MARCH 1995
$3.95

SCIENTIFIC AMERICAN

Can science "cure" crime?

Protein computers.

Genes that cause cancer.

RoboTuna, *the swimming robot, will someday mimic its fishy peers.*

SCIENTIFIC AMERICAN

March 1995 Volume 272 Number 3

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Second-class postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111; fax: (212) 355-0408 or send E-mail to SCAinquiry@aol.com. **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.**

DEPARTMENTS

THE COVER painting depicts an artist's conception of a free-swimming robotic fish. The forerunners of such devices are under development at M.I.T., where they are helping scientists understand how fish instinctively exploit the principles of fluid mechanics to achieve enviable speed and agility. One day robotic craft may use these same principles as they explore the ocean, maintain offshore platforms and perform military missions (see "An Efficient Swimming Machine," by M. S. Triantafyllou and G. S. Triantafyllou, page 64). Painting by Al Kamajian.

SCIENTIFIC AMERICAN®

Established 1845

EDITOR IN CHIEF: John Rennie

BOARD OF EDITORS: Michelle Press, *Managing Editor;* Marguerite Holloway, *News Editor;* Ricki L. Rusting, *Associate Editor;* Timothy M. Beardsley; W. Wayt Gibbs; John Horgan, *Senior Writer;* Kristin Leutwyler; Philip Morrison, *Book Editor;* Madhusree Mukerjee; Sasha Nemecek; Corey S. Powell; David A. Schneider; Gary Stix; Paul Wallich; Philip M. Yam

ART: Edward Bell, *Art Director;* Jessie Nathans, *Associate Art Director;* Johnny Johnson, *Assistant Art Director;* Nisa Geller, *Photography Editor;* Lisa Burnett, *Production Editor*

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Molly K. Frances; Daniel C. Schlenoff; Bridget Gerety

PRODUCTION: Richard Sasso, *Associate Publisher/Vice President, Production;* William Sherman, *Director, Production; Managers:* Carol Albert, *Print Production;* Janet Cermak, *Makeup & Quality Control;* Tanya DeSilva, *Prepress;* Silvia Di Placido, *Special Projects;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Ad Traffic: Carl Cherebin; Carey S. Ballard; Kelly Ann Mercado

CIRCULATION: Lorraine Leib Terlecki, *Associate Publisher/Circulation Director;* Katherine Robold, *Circulation Manager;* Joanne Guralnick, *Circulation Promotion Manager;* Rosa Davis, *Fulfillment Manager*

ADVERTISING: Kate Dobson, *Associate Publisher/Advertising Director.* OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager;* Randy James, Rick Johnson, Elizabeth Ryan, Timothy Whiting. CHICAGO: 333 N. Michigan Ave., Suite 912, Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager.* WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025; Lisa K. Carden, *Advertising Manager;* Tonia Wendt. 235 Montgomery St., Suite 724, San Francisco, CA 94104; Debra Silver. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

MARKETING SERVICES: Laura Salant, *Marketing Director;* Diane Schube, *Promotion Manager;* Susan Spirakis, *Research Manager;* Nancy Mongelli, *Assistant Marketing Manager*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; Vivienne Davidson, Linda Kaufman, Intermedia Ltd., Paris; Karin Ohff, Groupe Expansion, Frankfurt; Barth David Schwartz, *Director, Special Projects,* Amsterdam. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.; TAIPEI: Jennifer Wu, JR International Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie M. Beaumonte, *General Manager;* Constance Holmes, *Manager, Advertising Accounting and Coordination*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017-1111

CHAIRMAN AND CHIEF EXECUTIVE OFFICER: John J. Hanley

CO-CHAIRMAN: Dr. Pierre Gerckens

CORPORATE OFFICERS: John J. Moeling, Jr., *President;* R. Vincent Barger, *Chief Financial Officer;* Robert L. Biewen, *Vice President*

DIRECTOR, ELECTRONIC PUBLISHING: Martin Paul

PRINTED IN U.S.A.

LETTERS TO THE EDITORS

Living in the Beginning

Leslie E. Orgel's article "The Origin of Life on the Earth" [SCIENTIFIC AMERICAN, October 1994], along with virtually every other text on prebiotic chemistry, refers to Stanley Miller's WHAM (water, hydrogen, ammonia and methane) experiment. In common with other descriptions of Miller-type experiments, Orgel's article makes no mention of the fact that any amino acids are "generally minor constituents of tars" (according to "The First Organisms," by A. G. Cairns-Smith; SCIENTIFIC AMERICAN, June 1985) and quite unusable as the building blocks of a protein. Why do hundreds of texts and articles fail to refer to this damaging flaw in the edifice that has been built on Miller's work?

C. W. STAMMERS
Bath, England

Orgel writes that "before the mid-17th century most people believed God had created mankind." I would add that, at the present time, a considerable number still do. The probability of chance formation and the further creation of the multitude of biochemical products necessary to sustain life is minuscule. In *Genesis and the Big Bang,* Gerald L. Schroeder states that to create a single protein by chance, 10^{110} trials would have had to be completed each second since the start of time. Less than 10^{18} seconds have elapsed since the "big bang." These computations are a formidable barrier to accepting the chance formation of life.

JOSEPH M. MILLER
Timonium, Md.

Orgel replies:

The criticism of Stanley Miller's experiment made by Cairns-Smith and cited above expresses an opinion that most scientists would find too extreme. I agree that textbooks oversimplify. In the future I hope they will make it clear that cometary impacts, reduction of carbon dioxide and other mechanisms are also possible sources of organics on the primitive earth.

Whatever *Scientific American* readers believe about the creation of mankind, I doubt that many believe in the spontaneous generation of frogs. Does Mr. Miller? If not, why didn't he quote my full sentence fairly ("…and that insects, frogs and other small creatures could arise spontaneously in mud or decaying matter.")? Regarding his second point, very few scientists think functional proteins (or frogs) could arise by chance in a single step. Complex, adapted entities arise from simpler ones by natural selection. Molecular biologists have shown unequivocally that ribozymes, comparable in many ways to protein enzymes, can be obtained at will in the laboratory from completely random mixtures of RNA through natural selection.

Out with the Bang?

In "The Evolution of the Universe," by P. James E. Peebles, David N. Schramm, Edwin L. Turner and Richard G. Kron [SCIENTIFIC AMERICAN, October 1994], the authors write: "At present, there are no fundamental challenges to the big bang theory…." But on a regular basis some reports do challenge the theory's fundamental underpinnings. The uniform expansion of the visible universe is not so uniform: a great part of the local universe reportedly is moving en masse in a different direction than the rest of the cosmos. The distribution of galaxies is not close to homogeneous on a three-dimensional map. Furthermore, does not the big bang theory rest on the ad hoc theory of "dark matter"? I remain skeptical of any group of scientists saying that their interpretation of fact is unchallenged.

BILL BUTLER
Palm Desert, Calif.

I encountered a problem in "The Evolution of the Universe." It takes a little less than seven hours for light from the sun to reach the outermost planet, Pluto. According to Einstein's Special Theory of Relativity, nothing travels faster than the speed of light. Yet the article states that "all the matter we can measure filled a region the size of the solar system…. The universe [then grew] by another factor of 1,000…. All of this occurred within the first minute of the expansion." What happened that allowed matter and energy to travel thousands of times faster than light?

JACK A. WILLIAMSON
Penetanguishene, Ontario

Peebles, Schramm, Turner and Kron reply:

The recent reports of challenges to the big bang theory really refer to attempts to understand the history of the cosmos within the context of this theory. Observations that galaxies cluster and move together on very large scales, for instance, present problems for certain theories of galaxy formation but are good news for others. Some of these theories do postulate exotic dark matter, but the reality of such matter is being tested in laboratory experiments. The faster-than-light expansion of space in the young universe does not violate Special Relativity, which only says that *information* cannot be transmitted faster than light.

The current flood of observational and experimental results makes this an exciting time for cosmology; as in the past, we will no doubt need to refine or even to revise our theories as the data improve. Still, the basic picture of the big bang has proved remarkably robust when confronted with new puzzles.

Making Better Minds

In "Will Robots Inherit the Earth?" [SCIENTIFIC AMERICAN, October 1994], Marvin Minsky tells us that "as a species, we seem to have reached a plateau in our intellectual development." The brain's limited capacity as a storehouse is, however, a red herring. Progress in knowledge depends on the availability of information, not on its place of storage, and information is more widely available when it lies in public networks than when it lies inside a human skull. There is a lesson here: the pursuit of knowledge is a social enterprise. Should we turn ourselves into turbocharged machines? My comments point to a different strategy—that we develop better ways of connecting people, as they are, to one another.

MARC MOREAU
Philadelphia, Pa.

Letters selected for publication may be edited for length and clarity. Unsolicited manuscripts and correspondence will not be returned or acknowledged unless accompanied by a stamped, self-addressed envelope.

50 AND 100 YEARS AGO

MARCH 1945

Better loaves in less time is the claim advanced for a novel bread-baking machine that uses infra-red rays instead of oven heat. Since the infra-red rays are more penetrating than ordinary heat from ovens, baking begins in the heart of the loaf practically as quickly as it does on the surface, and the cooking proceeds more evenly throughout. Smoother crust is another advantageous feature of the process."

"The technology for exploring the upper atmosphere has been advanced considerably by the use of a device known as 'The Weather Broadcaster.' The Weather Broadcaster contains three weather 'feelers' that are sensitive to changes in temperature, humidity, and pressure. A simple alarm clock mechanism keeps a recording helix continuously revolving, and a radio transmitting device sends signals to a recorder on the ground. The whole instrument is carried aloft by a five-foot latex balloon to an average limit of ascension of between 50,000 and 60,000 feet."

"Everyone knows that a big building boom is due just over the horizon. Is John Q. Public going to be satisfied with a domicile that looks just like his neighbors', for as far as the eye can see? It is hardly necessary here to record the negative answer."

" 'Transatlantic rockets are unlikely in this war, but rocket planes making flights from London to Paris will materialize in the not too distant future,' according to Alfred Africano, former President of the American Rocket Society. 'While military applications of jet propulsion engines are now possible, the subject is still in its infancy as far as commercial applications of the technology are concerned.' "

"A life-size 'copper man' which reproduces the temperature response of the human circulatory system has been developed as a test machine for electrically warmed flying suits and other similar equipment. The copper man provides the perfect scientific answer to the problem of testing electrically warmed flying suits, gloves, shoes, and blankets at low temperatures without inflicting suffering and danger on human beings."

SCIENTIFIC AMERICAN

MARCH 1895

Lord Rayleigh startled the world by announcing the discovery of a new constituent of the atmosphere. The new gas is called 'argon'; and, so far as is at present known, it stands entirely unrelated to any other chemical substance in nature."

"It is a well known fact in chemistry that red phosphorus—one of the constituents of the safety match box rubber—combines with explosive violence with chlorate of potash; but the possibility of such a reaction taking place in a person's pocket has not been foreseen. However, several papers recently reported that the simultaneous occurrence of a safety match box and chlorate of potash lozenges in the same pocket led to a series of small-scale explosions, setting fire to the clothes of the unfortunate wearer and severely burning his legs."

"M. De Chateaubourg describes a new treatment of whooping cough, which consists in injecting, subcutaneously, two cubic centimeters and a half of a ten per cent solution of guaiacol and eucalyptol in sterilized oil. After the third injection, the fits of coughing diminish noticeably, the appetite returns, and, as the vomiting rapidly ceases and the general condition begins to feel the good effects of the treatment, the whooping cough disappears. The author reported five cases."

"M. Dieulafoy, who with his wife explored the ruins of Susa, has been elected to the French Academie des Inscriptions. Mme. Dieulafoy not only received the Legion of Honor for her share in the work, but also the right to wear men's clothes in public."

"Among the extraordinary passions for eating uncommon things must be reckoned that which some peoples exhibit for eating earth or clay. This practice appears to have once prevailed all over the world. In some places, the custom has degenerated into a ceremonial one, while in others the eating of this strange food still prevails as a kind of necessity to the lives of those who are addicted to it."

"In our illustration of the Teachers' College of New York City, it will be noticed that the chemical laboratory is occupied by children and grown people. The boys as well as girls are the students of the Horace Mann School, doing practical work in chemistry, while interspersed among them, either assisting, teaching or observing, are seen the adult students of the Teachers' College. At the College, students are not simply taught chemistry or physics, but study the most advanced methods of teaching these sciences in the school room and laboratory."

Chemical laboratory at Teachers' College

Endangered: One Endangered Species Act

Cutting resumes in the forest—and on the Hill

Last December, Federal District Judge William L. Dwyer gave a legal thumbs-up to the Clinton administration's compromise plan for logging and conservation in federally owned forests of the Northwest. The program allows logging operations—which the judge halted in 1991—to recommence at reduced rates. It also establishes a mosaic of six different forms of logging-controlled areas. The strategy aims to guard not only species listed as threatened under the Endangered Species Act, such as the northern spotted owl and the marbled murrelet, but hundreds of other life-forms.

The verdict marked a crucial turning point in the protracted, bitter struggle between environmentalists and timber interests. Yet, in all probability, it will not be the last word. Biologists, who generally accept the agreement as a step in the right direction, disagree over how much it protects imperiled species. In addition, the plan relies on federal agencies to monitor habitats—something critics say government has failed to do.

Moreover, Congress may rewrite the very laws underlying the plan. Representative Don Young of Alaska, chairman of the Resources Committee in the House (known until this year as the Natural Resources Committee), has declared his immediate intention to rework the Endangered Species Act. Young, who has considerable influence, openly disdains efforts to protect rare plants and animals. He denies there is good evidence that the spotted owl is threatened, a view shared by the timber industry. David S. Wilcove of the Environmental Defense Fund, who argues that the owl in question has been better studied than almost any U.S. bird, says Young's position represents "a degree of denial worthy of inclusion in a psychology textbook."

Young further maintains that the act should compensate property owners who refrain from development because of its provisions. To many environmentalists, the cost of such a change would make the act untenable.

Congressional and legal assaults on the act could affect the outcome of any future challenges to the way in which the forestry plan is implemented. And, according to scientists, there are many ways its enactment could be less than perfect. E. Charles Meslow of the Wildlife Management Institute in Washington, D.C., contends the "agencies have never been able to accomplish the monitoring that's been specified."

Meslow says he can now see flaws in the plan that he helped to design. The scientific groups involved drew up 10 different options, each allowing various amounts of logging. Today Meslow believes that from the outset the administration had an unstated goal for the amount of timber to be harvested. "If we'd known that, we would have spent our time more wisely," Meslow explains.

The compromise breaks new ground by employing "ecosystem management." Rather than catering to every species that might cause concern, the approach focuses on preserving entire habitats. The pragmatic philosophy acknowledges the paucity of hard data on most affected creatures. "My opinion has always been that if we persisted in a species-by-species approach, society was not going to have enough patience," says Jerry F. Franklin, a professor at the University of Washington and a key player in the drafting of the plan.

Franklin points out that the view of many biologists that the spotted owl is in accelerated decline "is not completely accepted." He thinks environmentalists should be content to have achieved most of their objectives. "Taking extreme positions on the amount of protection needed is a pretty dangerous thing to do," he states. "A number of environmental scientists have not got the message yet."

JOE McDONALD *Animals Animals*

***NORTHERN SPOTTED OWL'S** habitat is protected under the forestry plan, but its survival is not assured.*

The critics remain unconvinced. The plan estimates, for example, that the spotted owl and the marbled murrelet have more than a 80 percent chance of keeping a "well-distributed" population over the next 100 years. But Daniel Doak, a mathematical modeler at the University of California at Santa Cruz, disagrees. "There is a real question about whether there will be any owls left in the wild in 100 years' time to enjoy the nice landscape we're making for them," Doak declares.

Christopher A. Frissell, an aquatic ecologist at the University of Montana, complains that the Forest Service deliberately decided not to scrutinize fish data. According to Frissell, findings indicate that several species may have less than a 50 percent chance of surviving under the agreement.

"There should be better consideration of species and stocks," concurs James R. Karr, director of the Institute for En-

vironmental Studies at the University of Washington. The focus on habitats, he believes, means that threats to some species, particularly salmonid fish, have been overlooked because they occur outside the geographic scope of the compromise. The plan recognizes that hundreds of less well known terrestrial and aquatic species in the region may not survive the changes.

Despite the doubts, conservationists seemed ready in January to accept Dwyer's ruling. "It's a reasonable plan, and I want to be supportive," says Wilcove of the Environmental Defense Fund. But, among all the swirling uncertainties, one thing is sure: the legal and scientific scrutiny of wildlife and the government's actions in the Pacific Northwest is far from over. —*Tim Beardsley*

Science and Art on Stage

Poets and physicists grapple with models and metaphors

In some circles science and art are known as "the two cultures," seemingly separated by an impenetrable wall. Members from each side of the divide gathered recently for a roundtable discussion at the Mitzi E. Newhouse Theater at Lincoln Center in New York City to talk about how the two disciplines can enrich each other—particularly as science and technology increasingly affect people's lives.

The discussion was catalyzed by the appearance at Lincoln Center of two plays by Tom Stoppard. Both *Hapgood,* now finishing its run, and *Arcadia,* which opens this month at the Vivian Beaumont Theater, draw on images and ideas from physics and mathematics. In *Hapgood* the title character is a British intelligence officer who protects scientific secrets during the cold war; another character, a Soviet scientist and double agent, employs models from physics as metaphors for human experience. *Arcadia* deals with related concepts: in a recent essay, Stoppard described the piece as "a seasoning of chaos and a pinch of thermodynamics following a dash of quantum mechanics."

The panelists opened their discussion by noting that scientists and artists benefit from models and metaphors. Biologists image proteins, chemists use balls and sticks to show molecules, and physicists describe the atom with quantum-mechanical representation. Writers take advantage of metaphors to deepen meaning. "I think artists [and scientists] use models in very similar kinds of ways," remarked poet-naturalist Diane Ackerman. Both hope the representations will let them "see some aspect of the human condition that fascinates them from yet another vantage point." She compared the poem "Thirteen Ways of Looking at a Blackbird," by Wallace Stevens, to its scientific counterpart on video, which another speaker dubbed

BIOSYM TECHNOLOGIES, INC.

***MOLECULAR MODEL** of buckminsterfullerene utilizes artistic techniques to display the symmetry elements of the molecule, named after the famous architect.*

"36 Ways of Looking at One Enzyme."

But perhaps scientists and artists manipulate these tools for different ends, countered John Hollander, a poet from Yale University. "Models are built in order that they may be discarded when there are new data. Metaphors never get discarded," he stated.

Indeed, chemist and poet Roald Hoffmann of Cornell University—who represents a convergence of science and art—maintained that differences of language exist between the two worlds. One of those "is the positive valuation of ambiguity in art. A word may have two meanings that conflict with each other; it may mean the same and the opposite of itself. That's what makes poetry work, in part," Hoffmann explained. "In science we try, but we don't really succeed, to clean up the language and to get the concepts straight between us." Nevertheless, scientific duality persists: a character in *Hapgood* draws on the theory of light's being both waves and particles to justify how he can work as a spy for both the British and the Soviets.

Ackerman said she believes artists like "the pure fun of using metaphors and structures from science." She gave the example of Paul West's novel *Gala,* in which every paragraph begins with a letter of the genetic code: "For him, it was a form of organization and play. I think writers do that very often."

Although the application of scientific metaphor is obviously not limited to writing, some of the speakers cited examples from their favorite texts. Physicist Melissa Franklin of Harvard University mentioned Thomas Pynchon's book *V,* in which he describes the electronic circuitry of a stereo system. According to Franklin, Pynchon "understands it all perfectly, from the shuddering of the speakers to the music going into his [character's] head." Furthermore, Franklin noted that Pynchon "describes it as if he's describing a sunset. It is just one of the most beautiful things I've seen."

