitis	Single-stranded RNA	Complex, enveloped
Arenaviridae	Lymphocytic choriomeningitis	Single-stranded RNA negative strand	Complex, enveloped
Picornaviridae	Polio	Single-stranded RNA	Icosahedron, naked
Bunyaviridae	Bunyamwera	Single-stranded RNA negative strand, fragmented	Complex, enveloped

*After F. Fenner, The classification and nomenclature of viruses: Summary of results of meetings of the International Committee on Taxonomy & Viruses in Madrid, September 1975, *Virology,* 71:371, 1976.

deoxyribonucleases). The viral genome carries information which specifies all viral structural and functional components required for the initiation and establishment of the infectious cycle and for the generation of new virions. This information is expressed in the alphabet of the genetic code (sequence of nucleotides) and may be contained in a double-stranded or single-stranded (parvoviruses) DNA, and double-stranded (reoviruses) or single-stranded RNA. The viral DNA may be linear or circular, and the viral RNA may be a single long chain or a number of shorter chains (fragmented genomes), each of which contains different genetic information. Furthermore, some RNA viruses have the genetic information expressed as a complementary nucleotide sequence. These are classified as negative-strand RNA viruses. Finally, the RNA tumor viruses have an intracellular DNA phase, during which the genetic information contained in the virion RNA is transcribed into a DNA and integrated into the host cell's genome. The discovery of this process came as a surprise, since it was believed that the flow of genetic information was unidirectional from DNA to RNA to protein and could not take place in the opposite direction, from RNA to DNA. The transcription of RNA to DNA was termed reverse transcription, and the RNA tumor viruses are sometimes referred to as retroviruses. A classification based on the relationship between viral genomes and their messenger RNA (mRNA) is shown in **Fig. 3**. *See* GENETIC CODE; TUMOR VIRUSES.

When introduced into a susceptible cell by either chemical or mechanical means, the naked viral nucleic acid is in most cases itself infectious. Two exceptions are the negative-strand RNA viruses and the RNA tumor viruses. In these cases the RNA has to be first transcribed and reverse-transcribed, respectively, into the proper form of genetic information before the infectious process can take place. This task is carried out by means of an enzyme which is contained in the protein shell of the virion nucleocapsid.

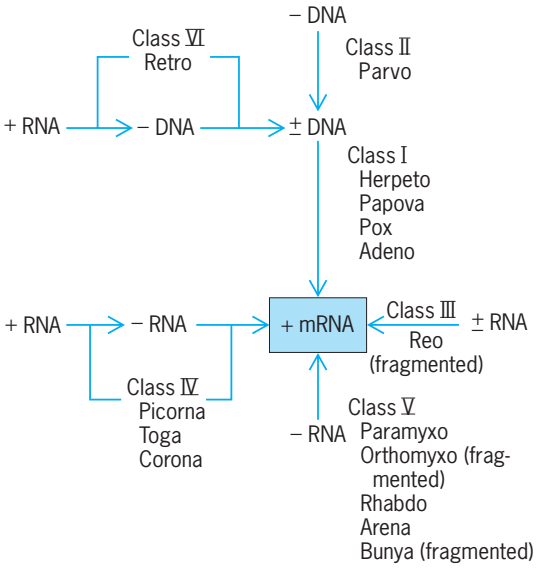


Fig. 3. Classification based on the relationship between the viral genome and its mRNA. (*After D. Baltimore, Expression of animal virus genomes, Bact. Rev., 35:235–241, 1971*)

The whole nucleocapsid is therefore required for infectivity.

Assay. Advances in virus research are closely linked with the utilization of new laboratory techniques. The introduction of tissue cultures eliminated the necessity to use experimental animals and revolutionized animal virus research. The application of plaque assays, first developed in studies of bacterial viruses, was of similar importance. In this assay, a monolayer of cells growing on solid support (usually in a petri dish) is infected with a dilute solution of virions and overlaid with a soft nutrient agar layer. The high viscosity of the agar layer prevents newly released virions from diffusing farther than the cells adjacent to the ones originally infected. After a sufficiently large number of surrounding cells are killed by newly released virions (after 1 day to 3 weeks, depending on the virus), a clear plaque becomes visible. Since each plaque was initiated by a single virus particle, the method allows a quantitative determination of the number of viable virions and offers a means of virus purification (the virus population isolated from a single plaque is relatively homogeneous). Plaque assays are frequently used in diagnosis and in the clinical isolation of viruses. Other assays are based on the ability of some viruses to hemagglutinate (aggregate) red blood cells. The highest dilution of a virus-containing extract which is still capable of hemagglutinating (end-point dilution) is used as a measure of virion concentration. Titrations with specific antibodies are similarly applied in the characterization and quantitation of viruses. *See* TISSUE CULTURE.

Infectious cycle. Viral infection is composed of several steps: adsorption, penetration, uncoating and eclipse, and maturation and release.

Adsorption. Adsorption takes place on specific receptors in the membrane of an animal cell. The presence or absence of these receptors determines the tissue or species susceptibility to infection by a virus. Enveloped viruses exhibit surface spikes which are involved in adsorption; however, most animal viruses do not possess obvious attachment structures.

Penetration. Penetration takes place through invagination and ingestion of the virion by the cell membrane (phagocytosis or viropexis). Enveloped viruses often also enter the cell by a process of fusion of the virion and cell membranes. Penetration is followed by uncoating of the nucleic acid, or in some cases by uncoating of the nucleocapsid. At this stage, the identity of the virion has disappeared, and viral infectivity cannot be recovered from disrupted cells (special methods are required to utilize the infectious nucleic acid). *See* PHAGOCYTOSIS.

Eclipse. The absence of infectious particles in cell extracts is characteristic of the eclipse period. During the eclipse the biochemical processes of the cell are manipulated to synthesize viral proteins and nucleic acids. The survival of viruses depends on this subversive ability. In the process of evolution they have developed an extraordinary efficiency and a remarkable repertoire of strategies. The eclipse period in infections with DNA viruses starts with the

transcription of the genetic information in the nucleus of the cell (except poxviruses), processing into mRNAs, and their translation into proteins (in the cytoplasm). This process is divided into early and late transcription. In the absence of newly synthesized viral components, the immediate early events must be entirely catalyzed by cellular enzymes. The early proteins, however, are virus-encoded functional proteins which will participate in the synthesis of viral DNA and of intermediate and late viral proteins, as well as in the shutoff of various cellular functions which might be detrimental to viral synthesis. The event that distinguishes early from late mRNA transcription and translation is the onset of viral DNA synthesis. The major late products are the structural proteins of the nucleocapsid. Almost as soon as these proteins are synthesized, they assemble with newly synthesized DNA molecules into virion nucleocapsids. The appearance of these particles signals the end of the eclipse period.

The events of the eclipse period in infections with RNA viruses are similar, except that they take place in the cytoplasm (influenza virus excepted), and a division into early and late transcription cannot be made. In the case of positive-strand RNA viruses, the viral RNA is itself the mRNA. In infections with negative-strand RNA viruses, the virion RNA in the nucleocapsid is first transcribed into positive mRNAs. Intracellular nucleocapsids are present throughout the entire infectious cycle, and the eclipse period cannot be defined in the classical sense. RNA tumor viruses reverse-transcribe their RNA into DNA, which enters the cell nucleus and becomes integrated into the cellular DNA. All viral mRNAs and genomic RNAs are generated by transcription of the integrated DNA (Fig. 3).

Maturation and release. The event characteristic of the maturation step is virion assembly and release. In many cases the protein shell is assembled first (procapsid) and the nucleic acid is inserted into it. During this insertion, processing of some shell proteins by cleavage takes place and is accompanied by a modification of the structure to accommodate the nucleic acid. Unenveloped viruses which mature in the cytoplasm (for example, poliovirus) often exit the cell rapidly by a reverse-phagocytosis process, even before the breakdown of the cell. In some cases, however, a large number of virus particles may accumulate inside the cell in crystalline arrays called inclusion bodies (**Fig. 4**). Viruses that mature in the nucleus are usually released slowly, and the damage to the cell is extensive. Enveloped viruses exit the cell by a process of budding. Viral envelope proteins (glycoproteins) become inserted at various sites into the cell membrane, where they also interact with matrix proteins and with nucleocapsids. The cellular membrane then curves around the complex and forms a bud which detaches from the rest of the cell. **Figure 5** shows virus particles budding from the cell surface of a mouse cell infected with murine leukemia virus.

Effect of viral infections. Two extreme types of effects are identified: lytic infections, which cause cell death by a variety of mechanisms with cell lysis as the most common outcome, and persistent infections, accompanied either by no apparent change in the host cell or by some interference with normal growth control, as in transformation of normal to cancer cells. The degenerative phenomena in tissue cultures during a lytic infection are called cytopathic or cytolytic effects. In animals, extensive destruction of tissue may accompany an infection by a lytic virus. *See* LYTIC INFECTION.

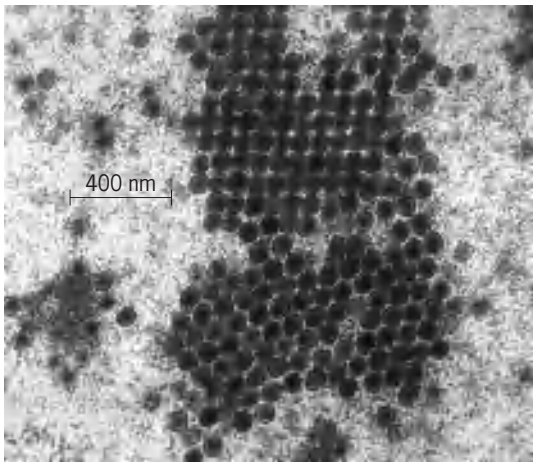


Fig. 4. Electron micrograph of thin section of chicken cells containing crystalline inclusion bodies formed by avian reovirus during infection. (*Courtesy of David Hill and John P. Sundberg*)

As a defense to certain conditions of infection, animal cells generate a group of substances called interferons which, by a complex mechanism, inhibit replication of viruses. They are specific to the cell species from which they were derived but not to the virus which elicited their generation. (Mouse interferon will protect mouse but not human cells from any viral infection.)

Genetics. Like other microorganisms, viruses mutate either spontaneously or following exposure to

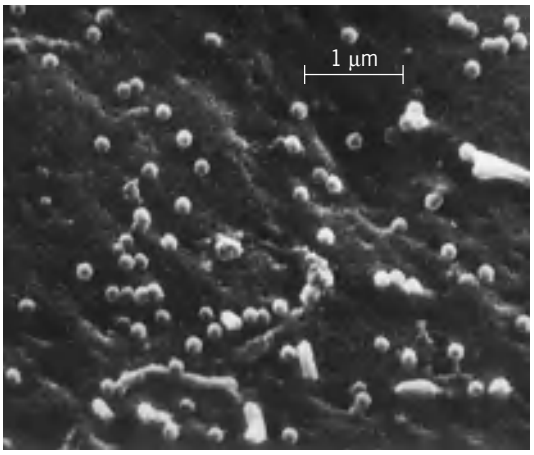


Fig. 5. Scanning electron micrograph of the surface of a mouse cell infected with murine leukemia virus. A large number of virus particles are shown in the process of budding. (*Courtesy of R. MacLeod*)

mutagenic agents. With the introduction of cell tissue cultures and with the development of plaque purification procedures, isolation and characterization of well-defined mutants became possible. In the past the most useful mutants in genetic studies of animal viruses were temperature-sensitive *(ts)* and host-range *(hr)* mutants. The *ts* mutants grew efficiently at lower than normal but not at normal temperatures (permissive and nonpermissive temperatures, respectively). The mutation caused a change of an amino acid in a viral protein which resulted in a biologically inactive configuration at the higher temperatures. Studies of the impaired proteins and their respective genes helped to elucidate the molecular mechanisms of various events in the infectious cycle. The *hr* mutants, which grew normally in certain host cells but not in others, were useful in studies of virus–host interactions. This approach to viral genetics was dependent on the isolation of the proper mutants, which was often fortuitous. With the advent of modern DNA technology of restricting, cloning, and sequencing, it became possible to generate site-specific mutants. Rather than depending on an accidental change of a nucleotide by a mutagen, the investigator could chemically alter a nucleotide at any selected site. In this manner, mutants were generated to answer very specific questions about viral gene expression and virus–host interactions. Since reverse transcription can be carried out in the test tube, these techniques could, in principle, be applied also to RNA viruses as long as their RNAs are infectious. DNA cloning, which permits the preparation of relatively large quantities of viral genomes, is making it possible to study viruses which were inaccessible to investigation either because of low yields obtained in living organisms or because of the lack of suitable tissue cultures. *See* GENETIC ENGINEERING; MUTATION; RECOMBINATION (GENETICS); RESTRICTION ENZYME.

This modern approach to animal virus genetics also rapidly advanced understanding of the complex molecular biology of the animal cell. Many structural and regulatory features of animal virus genes are similar to those of the animal cell, yet are easier to study because of the smaller size of viral genomes. *See* GENETICS; MOLECULAR BIOLOGY.

Pathology. Virus infections spread in several ways: through aerosols and dust, by direct contact with carriers or their excretions, and by bites or stings of animal and insect vectors. At the point of entry, infected cells undergo viremia. From there, the virus becomes disseminated by secretions. It is carried through the lymphatic system and bloodstream to other target organs, where secondary viremias occur (except in localized infections like warts). In most cases viral infections are of short duration and great severity. However, persistent infections are not uncommon (herpes, adeno, various paramyxoviruses like measles). *See* VIRUS INFECTION, LATENT, PERSISTENT, SLOW.

The afflicted organism mounts a variety of defenses, the most important of which is the immune response. Circulating antibodies against viral proteins are generated. Those interacting with virion surface proteins neutralize the infectious potential of the virus. Although the antibodies are specific against the virus which has elicited them, they will cross-react with closely related virus strains. The specificity of neutralizing antibodies obtained from experimentally injected animals is utilized for diagnostic purposes or in quantitative assays. In addition to the circulating antibodies, cell-mediated immune responses also take place. The most important of these is the production of cytotoxic thymus-derived lymphocytes, found in the lymph nodes, spleen, and blood. They destroy all cells which harbor in their membrane viral glycoproteins, regardless of whether these cells are actively involved in virus synthesis or acquired the viral proteins by membrane fusion with inactive virions or cell debris. The cell-mediated immunity has been demonstrated to be more important to the process of recovery than circulating antibodies. In spite of their beneficial role, immune responses often seriously contribute to the pathology of the disease. Circulating antigen–antibody complexes can lodge in organs and cause inflammation; cell-mediated responses have been known to produce severe shock syndromes in patients with a history of previous exposures to the virus. *See* ANTIBODY; AUTOIMMUNITY; CELLULAR IMMUNOLOGY; IMMUNITY; VIRUS INTERFERENCE.

Control. Viruses are resistant to the antibiotics commonly used against bacterial infections. The use of chemotherapeutic agents with antiviral activity is plagued by their toxicity to the animal host. However, the application of vaccines has been successful in the control of several viruses. The vaccines elicit immune responses and provide sometimes lifelong protection. Two types of vaccines have been applied: inactivated virus and live attentuated virus. Various inactivation procedures are available. For example, in preparations of polio vaccines (Salk), a combination of formalin and heat was utilized. Protection by attenuated live virus vaccines dates back as early as 1776, when Edward Jenner used infections with harmless cowpox to achieve protection from the devastating effects of smallpox. In modern times, an attentuated laboratory strain of smallpox has been applied so successfully that the disease is considered to be eradicated. A small probability of back mutations of the attenuated virus to a virulent strain makes applications of live vaccines somewhat riskier. On the other hand, protection is longer-lasting and, by virtue of spread to nonvaccinated individuals, more beneficial to the population group (herd effect). *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; POLIOMYELITIS; VACCINATION; VIRUS CHEMOPROPHYLAXIS.

In order to achieve full protection, it is important that the vaccine contain all the distinct antigenic types of the virus. Development of monoclonal antibodies led to a better characterization of these types in naturally occurring viruses. This information will undoubtedly lead to better vaccines. Moreover, monoclonal antibodies have aided investigations into the molecular structure of viral antigenic groups and brightened future prospects for synthetic vaccines.

See MONOCLONAL ANTIBODIES; VIRUS; VIRUS CLASSIFICATION. M. E. Reichmann

Bibliography. R. Dulbecco and H. S. Ginsberg, *Virology*, 1980; H. Fraenkel-Conrat and P. C. Kimball, *Virology*, 1982; H. Fraenkel-Conrat and R. R. Wagner (eds.), *Comprehensive Virology*, 1974–1983; R. A. Lerner, Synthetic vaccines, *Sci. Amer.*, pp. 66–74, February 1983; S. E. Luria et al., *General Virology*, 3d ed., 1978; D. Nathans, Restriction endonucleases, Simian Virus 40 and the new genetics, *Science*, 206:903–909, 1979; K. M. Smith and M. A. Lauffer (eds.), *Advances in Virus Research*, 1958–1983.

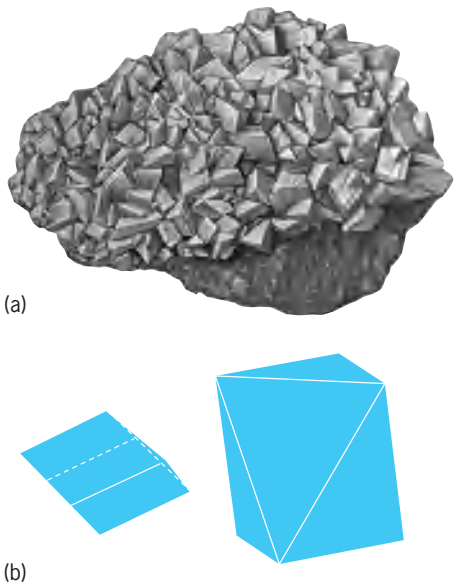
Anise

One of the earliest aromatics mentioned in literature. The plant, *Pimpinella anisum* (Apiaceae), is an annual herb about 2 ft (0.6 m) tall and a native of the Mediterranean region. It is cultivated extensively in Europe, Asia Minor, India, and parts of South America. The small fruits are used for flavoring cakes, curries, pastry, and candy. The distilled oil is used in medicine, soaps, perfumery, and cosmetics. *See* APIALES; SPICE AND FLAVORING.

Perry D. Strausbaugh; Earl L. Core

Ankerite

The carbonate mineral $Ca(Fe,Mg)(CO_3)_2$, also commonly containing some manganese. The mineral has hexagonal (rhombohedral) symmetry (see **illus.**) and has the cation-ordered structure of dolomite. The name is applied only to those species in which at least 20% of the magnesium positions are occupied by iron or manganese; species containing less iron are termed ferroan dolomites. The pure compound, $CaFe(CO_3)_2$, has never been found in nature and has never been synthesized as an ordered compound. *See* DOLOMITE.



Ankerite. (*a*) Specimen of the mineral (*specimen from Department of Geology, Bryn Mawr College*). (*b*) Crystal habits (*after C. Klein, Manual of Mineralogy, 21st ed., John Wiley & Sons, Inc., 1993*).

Ankerite typically occurs in sedimentary rocks as a result of low-temperature metasomatism. It is commonly white to light brown, its specific gravity is about 3, and its hardness is about 4 on Mohs scale. *See* CARBONATE MINERALS; METASOMATISM. Alan M. Gaines

Bibliography. L. L. Y. Chang, R. A. Howie and J. Zussman, *Rock-Forming Minerals*, vol. 58: *Non-silicates: Sulphates, Carbonates, Phosphates, Halides*, 1995; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 1a: *Orthosilicates* 2d ed., 1997.

Annelida

The phylum comprising the multisegmented, invertebrate wormlike animals, of which the most numerous are the marine bristle worms and the most familiar the terrestrial earthworms. The Annelida (meaning little annuli or rings) include the Polychaeta (meaning many setae); the earthworms and fresh-water worms, or Oligochaeta (meaning few setae); the marine and fresh-water Hirudinea or leeches; and two other marine classes having affinities with the Polychaeta: the Archiannelida (meaning primitive annelids) are small heteromorphic marine worms, and the Myzostomaria (meaning sucker mouths) are parasites of crinoid echinoderms. *See* HIRUDINEA; MYZOSTOMIDA; OLIGOCHAETA; POLYCHAETA.

General Characteristics

These five groups share few common characters and little resemblance except that most have a wormlike body. Typically they are bilaterally symmetrical, lack a skeleton, and have a short to long linear body divided into rings or segments, which are separated from one another by transverse walls or septa. The mouth is an anteroventral or anterior vent at the forward end of the alimentary tract, and the anus posterodorsal or posterior at the hind end of the gut.

Metamerism. The linear series of segments, or metameres, from anterior to posterior ends constitute the annelid body. These segments may be similar throughout, resulting in an annulated cylinder, as in earthworms and *Lumbrineris*. More frequently the successive segments are dissimilar, resulting in regions modified for particular functions. Each segment may be simple or uniannular, corresponding to a metamere, or it may be divided or multiannulate. The total number of segments varies from five to several hundred, according to kind. Segments may have lateral fleshy outgrowths called parapodia (meaning side feet), armed with special secreted bristles or rods, called setae and acicula; they provide protection and aid in locomotion. Setae are lacking in Hirudinea and some polychaetes. The body is covered by a thin to thick epithelium which is never shed.

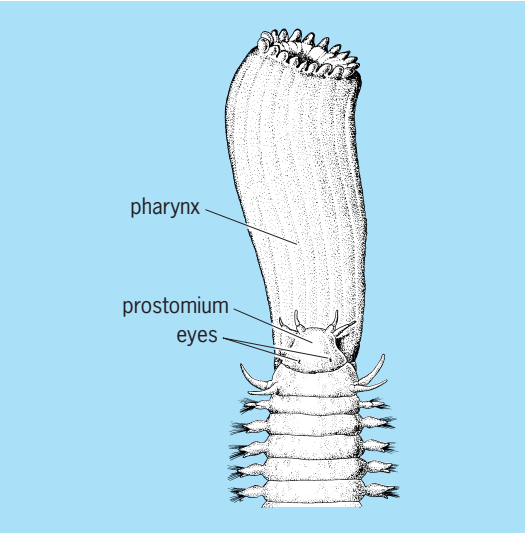


Fig. 1. *Eteone* (Phyllodocidae), everted pharynx, simple prostomial eyes, in dorsal view. (*After O. Hartman, The littoral marine annelids of the Gulf of Mexico, Publ. Inst. Mar. Sci., 2:7–124, 1951*)

Sense organs. Annelids have sense organs of many kinds. Most conspicuous are those on the anterior, or head end, and on parapodia. Eyes which function as photoreceptors may be variously developed, simple (**Fig. 1**) to complex, small to large, absent to numerous, and obscure to bizarre. Structurally they are simple light-absorbing pigment spots, usually paired on the head, but also on body segments or anal end. In some sabellids they are on the tentacular crown and may number in the hundreds, or they are on the pygidium (**Fig. 2**) and aid in backward movement.

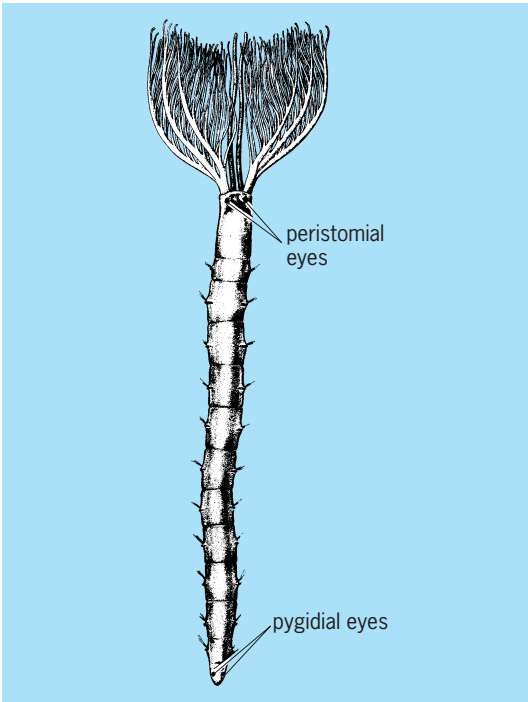


Fig. 2. *Fabricia* (Sabellidae), peristomial and pygidial eyes, in ventral view. (*After O. Hartman, Fabricinae feather-duster polychaetous annelids in the Pacific, Pac. Sci., 5:379–391, 1951*)

In some polychaetes they are highly developed, with lenses and retina (**Fig. 3**), and may be at the end of an erect stalk. Swarming stages of benthic annelids may acquire large modified eyes which are an attractant at sexual maturity. Some pelagic polychaetes have large eyes which make up most of the anterior end (**Fig. 4**). *See* EYE (INVERTEBRATE).

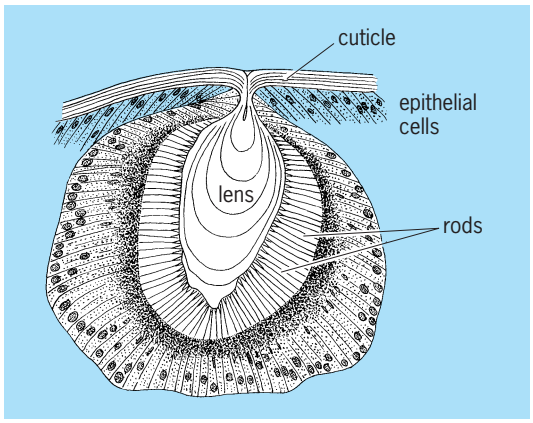


Fig. 3. Cross section of eye of *Eunice* (Eunicidae). (*After O. Pflugfelder, Ueber den feineren Bau der Augen freilebender Polychaeten, Z. wiss. Zool., 142:540–586, 1932*)

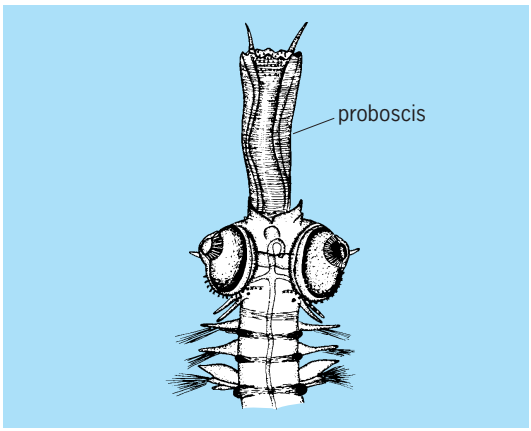


Fig. 4. Anterior end of *Torrea* (Alciopidae), large compound eyes and everted proboscis, in dorsal view. (*After E. Claparède, Les Annelides Chetopodes du Golfe de Naples, Mém. Soc. Phys. Genève, 20:1–225, 1870*)

Feelers or tactoreceptors are frequently on the prostomium as antennae; on anterior segments as long filiform or thick fleshy structures called tentacles; and on parapodia as cirri, papillae, scales, or tactile hairs. Cilia in bands or clusters or in grooves occur in specific patterns. The most conspicuous receptors are the large caruncles (nuchal organs) of amphinomid polychaetes, as simple (**Fig. 5**) or complex (**Fig. 6**) fleshy, folded, paired organs surrounding the cephalic structures. The nuchal organs of some syllids and phyllodocids, as well as the ciliated ridges of errantiate and sedentary polychaetes, illustrate the diversity and importance of these structures. Proprioreceptors are sense organs which respond to stimuli originating within the animal, such

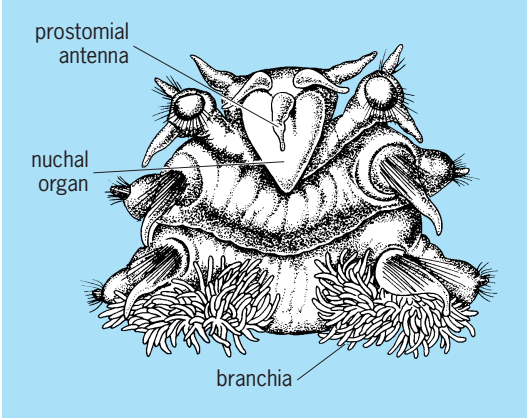


Fig. 5. Simple nuchal organ or caruncle of *Amphinome* (Amphinomidae), in dorsal view. (*After O. Hartman, The littoral marine annelids of the Gulf of Mexico, Publ. Inst. Mar. Sci., 2:7–124, 1951*)

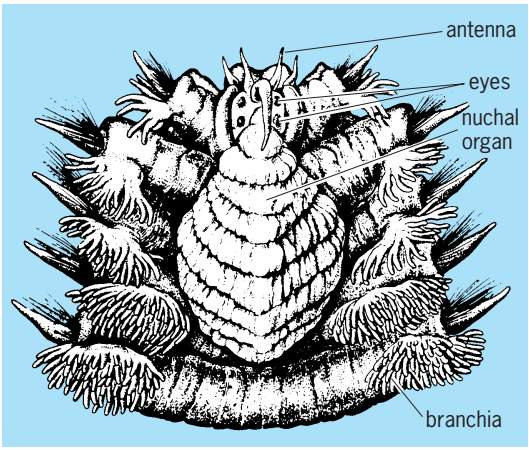


Fig. 6. Anterior end of *Hermodice* (Amphinomidae) with complex caruncle, in dorsal view. (*After O. Hartman, The littoral marine annelids of the Gulf of Mexico, Publ. Inst. Mar. Sci., 2:7–124, 1951*)

as hunger. Trophic reactions to sense food and repelling organs to deter enemies are well developed in many annelids.