The panelists concluded that people tend to feel intimidated by science but that artistic treatments might help convince them that science is interesting and accessible. At the close of the discussion, Hoffmann answered a question about how art and science can become more integrated into daily life. "I think we must get away from that 'high' thing," he responded. "I think it is important not to define theater as high theater. I think Bob Dylan writes poetry. And a lot of simple, everyday experiences are examples of physics, like cooking or watching a tire deflate," he said. "I think we can bring that to young people. I think to do that would be to accomplish a great thing." —*Sasha Nemecek*

Not Yet Elemental, My Dear Seaborg

The periodic table gains 110 and 111—but no names

The discovery of new elements can be cause for celebration, but lately it has become cause for argument. Researchers at the Center for Heavy Ion Research in Darmstadt, Germany, announced last November that they had created element 110. Then, in December, they presented 111. Choosing the right name for the new substances may prove more challenging than making them.

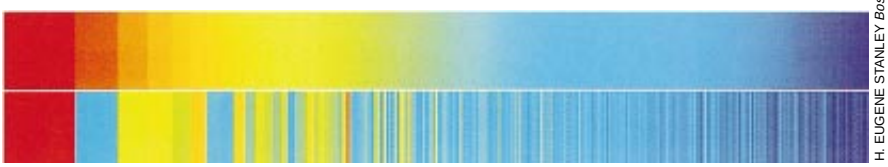
The findings come right on the heels of an intense fight over what to dub 101 through 109. Although the elements themselves do not endure—for instance, 110 lives for about two thousandths of a second—some of the researchers who made them would like to. The discoverers of 106 provisionally named it seaborgium, after Glenn T. Seaborg, a leading U.S. researcher. But the International Union of Pure and Applied Chemistry will vote this August on whether the name should be rutherfordium instead, claiming that an element should not be named after a living person. (British physicist Ernest Rutherford died in 1937.) Not only does this plan upset the parents of element 106, it makes naming other heavy elements more difficult: number 104 was previously known as rutherfordium, except by some Russian scientists who referred to it as kurchatovium. Now the recommended name is dubnium.

As for elements 110 and 111, they will have to wait their turn. Albert Ghiorso of Lawrence Berkeley Laboratory mentions a rumor he heard that the Darmstadt group might hold element 111 hostage until the other names are settled. And Seaborg has "no idea" what element 110 might be christened: "There's tremendous confusion right now."

Irrespective of the naming game, scientists continue to make these short-lived compounds to verify theoretical calculations and to satisfy basic curiosity. "You never know what will happen along the way," Ghiorso says. The just created elements promise to gratify researchers by demonstrating "that superheavy elements are within our grasp," he adds. Investigators are now aiming for element 114, which calculations suggest will be particularly stable. But 112 will probably be very difficult to make—as well as to name. —*Sasha Nemecek*

Talking Trash

Linguistic patterns show up in junk DNA



H. EUGENE STANLEY *Boston University*

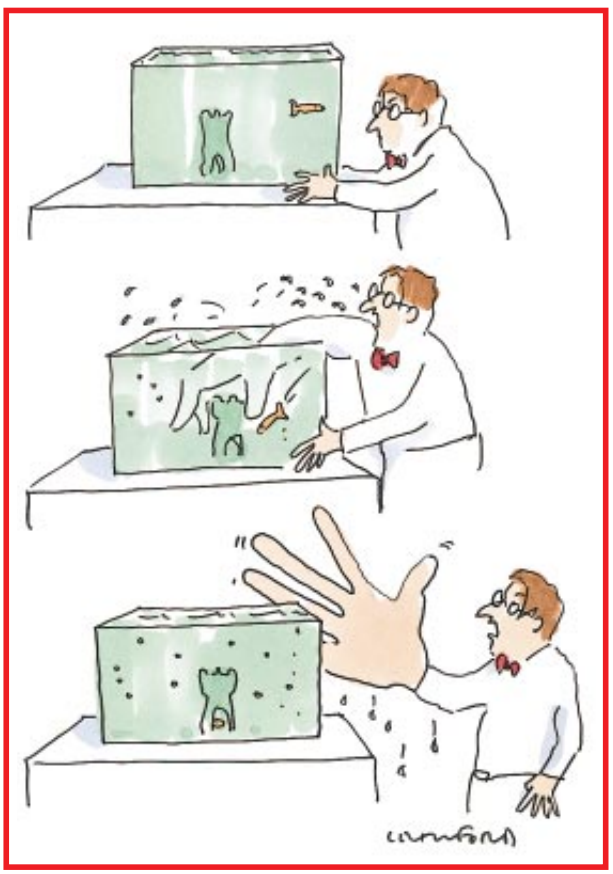
***"WORD" USE in yeast DNA, represented with spectral colors, changes in a regular way in noncoding segments* (top) *but not in coding parts* (bottom).**

What's in a word? Several nucleotides, some researchers might say. By applying statistical methods developed by linguists, investigators have found that "junk" parts of the genomes of many organisms may be expressing a language. These regions have traditionally been regarded as useless accumulations of material from millions of years of evolution. "The feeling is," says Boston University physicist H. Eugene Stanley, "that there's something going on in the noncoding region."

Junk DNA got its name because the nucleotides there (the fundamental pieces of DNA, combined into so-called base pairs) do not encode instructions for making proteins, the basis for life. In fact, the vast majority of genetic material in organisms from bacteria to mammals consists of noncoding DNA segments, which are interspersed with the coding parts. In humans, about 97 percent of the genome is junk.

Over the past 10 years biologists began to suspect that this feature is not entirely trivial. "It's unlikely that every base pair in noncoding DNA is critical, but it is also foolish to say that all of it is junk," notes Robert Tjian, a biochemist at the University of California at Berkeley. For instance, studies have found that mutations in certain parts of the noncoding regions lead to cancer.

Physicists backed the suspicions a few years ago, when those studying fractals noticed certain patterns in junk DNA. They found that noncoding sequences display what are termed long-range correlations. That is, the position of a nucleotide depends to some extent on the placement of other nucleotides. Their patterns follow a fractallike property called $1/f$ noise, which is inherent in many physical systems that evolve over time, such as electronic circuits, periodicity of earthquakes and even traffic patterns. In the genome, however, the long-range correlations held only for the noncoding sequences; the coding parts exhibited an uncorrelated pattern.



MICHAEL CRAWFORD

Those signs suggested that junk DNA might contain some kind of organized information. To decipher the message, Stanley and his colleagues Rosario N. Mantegna, Sergey V. Buldyrev and Shlomo Havlin collaborated with Ary L. Goldberger, Chung-Kang Peng and Michael Simons of Harvard Medical School. They borrowed from the work of linguist George K. Zipf, who, by looking at texts from several languages, ranked the frequency with which words occur. Plotting the rank of words against those in a text produces a distinct relation. The most common word ("the" in English) occurs 10 times more often than the 10th most common word, 100 times more often than the 100th most common, and so forth.

The researchers tested the relation on 40 DNA sequences of species ranging from viruses to humans. They then grouped pairs of nucleotides to create words between three and eight base pairs long (it takes three pairs to specify an amino acid). In every case, they found that noncoding regions followed the Zipf relation more closely than did coding regions, suggesting that junk DNA follows the structure of languages.

"We didn't expect the coding DNA to obey Zipf," Stanley notes. "A code is literal—one if by land, two if by sea. You can't have any mistakes in a code." Language, in contrast, is a statistical, structured system with built-in redundancies. A few mumbled words or scattered typos usually do not render a sentence incomprehensible. In fact, the workers tested this notion of repetition by applying a second analysis, this time from information theorist Claude E. Shannon, who in the 1950s quantified redundancies in languages. They found that junk DNA contains three to four times the redundancies of coding segments.

Because of the statistical nature of the results, the researchers admit their findings are unlikely to help biologists identify functional aspects of junk DNA. Rather the work may indicate something about efficient information storage. "There has to be some sort of hierarchical arrangement of the information to allow one to use it in an efficient fashion and to have some adaptability and flexibility," Goldberger observes. Another speculation is that junk sequences may be essential to the way DNA has to fold to fit into a nucleus.

Some researchers question whether the group has found anything significant. One of those is Benoit B. Mandelbrot of Yale University. In the 1950s the mathematician pointed out that Zipf's law is a statistical numbers game that has little to do with recognizable language features, such as semantics. Moreover, he claims the group made several errors. "Their evidence does not establish Zipf's law even remotely," he says.

But such criticisms are not stopping the Boston workers from trying to decipher junk DNA's tongue. "It could be a dead language," Stanley says, "but the search will be exciting." —*Philip Yam*

SCIENTIFIC AMERICAN

CORRESPONDENCE

Reprints are available; to order, write Reprint Department, Scientific American, 415 Madison Avenue, New York, NY 10017-1111, or fax inquiries to (212) 355-0408.

Back issues: $6 each ($7 outside U.S.) prepaid. Most numbers available. Credit card (Mastercard/Visa) orders for two or more issues accepted. To order, fax (212) 355-0408.

Index of articles since 1948 available in electronic format. Write SciDex®, Scientific American, 415 Madison Avenue, New York, NY 10017-1111, fax (212) 980-8175 or call (800) 777-0444.

***Scientific American*-branded products** available. For free catalogue, write Scientific American Selections, P.O. Box 11314, Des Moines, IA 50340-1314, or call (800) 777-0444.

Editorial correspondence should be addressed to The Editors, Scientific American, 415 Madison Avenue, New York, NY 10017-1111. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, Scientific American, 415 Madison Avenue, New York, NY 10017-1111, or fax (212) 754-1138.

Subscription correspondence should be addressed to Subscription Manager, Scientific American, P.O. Box 3187, Harlan, IA 51537. The date of the last issue of your subscription appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (mailing label, if possible) and your new address. E-mail: SCAinquiry@aol.com.

Pass the Plutonium, Please

International nuclear worries warm up

Getting rid of the long-lived, radioactive by-products of nuclear power is a problem that has stalked the industry for its 40-year history—and the pressure will intensify this year. No nation has established a permanent disposal site for its most dangerous products: used reactor fuel and the high-level waste created when it is reprocessed. Attempts to locate such sites have ignited public outcry in several countries. And next month is likely to see controversy over another nuclear by-product, plutonium.

What promises to be a stormy conference will decide whether, or for how long, to extend the Nuclear Non-Proliferation Treaty, which aims to prevent the spread of nuclear weapons. Global trade in plutonium, which is extracted from used fuel during reprocessing, makes it harder to enforce the treaty, reports the Nuclear Control Institute in Washington, D.C. Reprocessing whets appetites for plutonium in countries lacking nuclear weapons. It also increases the chances that a few kilograms of the metal, enough for a crude nuclear bomb, will find their way into the hands of terrorists.

Putting aside issues of safety, the Nuclear Control Institute argues that reprocessing spent fuel makes little economic sense. The fundamental problem for countries that reprocess is that plutonium is accumulating faster than it is being used. Nuclear power has not expanded as the industry expected it would, and the price of uranium, the primary nuclear fuel, is at a historic low. Thus, there is little industrial demand for plutonium, which can be used in some reactors as a supplemental fuel.

International commercial reprocessing means that high-level waste and spent fuel have to be shipped around the globe. France is steeling itself for protests when the state-owned reprocessing company COGEMA ships the first in a likely series of cargoes of high-level waste from its plant in La Hague to Japan. The material, produced during the reprocessing of Japanese fuel, has been solidified into glass blocks. The first shipment was scheduled for February. The obvious route runs through the Panama Canal, but Caribbean nations have objected to having the waste enter their waters. The Philippines, which lies on an alternative route, has also banned the cargo.

The French government seems to have little stomach for a fight. "It will in the future be very difficult to reprocess for other people," comments Daniel Leroy of the French Embassy in Washington, D.C. He predicts that more countries will begin to reprocess their own waste. Indeed, Japan, which has the largest nuclear program in the world after the U.S., the U.K. and France, is building a second reprocessing plant.

A drive toward nuclear self-sufficiency could be bad news for companies such as COGEMA and for British Nuclear Fuels Limited (BNFL), which also reprocesses foreign fuel. BNFL is commissioning a new plant at its site in Sellafield, despite doubts that have been raised about its health effects on the local population and its economic viability. The financial worries were underscored last December when two German utilities said they would break contracts with BNFL and pay the penalties.

Russia may also be competing for COGEMA's and BNFL's business. The country is expanding reprocessing operations as a means of earning foreign capital. A plant at Krasnoyarsk in Siberia, which has languished unfinished for several years, will apparently be completed, according to Russian press reports. South Korea is among the nations that might reprocess spent fuel there.

Opposition by environmentalists is mounting, however. Russia allowed high-level waste from reprocessed Finnish fuel to remain in the country—taking the problem out of Finnish hands. Damon Moglen of Greenpeace International states that the Czech Republic wants to establish a similar arrangement. Russia's preferred method of disposal has been to pump high-level waste in liquid form into the ground—but it has already started to leach out.

Although the U.S. refrains from reprocessing, utilities have encountered legal problems over accumulating spent fuel. Most of it is in "pools" where it is allowed to cool for some years. Some of these are now full, and local jurisdictions are objecting to the huge casks that the utilities want to use for ongoing storage. The Department of Energy remains under congressional mandate to decide by the year 2001 whether Yucca Mountain in Nevada is suitable as a location for a permanent repository. If certified, the site could commence operations around 2010.

Prospects for the facility, however, have had a recent setback. The DOE found water containing tritium from aboveground atomic tests—which therefore must be younger than 50 years—less than half a mile from the suggest-

ed site. The contamination was found at a depth of about 1,000 feet, the same depth of the planned facility. According to Robert R. Loux, head of Nevada's Agency for Nuclear Projects, which opposes the Yucca proposal, the deep presence of "young water" indicates that it can travel more freely in the mountain than the DOE assumed. The state intends to make the tainted water the basis of constitutional challenges to the program.

A congressional bill introduced on January 5 by Senator J. Bennett Johnston of Louisiana would solve many of the U.S. industry's problems by waiving some of the legal requirements that apply to the Yucca Mountain project—including some relating to water quality—thereby easing certification. But Johnston, a beneficiary of the nuclear power industry's political action committee, may find his bill facing rough passage. Its main provision is a proposal for a temporary, aboveground storage facility at Yucca Mountain that would be used until the final one is completed. The cost of the entire project runs into the billions of dollars. And with budget cuts being the political flavor of the month, any such proposal is likely to receive more than desultory scrutiny.

There could be an alternative by the time Johnston's bill is considered. The Mescalero Apaches of New Mexico may establish a temporary storage facility in that state. Despite political opposition, this private-sector initiative may find more support than the DOE can, although legal challenges could hold up the project for years. Still, the radioactive albatross around the nuclear power industry's neck shows no signs of going away anytime soon, nationally or internationally. *—Tim Beardsley*

Nuclear Empowerment

Just as nuclear power seems to be dying a slow, disreputable death comes a proposal for making reactors clean, safe and unattractive to those who might want to blow up New York City. If nuclear fission is induced not by neutrons from chain reactions but by neutrons from a new generation of powerful accelerators, the reactor cannot burn out of control, argue scientists at Los Alamos National Laboratory and CERN, the European laboratory for particle physics near Geneva. Radioactive waste could also be eliminated, and bomb makers could not divert the fuel.

Conventional reactors often use a mixture of uranium isotopes. When a neutron hits one such nucleus, it breaks into two halves, releasing energy and two or three other neutrons. These particles either escape or induce further fissions, producing a chain reaction. The reactor needs to maintain a delicate balance in the number of neutrons. Too few, and the reaction dies out. Too many—even by 1 percent—and there could be a Chernobyl. (Control rods, which absorb excess neutrons, help to tame the reaction.)

Imagine instead that the neutrons were produced by an accelerator—in practice, by a beam of protons hitting a heavy target. Then there would be no danger of a runaway. "If the lights go out, the cooler stops circulating, and nothing works," says Edward A. Heighway of Los Alamos. "The reaction will just stop." The neutrons transmute the fuel, which consists of thorium 232, into uranium 233, which fissions, producing heat. About 15 percent of the energy would be fed back to run the accelerator. The rest—1,000 megawatts in one scenario—can be purchased.

Because thorium 232, which is naturally abundant, is not fissile, the reactor's inventory is not attractive for military uses. And although uranium 233 can fission, it is contaminated with other uranium isotopes. A dedicated bomb builder would have to separate these out by an arduous process.

The thorium reactor would also produce about 100 times less radioactive waste than would other reactors. The worst of these elements—those that are radioactive for thousands of years or that might leak out of a container—could be transmuted by neutrons. Weapons-grade plutonium or waste from conventional reactors may also be consumed. (But if all the neutrons are used up during cleaning, none would be around to generate more energy, so burning all the waste might not be cost-effective. After about 40 years of running, some kilograms would probably have to be consigned to man-made repositories.)

"Our goal is similar to that of fusion," says Carlo Rubbia of CERN. "We want to find an environmentally acceptable fuel. But in this case, there are no significant technological barriers." Although no accelerator can yet deliver a proton beam of the required intensity, Charles Bowman of Los Alamos envisions building a prototype in seven years: "It's just a question of getting on with it." That is, if the Department of Energy or the European Union, as the case may be, sees fit to provide the funds. *—Madhusree Mukerjee*

CHIP SIMONS

***ACCELERATORS** such as this one can be used to drive clean nuclear fission, explains Edward A. Heighway of Los Alamos National Laboratory.*

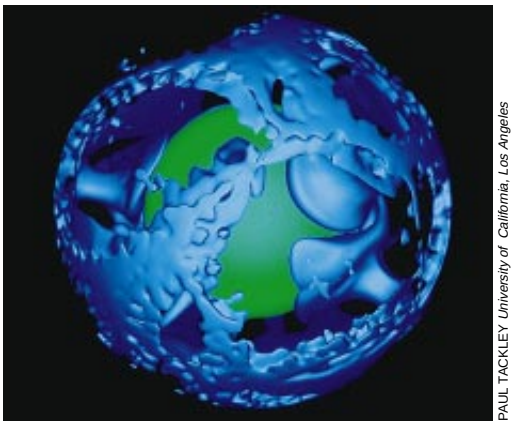
The Chaos Within

As the mantle turns, so do views on how

Nothing is constant—not the world, and certainly not the way scientists view it. For nearly two decades geophysicists have engaged in spirited disagreements about how the earth moves internally. Yet a consensus finally appears to be emerging on the nature of the deep motions that give rise to some of the globe's great changes: earthquakes, volcanoes and the roving of continents.

Opinions about the earth's internal structure have traditionally been rooted in one of two theories. The first posits that the mantle (the thick zone of hot rock situated between the metallic core and the surface) is divided into two layers. In this view, the sections circulate internally but do not intermingle. The alternative holds that convective churning dredges material from the core up through the mantle to the surface and back down again.

"As usual, the answer lies somewhere in between," says Paul J. Tackley of the University of California at Los Angeles. Tackley and his colleagues have developed elaborate computer models of the earth's interior, which have helped inspire the current scientific compromise. When simulating the passage of eons, these models show an earth that oscillates between layered and top-to-bottom circulation. Most of the time the upper and lower mantles convect independently of each other. But cooler material accumulates at the bottom of the upper mantle until a threshold is reached. At that point, the material crashes down through the boundary, forcing a surge of hot rock up from the lower mantle.



PAUL TACKLEY University of California, Los Angeles

COOL ROCKS* (blue) *may plunge to the core* (green) *and mix the mantle from top to bottom.

This vision jibes with recent studies of the behavior of the mobile plates that make up the earth's crust. As continents move, parts of the crust squeeze together, driving plates inward. Rob van der Hilst of the University of Leeds and his team have found that in some places plates seem to deform and pause at the boundary between the layers of the mantle. In other locations material appears to sink far lower, perhaps all the way to the top of the earth's core. Such deep divers may correspond to the catastrophic flows that show up on the computer screens. "We now see that the mantle convects in a very complex style," comments Hua-Wei Zhou of the University of Houston.

So is the debate closed? Not quite. Don L. Anderson of the California Institute of Technology, a longtime booster of layered convection, remains unconvinced by the evidence that the lower mantle interacts with the earth's outer layers. He argues that all the chaotic behavior inferred from the evidence could easily originate in the upper mantle. And Anderson expresses caution about computer models because they do not yet take into account the thermal effects of continents.

Gerald Schubert, also at U.C.L.A., concurs that, absent continents, the simulations are seriously hindered. Even so, he finds it encouraging that computer models of the earth yield the same kind of erratic behavior that seismologists and geochemists have been detecting for real.

The new synthesis has implications for understanding how the continents—and even living things—evolved. In a recent paper in *Nature* Mordechai Stein of the Hebrew University in Jerusalem and Albrecht W. Hofmann of the Max

Swing Wide of That One

Some business executives flying to the Far East last fall learned that not all activity in the Pacific Rim is economic. As the tectonic plate blanketing that basin shifts inexorably to the northwest and plunges into the earth's mantle, it engenders large earthquakes and explosive volcanism. On September 30, after three weeks of minor activity, the Kliuchevskoi volcano—one of many dotting Russia's Kamchatka peninsula—erupted in full glory.

The plume of ash that spewed out of Mount Kliuchevskoi disrupted North Pacific air traffic for nearly three days. During this time, the event was also witnessed by astronauts on board the space shuttle *Endeavour*. At the right is one space-based perspective of the 15,584-foot (4,750-meter) volcano, showing the column of ash being blown by high-altitude winds into the Pacific air lanes. *—David Schneider*

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Planck Institute for Chemistry in Mainz note that continents did not appear simultaneously but seem to have grown in fits and starts. They speculate that great pulses of geologic activity occurred when the mantle overturned. Such episodes could have affected life, both by creating dry land and by dumping carbon dioxide into the atmosphere, which would cause increased greenhouse warming. Indeed, rapid crust formation seems to have been associated with the balmy temperatures dinosaurs enjoyed 100 million years ago.

"This is an exciting time, because people from different disciplines can all finally work together," Zhou exclaims. In science as in nature, chaos makes the world go 'round. —*Corey S. Powell*

Ban That Embargo

Physicians advocate lifting sanctions against Cuba

The Cuban government has always pointed to its health care system as one of the triumphs of the 1959 revolution. But the end of subsidies from the former Soviet Union and an economy overly reliant on producing sugar have placed enormous financial strains on the island's network of clinics and hospitals. Basic supplies, from antibiotics to sterilizing detergents, are hard to find. And Cubans' health has worsened because they have had little to eat during what the government has dubbed the *periodo especial.*

The crisis has prompted a U.S.-based physicians' group to ask that the American government lift the embargo on sales of food and medical supplies. Last fall Jack P. Whisnant, president of the American Academy of Neurology, wrote to President Bill Clinton and every congressional representative to request that the ban on these items cease. The appeal is similar to pleas by the American Public Health Association and by the United Nations General Assembly, which have both called for the full lifting of the embargo.

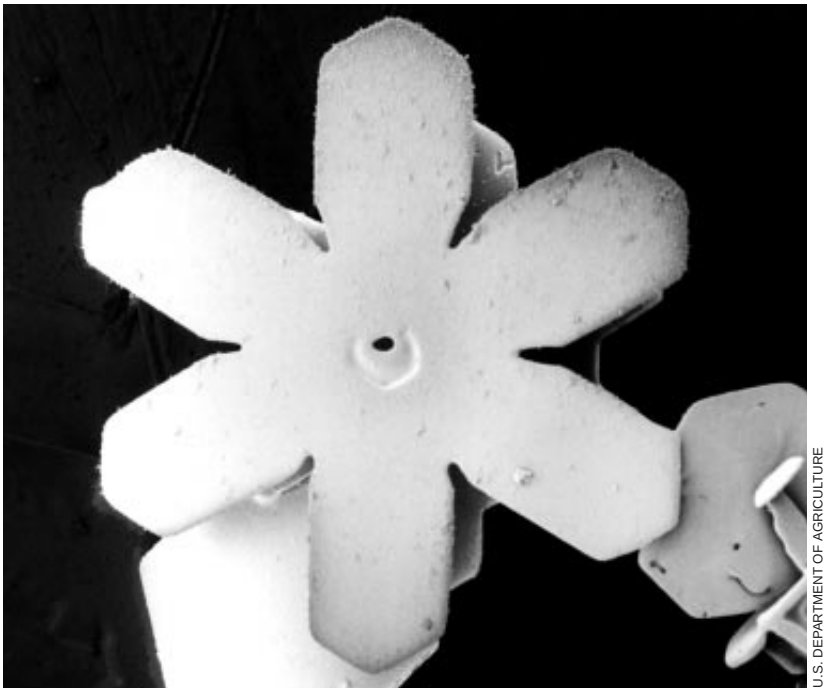
For its part, the academy decided to tread into the mire of Cuban-American politics when a member documented one of the worst neurological epidemics of the 20th century. At the behest of the Pan American Health Organization (PAHO), Gustavo C. Roman led a team of physicians on a visit to Cuba from May to September 1993. Roman, who was then chief of neuroepidemiology at the National Institutes of Health, and his colleagues confirmed the Cuban finding that more than 50,000 of the 11 million inhabitants were suffering from such maladies as optic neuropathy (visual loss), deafness, sensory neuropathy (loss of sensation in the hands and feet) and a spinal cord disorder that impaired walking and bladder control.

Cuban and PAHO physicians discarded a hypothesis that the illnesses resulted from toxins. They determined that a spare diet, along with great physical exertion because of the lack of transportation, had caused severe thiamine deficiency. (Thiamine, a B vitamin, is needed to metabolize sugar.) A comparable outbreak was seen among malnourished Allied inmates in Japanese prison camps during World War II. Distribution of B vitamins to the Cuban population curbed the neuropathies; 200 people did not fully recover.

Roman reported his findings in *Neurology* in 1994. Although details of the scourge had been chronicled elsewhere, Roman went beyond his capacity as a neutral statistician: he lay part of the blame on a 1992 law that tightens the 30-year-old embargo prohibiting U.S. companies from trade with Cuba.

The Cuban Democracy Act of 1992 blocks foreign subsidiaries of U.S. businesses from trading with Cuba. Until its passage, dealings with subsidiaries had allowed Cuba to import critical medicines and foodstuffs cheaply. "Although the U.S. economic embargo may not have been the primary cause of the epidemic in Cuba, it has contributed to its development, complicated its investi-

Scientists' Sense of Snow



U.S. DEPARTMENT OF AGRICULTURE

Seeing the world in a grain of sand may be passé. Researchers at the U.S. Department of Agriculture are endeavoring instead to find the future in a grain of snow. By analyzing the size, structure and water content of snowflakes, hydrologists there hope to predict the amount of spring runoff that will be available in areas where snow is a major source of agricultural water.

The approach originated in the USDA's Scanning Electron Microscopy Laboratory in Beltsville, Md. In December 1993, when the lab completed the installation of a low-temperature specimen holder, scientists scurried outside to find an insect to image. All they found was a fresh blanket of snow. The investigators collected flakes, dipped them in liquid nitrogen, coated them with a thin layer of platinum and scanned them. The pictures ultimately reached hydrologist Albert Rango and his colleagues, also at the USDA, who envisioned using them to improve estimates of the amount of water held in drifts. "We use microwave data from satellites to determine the area and number of snow grains in a snowpack," Rango notes. "But there is some confusion about the actual water content." The group is sampling snow in the Sierra Nevada Mountains and will combine these data with satellite information. —*Steven Vames*

gation and treatment, and continues to hamper its prevention," Roman wrote. He cited other factors related to the embargo, such as the increased costs of importing soybeans from as far away as China.

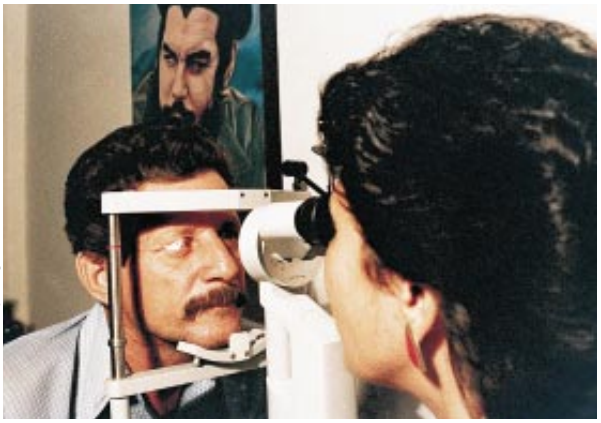
Roman's commentary has been echoed by other health workers. An article in the October 1994 issue of the *Journal of the Florida Medical Association,* "The Time Has Come to Lift the Economic Embargo against Cuba," noted that the country experienced substantially higher costs for health care imports and, in some cases, an inability to obtain these critical goods at all. "Some essential medicines and supplies are produced only in the United States and thus can no longer be purchased," the authors observed. "These include the only effective treatment for pediatric leukemia, x-ray film for breast cancer detection, U.S.-made replacement parts for European-made respirators, and Spanish-language medical books from a firm recently bought by a United States conglomerate."

Advocacy to ease the ban is a provocative gesture, given the opposition of many Cuban-Americans to Fidel Castro's regime. Editors at the *Journal of the Florida Medical Association* were flooded with negative calls; one of its authors, Anthony F. Kirkpatrick, says he has received violent threats. The editor of *Neurology* received a letter from a physician asking why Roman's "extremist political views" had made their way into a scientific journal.

Calls for allowing medical and food exports have not prompted any flurry of activity in Washington. The Clinton administration has not replied to the neurology academy's letter, and all but a few congressmen ignored it. The State Department says the law permits donations of food and medicine by pharmaceutical companies and charitable and religious organizations, which have totaled $50 million since 1992. Other countries can still trade with Cuba—although their ships cannot visit a U.S. port for six months after a stop there.

The revised embargo, officials explain, does not specifically exclude medicine. It states that the administration must be able to verify, through on-site inspections, that these goods get used for their intended purpose. The restriction was included because such items are said to be sold for foreign exchange. "Medical donations can be found in dollar stores for tourists," says Richard A. Nuccio of the State Department's Bureau of Inter-American Affairs.

A Johnson & Johnson subsidiary in Europe has, in fact, exported anesthetics and some other supplies to Cuba. But it was necessary to get the Belgian



ADALBERTO ROQUE *Agence France-Presse*

CLINICIAN inspects a patient for visual loss during a Cuban neurological epidemic in 1992 and 1993 that resulted from a diet poor in thiamine and other B vitamins.

Embassy in Cuba or the PAHO to ensure that they were used for their specified purpose—a requirement that has deterred the company from making routine shipments.

Critics of the U.S. policy maintain that the impact of the embargo is to discourage trade in medicine. They fail to see why Cuba must be subject to restrictions when the U.S. trades with China and Saudi Arabia, governments with one-party systems and their own share of human-rights problems.

The intractable positions may ultimately be swayed by the irresistible desire to buy and sell, however small the emerging market. If McDonald's has reached Moscow, Coca-Cola can surely make the 90-mile jump from Key West. A number of U.S. companies, from General Motors to Archer Daniels Midland, have reportedly begun to plan for the day Havana once again becomes a shuttle trip from Miami. If Cubans lace rum with Coke again, Tylenol and Clorox cannot be far behind. *—Gary Stix*

Treatment That Tightens the Belt

Is insurance part of America's obesity problem?

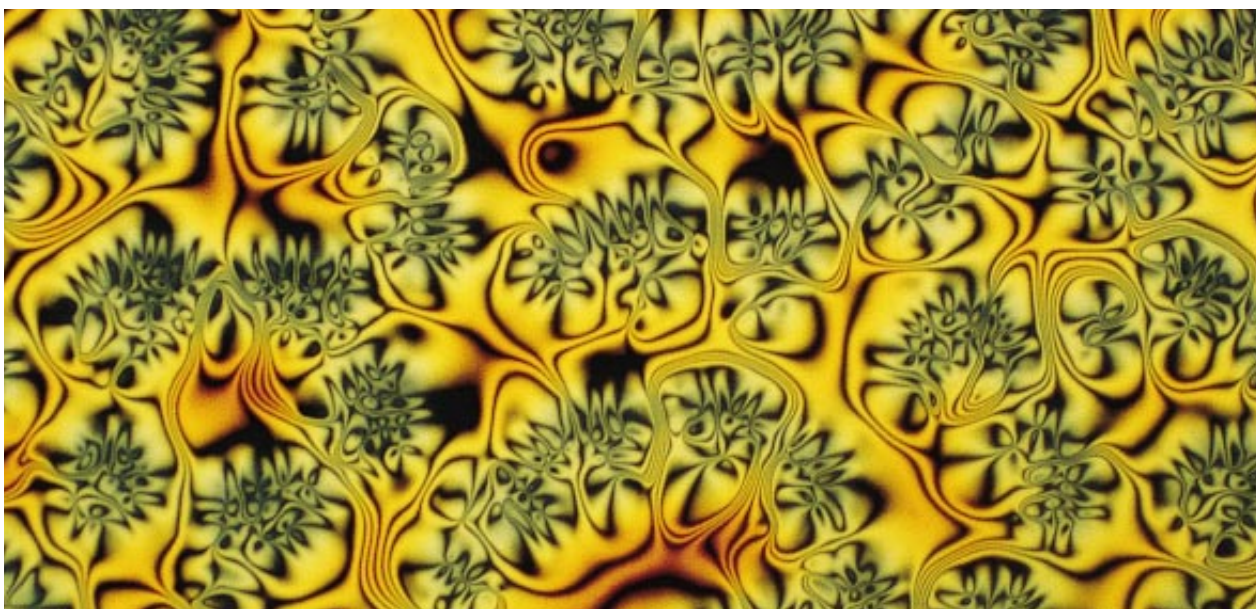
Although the American population is stabilizing in number, it is ballooning in mass. The epidemic in obesity continues to spread, now afflicting 35 percent of women and 31 percent of men, up from 27 and 24 percent just 10 years ago. As metabolic researchers clarify the links between excess fat and a host of diseases, the bill for our larger meals is also becoming clearer: more than $70 billion in health care costs, according to a recent report by the Institute of Medicine (IOM).

Fads, quacks and the proliferation of "lite" foods have failed to reverse the trend. Medicine, an increasingly vocal minority of doctors argue, can and should do better. But it faces two hurdles: physicians tend to dismiss their patients' obesity as a behavioral problem, and many insurance companies refuse to reimburse patients for clinical weight-loss treatments.

The first obstacle seems to be yielding to a new medical consensus that obesity is a physiological disease—mysterious, incurable, yet preventable and treatable. "There is a sea change happening in the treatment of obesity," asserts Albert J. Stunkard, a psychiatrist at the University of Pennsylvania who contributed to the IOM report, *Weighing the Options: Criteria for Evaluating Weight-Management Programs.*

The shift in attitude is the result in part of accumulating scientific evidence that brings both good and bad news. The bad news is that the obese—those whose weight in kilograms is more than 27 times the square of their height in meters—run a significantly higher risk of coronary heart disease, stroke, high blood pressure, sleep apnea, dia-

Liquid Crystals on Display



OLEG D. LAVRENTOVICH *Kent State University*

In the liquid-crystal display of a laptop computer or hand-held video game, glass sheets constrain the fluid's long molecules so that they lie in fixed directions. But suppose a film of liquid crystal floats freely on a glycerin substrate. Then the molecules lie in any direction they please on the glycerin and stand up vertically near the air, forming a new curiosity of condensed-matter physics.

This free film is between just 10 to 1,000 molecules thick, so that the orientation of these molecules has to change very quickly between the bounding surfaces. The fluid tries to parcel out the resulting stress by creating defects along the plane. In one such feature—called a boojum—the molecules lie in a star-shaped pattern on the lower surface and tilt up in successive layers.

Such defects show up when the film is placed between crossed polarizers. The first filter permits only light that has an electric field pulsing in one direction to pass through. The second polarizer stops this very light, unless the intervening liquid crystal has rotated the electric field. The fluid can perform this trick when the molecules lie at some angle (other than 0 or 90 degrees) to the polarizers. Thus, a boojum appears as a cross, and more complex defects appear as points with many dark lines radiating out, as in the micrograph above. *—Madhusree Mukerjee*

betes mellitus, gout, arthritis, gallstones, infertility, accidents and childbirth complications. Between 80 and 90 percent of the 10 million Americans affected by noninsulin-dependent diabetes are obese, for example.

The good news, says Arthur Frank, medical director of George Washington University's Obesity Management Program, is that "people don't have to lose all their excess weight in order to accomplish a great deal." In a recent study of severely obese diabetics, all the patients swallowing drugs and 87 percent of those injecting insulin who shed at least 22 percent of their initial weight were able to stop taking medication.

The discovery announced last December of a gene for a protein that in mice appears to act as a "satiety" hormone raises the hope that appetite might one day be controlled in people. Meanwhile Stunkard and others advocate increased use of drugs and surgery. "Unfortunately, there is a tremendous disinclination to prescribe drugs to treat obesity because of previous abuse of amphetamines, even though newer drugs are not addictive," Stunkard says.

The reluctance extends to state medical review boards, most of which prohibit prescribing antiobesity drugs for more than three months. The Food and Drug Administration has also dragged its feet, approving no new medicines for obesity since 1972. One application has languished six years so far under review.

Gastric surgery—stapling a corner of the stomach—is a riskier but much more effective option for dangerously heavy people. "There is compelling evidence," the IOM panel concluded, "that comorbidities [such as diabetes] are reduced or delayed in severely obese patients who have lost weight as a result of gastric surgery. Therefore, it is puzzling that this is not more widely used."

The devil may be in the dollars. Invasive surgery is expensive, and, the IOM authors note, it is not always reimbursable by insurance companies. In fact, according to F. Xavier Pi-Sunyer, director of the Obesity Research Center at St. Luke's–Roosevelt Hospital in New York City, "many managed-care programs and health insurance plans specifically exclude obesity treatment from their benefits package."

Richard Coorsh, a spokesman for the Health Insurance Association of America, denies that charge. "It is a fairly accepted practice to reimburse for physician-supervised weight-reduction programs as part of a doctor's care," he asserts. Stunkard and Frank raise their eyebrows at this remark. "I would go to the mat with them on that," Stunkard says. "I have never gotten a cent from any insurance company for any patient that I have treated, unless we put in a claim for hypertension by fudging the [blood pressure] numbers a bit."

There are other arguments insurers could make to justify their decision. Some studies have found that losing fat may be nearly as hazardous as keeping it: in one, patients who slimmed by 15 percent or more faced twice the mortality risk of those who dropped 5 percent or less. And evaluations of popular dieting programs reveal that even the stalwarts who complete them (most drop out) gain back an average of two thirds of their 10 percent loss within a year. For treatments to work, Stunkard concedes, "you have to continue them for years." That could get expensive.

Frank is cynical about the prospects for insurance coverage. "Insurance companies have never denied payments because a treatment doesn't work—look at the common cold," he says. "They are denying benefits because they can get away with it. And they will get away with it as long as society has a prejudice against obesity." —*W. Wayt Gibbs*

THE ANALYTICAL ECONOMIST

A Mystery Inside a Riddle Inside an Enigma

For the past 30 years the Indian state of Kerala has been one of economists' favorite anomalies. People there are very poor, and the economy is growing very slowly, but its respectable figures for child mortality, literacy and general life expectancy are what one might expect from a far richer region. Experts such as Amartya Sen of Harvard University have cited the state's accomplishments as proof that purely economic yardsticks, such as per capita gross national products, are doubtful indicators of an area's true prosperity [see "The Economics of Life and Death," by Amartya Sen; SCIENTIFIC AMERICAN, May 1993].

Kerala, which has pursued redistributive social programs since communists won elections in 1957, also scores high on equalizing income distribution and on improving the situation of women—a key factor in achieving the low population growth that sets Kerala apart from most impoverished regions. As a result, some development experts have suggested that Kerala may be a model for the low-consumption future that we all could face as rising living standards collide with limited resources.

Another set of noneconomic data, however, may be calling into question the rosy picture that Sen and others paint. Crime is up, the suicide rate is three times the national average and more than 600,000 of Kerala's 29 million citizens have left to find employment elsewhere. Indeed, the expatriate work force—about 300,000 of whom labor in Persian Gulf countries such as Saudi Arabia and Kuwait—sent home enough money in 1989 to expand Kerala's total income by 13 percent, says Richard W. Franke of Montclair State College. Most of these remittances went toward buying consumer goods or financing house construction.

RAGHUBIR SINGH

KERALAN WOMEN do relatively well economically and serve as a model for development. But the Indian state remains a financial and social puzzle.

As a result, according to government figures, Keralan consumption exceeded local earnings by about 4 percent. Expatriates also shipped VCRs and other items to their families, but the volume of that traffic is even more difficult to estimate than are the flows of money. Meanwhile local unemployment rates are officially about 25 percent and unofficially closer to 35 percent, states Franke, one of the few U.S. scholars actually studying Kerala.