Nuchal organs are ciliated, eversible or stationary pouches or grooves located in the anterior or neck region. They relay messages of the immediate environment of the individual. When partly or wholly enclosed as pouches, they may contain small grains either of secreted or foreign origin to form statocysts; they function in righting movements of the body. Setae are part of the sensory system for they detect changes in the environment through their basal attachments. Shallow-water species often have short, strong, resistant setae, whereas abyssal species have long, slender, simple setae. Each organ is unique and well adapted to its role in the development, growth, protection, and reproduction of the species involved.

Chromatophores are cell clusters which change their shape and size to conform to the shadows of the animal's background, responding to changes in light intensities, and therefore are generally protective. They are well developed in translucent pelagic larvae which exist at the surface of the sea; they screen damaging intensities of light from delicate tissues. Oligochaetes and hirudineans, which generally lack eyes or special light receptors, are sensitive to light changes through the surface epithelium. *See* CHROMATOPHORE.

Luminescence. Light production is known in many annelids. Some oligochaetes shed luminous mucus through the mouth or anus. Many polychaetes have special photocytes which luminesce. Some *Chaetopterus* colonies shed bright mucus from the palpi and parapodia. *Polycirrus* among the terebellids is highly luminescent. Pelagic spawning stages of *Odontosyllis* (a syllid) and *Tharyx* (cirratulid) flash bright light when disturbed. Some scale worms (polynoids) have light-producing cells on the undersides of elytra. Flashes are rhythmic and may continue for a minute or more. Luminescence may lure other individuals of the species or aid escape. The posterior, broken-off end continues to light up long after the anterior end has escaped. *See* BIOLUMINESCENCE.

Digestive system. The alimentary tract is a straight or sinuous tube, consisting of mouth, pharynx, esophagus, stomach, gut, and pygidium or anus. Its several parts may be further differentiated and named for structure and function. The mouth may be a simple anterior (oligochaetes) or anteroventral (many polychaetes) pore provided with highly complex organs or accessory parts. The upper lip is formed by the prostomium, and the lower lip by the ventrum of the first segment. Accessory organs include the short to long grooved palpi (**Fig. 7**) of many polychaetes, which direct food to the mouth or also select and propel nutrients along ciliated tracts. The building organs and cementing glands of some tubicolous annelids are associated with the mouth; they select inert particles for shape and size and attach them in specific patterns to a basic secreted mucoid membrane, resulting in tubes which are highly characteristic (**Fig. 8**). In sand-cementing or reef-building colonial species, the beach front may be covered with cemented bands of tubes highly

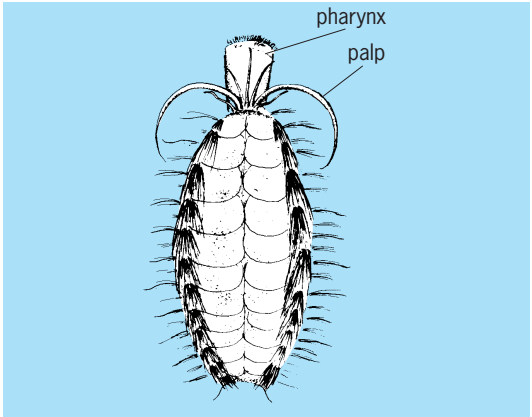


Fig. 7. *Laetmonice* (Aphroditidae), everted pharynx, long palpi, in dorsal view. (*After W. McIntosh, A monograph of the British annelids, Roy. Soc. London, 1(pt. 2):215–442, 1900*)

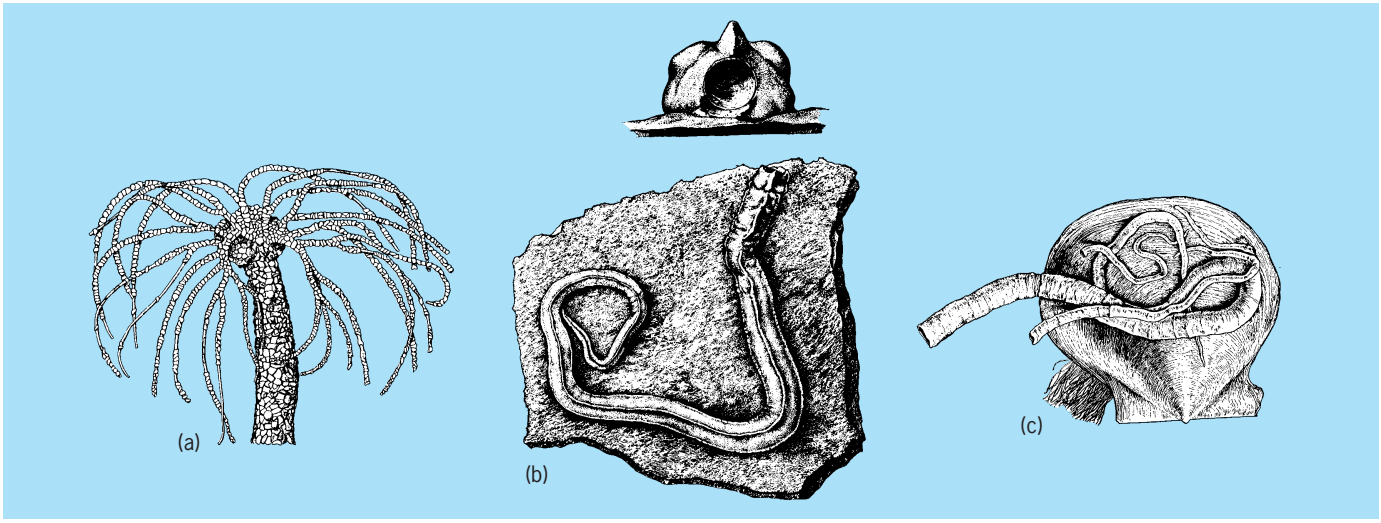


Fig. 8. Tubes of sedentary polychaetes. (*a*) Distal end of tube of *Lanice* (Terebellidae) (*after O. Hartman, Polychaetous annelids collected by the USNS Eltanin and Staten Island cruises, chiefly from Antarctic seas, Allan Hancock Monogr. Mar. Biol., 2:1–387, 1967*). (*b*) Tube of *Vermiliopsis* (Serpulidae), entire and in frontal view. (*c*) Several tubes of *Protula* (Serpulidae) on deep-water pelecypod (*after O. Hartman, Endemism in the North Pacific Ocean, with emphasis on the distribution of marine annelids, and descriptions of new or little known species, Essays in the Natural Sciences in Honor of Captain Allan Hancock, University of Southern California, Los Angeles, 1955*).

resistant to wave action. Colonial serpulid worms secrete calcium carbonate from special collar cells, formed into highly organized tubes with the aid of oral organs.

The mouth is followed by the buccal cavity which may be modified as a proboscis or saclike eversible pouch. In nonselective, mud-eating annelids the pouch may be voluminous when everted, its surface papillated or branched so as to increase its sweeping surface as it brushes over the surface muds; the adherent materials are swept into the oral cavity and swallowed.

The buccal cavity may be followed by a short to long muscularized eversible pharynx (**Fig. 9**) which captures and breaks up or compresses food particles. Its inner walls may be fortified by papillae (**Fig. 10**) or hard gnaths or jaws (**Fig. 11**) called para-, micro-, or macrognaths, maxillae, mandibles, chevrons, trepan, or other descriptive names. A short esophagus leads to the muscular stomach or digestive region. Lateral ceca or pouches may be present, along esophageal and stomach portions, to increase the amount of surface for secretion and digestion, especially in short-bodied worms. Peristaltic (clasping and compressing) movements are rhythmic and result in the food bolus being digested, and wastes separated and pushed into the gut. In nonselective-feeding annelids the gut may be distended with great amounts of inert materials; in selective feeders there are few remains but those of living animals. The proctodaeum, or region preceding the anus, expels the wastes as fecal pellets of characteristic form.

The proctodaeum with pygidium may be a simple, tapering end with posterior pore or may be highly modified as a plaque. In some pelagic swarming nereids the pygidium of the male acquires a circlet of papillae through which sperm cells are shed into the sea. Defecation, or disposal of wastes from the gut, is usually posteriorly. Tubicolous annelids which lie head down may reverse their position for this function; others unable to turn may have an anal ciliated groove directing fecal materials forward and dorsal, away from the mouth.

Nervous system. The nervous system consists of a dorsal, bilaterally symmetrical, ganglionic mass or brain within or behind the preoral region. The brain is connected to the ventral cord through the circumesophageal connectives which extend about the

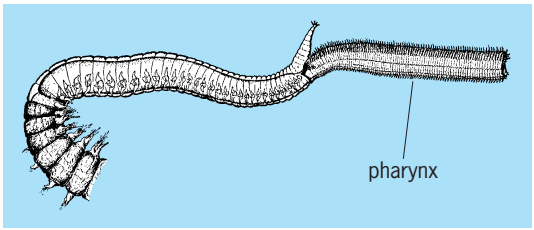


Fig. 9. Anterior end of *Glycinde* (Goniadidae), pharynx partly everted, in right lateral view. (*After O. Hartman, Goniadidae, Glyceridae, Nephtyidae, Allan Hancock Pacific Expedition, 15(1):1–181, 1950*)

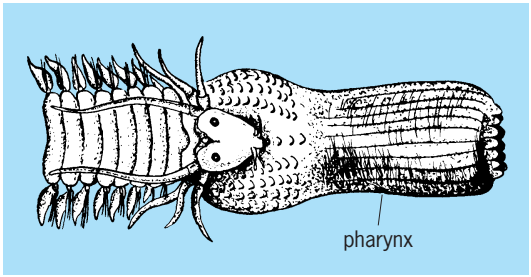


Fig. 10. Anterior end of *Anaitides* (Phyllodocidae) with pharynx fully everted, in dorsal view. (*After E. Rioja, Datos para el conocimiento de la fauna de Anélidos poliquetos del Cantábrico, Trab. Mus. Nac. Madrid, Zool. 37, pp. 1–99, 1918*)

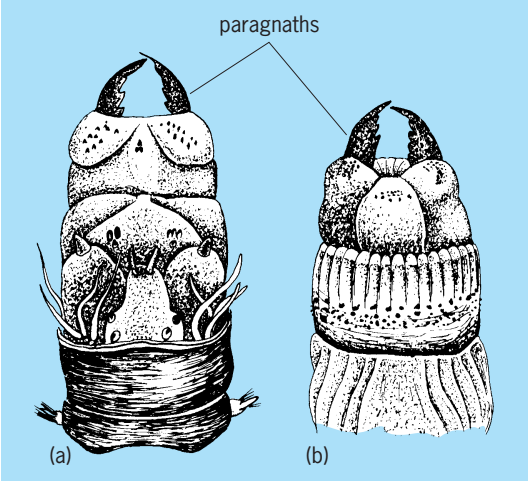


Fig. 11. Anterior end of *Nereis* (Nereidae) with everted pharynx showing distal jaws and paragnaths, (*a*) dorsal and (*b*) ventral views. (*After E. Rioja, Datos para el conocimiento de la fauna de Anelidos poliquetos del Cantábrico, Trab. Mus. Nac. Madrid, Zool. 37, pp. 1–99, 1918*)

oral cavity (**Fig. 12**). The ventral cord may be single or paired, nearly smooth or nodular, or ganglionated according to the segmental pattern of the body. The brain sends out lateral branches to the eyes, palpi, antennae, or other structures; the ventral cord has lateral branches to all fleshy parts which receive stimuli. A giant axon is an enlarged part of the ventral cord, present in many long, muscular, or actively moving oligochaetes and polychaetes. It permits rapid transfer of stimuli and muscular response, resulting in abrupt response. Messages are relayed directly from the source of stimulation to the muscles, without reaching the brain. These giant fibers were first discovered in oligochaetes in 1861, and have since been found in many other annelids. They vary in size and complexity, but are always larger than normal neurons to which they are connected. Their cells arise in the circumesophageal ganglion and extend throughout the length of the worm, or they arise much farther back and may extend longitudinally or crosswise. Each giant fiber may represent a single nerve cell or fusion of several nerve cells. Their diameters

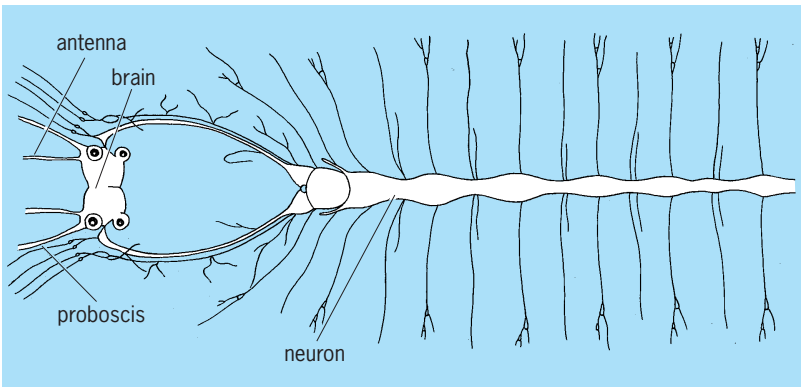


Fig. 12. Diagram of nervous system of *Nereis* showing some of the nerve branches. (*After S. Maxwell, Beiträ zur Gehirnphysiologie der Anneliden, Pfluegers Arch., Berlin, 67:263–297, 1897*)

are largest in *Myxicola* (Sabellidae) and much smaller in most other annelids. Their occurrence has been reported from widely related families.

Neurosecretion is a process of special cells located in the brain. Typically, nerve cells are polarized, consisting of an axon or head and a slender fiber. Secretory cells lack a fiber and secrete hormones. In some annelids they affect the time and rate of metamorphosis at maturity. In nereids the hormones inhibit the processes of metamorphosis, until the sexual cells have arrived at their full stage of development. The two processes, sexual development and transformation to an epitoke, are separate and each under its own control. *See* NEUROSECRETION.

Circulatory system. The circulatory system consists of dorsal and ventral longitudinal, median vessels located above and below the alimentary tract. Lateral branches extend to all parts of the body. Pulsating or propelling contractile portions, sometimes called hearts, are in the anterior dorsal vessel or also at intervals in the ventral vessels. In oligochaetes these are segmental vessels of varying number connecting the dorsal and ventral vessels and surrounding some portion of the alimentary tract.

The contained blood is red through the presence of a hemoglobinlike substance, which may be in corpuscles or granular in the fluid, or it is green, through the presence of chlorocruorin. These colors when diluted are yellow or colorless, as in many small annelids. *See* RESPIRATORY PIGMENTS (INVERTEBRATE).

The purpose of the blood is to transfer nutrients to all parts of the body and to exchange carbon dioxide for oxygen, either through the skin or through special organs called branchiae (Fig. 5). They are fingerlike fleshy prolongations of the epithelium, either simple, branched, or spiraled structures, at anterior or also posterior ends of the body. Each branchia contains a vascular loop through which the blood flows in one direction. The oxygen requirements of annelids may be very low, illustrated in worms existing in stagnant pools and in deep oceanic basins. Irrigation in the tube, burrow, or environment may be achieved through U-shaped structures which open to the surface or through undulations of the body or parapodia, to ensure a flow of water and nutrients.

Some annelids lack a closed circulatory system so that blood and coelomic fluids mix freely, resulting in a hemocoel; it may be partial or complete. Many annelids have a special organ called a cardiac or heart body surrounding the pulsating vessel and sometimes visible as a thick brown or red body of spongy tissue; its function is to dispose of circulatory wastes. The circulatory and coelomic fluids aid in maintaining turgidity of the annelid body, and with musculature control they act as a kind of skeleton.

Excretory system. The excretory system functions to dispose of wastes in the coelom, resulting from metabolism. The special organs for the function are called nephridia. In their simplest form they are protonephridia, consisting of a strand of cells connecting the coelom to the body wall, usually a pair to a segment. The inner end terminates in a flagellum which propels wastes outward; the distal end is a

pore at the surface of the body called a nephridiopore. Primitive nephridia occur in larval, small, or primitive annelids. Clusters of simple nephridia, solenocytes, basally joined to a tube, are present in many errantiate polychaetes. More complex organs, or metanephridia, have a ciliated nephrostome or funnel opening into the coelom and continued to the surface as a complex organ. They function to transport wastes and at sexual maturity may serve as gonoducts to release gonadial products. Complex nephridia are present in all oligochaetes and many polychaetes.

Nephridia may be few in number and conspicuous, or limited to anterior segments as in many sedentary families, or may be segmental, a pair to a segment, or dispersed and irregular. Some coelomic wastes, which may result through phagocytosis, or cell breakdown, cannot be eliminated through nephridia and may accumulate to form conspicuous dark blobs in posterior segments.

Muscular system. The muscular system consists of an outer circular and an inner longitudinal system of muscles, each varying in extent and density according to species. In addition, an oblique series, between outer and inner layers, is well developed in annelids performing complex lateral movements. Long, very active burrowers or crawlers have an extensive musculature, whereas short, sluggish forms may have diminished musculature development. The circular muscles form a sheath beneath the epithelium; they effect an increase or decrease in body diameter, and are best developed in tubicolous and burrowing forms. Longitudinal muscles are often in paired thin to thick bands in dorsal and ventral positions; they effect a shortening or lengthening of the body, and are most conspicuous in errantiate species. The oblique muscles occur in incomplete bands, between circular and oblique layers, connecting parts of the parapodia to the body wall and to other muscular layers.

Movements are achieved mainly by coordinated muscular contractions and expansions of the laterally projecting parapodia or setae, resulting in an undulating or meandering movement. Swimming species may move from side to side or by successive forward darts and stops. Some annelids with reduced parapodia and a long proboscis use the latter in progression by extension and withdrawal of the eversible part of the alimentary tract. Some annelids secrete mucus to move over irregular surfaces.

Regeneration. The ability to replace lost parts is highly developed in annelids. Most frequent is the replacement of tail, parapodia, and setae. The anterior end may be replaced provided the break is postpharyngeal. The torn end is first covered over with scar tissue, then differentiated into epithelial cells and all other tissues characteristic of the whole animal. The tendency is to restore the number of segments peculiar to the species. Least differentiated species regenerate more completely than highly modified ones. Younger individuals repair more easily or quickly than older ones. Fragmentation may result in as many new individuals as there are pieces.

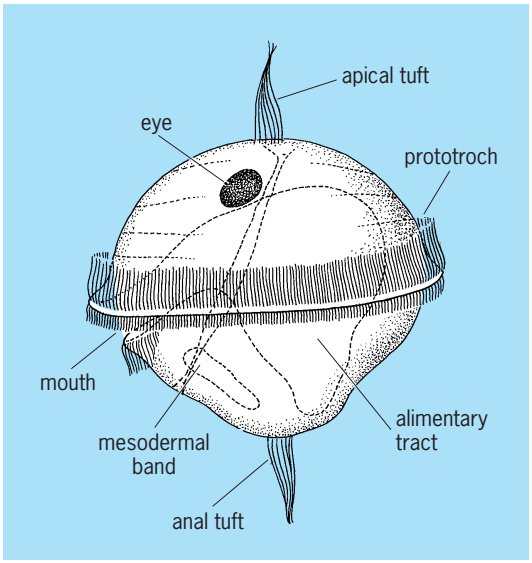


Fig. 13. Lateral view of larval trochophore with parts indicated and much enlarged.

The presence of a part of the ventral nerve cord is necessary to develop a new head. *See* REGENERATIVE BIOLOGY.

Reproduction. Depending on the species, reproduction may be sexual, asexual, or both. Sexual reproduction may be dioecious, in which male and female are similar, rarely dissimilar. Individuals may be hermaphroditic, both male and female, but with cross fertilization. Some annelids are protandric hermaphrodites, in which the sexual stages alternate. Epigamy refers to the transformation of mature individuals from a benthic to a planktonic form; changes occur in parapodial lobes and the alimentary tract may be absorbed. Fully metamorphosed individuals leave their tubes, swarm to the surface of the sea, sometimes in rhythmic or lunar patterns, deposit sperm and ova, and then disintegrate. In the palolo (Eunicidae) only posterior ends are released, and head ends retreat to regenerate new tails. The ova which are deposited at the surface of the sea develop into ciliated spheres called trochophores (**Fig. 13**), which develop into segmented metatrochs (**Fig. 14**) and then into larvae that begin to resemble adults (**Fig. 15**) into which they will metamorphose.

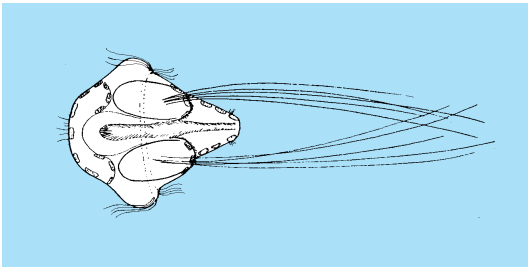


Fig. 14. Metatroch of planktonic sabellarian larva, showing early segmentation and long swimming setae. (*After O. Hartman, Polychaetous annelids, Paraonidae, Magelonidae, Longosomidae, Ctenodrilidae and Sabellariidae, Allan Hancock Pacific Expedition, 10(3):311–389, 1944*)

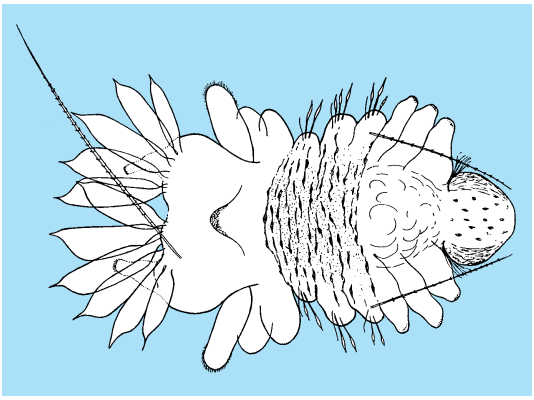


Fig. 15. Transitional stage of metatroch larva of sabellarian larva, with larval setae nearly replaced. (*After O. Hartman, Polychaetous annelids, Paraonidae, Magelonidae, Longosomidae, Ctenodrilidae and Sabellariidae, Allan Hancock Pacific Expedition, 10(3):311–389, 1944*)

At the appropriate time they sink to the bottom and assume a benthic development. *See* PROTANDRY.

Olga Hartman

Fossils

Fossil annelids, or segmented worms, are mostly soft-bodied, yet sufficiently common to indicate that this large and varied group of invertebrates has been abundant for more than 500 million years.

Origins. The bulk of the annelidan fossil record is represented by the polychaetes. The earliest definite occurrences are from the Lower Cambrian of Greenland (Sirius Passet fauna). Possible precursors to the polychaetes exist in the form of wiwaxiids and the more primitive halkieriids. Both are sclerite-bearing, and in the wiwaxiids the microstructure is identical to annelidan chaetae; in these animals the sclerites form a protective coat, but it is plausible to derive the notochaetae and neurochaetae from, respectively, an imbricated dorsal array and fanlike extensions of leglike parapodia. The coexistence of polychaetes and halkieriids in the Sirius Passet fauna indicates an earlier divergence, but a number of Ediacaran fossils previously assigned to the annelids, such as *Dickinsonia*, probably are not closely related.

Soft-bodied polychaetes. The soft-bodied record from the Cambrian onward is sparse, but polychaetes are known from a number of exceptionally preserved fossil biotas. A diverse assemblage occurs in the Middle Cambrian Burgess Shale of British Columbia. Not only do these polychaetes (**Fig. 16**) show a very wide morphological range, but with one possible exception it has not proved feasible to accommodate any of the six species in a known fossil or living family. *See* BURGESS SHALE.

Other landmarks in the polychaete fossil record are a series of deposits from the Carboniferous that include the Francis Creek Shale (Mazon Creek ironstone nodules; **Fig. 17**) of Illinois, the Bear Gulch Limestone of Montana, and the Granton shrimp bed of Edinburgh, Scotland. The first two units have yielded a wide variety of polychaetes, many with jaws. Although only one polychaete occurs at Granton, it is significant in apparently representing the earliest tomopterid. In younger deposits, soft-bodied polychaetes are known from horizons in the Triassic (Voltzia Sandstone of eastern France, also deposits in Italy and Madagascar), Jurassic (Osteno beds of Italy, Solnhofen Limestone of Germany), Cretaceous (Sahel Alma fish beds of Lebanon), and Eocene (Monte Bolca of Italy).

Scolecodonts. Polychaete jaws, or scolecodonts (**Fig. 18**), first appear in the lower Ordovician but seem to be conspicuously more abundant in the Paleozoic than in either the Mesozoic or Cenozoic. Many are attributed to the eunicids, the jaws of which seem to have a high preservation potential in comparison to those of many other families. Such variation is due to differences in original scolecodont mineralogy, which are still imperfectly understood. However, it seems clear that composition varies taxonomically, with different families having scolecodonts either composed of organic material (scleroproteins) or having this organic matrix mineralized to varying degrees by either aragonite or calcite. In Recent polychaetes, for example, eunicids and onuphids have calcite mineralization and lumbrinerids have aragonite, while arabellids and glycerids are

Fig. 16. *Canadia spinosa* from the Middle Cambrian Burgess Shale of British Columbia. Note the pair of anterior tentacles, broad dorsal chaetae extending across much of the trunk, and ventral chaetae extending from either margin.

unmineralized. It also appears that the composition of scolecodonts has changed during time. Paleozoic scolecodonts were predominantly organic and strongly sclerotized, whereas calcareous mineralization became widespread in the Mesozoic. This mineralogical shift may explain the scarcity of post-Paleozoic scolecodonts, given that aragonite is relatively unstable, and most scolecodonts are extracted in the laboratory by acid digestion. Work on preservation potential, including survival during digestion by predators and resistance to decay in the sediment, is improving understanding, but the correlation between expected survival according to composition and field observation is by no means exact.

Other fossil records. Apart from scolecodonts, the hard-part record of polychaetes consists of tubes, especially the calcareous serpulids and related spirorbids. Spirorbids first appear in the Ordovician, and both groups are a frequent component of many Mesozoic and Cenozoic assemblages. They may occur as encrusters on reefs, hard grounds, and shell debris, or as free-living on sandy substrates. Numerous groups of polychaetes inhabit tubes that are variously reinforced so as to give a wide spectrum of fossilization potential. Most important in the fossil record are the terebellids, sabellariids, and pectinariids, all of which construct an agglutinated tube bound together by mucous secretions. This may be of mud, sand grains, or other selected particles such as foraminiferal tests, bivalve shells, echinoderm ossicles, and fish scales. Assignment to particular groups largely depends on comparisons with living relatives. However, a remarkable specimen from the Miocene of New Zealand shows a soft-bodied sabellid within its tube. In the Paleozoic a variety of tubicolous fossils have been attributed to polychaetes, but many of these are of questionable status.