Franke contends, however, that those who attempt to portray Kerala as an egalitarian utopia gone awry are attacking a straw figure. Unemployment and the lack of industrial investment have been well-known sore points in the state's profile for decades. Despite a plethora of projects, neither the left-wing governments that ran the state on and off between 1957 and 1991 nor the current conservative regime has been able to do much about the situation. Indian industrialists have tended to avoid Kerala because of powerful unions and high wages relative to other states; in spite or, perhaps, because of this pro-worker reputation, local businesses lose fewer days to strikes than does the average firm elsewhere in India.

Kerala's mixed situation seems to call into question not only any simple measure of development or economic health but many of the assumptions on which such measures are based. The volume of remittances, for example, skews regional income figures in ways that are difficult to adjust. When economists speak of integrating a locality into the global economy, they usually think in terms of large, easily measured flows of imports and exports—not hundreds of thousands of accountants, nurses, maids and construction workers sending cash to their cousins.

Similarly, labor economics predicts that high unemployment will eventually cause wages to fall, but this adjustment does not appear to have happened in Kerala. The anomaly might be explained by the strong unions or by government intervention in the market or even by the notion that gaining the basic necessities of life makes work at low wages less attractive. Yet invoking such factors raises questions about whether such an unusual labor market is amenable to conventional analysis.

Kerala's lack of capital for investment is a final puzzle piece. It is not surprising that much of the money sent home by expatriates should go to home building and consumer goods, Franke says; despite large-scale government construction programs, much of the state's housing is still substandard. The question is where the money goes from there. Although remittances have spurred some growth in the local construction industry, much of the money has seemingly flowed out of Kerala as invisibly as it flowed in.

If the situation does not change, Keralans will have a long time to ponder these apparent paradoxes. The most recent census reports that life expectancy is now up to 69 years, longer than in any other poor region, and on a par with the formerly industrialized nations of eastern Europe. *—Paul Wallich*

A Recombinant Feast

New bioengineered crops move toward market

Jim P. Corrigan, the proprietor of Carrot Top, a premium produce store in the Chicago suburbs, says he talked to hundreds of reporters following his first sales last May of the genetically engineered Flavr Savr tomato. "I was surprised that a tomato could generate that much interest," Corrigan notes.

As the months passed, the number of calls dwindled—and so did the intense national interest in bioengineered crops. "All of a sudden biotechnology foods have become a nonissue for the media," remarks Thomas J. Hoban of North Carolina State University, who has conducted consumer surveys on food biotechnology.

ASSOCIATED PRESS

***GENETICALLY ENGINEERED TOMATOES** are offered by Roger Salquist of Calgene. These Flavr Savr tomatoes soften slowly, letting the produce ripen longer on the vine.*

Nevertheless, public-interest organizations are still trying to focus attention on the health and environmental implications stemming from more than 2,200 U.S. field trials of bioengineered crops. Many of these groups view the regulatory agencies as unwilling to assess the long-term risks of prospective products.

Last November the Food and Drug Administration raised no objections to marketing seven bioengineered crops—including a yellow crookneck squash resistant to viruses and a potato that produces its own insecticide. In December the U.S. Department of Agriculture gave its endorsement to the yellow squash. Because the gourd passed muster, it will be treated like any other fruit or vegetable.

Calgene, maker of the Flavr Savr, voluntarily slapped a "genetically engineered" label on its tomatoes, the world's first produce with that designation. But farmers who sow seeds this spring for the squash, known as Freedom II, or more cryptically ZW-20, will probably not label it. (The other six crops that passed through the FDA still need a stamp from the USDA or the Environmental Protection Agency, or both, but no major obstacles are expected.)

Some biotechnology companies hope the quiet will persist. Asgrow Seed Company, parent of ZW-20, was pleased that its progeny has not attracted a surfeit of media calls. If the squash had been the first gene-tinkered fruit or vegetable to reach the marketplace, it would have dramatically increased the company's public-relations costs, says Asgrow's Leo Zanoni. "Although we had the product available, it was not to our advantage to push it through that rapidly," he adds.

The relatively benign public reaction to a bioaltered vegetable might also have differed if ZW-20 had reached the headlines first. Some prominent scientists have questioned what they describe as an insufficient review of the squash. They called for further assessment of whether the squash could transfer genes to wild weeds that are its relatives, making them hardier.

Hugh D. Wilson, a taxonomist with Texas A&M University who analyzed the ZW-20 for the USDA, points to evidence that the yellow crookneck squash had wild ancestors whose origins are tied to the American continent north of Mexico. Few other economically useful plants are so classified. Scientists worry that this large and varied gene pool needed for crossbreeding might be diminished if a genetically engineered trait gets transferred to a domesticated plant's wild relatives. A wild plant containing the new gene might thrive and drive out close kin lacking the gene.

The USDA found no merit to any of these arguments. The transfer of the gene to a wild relative would not provide any selective advantage, the agency concluded, since there was no indication that the squash's cousins were infected by these viruses. So resistance to these viruses would not enhance the fitness of this plant.

Much of the USDA evaluation was based on an analysis of interactions between the squash and its relatives that was performed for the agency by Asgrow, ZW-20's creator. "We think the fact that the mere 14 wild plants they looked at didn't have a virus doesn't tell you anything about selective pressures," says Jane Rissler of the Union of Concerned Scientists.

The biotechnology industry's hopes for keeping a low profile may be short-lived. The Pure Food Campaign—a group that has squashed tomatoes and dumped milk—planned to begin a new series of protests in February. "We believe [the government] inadequately looked at the environmental issues, the health and safety issues and the labeling issues," says Jeremy Rifkin, the group's president.

The new campaign will have more of a focus on bioengineered crops. Much of the Pure Food Campaign's protests to date have centered on bovine somatotropin (BST), the genetically engineered hormone that increases cows' milk production. The FDA declared the hormone safe and requires no labeling for it, but critics voice concerns about associated udder infections and the economic impacts of increased production.

Some companies are grappling with these safety issues. Pioneer Hi-Bred International has stopped commercializing a soybean that incorporates additional methionine, an amino acid in which the legume is deficient. In its soybean, production of the amino acid is driven by a gene from the Brazil nut, to which people sometimes have allergic

reactions. University of Nebraska scientists found that blood serum from six of eight individuals allergic to Brazil nuts produced antibodies to the soybeans, suggesting the possibility of an adverse reaction. The company is now seeking another means of enhancing methionine levels.

Despite an incident or two, U.S. agricultural biotechnology flourishes. Rifkin would have an easier job if he moved to Stuttgart. Europe has not provided a similar welcome mat for bioengineered crops, even though it does not wish to be left behind in this emerging industry. In December the European Union voted to allow limited testing of BST but opted to continue a ban on commercial use of the compound through 1999.

Europe, too, may lack the American flair for public relations. In the U.S., Calgene has even plied members of Congress with bacon, lettuce and tomato sandwiches—all made with the Flavr Savr. Natural selection in the marketplace depends heavily on recombinant image making. *—Gary Stix*

GIF Us a Break

It was all over before the arguments really even began. On December 29, CompuServe announced that it would henceforth charge royalties on the Graphics Interchange Format, or GIF. Electronic yowls of protest surged over computer networks—at least among those not too busy drinking, digesting or otherwise holiday making to notice. The GIF file is the networked world's equivalent of the photographic print. It encodes millions of images on disk drives across the globe. Was the whole on-line community to be forced to scrape together its Christmas money from Santa and turn it over to CompuScrooge?

Not this time. Unisys, which owns the patent whose infringement had forced CompuServe to demand royalties in the first place (don't worry, these complications will be explained shortly), clarified in early January. It wanted royalties only from for-profit developers of software that encoded or decoded GIF files. Unisys had no intention of charging for GIF storage or transmission. It wanted small royalties, about 1 percent of the average selling price. And it would not charge anybody who had developed a program before 1995—that is, before CompuServe's announcement. Yowls dimmed to grumbles, and net arguments drifted back to sex and politics as usual.

Next time, however, things could be different. Although the circumstances of the GIF case were by no means typical, the application of slow-moving patent protection to fast-moving software development carries the risk that someday someone really will decide to hold the networked world to ransom. Certain net-watchers now argue that solving the intellectual-property problems created by software will require a third form of legal protection for ideas—not copyright, not patent, but something different.

The trouble with software patents is that they are sweeping and slow. A patent grants ownership of an idea. Full stop. It doesn't matter if somebody holed up in a cave in Tibet has reinvented your idea from thin air and incense smoke. If it's your idea, they have to pay. But a patent takes between a year and a half to two years to issue. Complicated software patents can take longer. In the U.S. patent applications are kept secret until granted.

But two years is also the life cycle of most software products. The Internet currently quadruples in size every two years (it now encompasses more than three million computers). Some parts of the networked world grow even faster. From 1992 to 1994, for example, the World Wide Web—linking text, pictures, video and sound—grew from hundreds of sites to hundreds of thousands. In 1994 the U.S. Patent and Trademark Office approved about 4,500 software patents. Now imagine what might have happened—what might still happen—if a patent submitted in 1992 covered a key component of the Web.

On the other side of the intellectual-property fence, copyright risks missing the point of software. It was designed to cover the text of a document or the look of an illustration rather than the function of a piece of machinery—or a piece of software. Yet it is precisely function that gives value to software. True, copyright is automatically granted as soon as the item is "published." But given a choice between the overweening market power of a patent and the lightweight speed of copyright, which would you think software developers would choose? (Hint: The U.S. Patent and Trademark Office granted only about half as many software patents in 1993 as it did in 1994.)

Some experts think there is a better way. Pamela Samuelson of the University of Pittsburgh Law School, Jerome H. Reichman of Vanderbilt Law School, Mitchell D. Kapor, founder of Lotus Development, and Randall Davis of the Massachusetts Institute of Technology recommend creating a middle ground for software. Ownership would begin immediately, without the long deliberations required for patents. It would also last for only a few years rather than the 19 years of a patent or the 75 years of a copyright. The interesting question, however, and the one expressly left open for debate by their proposals—published in the December 1994 *Columbia Law Review*—is what precisely the law should give ownership to.

Copyright grants control over the text of a document but not over the ideas expressed therein. Patents cover ideas for making things with a given function. Defining middle ground between idea and expression to fit the middle ground of intellectual property should be no mean feat. Presumably no one could simply copy a piece of software. But what if two researchers come up with the same idea independently? What if only part of the software is copied?

Not easy questions to answer. But they are well worth debating, given the legal carnage that patents could unleash in the software industry. Of course, it is worth remembering that bad laws are not the only reason bad things happen. People can just plain mess up. And while CompuServe's Christmas message certainly highlights the dangers of less than perfect law, the actual facts of the case involve a lot of just plain messing up.

Here are the facts. In 1985 Sperry, which merged with Burroughs to form Unisys, was granted a patent on a method of compressing data called the Lempel Zev Welch (LZW) algorithm. CompuServe did not seem to notice. In 1987 CompuServe began developing the GIF to store and transmit graphic images based on—you guessed it—the LZW algorithm. Unisys did not seem to notice. From 1987 to 1993 CompuServe blithely encouraged programmers to use the GIF. Many did. In 1994 when Unisys forced CompuServe to pay royalties on LZW for GIF files, everybody would have noticed had the two tried to pass the cost of those royalties on to the network. Fortunately, common sense and humility prevailed, at least this time. *—John Browning*

MRI Goes Back to the Future

New designs embrace simpler magnets—and lower costs

Instead of becoming higher tech with time, magnetic resonance imaging (MRI) appears to be going retro. Although clinical use of MRI scanners is nearly two decades old—and several million scans are now done annually—some doctors and patients remain frustrated with the machines' low-temperature superconductive magnets and their expense. Several manufacturers are answering these concerns, not with sophisticated high-temperature superconductors but with straightforward solutions reminiscent of some early versions.

From the outset, many MRI designers chose to employ technically elegant superconductive magnet systems; these "supercons" make up the majority of the 4,000 or so scanners found in the U.S. The advantage of these magnets is their ability to maintain high currents and intense fields while expending negligible energy. Their enormous fields give bigger signals and yield clearer images than those of lower-field machines.

High-field superconductive MRIs do, however, have drawbacks. They usually require extensive magnetic shielding to be erected around them. To keep their coils chilled to within a few degrees of absolute zero, the magnets also demand regular feedings of liquid helium—a commodity both expensive and inherently perishable because it boils away quickly. The liquefied gas can also be dangerously explosive and must be handled by trained technicians.

Superconductive designs further suffer from restrictive geometry. The patient must lie in a narrow tunnel within a helium-filled vacuum dewar, an experience—not unlike being trapped in a thermos bottle—that can sometimes provoke claustrophobia. J. Carlos Meléndez and Ernest McCrank of the University of Western Ontario reported in the August 11, 1993, issue of the *Journal of the American Medical Association* that "10 percent of patients undergoing an MRI examination at our institution experience anxiety to the point that the procedure has to be modified, postponed, or canceled."

Another disappointment with high-field designs is the expense: such MRI facilities can require an outlay of $2 million or more. These expenditures are, of course, reflected in patient fees, which range from $500 to more than $1,000 for a standard brain scan.

Many of these problems can be avoided with lower-field MRIs that utilize permanent magnets or electromagnets. These installations are relatively easy to shield and hence have less demanding siting requirements. Maintenance is simpler because the magnets do not use liquid helium. Moreover, such MRIs now offer anticlaustrophobic architecture. Spacious whole-body scanners can, for instance, permit the imaging of trauma patients or scanning during surgery.

These systems also cost a lot less. Open whole-body scanners can, for example, be had for well under $1 million from Fonar Corporation, a manufacturer that has consistently shunned the complications of cryogenic magnets, or from Toshiba, a company that sells both permanent-magnet and supercon types.

The key question is whether the use of lower magnetic-field intensities compromises the technique. Jeffrey C. Weinreb, a radiologist at New York University Medical Center, cautions that the answer is complex: "If your major interest is in the quality of the images and the range of capabilities, there may be some advantage to the high-field systems." Weinreb adds that "midfield machines maybe don't have all the bells and whistles, but they can do a nice job for 90 percent of the things you need to do."

Raymond Damadian, a pioneer of magnetic resonance imaging who is now president of Fonar, says price is the real issue. "My perception is that the difference in quality of the images is small: they get to looking a lot different when you haven't got the money to buy one," he notes. Catering to customers who have not previously had the money seems to be a growing trend.

General Electric, long prominent in manufacturing superconductive MRI systems, will soon add a permanent-magnet model to its line of supercons. Smaller firms are marketing compact, permanent-magnet scanners designed specifically to image arms and legs. These recent shifts toward more economical designs suggest that the MRI industry is responding to cost-consciousness in the health care marketplace, which of late has been particularly disappointing for makers. Sales of MRI systems have declined by 50 percent in the past two years.

Although it is tempting for manufacturers to portray their new magnet configurations as remarkable technological advances, permanent magnets and relatively open designs have been available since the inception of commercial MRI in the early 1980s. What is fundamentally new is that the economy of such systems now really matters.

The money saved with more manageable magnets should translate to reduced capital and operating expenses for these spectacular but often prohibitively pricey diagnostic tools. So it is not unreasonable to hope that the high fees associated with MRI will, like the magnetic-field intensities used to generate the images, defy expectation and head downward. —*David Schneider*

HANK MORGAN *Science Source, Photo Researchers, Inc.*

***NARROW ACCESS** for patients and massive shielding are typical of superconductive MRIs.*

Phone Fight

Regulators may be in the dark about telecommunications

To hear the world's telephone companies tell it, providing basic communications to homes and small businesses requires high-tech noblesse oblige. Customers are demanding, the technology is uninteresting and regulations keep the prices of such services below the cost of providing them. Who, they muse, could afford to shoulder this load without a regulatory fillip here, a cross-subsidy there or a market privilege elsewhere? Enter Ionica, a small British company that reckons technology has transformed telephone service from burden to opportunity. If Ionica is right, many assumptions about telecommunications markets are wrong.

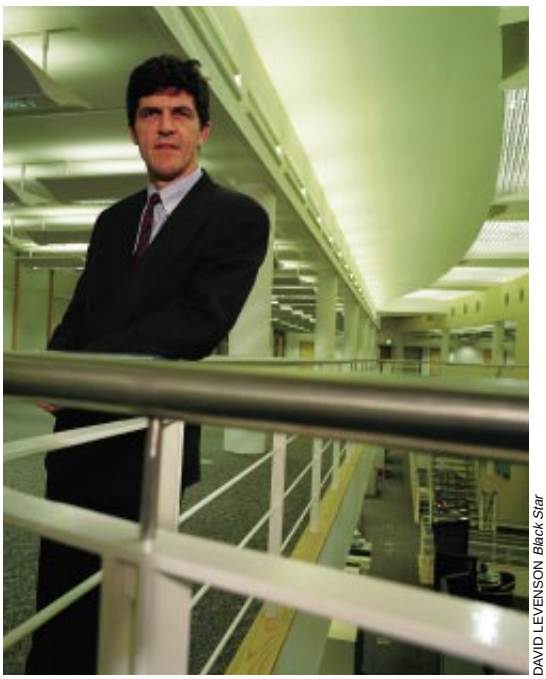
For now, this experiment is unfolding in Britain because Britain is the only country whose regulators will allow it to happen. But given a Republican Congress talking about reviving legislation to deregulate U.S. telecommunications—and other countries moving in this direction as well—the British case merits attention. Unlike the U.S., which has maintained regulated monopolies for local services while promoting competition in long-distance and other lucrative arenas, Britain has encouraged rivalry in all telecommunications sectors for a decade. In 1991, when Britain's market was opened to all, most entrants avoided local service and grabbed for higher profit margins. Nigel Playford, however, founded Ionica to compete solely for low-volume, local services.

A key problem for telecommunications contestants is that the cost of wiring homes and businesses is huge compared with the amount of revenues collected. Cable television companies—the chief competition in local British telecommunications—spent more than £600 ($900) for each of the some four million homes reached by their cables. Only the sale of video entertainment could justify such an outlay. But Ionica figures that by using radio-based systems it can reach homes for less than £10 apiece—so it can boldly compete where none have ventured before.

Ionica will reach customers with a stripped-down version of the technology used to create digital mobile telephones. Ionica's device, developed jointly with Northern Telecom in Canada, is not mobile: its transceiver is screwed to the side of a buyer's house. Thus, Ionica need not worry about miniaturization, battery life, reception within buildings or tracking roaming callers.

This spring Ionica will roll out its options. It promises prices well below current ones and quality at least as high. Rivals sniff a bit at Ionica's contention, arguing radio interference could prove more of a problem than the company now admits—particularly in the English rain. Ionica retorts that its system is amenable to error-correction techniques that its competitors' analog ones are not.

The interesting twist to Ionica's proposition is that the very technology that enables it to compete limits it to basic services. Ionica's bandwidth of tens or hundreds of thousands of bits a second is great for voice, fax, data and Internet access. But video requires at least a million bits a second. One of the ironies of the struggle for telecommunications markets has been that competition has been made fiercer by the assumption that most companies will eventually offer similar services. Because information is increasingly transmitted in digital form, runs conventional wisdom, it will all increasingly flow together across the same lines. Given new investments in telephone and cable networks, the two will "converge" into a web that carries voice, data and video side by side. Ionica presents another option. Choosing Ionica is not merely choosing a different route to a common goal but choosing a different goal—one that excludes video. Ionica is just the beginning.



DAVID LEVENSON *Black Star*

NIGEL PLAYFORD, founder of Ionica, is tackling assumptions about telecommunications markets.

So-called personal communications networks (PCNs), currently being built on both sides of the Atlantic, will use digital, wireless systems for telephones that travel everywhere. (Nearly every U.S. telecommunications company put in a bid at the government's recent auction of spectrum so they can set up PCNs.) Peter W. Huber, a telecommunications expert in Washington, D.C., has coined the phrase "geodesic network" to describe the interconnected network of networks that is beginning to follow deregulation. These choices for consumers bring quandaries for regulators.