It has also been customary to attribute many burrows and other trace fossils to polychaete activities, but in few cases are such assignments secure because a variety of vermiform phyla are capable of making effectively indistinguishable traces, but trace fossils (*Walcottia*) from the Upper Ordovician of Cincinnati are so similar to polychaetes that an assignment to this group seems secure. Characteristic borings into hard substrates, such as by spionid polychaetes, are also diagnostic. A unique specimen from the Middle Devonian of New York has a spionidlike polychaete still preserved in its boring. *See* TRACE FOSSILS.

Oligochaetes and leeches. The fossil record of oligochaetes is lamentable, in part because their predominantly terrestrial and fresh-water habitats have a relatively poor rock record. Nevertheless, the reproductive cocoons characteristic of the clitellates (oligochaetes and leeches) have been recognized as far back as the Triassic and may be more common as fossils than is generally realized. In addition, there is a record of oligochaete trace fossils in the form of burrows and associated fecal pellets from ancient soils (paleosols). The most convincing of these come from Cenozoic paleosols, especially in the western interior of the United States. However, the oldest records appear to date back to the Triassic.

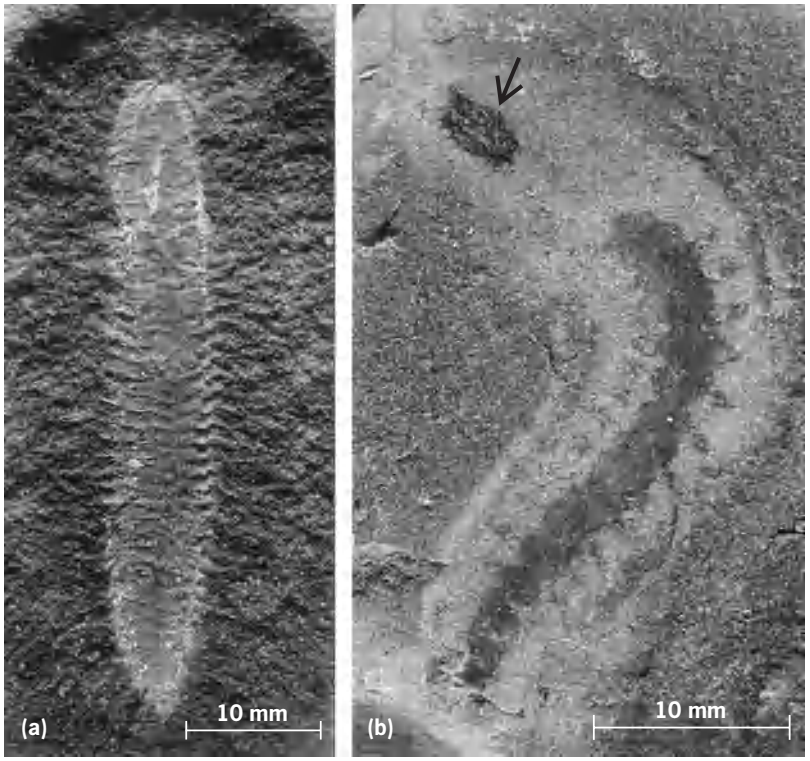


Fig. 17. Polychaetes from the Francis Creek Shale (Mazon Creek ironstone nodules). (*a*) *Astreptoscolex anasillosus*, a nephytid; note the marginal chaetae and cuticular creases. (*b*) *Esconites zelus*, a glycerid; the body is comparatively poorly preserved, but the anterior scolecodont apparatus (arrow) is in place.

Fig. 18. Scolecodont element (*Kettnerites martinssoni*) from the Silurian of Gotland, Sweden. (*Courtesy of Claes Bergman, Kristianstad, Sweden*)

The leeches are best known for their blood-sucking abilities, but they also include a number of predators. Apart from the cocoons, they are poorly represented in the fossil record. A tentative record from the Lower Silurian of Wisconsin is based on the putative terminal sucker of an annulated soft-bodied worm. Much younger worms from the Upper Jurassic Solnhofen Limestone are also claimed, on slender evidence, to represent leeches.

Echiurans. Molecular data indicate that the echiurans, long thought to be related to the Annelida, may nest within the Polychaeta. The fossil record, however, is restricted to examples from the upper Carboniferous (Mazon Creek) and, more speculatively, various traces attributed to the activity of the characteristic proboscis.

Pogonophorans. Pogonophorans were once regarded as a separate phylum, but evidence (including molecular) now indicates that they nest within the annelids. Soft-bodied remains are not known, but certain tubicolous fossils have been assigned with varying degrees of confidence to the pogonophorans. Some are associated with hydrothermal deposits, ancient analogues of the midocean ridge "black smokers." *See* FOSSIL. S. Conway Morris

Bibliography. D. E. G. Briggs and E. N. K. Clarkson, The first tomopterid, a polychaete from the Carboniferous of Scotland, *Lethaia*, 20:257–262, 1987; G. K. Colbath, Jaw mineralogy in eucinean polychaetes (Annelida), *Micropaleontology*, 32:186–189, 1986; S. Conway Morris, Middle Cambrian polychaetes from the Burgess Shale of British Columbia, *Phil. Trans. Roy. Soc. London*, B285:227–274, 1979; S. Conway Morris and J. S. Peel, Articulated halkieriids from the Lower Cambrian of North Greenland and their role in early protostome evolution, *Phil. Trans. Roy. Soc. London*, B347:305–358, 1995; D. Jones and I. Thomson, Echiura from the Pennsylvanian Essex fauna of northern Illinois, *Lethaia*, 10:317–325, 1977; S. B. Manum, M. N. Bose, and R. T. Sawyer, Clitellate cocoons in freshwater deposits since the Triassic, *Zool. Scripta*, 20:347–366, 1991; N. Kotake, Deep-sea echiurans: Possible producers of *Zoophycos*, *Lethaia*, 25:311–316, 1992; I. Thompson, Errant polychaetes (Annelida) from the Pennsylvanian Essex fauna of northern Illinois, *Palaeontographica*, Abteilung A, 163:159–199, 1979.

Anomopoda

An order of fresh-water branchiopod crustaceans, formerly included in the Cladocera. Exceptionally these organisms may be as much as 6 mm (0.24 in.) in length, but often are less than 1 mm (0.04 in.). So-called water fleas of the genus *Daphnia* are the most familiar anomopods. *Daphnia* and its close relatives swim freely in open water; however, many anomopods are benthic, predominantly crawling species.

The short trunk is protected by a functionally bivalved carapace that covers the five or six pairs of trunk limbs and the terminal postabdomen that bears two claws. The trunk limbs are extremely diverse in structure, both among themselves and between species. They collect food, usually particulate, by scraping or filtration or both, and are used by many benthic species for crawling. The postabdomen is often used for pushing. Swimming is by means of antennae. The short antennules (first antennae) are sensory. There is a single median sessile eye, derived by fusion of two eyes.

Most species feed on minute particles, but *Pseudochydorus* is a scavenger, and *Anchistropus* has the remarkable habit of living parasitically on *Hydra*, against whose stinging cells it is protected. The mandibles are of the rolling, crushing type.

Reproduction is mainly by parthenogenesis, eggs and young being carried dorsally in a brood pouch beneath the carapace. Some races of some species of *Daphnia* are obligate parthenogens. Cyclical parthenogenesis is common in temperate zone species; summer parthenogenesis alternates with sexual reproduction in autumn. Fertilized eggs are cast off, protected by a modified portion of the carapace, which is often called an ephippium, and the eggs can overwinter or resist drought. Worldwide in distribution, anomopods are among the most abundant and successful of all fresh-water invertebrates. *See* BRANCHIOPODA. Geoffrey Fryer

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Anopla

A class of the phylum Rhynchocoela which is divided into the orders Palaeonemertini and Heteronemertini. The simple tubular proboscis lacks stylets and resembles the body wall in structure. The mouth is posterior to the brain. The nervous system lies either immediately below the epidermis or among the musculature of the body wall. The vascular system is well developed.

Anoplan rhynchocoelans are common on rocky shores, living beneath stones or interstitially in shell debris or coarser sands. They are carnivorous, preying almost entirely on annelids, but will also take carrion if this is not too decomposed. Acid secretions kill or immobilize prey in the foregut, and digestion occurs rapidly in the intestine by extra- and intracellular processes. *See* ENOPLA; HETERONEMERTINI; PALAEONEMERTINI; NEMERTEA. J. B. Jennings

Anoplura

A small group of insects usually considered to constitute an order. They are commonly known as the sucking lice. All are parasites living in the covering hair of mammals. About 250 species are now known, and these comprise probably about half of the species in the world, but the group is known in detail to few persons.

Morphology. These lice are distinguishable from the Mallophaga by their mouthparts and manner of feeding. The mouthparts consist of three very slender stylets which form a tube. When at rest, they are retracted in a pocket which lies just behind the mouth. The antennae are usually five-segmented, rarely three-segmented. The thoracic segments are always very closely fused and lack wings. The claw is one-segmented, and on at least one pair of legs this claw is enlarged and can be folded into a process from the tibia to grasp a hair of the host. In some species two and in others all of the legs are thus modified. The ovipositor is very much reduced. It consists of but little more than two flaps which are able to close around a hair. Eyes are commonly lacking, and in those species in which they do occur, they are reduced to a pair of simple lenses or two light-receptive spots. All the species are quite small;

the largest scarcely exceeds 0.2 in. (5 mm) in length and the smallest does not attain 0.04 in. (1 mm).

Parasitic activity. The eggs are attached to the hairs of the host by a glue which issues from the base of the ovipositor and envelops a hair and one end of the egg. The young escape from the egg through a lid or operculum at the free end. In general, they are much like the adult and immediately begin feeding upon the blood of the host. Feeding is accomplished by thrusting forth the tube formed by the stylets, piercing the skin, and sucking blood by a pump in the throat of the insect. Very few species have been studied closely enough to determine their life-span, but in the human louse this is about a month.

Disease transmission. This habit of sucking blood gives the Anoplura a special importance. They take up any disease-producing organisms in the host's blood and may transfer these organisms to another individual. Thus, they transfer the organisms which cause two of the most important diseases of humans, epidemic typhus and relapsing fever. A third disease, trench fever, which gained great prominence during World War I, has also been of very great importance. The transmission of these diseases depends entirely upon certain habits of the insects, which are coincident with certain habits of their hosts. *See* RELAPSING FEVER.

Host specificity. Indeed, these lice have become so closely adapted to life upon a particular kind of mammal, both physiologically and physically, that they can scarcely move about except when they are surrounded by its hair or body covering, and they can subsist only on its blood for any length of time. The transfer of a louse, from one host to another, can occur only when the hosts are in close bodily contact, as in the nest or at the time of mating. In the case of humans, the lice may become detached and enter the clothing or the bedding. In times of social disturbance, such as war or when people are crowded together as formerly occurred on ships, or in jails or slums, the opportunities for an exchange of these parasites are enormously increased. Under these conditions, in certain parts of the world louse-borne diseases may become epidemic and produce tremendous mortality. The close restriction of each species of lice to its special host is sufficient to account for the fact that diseases are not transferred from one species of mammal to another. This has made it possible to prevent the epidemic spread of human diseases transmitted by lice by employing methods of cleanliness and destroying the lice. One epidemic disease of rodents that is carried by lice is known from India, and there are in all probability others.

Human species. The degree to which lice are restricted to one kind of mammal or to a few closely related kinds may be indicated in the following discussion. There are two species of lice which occur as common ectoparasites upon humans. One of them is *Pediculus humanus*, the head and body louse which transmits the diseases mentioned above (**Fig. 1**). Apparently it sometimes occurs upon gibbons in zoos and has been recorded from certain

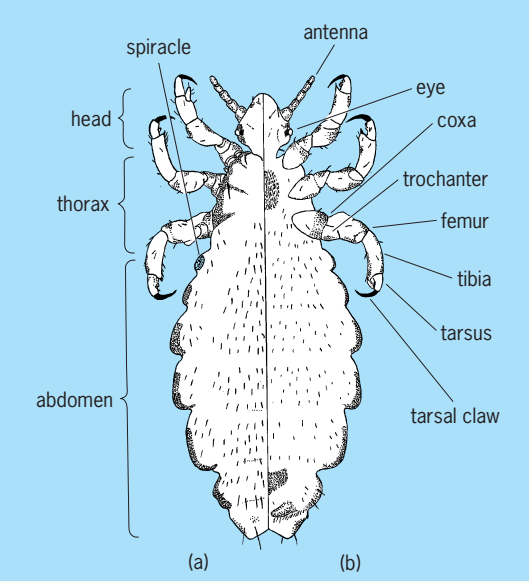


Fig. 1. The louse *Pediculus humanus*, female, shown in (*a*) dorsal and (*b*) ventral views.

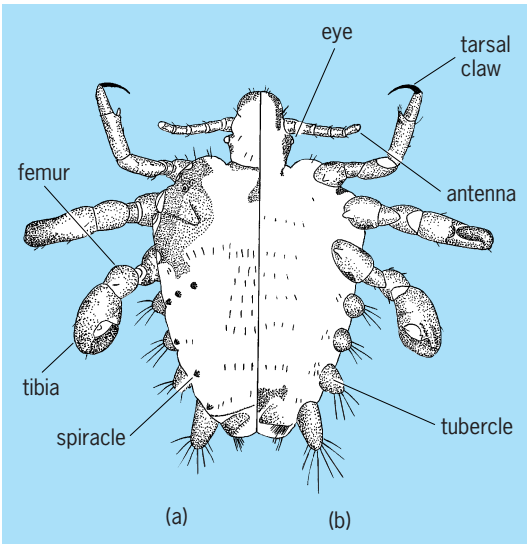


Fig. 2. The louse *Phthirius pubis*, female, shown in (*a*) dorsal and (*b*) ventral views.