The first concerns pricing. Regulators often tax long-distance and business customers so as to subsidize low-cost services. Inasmuch as Ionica and its ilk live up to their promise, they will push consumers toward a less than optimal world: one in which artificially high prices for long distance are used to make the price of perhaps less efficient local services artificially competitive.

More intriguing problems concern diversity. The official job of regulators has been to establish minimum levels of performance, thereby creating common ground for communications. The assumption of this role seems to live on in the American politicians' promise that competition will foster an "information superhighway"—as if that single result were a foregone conclusion. But what if some people simply opt to travel in the slow lane? Are they "information have-nots," to take the phrase of Vice President Al Gore, or merely smart consumers?

Diversity is approaching everywhere. In remote areas of Texas, GTE is using wireless technologies to cut the cost of reaching people far from wires. In the developing world, many countries are wondering if fixed-cellular services might offer a quick, cheap way of creating a telecommunications infrastructure.

Ionica and Northern Telecom have licensed their technology to Mexico and Indonesia, and they are talking to, among others, China. Today in the developed world the question is how the low cost and convenience of wireless will compete with the huge bandwidth of wired services. Tomorrow in the developing world the question could be quite the opposite. In either case, local markets may yet be transformed from erstwhile public service to competitive battleground. —*John Browning*

PROFILE: FRED HOYLE

The Return of the Maverick

In the dead of night, when the demons come, a special fear may creep into the hearts of scientists: What if Fred Hoyle is right? Then astronomy is a sham, biology a house of cards and modern medicine an illusion.

Those who adhere to the paradigms that be have more reason than usual to harbor such worries these days. The media, no doubt bored by the glacial pace of mainstream research, have acquired a sudden fondness for Hoyle, whom they had long vilified or ignored. Journalists are treating the British iconoclast's attacks on the big bang theory and other pillars of modern science with newfound respect. "A second dawn for the universal rebel," proclaimed the London *Times* last fall.

Peer-reviewed journals have also clasped Hoyle's autobiography, *Home Is Where the Wind Blows,* to their bosoms. "What good fortune to have this beautifully written autobiography of one of this century's leading scientists," gushed *Nature,* which in 1993 had harrumphed that Hoyle's recent writings provided "full documentation of the way in which a brilliant mind can be turned to the pursuit of bizarre ideas."

Hoyle's autobiography recalls just how much he accomplished in his prime. In the 1950s he helped to show that we are made, literally, of stardust: elements forged in the cores of stars and flung into space by supernovae. He founded the prestigious Institute of Theoretical Astronomy at the University of Cambridge in the early 1960s and served as its first director. For these and other achievements he was knighted in 1972. Yes, Hoyle is Sir Fred.

Hoyle's personality no doubt adds to his appeal. With his pug nose, jutting jaw and penchant for slang—colleagues are "chaps" and a flawed theory a "bust flush"—he exudes a kind of blue-collar integrity. Hoyle strikes one as a man doggedly pursuing the truth, to hell with the consequences. And he has a knack for sounding reasonable. Arguing that the seeds of life must have come from space, he points out that asteroid impacts rendered the earth uninhabitable until at least 3.8 billion years ago and that cellular life had almost certainly appeared by 3.7 billion years ago. If one thinks of the entire 4.5-billion-year history of the planet as a 24-hour day, Hoyle elaborates, then life appeared in about half an hour.

"You've got to discover DNA; you've got to make thousands of enzymes in that half an hour," he explains to me in an interview at his home in Bournemouth. "And you've got to do it in a very hostile situation. So I find when you put all this together it doesn't add up to a very attractive situation." The spontaneous generation of life on the earth, Hoyle once remarked, would have been as likely as the assemblage of a 747 aircraft by a tornado passing through a junkyard.

As Hoyle spoke, I found myself nodding in agreement. Yes, of course life could not have originated here. What could be more obvious? Only later did I realize that according to Hoyle's timetable, apes were transmogrified into humans some 20 seconds ago, and modern civilization sprang into existence in less than one tenth of a second.

Hoyle's persistence in promoting his views has not been in vain. Last summer radio astronomers at the University of Illinois found the spectral signature of amino acids—the building blocks of proteins—in interstellar space. *New Scientist* concluded, rather generously, that the observations lent credence to the claims of Hoyle and his collaborator N. Chandra Wickramasinghe that space is teeming with microbes. Just one year earlier the British journal had compared one of their books with Erich Von Daniken's pseudoscientific best-seller *Chariots of the Gods.*

DAVID LEVENSON *Black Star*

SIR FRED has been honored and reviled by the establishment.

Hoyle has accepted the recent flurry of favorable publicity with a grain of salt. "If a chap lives to 80, he deserves a pat on the back." (Actually, Hoyle turns 80 this June.) He seems pleased at his continuing ability to provoke controversy. "When I was young, the old regarded me as an outrageous young fellow," he says with a rakish grin, "and now that I'm old the young regard me as an outrageous old fellow."

In his autobiography Hoyle depicts himself as a bright but restless boy. The son of a teacher and a cloth salesman, he often skipped classes in favor of watching barges lumber through canal locks near his home in Yorkshire. Still, he was smart enough to gain entrance to Cambridge. His doctoral supervisor, the physicist P.A.M. Dirac, helped him win a position there shortly after World War II.

Hoyle swiftly moved to the forefront of astronomy, showing how nuclear physics could illuminate such celestial phenomena as white dwarfs, red giants, supernovae and the brilliant radio sources that came to be known as quasars. In 1953 Hoyle's investigations of how stars generate heavy elements led him to predict the existence of a previously unknown state of the isotope carbon 12. Shortly thereafter the physicist William A. Fowler performed experiments that confirmed Hoyle's prediction. Hoyle's work on stellar nucleosynthesis culminated in a 1957 paper, written with

Fowler and Geoffrey R. and E. Margaret Burbidge, that remains a milestone in modern astrophysics. Even Hoyle's critics think he deserved to share the Nobel Prize that Fowler received in 1983 for this research.

Hoyle's speculations on the universe's origin, or lack thereof, grew out of his friendship with Thomas Gold and Hermann Bondi, physicists with whom he had designed radar systems during the war. "Bondi had a relative somewhere—he seemed to have relatives everywhere—and one sent him a case of rum," Hoyle recounts. While imbibing Bondi's elixir, the three physicists turned to a perennial puzzle of the young and intoxicated: How did we come to be?

The finding that all galaxies in the cosmos are receding from one another had already convinced many astronomers that the universe had exploded into being at a specific time in the past and was still expanding. Hoyle's fundamental objection to this model was and is philosophical: he believes it does not make sense to talk about the creation of the universe unless one already has space and time for the universe to be created in.

"You lose the universality of the laws of physics," Hoyle explains. "Physics is no longer." The only alternative to this absurdity, Hoyle then decided, was that space and time must have always existed. He, Gold and Bondi thus invented the steady-state theory, which posits that the universe is infinite both in space and time and constantly generates new matter through some unknown mechanism.

Together with his longtime collaborator Geoffrey Burbidge of the University of California at San Diego and the Indian astronomer J. V. Narlikar, Hoyle has recently developed a new and improved version of the steady-state theory. This "quasi–steady state" theory replaces one big bang with many little bangs, perhaps occurring within quasars and other so-called active galaxies. These little bangs generate light elements such as helium and lithium as well as the local expansion of galaxies.

Although Hoyle and his colleagues have outlined their quasi–steady state theory in such reputable outlets as the *Astrophysical Journal,* they have won few converts. Astronomers contend that the theory merely substitutes many little miracles for a single large one. Moreover, Hoyle's group offers no plausible explanation for the microwave radiation, discovered in 1965, which most cosmologists believe is the afterglow of the big bang. Hoyle insists that recent versions of the big bang theory—which can account for observations only by invoking vast amounts of dark matter—are much more deeply flawed. "It's like medieval theology," he snaps.

One of the great ironies of modern science is that Hoyle, the big bang's most notorious basher, coined the term in 1950 while he was doing a series of radio lectures on astronomy. Hoyle did not intend to disparage the theory, as many accounts have suggested, but merely to make it vivid. Last year *Sky & Telescope* magazine held a contest to rename the theory. After mulling over thousands of suggestions, the judges announced they could find none worthy of supplanting "big bang." Hoyle was not surprised. "Words are like harpoons," he comments. "Once they go in, they are very hard to pull out."

In the 1960s Hoyle curbed his theoretical speculations somewhat as a result of his involvement in Cambridge's Institute of Theoretical Astronomy. Although he helped to make the institute a potent force in astronomy, he grew tired of fighting with university officials over administrative matters and resigned from the directorship in 1972. The hatchet was not buried until 1992, when the institute held a ceremony honoring Hoyle and unveiled a statue of him.

DAVID LEVENSON *Black Star*

HOYLE sees purpose everywhere.

Hoyle claims to have no regrets about abandoning his lofty position. "From the point of view of ideas it was a stultifying period," he says. Hoyle soon began collaborating with Wickramasinghe, now at the University of Wales, on a study of complex molecules in space. They eventually concluded—based on their interpretation of data from radio and optical telescopes—that space is filled not only with organic compounds but with bacteria and other organisms. Hoyle had actually first broached this possibility in his 1957 book *The Black Cloud,* which remains the best known of his dozen science fiction novels.

Hoyle and Wickramasinghe now assert that space-faring microbes spawned life on the earth and spurred evolution thereafter. Perhaps their most controversial hypothesis is that epidemics of influenza, whooping cough and other diseases are triggered when the earth passes through clouds of pathogens.

Discussing the biomedical establishment's belief in the more conventional, person-to-person mode of disease transmission, Hoyle shows a rare flash of anger. "They don't look at those data and say, 'Well, it's wrong,' and stop teaching it. They just go on doping out the same rubbish. And that's why if you go to the hospital and there's something wrong with you, you'll be lucky if they cure it."

But if space is swarming with microorganisms, why haven't they been detected? They probably have been, Hoyle replies. He suspects that U.S. experiments on high-altitude balloons and other platforms turned up evidence of life in space in the 1960s, but officials hushed it up. Why? Perhaps for reasons related to national security, Hoyle suggests, or because the results contradicted received wisdom. (Scientists familiar with the high-altitude experiments deny Hoyle's allegations.)

"Science today is locked into paradigms," he declares. "Every avenue is blocked by beliefs that are wrong, and if you try to get anything published by a journal today, you will run up against a paradigm, and the editors will turn it down." Hoyle emphasizes that he has never asserted, as some reports have stated, that the AIDS virus has an extraterrestrial source. It "is such a strange virus I have to believe it's a laboratory product," he comments. Is Hoyle suggesting that the pathogen might have been produced by a biological warfare program that went awry? "Yes, that's my feeling," he responds.

Purpose pervades Hoyle's universe. He has long felt that natural selection alone could not account for the appearance and rapid evolution of life on the earth. Some supernatural intelligence must be directing the evolution of life and indeed of the entire cosmos—although to what end Hoyle does not know. The universe is an "obvious fix," he remarks. "There are too many things that look accidental that are not."

Sensible scientists will dismiss such talk as preposterous. But every now and then, in their inevitable moments of doubt, they may wonder: Could Sir Fred be right? —*John Horgan*

SCIENTIFIC AMERICAN

Faster Evaluation of Vital Drugs

Traditional clinical trials may delay the availability of lifesaving therapies. Regulators now attempt to balance speed against the risk of errors

by David A. Kessler and Karyn L. Feiden

In the spring of 1994 a group of AIDS activists gathered on the third floor of the Parklawn Building, headquarters of the Food and Drug Administration in Rockville, Md. The activists made an extraordinary plea to top agency officials: don't approve drugs to treat disease caused by the human immunodeficiency virus (HIV) too quickly. No one familiar with events of the previous decade could have predicted this turnabout. Only a few years earlier angry activists had besieged Parklawn, demanding access to compounds that had barely moved out of the test tube. Faced with imminent death, people with AIDS clamored for the right to take therapeutic risks. Experimental drugs were a ray of hope on a bleak treatment landscape, and many patients were unwilling to accept any restraint on access.

Promising drugs generally take several years to test; an additional 18 months or more may elapse from the time a sponsor requests product approval until a compound is widely available. In response to concerns about the length of time needed to develop and evaluate new drugs, the FDA made some dramatic changes in the way it conducts business. In the late 1980s it introduced rules that expand access to unapproved but promising therapies. In 1992 the agency adopted a regulation that allows it to approve drugs before complete data on their safety and efficacy have been collected. These initiatives, which apply only to drugs for serious and life-threatening diseases, attempt to balance the urgent needs of desperately ill patients with the FDA's responsibility to determine whether the drugs work.

Have these policies been too successful in speeding the availability of new drugs? AIDS activists at the FDA last spring feared that answers had begun to take a back seat to access. They were concerned that patients and their physicians did not know how to make optimal use of the existing antiviral AIDS drugs. They wanted to discuss ways to learn more about the value of experimental therapies and to learn it sooner—when each drug should be administered, to whom and in what dose. Without such information, novel drugs might be of little use.

Although these activists' caution reflected only one perspective within the AIDS communities, there is no doubt that making a drug widely available early means that it will initially be used without full understanding of its characteristics. The FDA stands behind the tools that have put therapies into patients' hands more swiftly, but it also remains committed to the time-honored legal standards for evaluating new products.

DONNA BINDER *Impact Visuals*

The Usual Approval Threshold

Under law, the FDA is charged with ensuring that therapeutic agents used in the U.S. are safe and effective. A therapy's known and potential benefits must outweigh its risks, under the proposed conditions of use. Most new

therapies move through a series of clinical trials. At each stage, researchers recruit progressively larger patient populations and accumulate more data about safety and efficacy. Once the company sponsoring a product believes it has enough information, it will submit a new product application to the FDA; the agency receives about 100 original drug applications every year.

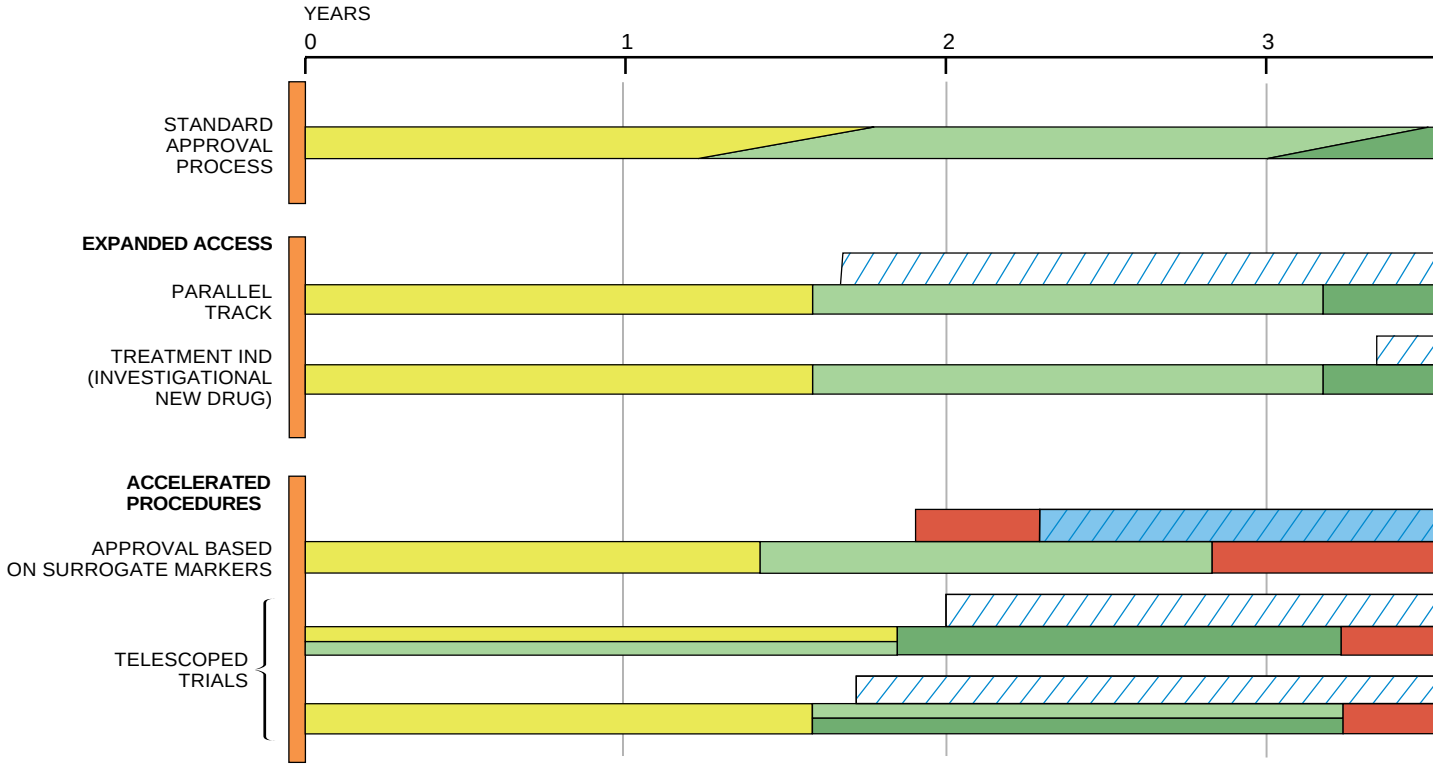
The agency usually requires at least two controlled (comparative) trials in humans as evidence that a medicine works. Although it may become apparent that a drug is useful relatively early, it generally takes longer to gather other important information, such as response at different doses, rates of less frequent but important adverse effects and the need to modify doses for particular classes of patients, such as people with kidney failure or other diseases. Certain rare side effects may become apparent only after a product is in widespread use.

For these reasons, pharmaceutical companies usually study a product in three phases. Phase 1 trials examine a drug in a small number of subjects to determine how well it is tolerated and how the drug is metabolized. They also look for short-term effects such as changes in blood pressure. (At this stage, evidence of activity is only a secondary concern.) In phase 2, researchers study patients with the condition that the drug is intended to treat to assess its effectiveness and the rate of more common adverse events. Depending on circumstances, the treatment group may be compared with a group receiving either a placebo—an inert substance without medical effect—or therapy known to be helpful. In general, placebos cannot be used when there is

DAVID A. KESSLER and KARYN L. FEIDEN are, respectively, commissioner of the Food and Drug Administration, and a writer specializing in health and science. Kessler was appointed commissioner in 1990. He received his M.D. from Harvard University and his J.D. from the University of Chicago and completed his pediatric training at Johns Hopkins University. Before becoming FDA commissioner, he ran a hospital in the Bronx. The authors would like to acknowledge the extensive contributions of these FDA experts: Rachel E. Behrman, Susan S. Ellenberg, David W. Feigal, Renie Schapiro, Jay P. Siegel, Robert J. Temple, Janet Woodcock and Randolph F. Wykoff.

PROTESTERS accuse the Food and Drug Administration of dragging its feet in reviewing potentially lifesaving medicines in this 1988 demonstration at the agency's headquarters. Since then, the FDA has streamlined its procedures and increased access to experimental drugs. Some activists have now suggested that new compounds are being made available too quickly.

Timetable for Drug Evaluation



DRUG DEVELOPMENT PROCESS can span years as a compound passes through a series of successively larger clinical studies (*top*). The FDA's expanded-access rules may make a drug for a serious illness available to patients while it is still under investigation (*middle*). The agency has also reduced the time it takes to approve new drugs for sale, and it may issue provisional approval for widespread marketing of a compound on the basis of significantly fewer data than it once required (*bottom*). A drug may then be removed from the market if later evaluations show that it is not beneficial.

a treatment known to improve survival or prevent irreversible damage.

Phase 3 studies, whose design is guided by the results of earlier trials, may involve as many as several thousand patients and are intended to provide additional definitive data for approval. These controlled trials assess response to different doses, long-term safety, drug interactions and responses that differ by sex. They also gather other data needed to label the drug properly and define the conditions for its use. Ideally, phase 3 trials will also investigate any differences in a compound's effect on subpopulations that differ by age, race, concurrent illness or other factors. Many of the studies are double-blind—neither the patients nor the investigators know who is getting which therapy—to safeguard against bias, and patients are randomly assigned to different treatments. Long-term safety data, however, may come from unblinded, uncontrolled studies.

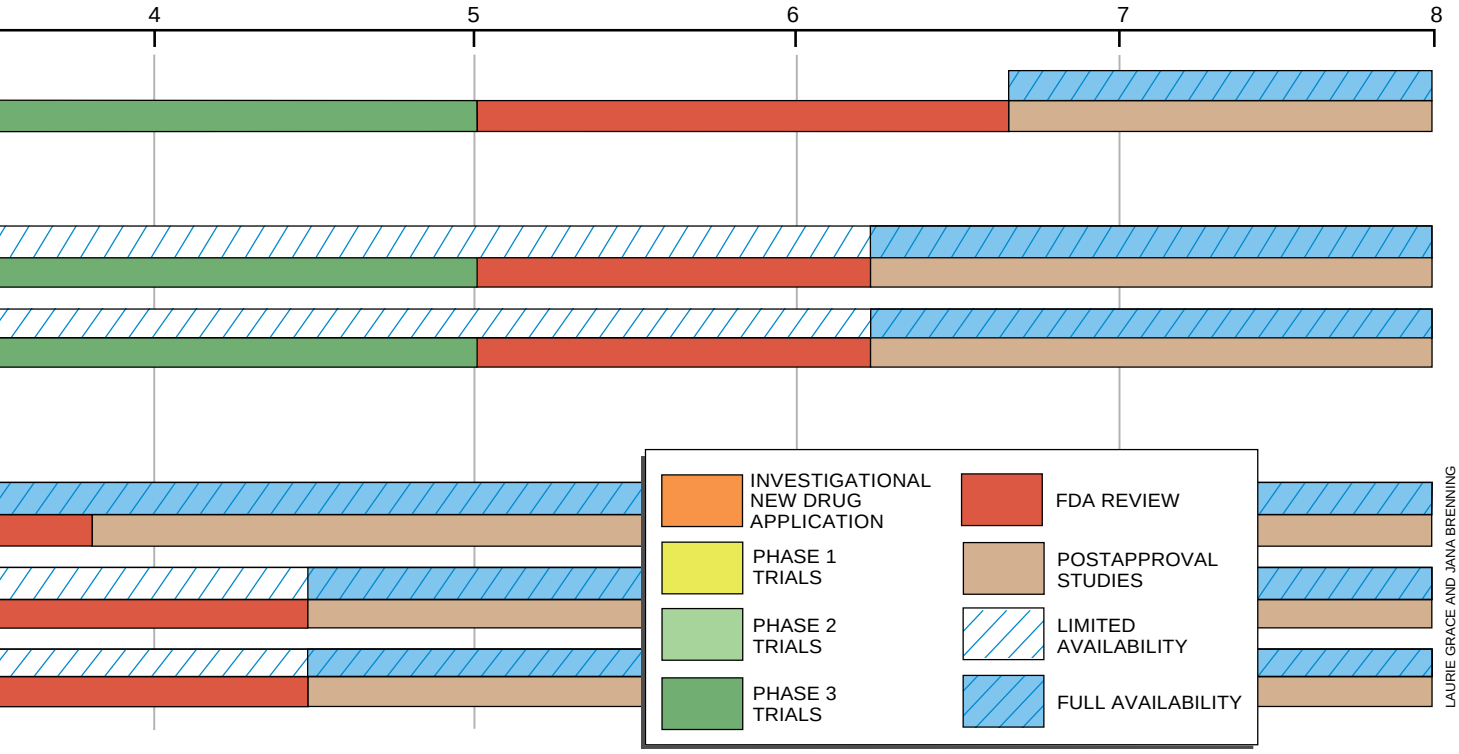
Trials may be conducted by drug companies or by researchers affiliated with hospitals, government laboratories or universities; depending on the drug being tested, sponsors may advertise for patients directly or wait for referrals from physicians. Sources of funds for trials are similarly variable: in the vast majority of cases the manufacturer covers the cost of the trial; the National Institutes of Health also supports studies, and patients or their insurers may cover the cost of other medical care associated with the trial. In some cases, the availability of a drug or of research support may be a crucial factor in determining where and how a trial is carried out.

For a product (such as a mild painkiller) that provides at most a marginal clinical advantage or treats a relatively innocuous condition, there is seldom any need to depart from the usual procedures. Under certain urgent circumstances, however, the FDA may allow access to the therapy before approval or approve promising therapies, on the basis of more limited data. The agency considers many factors, including the severity of the disease and current options for treating it, the information available on the new treatment, the adequacy of studies designed to secure more conclusive answers and the risk that increasing access might make it more difficult to find out whether the drug works or not. Drug companies must still conduct well-controlled trials before final approval.

Inevitably, there is a trade-off between speed and certainty. Judgments based on preliminary or incomplete data can increase the chance that important adverse effects will emerge to undermine the drug's value or that errors in the recommended dosage will lead to suboptimal use. Moreover, once a drug has been made widely available, controlled studies to determine appropriate dosages, duration of use or interactions among drugs can become more difficult to perform. Sponsors have less incentive to conduct further studies, and patients who can get medication without signing up for a trial may be less willing to participate. These risks make it imperative that research continue until an experimental drug has met the same standards of safety and efficacy as other therapeutic products. The FDA is responsible for ensuring that sponsors complete those studies.

Early Access to Therapies

For many years, the agency has permitted cancer patients to receive certain promising experimental therapies. Since 1987 the Treatment Investigational New Drug (IND) regulations allow promising drugs for other serious illnesses to be widely distributed on the basis of considerably fewer data than are required for full marketing approv-



al. Typically the clinical trial process is nearing completion; the sponsor must be pursuing marketing approval diligently. More recently "parallel track" regulations, issued in 1992, have let patients with HIV disease who do not qualify for clinical trials and have no therapeutic alternatives receive experimental therapies through their private physicians very early in testing.

Expanded access can provide treatment to people with otherwise untreatable, life-threatening diseases. Yet it can also discourage enrollment in controlled studies and so prevent definitive data from ever being collected.

The use of ganciclovir to prevent cytomegalovirus (CMV) retinitis, a sight-threatening opportunistic infection associated with AIDS, offers a sobering example of what can go wrong. The widespread, uncontrolled distribution of unapproved ganciclovir convinced both patients and the scientific community that the therapy worked; researchers thus could not conduct controlled trials in which patients risked receiving a placebo. After much discussion, the sponsor, Syntex, gained approval on the basis of a trial using a historical control (that is, by comparing the course of disease in patients who used ganciclovir with records of patients who had had CMV retinitis before ganciclovir became available). In retrospect, it seems clear that small randomized trials conducted before the drug was widely distributed would have collected more rigorous efficacy data and allowed the drug to be approved several years earlier.

Exosurf, used to prevent or treat hyaline membrane disease, a leading cause of death among premature infants, yields a more reassuring example of how expanded access mechanisms should be applied. After Burroughs Wellcome shared interim data from placebo-controlled trials with the FDA, agency reviewers concluded that marketing approval was likely, assuming that the findings were confirmed by a closer review. The agency then agreed with Burroughs Wellcome's decision to halt the trial and allowed Exosurf to be distributed while the company wrote the marketing application.

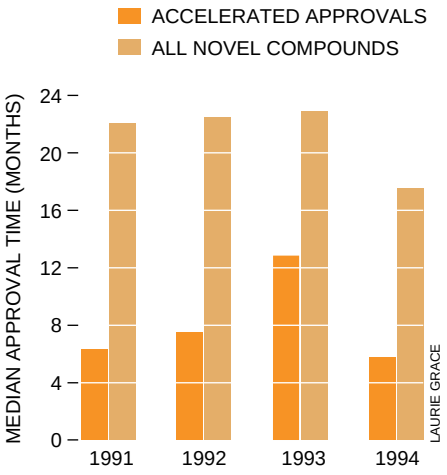
From the FDA's perspective, participation in the analysis of interim data was potentially risky—if the definitive review had led to rejection, early termination of the Exosurf study would have resulted in a lost opportunity to secure more complete information. Yet the drug was important enough that the chance seemed worth taking. Fortunately, full analysis resulted in approval just five months after the company submitted its application, well before the drug would have been available otherwise.

The standards of evidence that allow a therapeutic product to be provided on an expanded-access basis become more rigorous as a product approaches final approval. Drugs distributed on a parallel track must be "promising." Treatment INDs, named nearer to the time of an approval decision, must show evidence that they "may be effective." If the agency actually approves a drug for marketing before trials are completed, the data must demonstrate a "reasonable" likelihood of benefit. In every case, information gathering continues. No matter who gets to use a drug as it winds through the testing process, it must eventually meet the most rigorous regulatory standard, providing "substantial evidence" that it works.

Accelerated Approval

What indicators should researchers measure to determine a drug's efficacy? In the case of a life-threatening disease, mortality is the most conclusive indicator of a drug's impact, but when a fatal disease progresses over the course of years, waiting for a discernible effect on mortality may lead to unacceptable delays. Instead researchers sometimes look at surrogate end points, laboratory or clinical measurements they believe to be correlated with a real benefit, even though they do not help a patient per se. For example, decreased blood pressure or cholesterol levels do not make a patient feel better but may reduce the risk of a heart attack or stroke.

In December 1992 the FDA formalized its procedures for accelerating the approval of drugs for serious diseases with no good alternative treatment on the basis of surrogate end points that are reasonably likely to be valid. In these cases, the sponsor must commit to con-



APPROVAL for drugs to treat some life-threatening diseases has been speeded up in recent years. The five drugs approved under accelerated procedures during the past four years typically passed FDA review in less than half the time required of other new therapies.

ducting controlled trials to verify clinical benefits after approval. (Usually such trials will be under way at the time of approval.) If those benefits fail to materialize, the FDA can remove the drug from the market quickly. This regulation accepts a modest risk of approving a drug that eventually proves ineffective in return for the gain of making a drug for a serious illness available significantly faster.

The review and approval of dideoxyinosine (ddI, an antiviral drug that acts much like zidovudine, also called AZT) was the model for the accelerated approval regulation. In April 1991, Bristol-Myers Squibb, ddI's sponsor, submitted a marketing application to the FDA on the basis of data from the initial phase of clinical studies, which included just 170 adults and 98 children. Like most early trials, these had not been designed to provide definitive efficacy data. There was no control group, and as a surrogate for clinical benefit, researchers had measured patients' CD4+ lymphocyte count, a barometer of immune system function. The available information was unlikely to permit drug approval because of the uncertainty of the CD4+ end point, but the absence of therapeutic alternatives for patients not responding to zidovudine argued against waiting nearly a year for definitive data from three controlled trials then under way. The agency therefore launched an intensive effort to obtain more information about the action of ddI on CD4+ cells and the relation between CD4+ cell counts and clinical outcome. In addition, investigators took an unusual, unscheduled look at CD4+ data from patients in an ongoing trial.

On October 9, 1991, the FDA approved ddI for patients with no satisfactory options, primarily because the compound had a positive effect on CD4+ cell count, and with the understanding that the trials would be completed. These ultimately produced persuasive clinical data that confirmed the early approval and also led to labeling changes broadening ddI's indication. Even so, CD4+ cell counts have not been so closely correlated with the progression of HIV disease and survival as many had hoped: a reminder that surrogate end points add uncertainty to drug evaluation.

The National Heart, Lung and Blood Institute's Cardiac Arrhythmia Suppression Trial (CAST) shows how treacherous surrogate end points can be. Researchers had found that individuals with ventricular premature contractions (an electrically abnormal heartbeat) after a heart attack were more likely to die than patients without those contractions. Because it seemed intuitively reasonable to believe that suppressing those contractions would reduce mortality, the results of the CAST trials were a major surprise. The drugs under study suppressed the abnormal beats, but the patients who used them died at 2.5 times the rate of the control group. Whether this outcome reflects the toxicity of the drugs studied or some other factor is not known.

Another risk, not so well appreciated, is that researchers may abandon a drug prematurely because it has no effect on a surrogate. Workers initially investigating gamma-interferon expected it to affect superoxide production and the destruction of bacteria by blood cells and thus to alleviate the life-threatening infections that characterize chronic granulomatous disease, a rare immune disorder seen in children. Although the drug's lack of influence on these laboratory findings was determined within a few weeks, the trial continued as scheduled. The results were surprising: after a year, gamma-interferon had clearly reduced the rate of infection but had no detectable effect on either surrogate. Only the clinical end point allowed the drug's value to be identified.

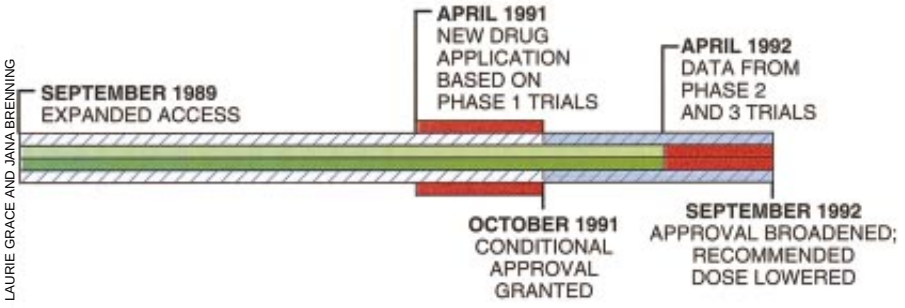
Even with their drawbacks, surrogate data can be a valuable way to bridge the gap between access (provided on the basis of a promising surrogate) and answers (secured when clinical benefit is established). In AIDS, as in other life-threatening diseases, the search continues for the best surrogates possible.

Flexible Trial Designs

The FDA is trying to encourage efficient drug development strategies by increasing the range of trial designs that may be acceptable. Researchers may be able to expedite data gathering without compromising quality. The agency encourages investigators to meet with FDA representatives so that research plans can be reviewed; it is often possible to depart from traditional practices and still maintain rigor.

One approach to speeding data collection is the "large, simple trial." Such trials have relaxed criteria for eligibility, enroll patients in a variety of care settings and collect considerably less extensive baseline and outcome data than do traditional trials. As a result, more patients with a broader range of characteristics can participate, and researchers can conduct very large studies in a remarkably short time.

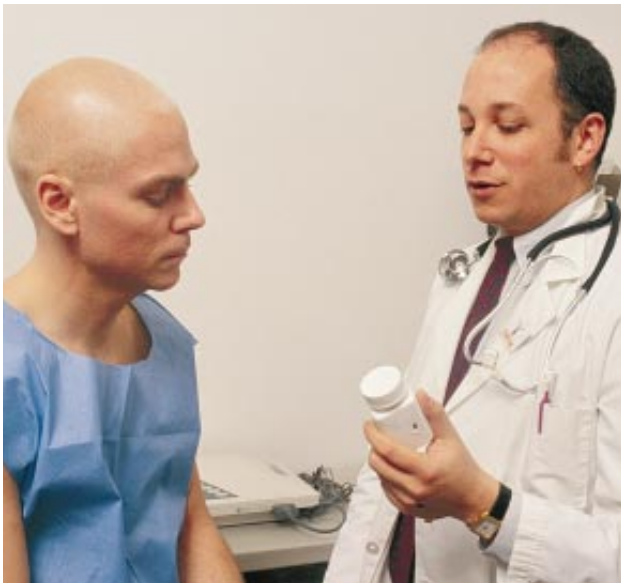
An example is the first International Study of Infarct Survival (ISIS), in which 16,000 patients in 10 countries were given an intravenous beta blocker or matching placebo after a heart attack. Investigators found a 15 percent reduction in mortality after seven days (4.6 percent in patients who did not receive the drug versus 3.9 percent among



CASE HISTORY of dideoxyinosine (ddI) shows how efforts to streamline the regulatory process have paid off. In September 1989 the drug was made available to many AIDS patients on an expanded-access basis. The FDA approved the drug for sale after reviewing preliminary results from ongoing studies and then expanded the approval once final results came in.

those who did), an effect that had not been detected in previous, smaller studies. The trial was simple because it required relatively little data collection, but its huge size allowed detection of a small, but clearly valuable, effect.

Such trials are generally most feasible when a drug's toxicity profile and mechanism of action are already well understood. They can identify effects of moderate size that could not have been discovered in small populations. Large, simple trials can also be used to collect important data about how a drug is best administered in clinical settings and can be a tool for making promising experimental therapies available while research is completed. Recently some observers have proposed that the FDA consider requiring such trials either before granting accelerated approval or afterward, as a means of completing postmarketing studies.



ERIC NELSON *Custom Medical Stock*

CLINICAL TRIAL of zidovudine in HIV-positive patients gathers data that could slow the incidence of AIDS. Conducting well-controlled trials can be difficult once a drug is available from other sources.

The criteria for entry into a study are another area where investigators may be more flexible. Historically, researchers have defined entry criteria narrowly and excluded, as far as possible, patients who use concomitant therapy. These restrictions have some merits, because they decrease variability and make it easier to show an effect of the study drug. Nevertheless, they also have scientific downsides. For example, it may be ethically difficult to deny seriously ill patients potentially beneficial treatments as a condition of trial participation. Moreover, these restrictions may slow enrollment, encourage the unrecorded use of other therapies and result in trials that do not mirror the actual conditions under which a drug may eventually be used. With acceptable increases in sample size, it is possible to apply much broader entry criteria and allow patients access to other treatments. This relaxation of entry criteria has been especially important in trials of AIDS drugs.

Where patients are using multiple therapies, investigators may have to study the toxicity and benefit of different drug combinations, rather than of a single agent, to determine whether a new compound adds measurably to an existing arsenal. Factorial designs that test multiple interventions are emerging as a fast, effective way of accumulating knowledge. In the simplest form, one group of patients receives therapy A, one receives therapy B, one receives both A and B and the fourth group receives neither. For example, different antiviral agents, or antiviral agents and agents to prevent opportunistic infections, can be evaluated concurrently in AIDS patients.

A factorial design can be more efficient than two separate trials because it can answer multiple questions through a single process. Results may be difficult to interpret, however, if medications interact in complex ways. For example, one therapy might have no effect when used alone, but a strong one in combination. Large sample sizes and careful statistical design will moderate such difficulties. Factorial designs also may alert investigators to important drug interactions.

Under certain circumstances, the FDA may base approval of a drug on results from a single study rather than two independent trials. The agency makes such decisions when the trial shows a particularly persuasive difference in survival or serious morbidity, as was the case in the initial multicenter, double-blind, placebo-controlled trial of zidovudine in patients with AIDS. Among patients receiving a placebo, 16 died, whereas only one death was recorded among those taking zidovudine. As a result, researchers stopped the trial, and the drug met with prompt approval.

Investigators now routinely conduct interim analyses of clinical results in studies examining survival or major morbidity and may halt a trial ahead of schedule as a result. For example, two key zidovudine trials ended after an interim analysis showed significantly slower disease progression among those taking the drug. The findings led to the March 1990 relabeling of zidovudine to include a broader spectrum of the HIV-infected population.

Decisions based on interim analyses must be made cautiously, however, because the data are incomplete and may lack sufficient long-term follow-up. In April 1993 British and French investigators reported that three years of follow-up time showed no significant benefit to the long-term use of zidovudine by asymptomatic individuals. Although that verdict in turn is not definitive—some lasting value to early zidovudine use has been demonstrated—it underscores the importance of continued follow-up when relying on interim analyses.

Telescoping Trials

A single phase 1 trial may sometimes be used to collect both toxicity data and the preliminary, short-term efficacy data ordinarily collected in phase 2. Such trials are usually larger than traditional phase 1 studies, should be randomized and must encompass a reasonable range of dosages. Under some circumstances, it may also be possible to eliminate phase 3. If the phase 2 trial of a drug for a serious disease shows an influence on survival or irreversible morbidity, the FDA may be able to conclude that the benefits of treatment outweigh risks, even if those risks need further evaluation.

There are hazards in telescoping trials, however; of particular concern is the absence of adequate dose-response information and reliance on a small sample of patients, often followed for only a short time. Had the work that led to the initial approval of zidovudine included a low-dose cohort, doctors might have learned earlier that the clinical benefit of zidovudine could be attained at a lower, much less toxic, dose and patients would have endured fewer side effects. Nevertheless, the early demonstration of increased survival made approval appropriate, with refinements to be made later. Fortunately, the pivotal studies for two other antiviral AIDS drugs, ddI and ddC, were designed with the zidovudine lesson in mind, and both included high- and low-dose arms.