New World monkeys. Other quite distinct species of the genus *Pediculus* are known from the chimpanzee and from New World monkeys. The other species occurring on humans is *Phthirius pubis*, known as the crab louse, and it is restricted as far as is known to humans, although a second species has been described from the gorilla (**Fig. 2**). It transmits no known disease. The lice of the Old World monkeys are referred to another genus, *Pedicinus*. *See* INSECTA. Gordon F. Ferris; Dwight M. DeLong

Anorexia nervosa

A psychiatric disorder in which a dramatic reduction in caloric intake consequent to excessive dieting leads to significant bodily, physiological,

biochemical, emotional, psychological, and behavioral disturbances.

Clinical presentation. Anorexia nervosa is typically an illness of adolescent females: 90% of all cases begin in girls who are between 12 and 20 years of age. Nevertheless, this disorder can also occur in males, in prepubertal girls, and in women well into their third decade. Moreover, if the illness becomes chronic, it can persist into mid-life and beyond. Estimates of the incidence of anorexia nervosa are limited but indicate that about 1–5% of adolescent girls living in industrialized cultures are affected, and that the rate is rising.

Anorexia nervosa literally means "nervous loss of appetite" but appears to have little to do with such. Rather, there is usually a conscious decision made by a teen-age girl, most commonly around the ages of 14 or 18 years, to embark upon a diet. This decision often occurs in the context of some stressful life change situation. In addition, there are usually two personality characteristics: a morbid fear of losing control over one's weight (the majority of girls who develop anorexia nervosa usually are slightly overweight prior to the illness' onset) and a distorted body image (that is, the perception of oneself as larger than one is in reality).

The amount of weight lost can vary considerably. The usual criterion for making a diagnosis of anorexia nervosa is a weight change of at least 25% from premorbid weight (to at least 15% below ideal weight in persons who were overweight at the onset). But this figure should be viewed as only a rule of thumb; some persons may be affected by the disorder and do not reach the 25% weight loss. In the most extreme cases, weight losses of up to 50% and more can occur.

In addition to these core disturbances, there is an array of associated symptoms that characterize most persons with anorexia nervosa. Amenorrhea, or absence of menstruation, occurs in 95–100% of anorectic women; about one-third of the cases of amenorrhea actually precede the onset of substantial weight loss. Increased physical activity, such as excessive running, swimming, or calisthenics, is noticeable in a majority of cases; early in the course of the illness it is associated with a sense of well-being and is goal-directed, but later it becomes more obsessional and frenetic. Insomnia, whether secondary to the starvation or the emotional turmoil characteristic of the disorder, is a frequently reported symptom; in some cases, it can lead to an abuse of sedatives or tranquilizers. Use of emetics, cathartics, and diuretics is not uncommon in women who suffer from anorexia nervosa for at least several months. These practices are employed in an attempt to minimize the absorption of food, to rid the body of weight associated with body fluids, and to symbolically cleanse the self of "alien" substances. Difficulty in recognizing satiation is a common complaint, and if the illness becomes chronic, sensations of hunger may also become difficult to discriminate. Obsessional thinking and depression are not unusual in anorectic women; whether these symptoms precede the illness is not always clear, but they become increasingly apparent with chronicity.

The course and prognosis of anorexia nervosa is highly variable. While a high recovery rate, perhaps above 67%, is found in those persons whose illness begins acutely in their early teens and who quickly receive treatment, the outlook is considerably bleaker in those persons who develop the disorder later, who do not receive early treatment, and who develop bulimia. For all individuals with anorexia nervosa, the recovery rate is probably around 40%, with another 30% showing moderate improvement and the remaining 30% running a debilitated, chronic course. Included in the last group is the 5–10% of those who actually die from such complications of the disorder as cardiovascular collapse, infection, and suicide.

Bulimia. Perhaps as many as 40–50% of anorectic patients whose illness persists for more than 1–2 years will develop the additional eating disturbance known as bulimia. Persons whose anorexia starts relatively late, for example, in their twenties, may be particularly vulnerable to becoming bulimic. Literally meaning "oxhunger," bulimia refers to binge eating or compulsive overeating wherein thousands of calories are consumed in a relatively brief period of time (for example, 2 h). The episodes usually occur in the evening in private and may be no more frequent than a few times per month. However, they can occur several times per day; the average rate is around several times per week. The binge characteristically involves carbohydrates—the very type of food that the anorectic so rigidly avoids the rest of the time—and will usually culminate in self-induced vomiting. There is usually enormous shame and self-loathing associated with bulimia; whereas dieting is equated with control and strength, binging has the opposite implication. Patients characteristically wake up the next morning with even greater resolve to diet, initiating a fast–feast cycle.

The precise nature of the relationship of bulimia to anorexia nervosa remains unclear. Not all anorectics become bulimic and not all bulimics were anorectic. But dieting is common as a precursor to bulimia, and high premorbid weight and chronicity of weight loss seem to predispose the anorectic to developing bulimia. There is some indication that anorectic persons who develop bulimia are also more prone to depression (and often have a family history of depression), impulsive behavior (for example, shoplifting), and more outgoing but chaotic social relationships than those persons who remain "pure restricting" anorectics.

A major concern about bulimia relates not to the overeating per se but to the subsequent vomiting. The chronic loss of gastric juices can lead to a depletion of the body's potassium, which in turn can seriously affect cardiac function (to the point of death). In general, the onset of bulimia is believed to signify a poor prognosis.

Psychological aspects. Although there is a broad range of symptoms, personality styles, precipitants, and outcomes that characterize anorexia nervosa, there is no simple explanation of its origins. In one

widely accepted conception, the illness is viewed as a desperate struggle by the vulnerable female adolescent to establish a sense of identity separate from that of her domineering, overbearing, controlling, and intrusive mother. By virtue of the mother's insensitivity to the cues of her child, such as hunger as an infant, the future anorectic has not received clear-cut confirmation of her internal experiences. This, coupled with the mother's overintrusiveness and domination, leaves the developing child with serious deficits in her sense of self and in her ability to identify and discriminate feelings and bodily sensations. Her body image is disturbed, and she feels a potentially paralyzing sense of ineffectiveness. In the context of the typical stresses of puberty and adolescence, for example, bodily changes of portending adult responsibility and sexuality, separation from mother and home, the vicissitudes of relations with peers, social and sexual encounters with the opposite sex, and the demands of school or work—the preanorectic, who is already sensing high expectations from her achievement-oriented family, turns toward control of her body weight as the only area left in her life in which she can achieve fully independent and effective control.

There is considerable emphasis on viewing the family system as the matrix in which the illness develops and for which the illness must have significance. Four characteristics of the families of anorectics are distinguishable: excessive mutual enmeshment, mutual overprotectiveness, rigidity in relation to internal problems and the external world, and an inability to achieve resolution of conflict. Anorexia is viewed as both a response to the lack of "living space" that the adolescent experiences and as a defense against the threats to the stability of the family system that the girl's normal development implies. Development of anorexia is thus a function of both individual and family.

It should be noted, however, that these formulations suffer from a common problem. They are based on assessments of anorectic patients—and their families—after the illness has become established. Thus, these theories cannot very well differentiate among predisposing, precipitating, and sustaining factors. Moreover, many of the problems and conflicts emphasized in anorectics—such as those concerning separation and independence, appearance, adult heterosexual relations—are no different from those of normally developing adolescents. And the more pathological features described in these individuals and their families—for example, feelings of emptiness, precarious sense of self, family enmeshment and rigidity—have been reported in other pathological but nonanorectic psychiatric conditions such as schizophrenia, borderline personality, and psychosomatic disorders. *See* SCHIZOPHRENIA; STRESS (PSYCHOLOGY).

Biological aspects. There are numerous physical, physiological, and biochemical changes that reflect primarily, although not solely, the ravages of starvation. In addition to the general bodily emaciation they manifest, anorectic individuals show brittle nails and thinning hair. Cold extremities, a slow pulse, a small heart, and a hypometabolic state, in compensation for the starvation, are also characteristic. In advanced cases, edema may occur. In anorectic women who also binge and vomit, tooth decay and enlargement of the salivary glands are common. A mild-to-moderate anemia and a diminished white blood cell count develop with progressive malnutrition. Diabetes insipidus, consequent to disturbed hypothalamic function, can also occur in advanced cases, producing large amounts of dilute urine and constant thirst. In addition, abnormalities in glucose tolerance and blood electrolytes have been observed. Particularly in women who vomit, the blood potassium level can be significantly low, possibly producing abnormalities on the electrocardiogram. A small percentage of anorectic subjects show vague abnormalities in the electroencephalogram. Interestingly, these abnormalities of electroencephalogram are not always clearly correlated with the extent of emaciation. *See* MALNUTRITION; PITUITARY GLAND DISORDERS.

Because of the prominence of amenorrhea in its symptomatology, there has been a long-standing interest in the endocrinology of anorexia nervosa. A number of reliable hormonal abnormalities have been documented. Most prominent among these are diminished hypothalamic–pituitary–gonadal axis function and elevated hypothalamic–pituitary–adrenal axis function. While urinary and plasma gonadotropin levels are low and their circadian and monthly cycles are deficient, plasma and urinary levels of cortisol are high and usually will not be suppressed by dexamethasone (a synthetic adrenal steroid) administration. These abnormalities point to hypothalamic dysfunction, presumably consequent to weight loss and starvation. Possibly they also reflect functional changes secondary to affective disturbances (for example, depression), behavioral changes (for example, excessive exercise, altered sleep patterns), or some preexisting biological vulnerability. Reversal of these endocrine aberrations usually occurs with clinical improvement, although considerable time may be required for full normalization. *See* ENDOCRINE MECHANISMS.

Treatment. As with virtually all psychiatric conditions for which the etiology is unknown and where no single empirically effective treatment exists, the therapeutic approaches to anorexia nervosa are diverse and reflect the different disciplines, training biases, and experiences of their proponents.

Nevertheless, there are some general considerations in the management of this illness. First, there are probably mild cases that resolve without any intervention whatsoever. Second, despite this, anorexia nervosa is a potentially pernicious illness and, once established, can be strikingly resistant to all treatments; hence, early intervention is called for once the diagnosis is made. Third, the approach to treatment must be tailored to the stage of the illness.

In the earliest cases and youngest children, the pediatrician's involvement in a supportive, gently inquiring, and educational fashion may be sufficient.

If no improvement is seen within a few weeks, the child may be referred to a psychiatrist, but the pediatrician should remain involved. Individual, insight-oriented psychotherapy directed toward increasing confidence in identifying and accepting bodily feelings, understanding the sources of one's low self-esteem and poor sense of self, and "trying out one's wings" in the context of a reliable, supportive, respectful, and inquiring, but nonintrusive, relationship has generally been the essence of treatment for nonhospitalized individuals, particularly before chronicity has set in. Individual psychotherapy still remains a critical part of any approach to treatment, but the recovery rate can be increased, perhaps beyond 80%, by the inclusion of regular family therapy as part of the treatment approach. This implies the willingness of all the family members to participate and requires the individual to be young enough to still be involved with her family system. Hospitalization must be seriously considered whenever weight loss becomes substantial, exercising or food-related rituals begin to interfere with day-to-day functioning, or the individual begins to appear physically weak or emotionally hopeless. *See* PSYCHOTHERAPY.

In the hospital, the first priority of treatment is directed toward correcting the biological abnormalities created by the extreme dieting (and, when present, the vomiting). If the individual appears unable or unwilling to resume adequate caloric intake, despite firm but supportive nursing and concomitant individual and family psychotherapy, more extreme measures may have to be instituted, including behavior modification or intravenous hyperalimentation.

The behavior modification approach assumes that the food avoidance, whatever its original determinants, has become a habit that can best be treated by appropriate positive and negative reinforcement. Thus, a schedule is mapped out which progressively "punishes" weight loss with confinement to bed, loss of visitors, and so on, while "rewarding" weight gain with off-ward privileges, access to exercising, and such.

While intravenous fluids can correct electrolyte imbalance, they cannot provide sufficient calories and nutrition to achieve weight gain. However, the availability now of total parenteral nutrition, that is, hyperalimentation with concentrated nutrients via a peripheral or a central large intravenous catheter, has provided a technique for reversing the severe emaciation caused by advanced anorexia nervosa.

There is some evidence that, particularly in anorectics who are also characterized by depression and bulimia, antidepressant or possibly anticonvulsant medication may be helpful in damping down the binging and thereby gradually normalizing eating behavior in general. Neuroleptics (such as chlorpromazine) can also be helpful in the acute hospital management, particularly in the anxious or excessively exercising patient, but their beneficial effects are probably only indirect, that is, secondary to general tranquilization. *See* PSYCHOPHARMACOLOGY; TRANQUILIZER.

Anorexia nervosa can become a chronic illness for many of its victims. Improvement in the hospital does not mean that care need not be provided after discharge. Moreover, while a certain percentage of individuals do make a complete recovery, the real therapeutic challenge is in working with the chronic anorectic victim over many years in an attempt to minimize the morbidity imposed by the emaciation, medical complications, depression, and social isolation which often occur. *See* AFFECTIVE DISORDERS.

Jack L. Katz

Bibliography. H. Bruch, *Eating Disorders: Obesity, Anorexia Nervosa, and the Person Within*, 1979; H. Bruch, *The Golden Cage: The Enigma of Anorexia Nervosa*, 1978; P. E. Garfinkel and D. M. Garner, *Anorexia Nervosa: A Multidimensional Perspective*, 1982; S. Minuchin, B. Rosman, and L. Baker, *Psychosomatic Families: Anorexia Nervosa in Context*, 1978; J. A. Sours, *Starving to Death in a Sea of Objects: The Anorexia Nervosa Syndrome*, 1980.

Anorthite

The calcium-rich plagioclase feldspar with composition Ab_0An_{100} to $Ab_{10}An_{90}$ (Ab = $NaAlSi_3O_8$; An = $CaAl_2Si_2O_8$), occurring in olivine-rich igneous rocks and rare volcanic ejecta (for example, at Mount Vesuvius, Italy, and Miyakejima, Japan). Hardness is 6 on Mohs scale, specific gravity 2.76, melting point 1550°C (2822°F). The crystal structure of Ab_0An_{100} (triclinic space group $P\bar{1}$) consists of an infinite three-dimensional array of corner-sharing [AlO_4] and [SiO_4] tetrahedra, alternately linked together in a framework of $[Al_2Si_2O_8]_\infty^{2-}$ composition in which charge-balancing calcium (Ca^{2+}) cations occupy four distinct, irregular cavities. The crystal structure of anorthite inverts to higher symmetry ($I\bar{1}$) with each of the following: (1) heating above 242°C (468°F), (2) adding ~10 mole % Ab component (substituting $Na^+ + Si^{4+} \leftrightarrow Ca^{2+} + Al^{3+}$), (3) increasing hydrostatic pressure above ~2.6 gigapascals (26,000 atmospheres), and (4) effectively disordering the aluminum (Al) and silicon (Si) atoms at very high temperatures. Each inversion is evidenced by certain types of structurally out-of-step (antiphase) domains that can be imaged by transmission electron microscopy. Natural anorthite has no commercial uses, but the synthetic material [$CaO \cdot Al_2O_3 \cdot 2SiO_2$ (known as CAS2)] is important in the ceramic industry and in certain composite materials with high-temperature applications. *See* CRYSTAL STRUCTURE; FELDSPAR; IGNEOUS ROCKS.

Paul H. Ribbe

Anorthoclase

The name usually given to alkali feldspars which have a chemical composition ranging from $Or_{40}Ab_{60}$ to $Or_{10}Ab_{90}\pm$ up to approximately 20 mole % An (Or, Ab, An = $KAlSi_3O_8$, $NaAlSi_3O_8$, $CaAl_2Si_2O_8$) and which deviate in one way or another from monoclinic symmetry tending toward triclinic symmetry. When found in nature, they usually do not consist

of a single phase but are composed of two or more kinds of K- and Na-rich domains mostly of submicroscopic size. In addition, they are frequently polysynthetically twinned after either or both of the albite and pericline laws. It appears that they originally grew as the monoclinic monalbite phase inverting and unmixing in the course of cooling during geological times. They are typically found in lavas or high-temperature rocks. *See* FELDSPAR; IGNEOUS ROCKS.

Fritz H. Laves

Anorthosite

A rock composed of 90 vol % or more of plagioclase feldspar. Strictly, the rock is composed entirely of crystals discernible with the eye, but some finely crystalline examples from the Moon have been called anorthosite or anorthositic breccia. Two principal types of anorthosite are based on field occurrence: layers in stratified complexes of igneous rock, and large massifs of rock up to 12,000 mi^2 (30,000 km^2) in area. Scientists have been fascinated with anorthosites because they are spectacular rocks (dark varieties are quarried and polished for ornamental use); valuable deposits of iron and titanium ore are associated with anorthosites; and the massif anorthosites appear to have been produced during a unique episode of ancient Earth history (about 1–2 billion years ago).

Definition, occurrence, and structure. Pure anorthosite has less than 10% of dark minerals—generally some combination of pyroxene, olivine, and oxides of iron and titanium; amphibole and biotite are rare, as are the light minerals apatite, zircon, scapolite, and calcite. Rocks with less than 90% but more than 78% of plagioclase are modified anorthosites (such as gabbroic anorthosite), and rocks with 78–65% of plagioclase are anorthositic (such as anorthositic gabbro). *See* GABBRO.

The structure, texture, and mineralogy vary with type of occurrence. One type of occurrence is as layers (up to 10 ft or 3 m thick) interstratified with layers rich in pyroxene or olivine. The bulk composition of the layered rock samples containing anorthosite layers is gabbroic. The Stillwater igneous complex in Montana is an example where the lower part of the complex is rich in olivine, pyroxene, and chromite, and the upper part is rich in plagioclase. The second type of occurrence is the massif type. Commonly, the massifs are domical in shape and weakly layered. There appear to be two kinds of massif anorthosites: irregular massifs typified by the Lake St. John body in Quebec, and domical massifs similar to the Adirondack (New York) massif made classic through the studies of A. Buddington. The irregular massifs are older than the domical massifs and commonly contain olivine. The domical massifs are associated with silicic rocks of uncertain origin. *See* MASSIF.

Possibly there is a third group of anorthosite occurrences: extremely ancient bodies of layered rock in which the layers of anorthosite contain calcium-rich plagioclase and the adjacent layers are rich in chromite and amphibole in addition to pyroxene. There are only a few examples of these apparently igneous complexes, in Greenland, southern Africa, and India. However, they appear to be terrestrial counterparts of lunar anorthosites. Recrystallization and deformation have pervasively affected these rocks, and it is difficult to discover their original attributes.

Layered anorthosites. Zones of anorthosite in layered complexes are composed mostly of wellformed, moderately zoned crystals of plagioclase (labradorite to bytownite) up to several millimeters long. The tabular crystals of plagioclase commonly are aligned subparallel to the layering. Irregular crystals of pyroxene, olivine, and, rarely, magnetite are interstitial to the plagioclase crystals. These features are best explained by the hypothesis of accumulation of plagioclase crystals onto the floor or roof of the reservoir of melt, where gravity and perhaps motion of melt caused them to lie with their centers of gravity in the lowest position (like pennies spilled on the floor). After accumulation, the intergranular melt gradually crystallized as overgrowths of plagioclase and interstitial pyroxene. The separation of accumulated plagioclase crystals in anorthosite layers from pyroxene or olivine in pyroxenite and peridotite layers is a puzzle to scientists, for it is expected that plagioclase and pyroxene crystallize together in sub-equal proportions from basaltic magmas, not first one mineral, then the other and back again as suggested by the layering.

Compositional zoning of plagioclase within layers and other structural and textural features suggest that plagioclase floated in certain intrusions. This is consistent with calculated densities of some postulated liquids, if the liquids are poor in water. Evolution of the magma body may be dynamically complex at the stage of plagioclase flotation and anorthosite formation, particularly if new batches of magma arriving from below are less dense than the old differentiated magma. Certain unusual rocks, including the platinum-rich Merensky reef horizon of the Bushveld intrusion located near the transition between mafic layers and anorthositic layers, may be related to such dynamical complexities. *See* MAGMA.

Massif anorthosites. In massifs the plagioclase has an intermediate composition, mostly labradorite. The plagioclase crystals are poorly formed and practically unzoned and may be up to around 10 ft (3 m) long, but crystals about a centimeter long constitute the bulk of most massifs. Rare dark varieties have moderately well-formed plagioclase crystals (see **illus.**) filled with tiny specks of dark minerals. In many specimens the large labradorite crystals display a play of colors (blue and green are common) in a particular orientation, an attractive quality related to very slow cooling and sought after in ornamental stone. Lighter varieties have smaller crystals of plagioclase lacking inclusions of dark minerals, which instead occur as millimeter-sized interstitial grains. (Thus the various shades of anorthosite colors result from variations in the size or granularity of the pigmenting dark minerals.) In transitional varieties large dark crystals of plagioclase are set in a fine-grained

Outcrop of massif anorthosite on the shore of Pipmuacan Reservoir, Quebec. Exceptionally well-preserved texture shows well-formed relict plagioclase crystals (now recrystallized) outlined by irregular black pyroxene grains. Two dark cores of twinned plagioclase are shown (circles).

sugarlike white matrix. Many of the plagioclase grains are bent and broken. The associated dark minerals are pyroxene, olivine, magnetite, and ilmenite. Deposits of almost pure magnetite occur and have been used for iron ore, although most deposits have too much titanium to be highly desired for iron ore. Other deposits are rich in olivine, spinel, and apatite as well as magnetite. The massifs are surrounded by gneissic rocks, and the original age relationships between anorthosite and gneiss are commonly uncertain. The textural and mineralogical features of massifs are indicative mainly of extensive recrystallization and deformation at high temperature. The dark varieties have textures similar to those of layered anorthosites and presumably have a similar igneous origin, but this cannot be determined with certainty.

Many anorthosite massifs formed about 1.1–1.7 billion years ago. Possibly they mark a unique episode in the history of the differentiation of the Earth. The restricted age of massif anorthosite still awaits a convincing explanation.

Lunar anorthosites. By comparison with terrestrial occurrences, most lunar anorthosites are very fine grained, although one rock has crystals up to a centimeter long. Much of the fine grain size results from comminution by meteorite impact, and some of it probably results from rapid crystallization of impact melts. The unbrecciated lunar anorthosites have well-formed crystals of calcium-rich plagioclase (anorthite) intergrown with traces of olivine, spinel, pyroxene, ilmenite, metallic iron, and a few other exotic minerals. Associated rocks and minor minerals suggest that lunar anorthosites are of the layered type formed by accumulation of crystals of plagioclase from basaltic melts. Evidently, lunar plagioclases floated in iron-rich lunar basaltic melts, and most scientists agree that rocks rich in plagioclase and including anorthosite compose a major part of the skin of the Moon. *See* ANDESINE; LABRADORITE; MOON. Alfred T. Anderson, Jr.

Role in growth of continental crust. Large occurrences of Archean anorthosite are formed principally in back-arc basin or oceanic settings. Although intrusions of Archean anorthosite into continental crust are known, they seem limited in size, and lack of exposure further hampers evaluation of their contribution to continental growth. Based upon what is known of these rocks, it may be concluded that their contributions to Archean continents were vastly less than that of gray tonalitic gneiss and komatiite-basalt associates of the greenstone belts. *See* ARCHEAN.

In contrast to the Archean, Proterozoic anorthosites were emplaced into continental crust in large massifs usually localized within belts such as the Grenville Province of Canada or the Rogaland-Ergusund region of southern Norway. In these belts anorthositic massifs may account for as much as 20% of the local crust. Most of these bodies were emplaced during the time period 1650–1000 million years ago (Ma) and appear to be related to the slow, anorogenic break-up of a mid-Proterozoic supercontinent. Within the Grenville Province the anorthosite massifs appear to span the 500-million-year interval from 1650 to 1150 Ma, and it is probable that the interval consists of anorthositic pulses at 1650, 1450, and 1150 Ma. These rocks account for a minimum surface area of close to 39,000 mi^2 (100,000 km^2), or about 13% of the crustal area of the Grenville Province. If this same concentration is assumed downward throughout the continental crust, a maximum value for anorthositic additions to continental crust may be computed. For a crust 24 mi (40 km) thick this results in 1×10^6 mi^3 (4×10^6 km^3) of anorthositic addition to the regional crust over 500 million years, or an average rate of crustal addition of 0.2 mi^3/year (0.8 km^3/year). This is about 150–250 times less than the 0.3 mi^3/year (1.2 km^3/year) rate estimated for present-day arc accretion or the 0.2–0.4 mi^3/year (1–2 km^3/year) estimates for crustal accretion during the interval 1900–1700 Ma. The belt of anorthosite massifs in the Grenville Province is about 900 mi (1500 km) in length. This gives an accretion rate per mile (kilometer) of magmatic belt of about 2 $mi^3/10^6$ years (5 $km^3/10^6$ years) as compared with about 12 $mi^3/10^6$ years (30 $km^3/10^6$ years) per mile of magmatic belt for modern arcs. If the anorthosite massifs are continued downward for only 3 mi (5 km) into the crust, their rate of crustal addition becomes about 0.2 $mi^3/10^6$ years (1 $km^3/10^6$ years) per mile of magmatic belt. This is a small fraction of other crustal growth mechanisms, but it is far from insignificant. It may be concluded that within the mid-Proterozoic the emplacement of anorogenic, massif anorthosites constituted a small but significant means of continental crustal growth. *See* CONTINENTS, EVOLUTION OF; IGNEOUS ROCKS; METAMORPHISM; PROTEROZOIC. James McLelland

Bibliography. L. D. Ashwal, *Anorthosites, Minerals and Rocks*, vol. 21, Springer-Verlag, 1993; R. B. Hargraves (ed.), *Physics of Magmatic Processes*, 1980; Y. Isachsen (ed.), *Origin of Anorthosite and Related Rocks*, N.Y. State Mus. Sci. Serv. Mem. 18, 1969; A. R. McBirney, *Igneous Petrology*, 2d ed., 1992; J. McLelland, Crustal growth associated

with anorogenic, mid-Proterozoic anorthosite massifs of northeastern North America, *Tectonophysics*, 161:331–334, 1989; B. Mason and W. G. Melson, *The Lunar Rocks*, 1970; A. Reymer and G. Schubert, Phanerozoic addition sites to the continental crust and crustal growth, *Tectonics*, 3:63–77, 1984.

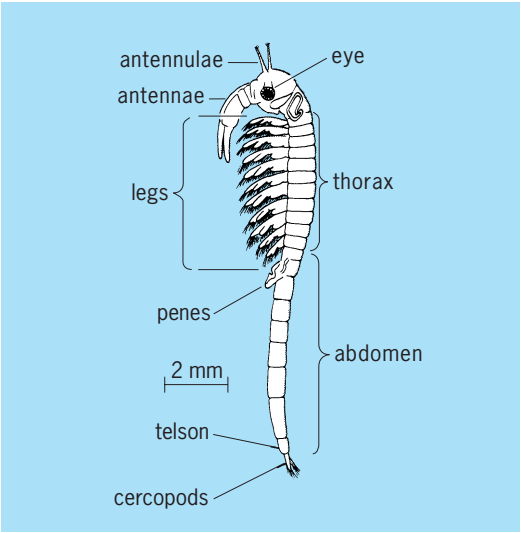
Anostraca

An order of branchiopod crustaceans, known as fairy shrimps and, in some cases, brine shrimps (Artemia). These organisms range up to about 4 in. (100 mm) in length, but usually are much smaller. The trunk consists of 19 to 27 segments, of which usually the first 11, sometimes the first 17 or 19, bear limbs, plus a telson bearing flattened furcal rami (see **illus.**). The trunk limbs are foliaceous, and all are of the same basic type but differ a little among themselves; each limb is differentiated into a series of endites and is usually filtratory. By beating in metachronal rhythm they propel the animal forward and draw food particles toward it. Anostracans usually swim with their ventral surface uppermost.

There is no carapace. The eyes are pedunculate. The antennae are large and modified as claspers in the male. The large mandibles are of the rolling, crushing type. The other mouthparts are very small.

Most anostracans subsist on minute particles that they sieve from the water with their trunk limbs and pass forward to the mouthparts. A few large species become carnivores as they grow, seizing small crustaceans, including in some cases smaller species of fairy shrimps.