Depending on the circumstances, one of many trial designs may be most appropriate for studying a new drug. As the FDA and researchers review plans

for clinical trials, we will continue to refine our understanding of what methods will yield the best possible data in a reasonable time.

Lives in the Balance

In the face of a serious or life-threatening disease, the need to obtain optimal information about promising therapies before adoption must be tempered by the imperative of making useful products available quickly. Ideally, the agency would approve a new product only on the basis of mature, solidly reproducible data demonstrating real benefit to patients. Nevertheless, there are inevitably times when a drug is so important that it is approved or widely distributed on the strength of a limited, but still sufficient database.

When this early approval occurs, investigators, clinicians and patients must understand the risks. Although the FDA has not yet had to revoke an accelerated approval, it is likely that we will eventually allow widespread access to a drug that does not work. Patients will have wasted precious time taking an ineffective product; they may even be harmed. This risk underscores the need for clinical trials to continue after accelerated approval has been granted. Incomplete or misleading data serve neither patients nor the cause of knowledge. For all the mechanisms that may be put in place to speed access to new therapies, the FDA's commitment to safety and efficacy remains unchanged—all approved drugs, even those for potentially fatal diseases, must ultimately meet the same rigorous standards.

FURTHER READING

CHANGES IN NORMAL DRUG APPROVAL PROCESS IN RESPONSE TO THE AIDS CRISIS. Ellen C. Cooper in *Food, Drug and Cosmetic Law Journal,* Vol. 45, No. 4, pages 329–338; July 1990.

DESIGN CONSIDERATIONS FOR AIDS TRIALS. David P. Byar et al. in *New England Journal of Medicine,* Vol. 323, No. 19, pages 1343–1348; November 8, 1990.

STATISTICAL ISSUES ARISING IN AIDS CLINICAL TRIALS. Susan S. Ellenberg, Dianne M. Finkelstein and David A. Schoenfeld in *Journal of the American Statistical Association,* Vol. 87, No. 418, pages 562–583; June 1992.

EVALUATING THERAPEUTIC INTERVENTIONS: SOME ISSUES AND EXPERIENCES. Thomas R. Fleming in *Statistical Science,* Vol. 7, No. 4, pages 428–456; November 1992.

TRENDS IN PHARMACEUTICAL DEVELOPMENT. Robert Temple in *Drug Information Journal,* Vol. 27, No. 2, pages 355–366; April 1993.

The Many Costs of Drug Testing

When researchers test a new drug, pressures from many different constituencies converge on the clinical trial. The Food and Drug Administration, for example, must worry about whether the trial will adequately answer questions about the drug. A drug company must decide whether sales will repay the cost of studies. Patients worry about getting the best treatment for their illness. Insurers attempt to avoid footing the bill for investigations that are not part of ordinary care. As the FDA rethinks how drugs should be studied, the balance among other parties may also shift.

HANK MORGAN *Science Source, Photo Researchers, Inc.*

ARTIST WITH AIDS receives the antiviral drug ganciclovir by intravenous infusion. New drugs now in clinical trials may control sight-threatening infections better than does ganciclovir—and with fewer side effects.

The cost of clinical trials—indeed, the cost of drug development as a whole—is a contentious issue. Consumer activist James P. Love of the Center for Study of Responsive Law contends that testing a typical compound in people consumes about $12.5 million. Economist Joseph DiMasi of Tufts University puts the figure at $93 million instead; the difference is that DiMasi counts the money spent testing other drugs that do not pan out (only about one in four generally does), animal testing and fixed R&D costs. He also adds "opportunity cost": the profit that could have been made by investing drug development dollars elsewhere. Opportunity costs account for approximately half of his figure, DiMasi says.

Regardless of whose number is closer, drug companies definitely do abandon some compounds for lack of a market rather than lack of therapeutic promise, DiMasi says. Orphan drug legislation enacted in 1983, which gives companies seven years of exclusive marketing rights, has had some impact on bringing useful but economically marginal products to market, he notes. At the same time, some companies have reaped windfalls by winning orphan status for drugs that later turned out to be immensely popular, among them human growth hormone.

Meanwhile as companies track opportunity costs, patients and insurers skirmish over the hidden extra fees of clinical trials. Abbey S. Meyers of the National Organization for Rare Disorders says that researchers do not charge for administering experimental therapy and recording clinical data but that everything else is up for grabs: a hospital stay while the drug is being administered, blood tests to check for side effects, even long-term care if the patient has a disabling reaction. In addition, participation in a study generally precludes patients from suing for injuries a new therapy may cause. Meyers recalls one proposal for a study of gene therapy for cancer in which investigators offered to alter the DNA of bone marrow cells free of charge as long as patients paid the $200,000 needed to have the cells extracted and reimplanted.

At one time, insurance companies covered parts of experimental therapies, but lately they have shied away from paying charges that patients receiving ordinary treatment would not incur. (Paradoxically, for some diseases, cancer and AIDS among them, research progresses so rapidly that clinical trials are an almost routine avenue for care.) Meyers expresses concern not only for the potential inequity of limiting advanced treatments to the few who can afford them but also for the way that scientific results may be skewed by testing drugs in a tiny, unrepresentative slice of the population. As pressure to contain health care spending increases, however, there are no obvious alternative candidates willing to foot the bill. *—Paul Wallich, staff writer*

Laser Control of Chemical Reactions

For years, chemists have sought to control reactions with lasers—and have mostly failed. Success may come from exploiting subtle quantum effects resulting from the interaction of light and matter

by Paul Brumer and Moshe Shapiro

Although the science of chemistry has made considerable progress in the past century, the main principles behind the industrial practice of chemistry have remained basically unchanged. Methods for breaking and re-forming chemical bonds still often rely heavily on altering the temperature and pressure of the reaction or adding a catalyst. This approach is often ineffective because it takes no account of our understanding of the motions of molecules. As a result, bulk reactions are often inefficient, generating large quantities of useless by-products in addition to the desired materials.

Recently investigators have devised new techniques, based on illuminating chemical compounds with lasers, that can potentially control the paths taken by reactions. These methods promise to alter the yields in selected ways by exploiting an essential feature of quantum mechanics—namely, the wavelike properties of both light and matter. The latest calculations show that lasers can be enormously effective in steering a reaction down a preferred pathway.

The idea of using lasers to drive reactions is not revolutionary—in fact, attempts to do so began shortly after the first laser was invented about 35 years ago. These devices emit radiation of a precise frequency, or color, and thus can impart a well-defined parcel of energy to a given target. Chemical bonds were seen as individual springs of different strengths, each of which vibrated when given a certain amount of energy. The hope was that lasers could be tuned to attack a particular bond, weakening or breaking it and thereby encouraging one product to form in lieu of another.

But this approach, called mode-selective chemistry, has seen only limited success. In fact, it is doomed to fail for the vast majority of molecules because it assumes that chemical bonds are largely independent of one another. Only a few compounds have bonds that meet this requirement. Rather most bonds are strongly interdependent: energy readily flows between them. As a consequence, the energy imparted by the laser is distributed throughout the molecule in a fashion similar to that resulting from traditional, and far cheaper, sources.

The latest approach, called coherent control, has grown out of research begun in the mid-1980s, when investigators began to take a look at properties of laser light that had been ignored in earlier considerations of chemical control. One such property is the coherence of laser light. "Coherence" relates to the way that atoms emit light. In conventional light sources, such as household bulbs, electricity heats up the filament, exciting tungsten atoms. The atoms spontaneously lose this energy by giving off a bit of light. Each atom, however, emits light independently of the other atoms. The total light from the filament therefore consists of a jumble of waves from individual atoms. In a way, light from ordinary sources resembles a column of soldiers marching out of step with one another. Physicists say that such light is incoherent, or out of phase. In contrast, atoms in lasers act together. Consequently, all the waves that make up the entire laser beam match perfectly, behaving like soldiers marching in step.

Coherent light readily displays an important property—constructive and destructive interference. The term "interference" does not necessarily mean that the waves disturb one another's trajectories; it refers to the way all waves (including those on water) combine. In constructive interference, the crests or troughs of two (or more) waves meet so that the wave heights or dips add together. The wave therefore becomes bigger and deeper (in other words, the amplitude of the wave increases). In destructive interference, crest meets trough, and the wave is extinguished.

Interfering Light Waves

The phenomena of constructive and destructive interference become apparent if we shine a laser beam on an opaque plate that has two narrow slits cut into it. Each slit acts as a new source of light waves. The waves emanating from each slit, which fan out, can then interfere with each other. The interference can be rendered visible if we place a viewing screen just beyond the slits. The screen will show a pattern of dark and light stripes, or fringes. The light areas correspond to constructive interference; the dark areas, destructive interference.

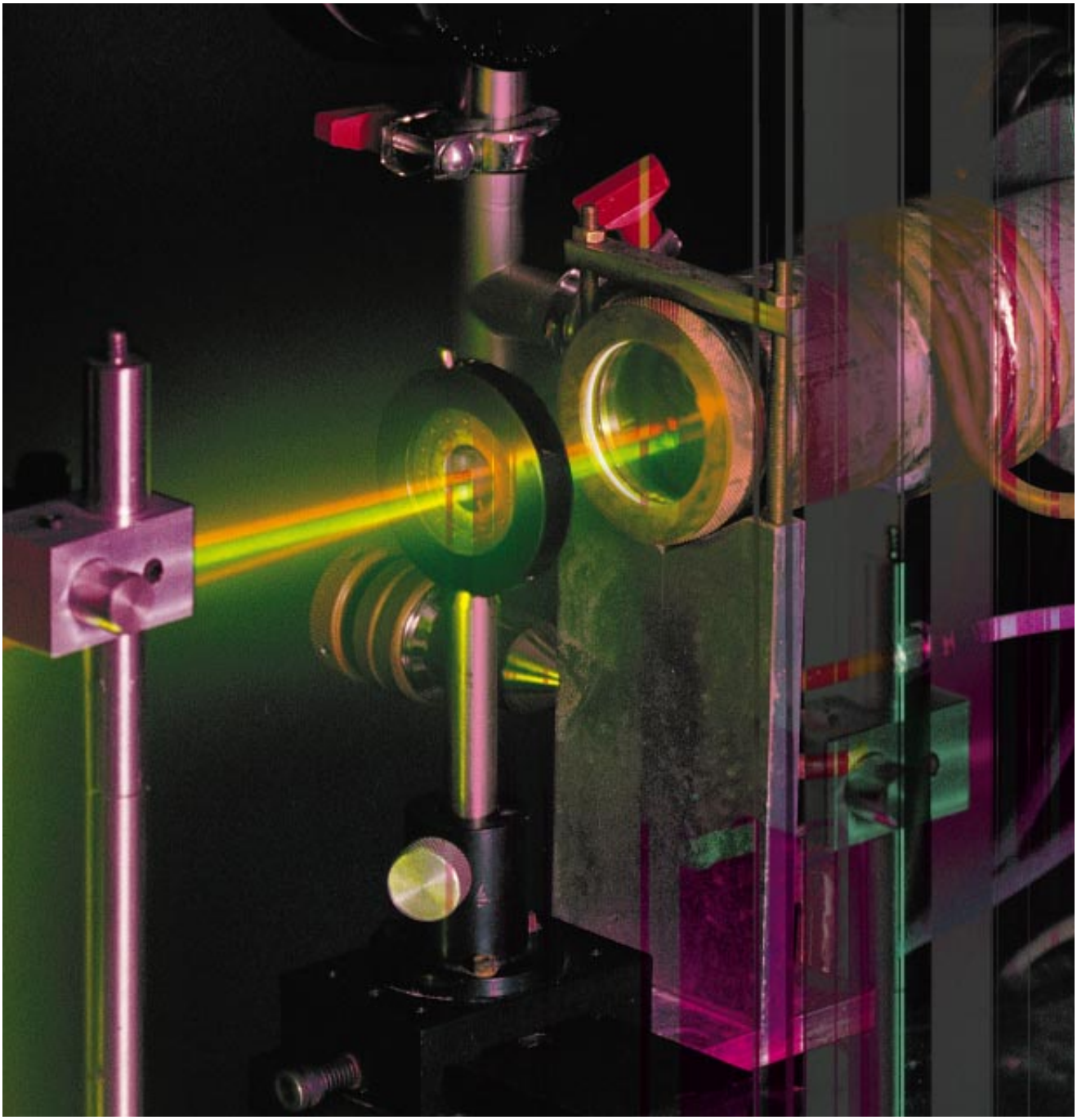
The latest technique to control molecules takes advantage of the fact that light is not the only entity that displays interference. In accordance with a basic

PAUL BRUMER and MOSHE SHAPIRO have been research collaborators since they met at Harvard University more than 20 years ago. Brumer is a professor in the chemical physics theory group at the University of Toronto and is currently a Killam Research Fellow. He recently received the Chemical Institute of Canada's Palladium Medal, one of the organization's highest awards. He earned his Ph.D. from Harvard. Shapiro obtained his Ph.D. from the Hebrew University of Jerusalem. The Jacques Mimran Professor of Theoretical Chemical Physics at the Weizmann Institute of Science in Rehovot, Israel, he has also held several visiting appointments throughout the U.S., Europe and Canada. The authors gratefully acknowledge funding and support from the U.S. Office of Naval Research.

principle of quantum mechanics, particles such as electrons, atoms and molecules can also behave as waves that can interfere with one another [see "The Duality in Matter and Light," by Berthold-Georg Englert, Marlan O. Scully and Herbert Walther; SCIENTIFIC AMERICAN, December 1994]. Experiments have fully confirmed the existence of such matter waves and of the associated interference of particles.

How can chemists exploit interference phenomena to control reactions? Numerous techniques have been developed, but the simplest one trains two different lasers beams on molecules. Each beam excites the wavelike aspect of the molecules in a particular way. These two matter waves can then interfere with each other. Constructive interference, in turn, may result in the formation of a particular product. Destructive interference would result in the absence of that product or the enhanced formation of another. It turns out that we can control the interference pattern—that is, the yield of the reaction products—by adjusting the coherence and intensity properties of the two laser beams.

A more specific explanation incorporates an abstraction known as a wave function, which physicists use to de-



RICK I ROSEN *SABA*

LASER BEAMS fired into a vessel filled with paired sodium atoms, or dimers, can control the breakup of the molecules. The beams, each of a different frequency (appearing as yellow and red), place the molecules into quantum states that interfere with one another. Each molecule subsequently breaks up into one sodium atom in its normal state and another sodium atom in any of several excited states. The particular state formed is governed by the degree of interference.

scribe atomic and molecular systems. Consider a molecule made of three components connected in a line—call them *A, B* and *C*. Because particles behave as waves, physicists can describe the initial state of the molecule with a mathematical entity—a wave function. It embodies all available information about the state of a particle and its motions, or dynamics. For example, the square of the wave function gives the probability of finding the molecule in a particular geometry. Wave functions serve as the basic descriptive tool of quantum mechanics.

Now imagine that we can break the *ABC* molecule apart in two ways. Either the bond between *A* and *B* can break, or the one between *B* and *C*. Thus, two pairs of compounds are possible: *A* and *BC* or *AB* and *C*. To produce either possibility, we must add energy to the molecule. We can irradiate it with a photon of a particular frequency (and thus impart a known amount of energy to it), which the molecule absorbs. If the photon is sufficiently energetic, it brings the *ABC* molecule to the final state—that is, to the energy level at which *ABC* dissociates into two products. The wave function describing the *ABC* molecule in this state incorporates two kinds of information. One is the wave nature of the laser light impinging on the molecule; the second is the quantum-mechanical wave nature of the molecule itself.

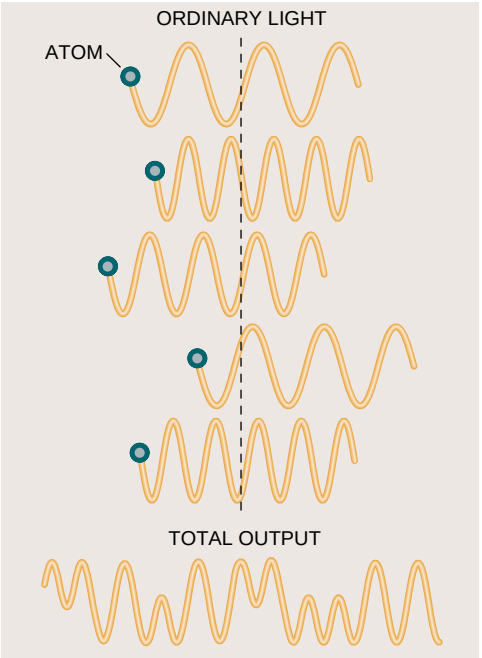
But what if we irradiate the molecule at one third the frequency (that is, with light carrying one third the energy)? In this case, the molecule would have to absorb three photons instead of one in order to dissociate. The molecular wave function at the point of dissociation would then be different, reflecting the fact that the molecule absorbed three photons rather than just one.

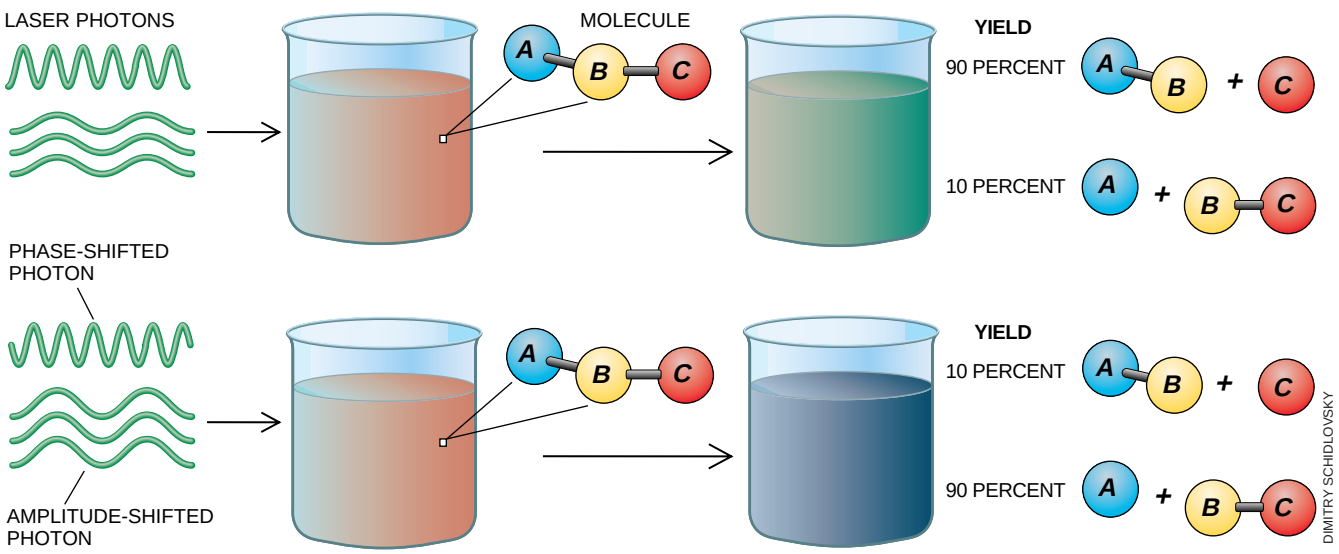
We can control molecular motion if we simultaneously irradiate *ABC* with both light fields. Doing so produces two distinct wave functions at the energy of dissociation. The two modes of excitation interfere, just as coherent light does when it passes through two slits [*see box on page 62*]. It may seem strange that two reaction pathways can interfere, but such interference is the essence of quantum mechanics.

Fortunately for chemists seeking to control molecules, the mathematical term describing the interference differs for the two possible outcomes of the reaction. We can create predominantly *A* and *BC* or mostly *AB* and *C* by adjusting the interference term. The interference, and hence the amount of each product, depends on the relative amplitude and phase of the two original laser beams, and so it can be altered by adjusting these characteristics. Note that because control relies on interference, the method does not require the use of intense lasers. In other words, weak light can have a substantial effect on the dynamics of molecules.