Reproduction is usually bisexual, but some brine shrimps are parthenogenetic. Females pass eggs into a ventral brood pouch and then shed them. Eggs are drought-resistant. They hatch as nauplii which develop by a series of molts, gradually adding segments posteriorly.



***Branchinecta paludosa*, male, small specimen, lateral aspect.**

Fairy shrimps frequent fresh water, usually temporary pools, in all regions of the world, but are also found in predator-free waters in Arctic and Antarctic regions. Brine shrimps also escape predation by living in hypersaline habitats such as salt lakes and salt pans to which they have become physiologically adapted. *See* BRANCHIOPODA.

Geoffrey Fryer

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Anoxic zones

Oxygen-depleted regions in marine environments. The dynamic steady state between oxygen supply and consumption determines the oxygen concentration. In regions where the rate of consumption equals or exceeds the rate of supply, seawater becomes devoid of oxygen and thus anoxic. In the open ocean, the only large regions that approach anoxic conditions are in the equatorial Pacific between 50 and 1000 m (165 and 3300 ft) depth and in the northern Arabian Sea and the Bay of Bengal in the Indian Ocean between 100 and 1000 m (330 and 3300 ft) depth. The Pacific oxygen-depleted region consists of vast tongues extending from Central America and Peru nearly to the middle of the ocean in some places. In parts of this zone, oxygen concentrations become very low, 15 micromoles per liter (atmospheric saturation is 200–300 μmol/L). Pore waters of marine sediments are sometimes anoxic a short distance below the sediment–water interface. The degree of oxygen consumption in sediment pore waters depends upon the amount of organic matter reaching the sediments and the rate of bioturbation (mixing of the surface sediment by benthic animals). In shallow regions (continental shelf and slope), pore waters are anoxic immediately below the sediment-water interface; in relatively rapid sedimentation-rate areas of the deep sea, the pore waters are usually anoxic within a few centimeters of the interface; and in pore waters of slowly accumulating deep-sea sediments, oxygen may never become totally depleted. *See* MARINE SEDIMENTS.

Restricted basins (areas where water becomes temporarily trapped) are often either permanently or intermittently anoxic. Here, the anoxia can be a consequence of natural conditions or anthropogenic causes such as eutrophication, or a combination of them. Classic examples are the Baltic Sea, Black Sea, Gulf of Mexico, Carioca Trench off the coast of Venezuela, and fiords on the Norwegian and British Columbia coasts. Lakes that receive large amounts of nutrients or organic matter (either from natural or human-produced sources) are often anoxic during the period of summer stratification. *See* BALTIC SEA; BLACK SEA; EUTROPHICATION; FIORD; GULF OF MEXICO.

The chemistry of many elements dissolved in seawater (particularly the trace elements) is vastly changed by the presence or absence of oxygen. Since large areas of the ocean water mass are in contact with oxygen-depleted pore waters, the potential

exists for anoxic conditions to have a marked effect on the chemistry of the sea. *See* SEAWATER.

Reduction-oxidation (redox) reactions. Organic matter is formed directly or indirectly through photosynthesis, which is driven by the energy of the Sun. Carbon dioxide, CO_2, the stable form of combined carbon, is here converted to energy-rich organic carbon compounds. The required electrons for this reduction are provided by water (H_2O) through the release of elemental oxygen. Organic carbon is a very unstable form of carbon and reacts with various oxidants, yielding back the energy required for its photosynthesis. The most common of these reactions is the oxidation of organic carbon. The electrons are transferred back from carbon to oxygen, yielding CO_2 and H_2O as final products. This reaction fuels animal life on land and in the sea. In the absence of oxygen, organic matter reacts with various other oxidants (that is, electron acceptors) to form CO_2 and to gain energy. *See* PHOTOSYNTHESIS.

The major redox reactions that occur in natural systems are given below. They are listed in the order of decreasing gain of free energy (2 and 3 are interchangeable). For simplicity, CH_2O is used to represent organic carbon.

1. Oxygen reduction:

$$O_2 + CH_2O \longrightarrow CO_2 + H_2O$$

2. Nitrate reduction (denitrification):

$$4NO_3^- + 5CH_2O + 4H^+ \longrightarrow 5CO_2 + 2N_2 + 7H_2O$$

3. Manganese(IV) reduction:

$$2MnO_2 + CH_2O + 4H^+ \longrightarrow CO_2 + 2Mn^{2+} + 3H_2O$$

4. Iron(III) reduction:

$$2Fe_2O_3 + CH_2O + 8H^+ \longrightarrow CO_2 + 4Fe^{2+} + 5H_2O$$

5. Sulfate reduction:

$$SO_4^{2-} + 2CH_2O + 2H^+ \longrightarrow 2CO_2 + H_2S + 2H_2O$$

6. Fermentation:

$$2CH_2O \longrightarrow CH_4 + CO_2$$

The reaction with oxygen is the most energy-yielding and continues until O_2 is nearly entirely consumed. Nitrate, NO_3^-, and manganese, Mn(IV), reduction follow the consumption of oxygen. NO_3^- generally exists in low concentrations, $<100\ \mu$mol/L, in seawater depleted in oxygen, but Mn(IV) is insoluble and exists mainly as a solid manganese oxide, MnO_2. The importance of the latter redox reaction thus depends upon the proximity of organic carbon and MnO_2 and, as a result, the reaction will be more important in sediments than in water. The same is true for the iron–organic matter reaction, because Fe(III) is insoluble in oxic waters and exists as an oxide or hydroxide (Fe_2O_3 or FeOOH). The sulfate reaction is very important in seawater because of its high concentration (28,000 μmol/L). In a system closed from the atmosphere, sulfate represents the largest oxidant reservoir in seawater (without sediments). In anoxic parts of the water column, such as in the Baltic Sea or Black Sea or as shown in **Fig. 1***a* for a fiord in British Columbia, the concentrations of hydrogen sulfide, H_2S, increase soon after oxygen and nitrate are depleted. In anoxic water or sediment

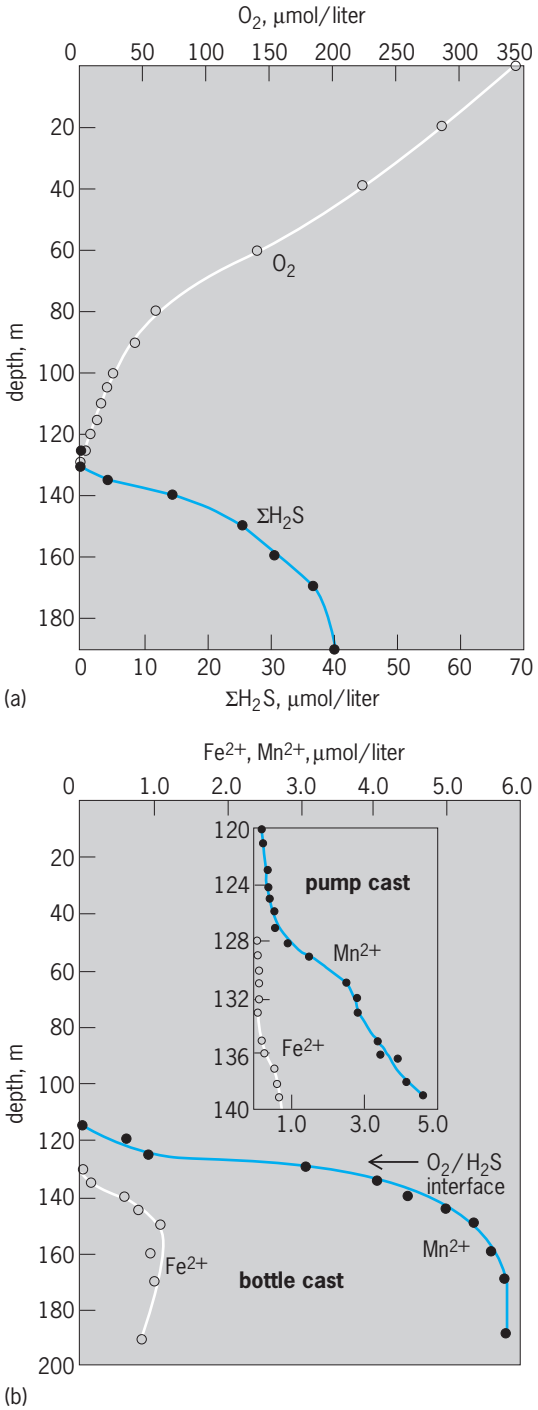


Fig. 1. Water column profiles from Saanich Inlet, an intermittent anoxic fiord in British Columbia, Canada. (*a*) Oxygen and hydrogen sulfide profiles showing the interface at 130 m (430 ft). (*b*) Dissolved iron and manganese at the same location. Pump cast indicates a more closely spaced profile at the location. 1 m = 3.3 ft. (*After S. Emerson, R. Cranston, and P. Liss, Redox species in a reducing fiord: Equilibrium and kinetic considerations, Deep Sea Res., 26A:859-878, 1979*)

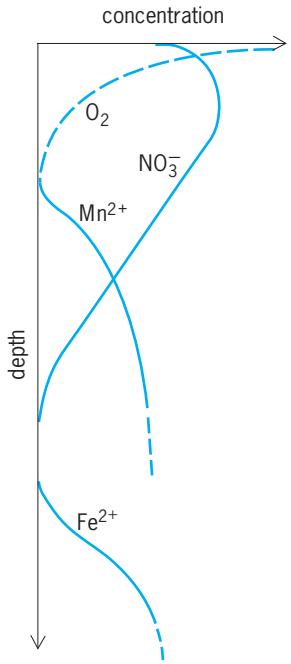


Fig. 2. Schematic representation of trends in pore water profiles. Depth and concentration are arbitrary units but represent concentrations measured or inferred (in the case of O_2) in equatorial Atlantic pore waters. (*After P. Froelich et al., Early oxidation of organic matter in pelagic sediments of the eastern equatorial Atlantic: Suboxic diagenesis, Geochim. Cosmochim. Acta, 43:1075-1090, 1979*)

samples, H_2S often can easily be smelled. After sulfate is consumed, organic matter fermentation results in the formation of CO_2 and methane, CH_4.

Chemical analysis of anoxic basins and marine-sediment pore waters reveals that the sequence of reactions shown above generally holds in nature. **Figure 2** is a schematic summary of the depth distribution of oxidants and oxidation products observed in pore waters of the equatorial Atlantic Ocean. The depth interval represents the top half meter of sediment. In these pore waters, oxygen consumption is followed by nitrate depletion and Mn^{2+} and Fe^{2+} production. In shallow-water marine sediments, rich in organic matter, the above sequence is confined to the top few centimeters, and sulfate reduction followed by methane formation dominates the pore water chemistry.

The fate of nutrients and transition metals in these systems is largely dependent upon the degradation of organic matter (nutrients and many transition metals are incorporated into organic matter at the sea surface and released during degradation); manganese and iron remobilization (both iron and manganese oxides are known to be excellent absorbers of transition metals); and sulfate reduction (some metals form very insoluble sulfides). Thus, redox processes are important to the control of the chemistry of oxidants, oxidation products, and transition metals. *See* HYDROSPHERE; OXIDATION-REDUCTION.

Reaction kinetics. Nearly all organic matter degradation reactions, which occur in natural systems, are mediated by bacteria. The above reactions are energetically favorable; however, activation energies are high enough that the rates are very slow at earth surface temperatures without bacterial catalysis. The rates of reaction are dependent upon the concentrations of organic carbon, oxidant concentration, and the number of bacteria present. In many cases, the last of these is large enough that it is not limiting, and in some cases both number of bacteria and oxidant concentration are abundant (as in some marine sediments), and the organic matter degradation reaction can be considered pseudo first-order.

The environmental oxidation rates of dissolved reduced species (such as Mn^{2+}, Fe^{2+}, NH_4^+, and H_2S), when they enter oxygenated water, are generally not well known. A case study, however, is the oxidation rate of manganese. Figure 1*a* shows the O_2 and H_2S distributions in Saanich Inlet, an intermittently anoxic fiord in British Columbia. The dissolved iron and manganese profiles are plotted in Fig. 1*b*. The difference in the rates of iron and manganese oxidation is revealed by the disparity in distance that they extend into the oxic region. The rate of iron oxidation is much faster than that of Mn^{2+}. The rate of manganese oxidation has been estimated in Saanich Inlet from the profile in Fig. 1*b* and information about water mixing at the O_2/H_2S interface. The mean lifetime for Mn^{2+} with respect to oxidation in the water, which contains O_2, is on the order of a few days. Although this is much slower than the Fe^{2+} oxidation rate, which is probably a few minutes, it is many orders of magnitude faster than rates of manganese oxidation based on abiotic laboratory experiments. Studies of the particulate matter at the O_2/H_2S interface in the water column of Saanich Inlet indicate that it is rich in bacteria, some of which have been identified as manganese oxidizers. This finding supports the hypothesis that the Mn^{2+} oxidation rate is bacterially catalyzed in natural systems. *See* CHEMICAL DYNAMICS.

Summary. The reasons that anoxic regions exist in the marine environment and the sequence of reactions, which occur during organic matter degradation, are reasonably well understood. To a first approximation, thermodynamic (energetic) predictions about the major redox reactions have been confirmed in pore waters of marine sediments. However, we have only begun to understand the kinetic aspects of organic matter degradation and oxidation of reduced species as they enter oxygenated waters. Reaction rates caused by bacterial catalysts in the environment have not been studied to the extent that quantitative generalizations can be made. This is one of the major challenges facing marine geochemists in the understanding of anoxic processes. *See* SEAWATER FERTILITY. Steven R. Emerson; Helmuth Thomas

Bibliography. F. M. M. Morel and J. C. Hering, *Principles and Applications of Aquatic Chemistry*, 1993; J. F. Pankow, *Aquatic Chemistry Concepts*, 1991; J. L. Sarmiento and N. Gruber, *Ocean Biogeochemical Dynamics*, 2006; W. Stumm (ed.), *Aquatic Surface Chemistry: Chemical Processes at the Particle–Water Interface*, 1987; W. Stumm and J. J. Morgan, *Aquatic Chemistry*, 3d ed., 1996.