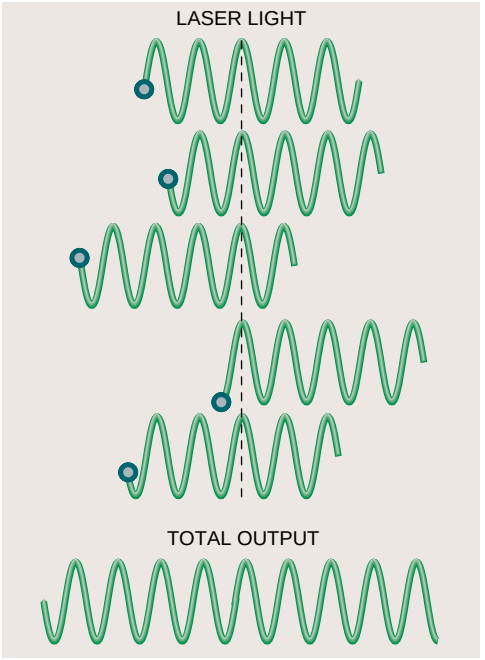
Working with Chi K. Chan, then at the University of Toronto, we studied the dissociation of the diatomic molecule iodine monobromine (IBr). This molecule can separate into I + Br or I + Br*, where the asterisk indicates that the bromine atom has excess energy. Our calculations predict that varying the intensities and relative phase of two la-

PHASE COHERENCE refers to the way atoms emit radiant energy. In ordinary sources such as lightbulbs, excited atoms spontaneously give off light independently of one another. Hence, the light waves from each do not "line up," and the total output over time consists of random emissions (*left*). In laser light (*right*), all the atoms emit in phase, so the output is coherent.





CONTROLLED CHEMICAL DISSOCIATION can be accomplished if a compound is irradiated simultaneously by two types of laser photons, one of which is more energetic than the other. For instance, the molecule *ABC* can dissociate if it absorbs a photon of a certain energy, or three photons, each with one third the energy. The wave functions associated with the excitation by each type of photon quantum-mechanically interfere with one another. The extent of the interference—and hence, the yield—is controlled by adjusting the relative amplitude (the height of the wave) and the phase of the laser light.



DIMITRY SCHIDLOVSKY

ser beams would provide an enormous range of control over the reaction. We would be able to vary the amount of energetic bromine produced to ensure that it accounted for between 25 and 95 percent of the total product formed. This degree of control far exceeds the hopes of conventional industrial chemistry, which seeks to improve reaction selectivity by about 10 percent. Experiments by Daniel S. Elliott of Purdue University have demonstrated the success of this approach for controlling the ionization of atoms. Robert J. Gordon of the University of Illinois at Chicago has also obtained beautiful results in the control of diatomic and polyatomic molecules.

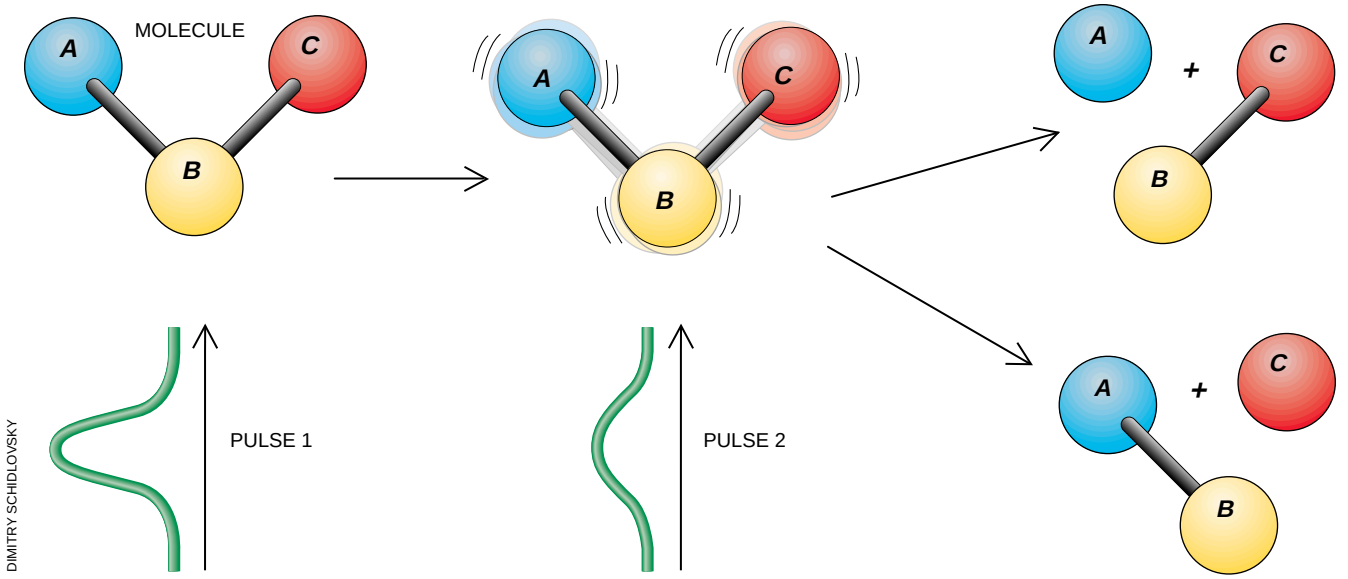
This form of coherent control is not restricted to the use of one photon and three photons at one third the energy. Rather quantum mechanics permits control over the probability that a particular reaction will occur if each of the two paths corresponds to the absorption of either an even number or an odd number of photons. For example, Andre Bandrauk of the University of Sherbrooke in Quebec has computationally displayed extensive control over the photodissociation of a chlorine molecule into energetic and nonenergetic atoms when the molecule absorbs two photons from one laser beam and four photons, each of half the energy, from the second laser beam. Similarly, we have shown that using one photon in conjunction with two photons, each of half the energy, can be used to control the direction that molecules take when they leave the reaction region. Such an ability may render the separation of products easier and thereby boost the efficiency of the reaction.

Current Limitations

Although the concepts behind controlling reactions with coherent light apply to a wide range of isolated molecules, there are at least two obstacles to their immediate broad application. One is that the efficiency of laser control drops substantially when the light waves or the molecular wave functions have ill-defined phases. Loss of phase definition occurs mainly because of collisions between molecules, which increase with higher temperature and pressure and which commonly occur in industrial environments. Further work is necessary to be able to introduce coherent control into modern commercial settings. Consequently, at present, the schemes must be applied in cleverly designed environments or must be restricted to a limited class of reactions. For instance, current technology works well on dilute gases, where the molecules are far apart and so collide less frequently.

The second major obstacle, now beginning to seem surmountable, involves the phase of the laser light. Given two arbitrary laser sources, we generally do not know the extent to which the light from one will be in phase with the other. In addition, the phase difference is affected by any instabilities in the equipment. An unstable phase difference between the two lasers reduces the degree of interference and control.

Sophisticated optical techniques have the potential to eliminate such phase problems. For example, photons may be generated by passing light at one frequency through a particular material that is thereby induced to emit light at another frequency. This process yields two light fields whose phase relations are well defined. We can further control the phase differences between the two light sources by temporarily slowing one light beam relative to the other.



DIMITRY SCHIDLOVSKY

TWO-PULSE LASER STRATEGY can also govern the yield of a chemical reaction. The first pulse places the molecule in a superposition state—in other words, it sets the molecule vibrating and rotating in a particular way. The motion depends on the characteristics of the molecule and the laser. The second pulse causes the molecule to break up. The differences in yield are controlled by varying the time between pulses and by changing the frequencies that make up the pulses.

Interference and Coherent Control

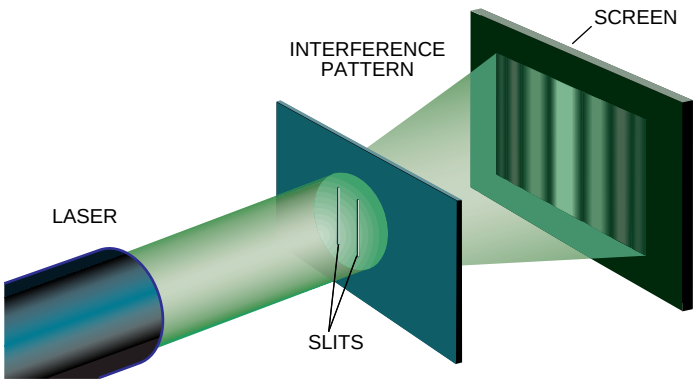
Perhaps the most unusual aspect of quantum mechanics is that, under certain circumstances, matter behaves just as waves do. In particular, it displays interference. Because this property is essential to the ability of lasers to control chemical reactions, some of the mathematics and more technical quantum-mechanical concepts underlying the phenomenon may be of interest.

Interference arises from the way waves add together. The rule for combining many different waves is first to sum their amplitudes (the height of the wave) and then square the result. Consider the interference of two waves whose amplitudes at a given position and time are a and b. The intensity of each wave is a^2 and b^2. The combined amplitude, c, is the sum $c = a + b$, and the combined intensity is $c^2 = (a+b)^2 = a^2 + b^2 + 2ab$.

Note that the combined intensity is not merely the sum of the intensities of each wave (which would be $a^2 + b^2$), but that an additional interference term, $2ab$, contributes. If a and b are both positive or both negative, this interference term is positive. The resulting intensity, c^2, is thus greater than the simple sum of intensities of each wave. The interference in this case is said to be constructive. If a is positive and b is negative (or vice versa), the interference term is negative, and the resulting intensity is smaller than the simple sum of the individual intensities. Interference in this case is said to be destructive.

The variation in intensity produced by interfering light waves can be seen in the famous double-slit experiment, in which a beam of coherent light passes through two slits and onto a screen (*diagram*). The bright regions on the screen arise from the constructive interference between the two beams of light passing through the two slits; dark regions result from the destructive interference. Intensities in between these two extremes arise from the combinations of maxima and minima meeting at the screen.

Now consider replacing the light with particles such as electrons, atoms or molecules. Classical intuition would suggest that the emergent pattern on the screen would resemble two nearly rectangular blobs, resulting from the direct passage of particles through the slits. In fact—and this is the essential feature of quantum mechanics—the observed pattern can clearly display an interference pattern. That is, the particles can show behavior characteristic of waves (as long as the particles are themselves "coherent"—that is, prepared so that they have well-defined wave functions). There is simply no way to explain these observations on the basis of a classical theory of particles.



DIMITRY SCHIDLOVSKY

INTERFERENCE FRINGES are made if coherent light passes through two slits. Particles such as molecules also interfere in this way.

For this reason, the proper description of particle dynamics in quantum mechanics entails a wave function that, as in the case for waves, is described by both an amplitude and a phase. In the double-slit experiment applied to particles, quantum wave functions arriving at the screen via the slits interfere with one another. Most significant is another unusual feature of the quantum world: interference arises because we do not know through which of the slits the particles passed. If we do record such information, the interference pattern disappears.

The two-slit experiment embodies the fundamental quantum principle that two (or more) phase-preserving routes that a molecule can take to some final state can be made to interfere. In the case of the one-photon plus three-photon experiment mentioned in the main body of the article, we do not know which of the two possible excitation routes leads to the observed final state. Hence, the two routes interfere. This optically induced interference forms the foundation for coherent radiative control of molecular processes.

Another way to control phase problems relies on intense laser beams, the focus of many recent investigations. With Zhidang Chen of the University of Toronto, we have shown that such lasers make it possible to bypass the need for radiation having well-defined and carefully controlled phases. In addition, such strong-field methods have the potential to increase substantially the absolute yield of the reaction and to overcome unwanted collisional effects. A particular scheme has just been used experimentally at the Weizmann Institute of Science in Israel by Irit Sofer, Alexander Shnitman, Ilya Golub, Amnon Yogev and us to demonstrate control over the various products formed in the dissociation of paired sodium molecules.

Using Pulses

Because interference of molecular pathways is the key to governing reactions, any laser scenario that induces such interference may serve as a means of controlling reactions. Instead of shining two steady beams on a target, one might use ultrashort pulses of laser light. Modern lasers can generate bursts as short as 10^{-14} second. Unlike continuous-wave radiation, a light pulse is made up of a collection of distinct frequencies and, hence, of a collection of photons with different energies. Such light also has a perhaps counterintuitive property. The briefer the pulse, the broader the range of energies within it.

This property plays a major role in pulsed-laser methods for controlling the outcomes of chemical reactions. By delivering a range of energies, a pulse can induce motion (such as vibration or rotation) in a molecule, which in turn affects the way it interacts with other light pulses. Ordinarily a molecule such as *ABC* exists at a specific (that is, quantized) energy value. A system at one of these fixed energies resides in a so-called stationary state and does not move over

time. For a molecule to undergo the dynamics, it must live in several energy levels at once. Such an assemblage of energy levels is called a superposition state. The wave function describing the superposition state is the sum of wave functions representing stationary states of different energies. To construct it, researchers shine a pulse of coherent laser light on the molecule. The way the molecule then moves depends on the nature of the light pulse and its interaction with the molecule. Thus, we can effect dynamic changes in the molecule by shifting the relative contribution of the frequencies that compose the pulse—that is, by shaping the pulse.

Several researchers have developed these ideas. They include Stuart A. Rice of the University of Chicago, David J. Tannor of Notre Dame University, Herschel Rabitz of Princeton University, Ronnie Kosloff of the Hebrew University of Jerusalem and Kent R. Wilson of the University of California at San Diego. Their results show that pulses built out of a complicated mixture of frequencies are required to control molecular dynamics optimally, but simple approximations often suffice to break apart molecules in a controlled way.

Although a single light pulse can alter the dynamics of a molecule, it does not by itself afford an active means of controlling the yield of a chemical reaction. Rather an idea originally introduced by Rice and Tannor and subsequently extended by us in collaboration with Tamar Seideman, now at the National Research Council of Canada, does allow for control with pulses. Specifically, one needs to employ a two-pulse sequence. The first pulse places the molecule in a superposition state that dictates how that molecule will later respond to the follow-up pulse. The second pulse breaks up the molecule into different products.

Although not apparent, this scenario is similar to that using continuous-wave lasers in that quantum interference between wave functions is responsible for the control. Interference between molecular wave functions, however, is now created by the various frequencies within the two light pulses incident on the molecule. The interference, and hence the yield of the products formed, can be altered easily by varying the interval between the two pulses and the frequencies that make up the first laser pulse. Hence, unlike the continuous-wave laser experiments in which a steady stream of products is formed, pulsed lasers let us take advantage of time as an experimental variable.

Indeed, computational studies by Izhak Levy, formerly at the Weizmann Institute, and us show that the range of control is potentially quite extensive. In an analysis of the dissociation of diatomic molecules, we showed that the yield can be varied so that the desired product accounts for between 3 percent of the total yield and 95 percent, depending on the laser settings selected. One can basically turn a process completely on or off. For polyatomic molecules, control is not as extensive but is considerable nonetheless. We have, for instance, successfully applied this approach to the dissociation of monodeuterated water (HOD), made of hydrogen, oxygen and deuterium, to produce controllable amounts of H + OD or D + OH.

Help for Pharmaceuticals

One of the first real-world applications of laser control may well be enjoyed by the pharmaceutical industry. Currently chemists must take care to ensure that reaction products adopt a specific conformation. Often the same molecule can exist in two forms, known as enantiomers. Like our right and left hands, enantiomers are mirror images of each other. Indeed, such molecules are often referred to as right- or left-handed. Drug companies expend considerable effort to form compounds with the correct handedness because often one enantiomer is biologically active, and the other is either inactive or harmful.

Laser control could be a solution for achieving the correct outcome. We examined the dissociation of a compound that can break up into either right- or left-handed forms—call the substance *ABA′*, where *A* and *A′* are enantiomers. The reaction can produce *A* and *BA′* or *A′* and *BA*. Because *ABA′* is highly symmetrical, traditional dissociation of *ABA′* by absorption of light does not push the reaction in any particular direction; the result is an equal yield of *A* and *A′*. But our studies show that under certain conditions (in particular, in the presence of a weak magnetic field) the two-pulse scheme can be used to control enantiomeric yield so that we can produce *A* rather than *A′*.

Procedures based on quantum interference can do more than control chemical reactions. They can be used to produce entirely novel kinds of technology. The methods can allow workers to select the particular energy state of the products of a chemical reaction. These products could in turn generate laser light at frequencies not obtainable with current equipment. Even more interesting, Paul B. Corkum of the National Research Council of Canada has proposed using interference effects to build lasers that emit supershort bursts of light, on the timescale of 10^{-16} second. This interval is about one tenth the length of pulses from the best lasers in use today.

With Gershon Kurizki of the Weizmann Institute, we have proposed using quantum interference to regulate the flow of electrons in semiconductors. One can design two pathways by which a donor atom loses an electron on absorption of light. These paths can be made to interfere. Controlling this interference means that the direction of the ejected electrons—and hence the direction of the electric current—can be regulated. The result would be a fast optical switch, perhaps on the order of 10^{-12} second, many times faster than present-day switches. Experimental evidence for such directional control has now been obtained by Boris Zeldovich, now at the University of Central Florida, in photoelectric detectors, by Corkum in semiconductor devices and by Elliott in atoms that are ionized by light.

The advent of quantum mechanics introduced new concepts in the understanding of nature. But we are now moving past the role of passive observer. As we approach the 21st century, it is clear that we can extend quantum-mechanical ideas to open up unprecedented possibilities for gaining further control over atomic, molecular and electronic processes.

FURTHER READING

COHERENT PULSE SEQUENCE CONTROL OF PRODUCT FORMATION IN CHEMICAL REACTIONS. D. J. Tannor and S. A. Rice in *Advances in Chemical Physics,* Vol. 70, Part 1, pages 441–523; 1988.

INTERFERENCE BETWEEN OPTICAL TRANSITIONS. C. Chen, Y.-Y. Yin and D. S. Elliott in *Physical Review Letters,* Vol. 64, No. 5, pages 507–510; January 29, 1990.

CONTROLLED PHOTON INDUCED SYMMETRY BREAKING□ CHIRAL MOLECULAR PRODUCTS FROM ACHIRAL PRECURSORS. M. Shapiro and P. Brumer in *Journal of Chemical Physics,* Vol. 95, No. 11, pages 8658–8661; December 1, 1991.

COHERENT LASER CONTROL OF BOUND-TO-BOUND TRANSITIONS OF HCL AND CO. S.-P. Lu, S. M. Park, Y. Xie and R. J. Gordon in *Journal of Chemical Physics,* Vol. 96, No. 9, pages 6613–6620; May 1, 1992.

COHERENT AND INCOHERENT LASER CONTROL OF PHOTO CHEMICAL REACTIONS. M. Shapiro and P. Brumer in *International Reviews in Physical Chemistry,* Vol. 13, No. 2, pages 187–229; September 1994.

COHERENCE CHEMISTRY: CONTROLLING CHEMICAL REACTIONS WITH LASERS. P. Brumer and M. Shapiro in *Accounts of Chemical Research,* Vol. 22, No. 12, pages 407–413; December 1994.

An Efficient Swimming Machine

Instinctive control of vortices lets fish swim the way they do. A robotic tuna has also managed it; boats and submarines may be next

by Michael S. Triantafyllou and George S. Triantafyllou

Over millions of years in a vast and often hostile realm, fish have evolved swimming capabilities far superior in many ways to what has been achieved by nautical science and technology. Instinctively, they use their superbly streamlined bodies to exploit fluid-mechanical principles in ways naval architects today can only dream about, achieving extraordinary propulsion efficiencies, acceleration and maneuverability.

Dolphins, for example, dart through water with impressive grace and apparent ease, playfully bursting through the waves as they follow ships cruising at 20 nautical miles per hour (knots), or about 23 mph. Whereas records of the maximum speeds of fish are not always reliable and are often quite contentious, marine biologists have reported that yellowfin tuna caught on a fishing line can swim at speeds of at least 40 knots. The aggressive pike overcomes its prey with short bursts of acceleration that can exceed that of gravity by about 20 times.

Similarly, detailed observations have shown that fish that depend on aquatic agility for their survival can reverse direction without slowing down and with a turning radius only 10 to 30 percent of the length of their bodies. For comparison, maneuvering ships must reduce their speed by more than 50 percent, and their turning radius is at least 10 times larger than the corresponding value for fish.

Nevertheless, and despite huge potential payoffs, relatively little work has

ROBERTO OSTI

been done to identify and apply the specific features of piscine propulsion that might benefit underwater and surface ships. Certainly the obstacle has not been lack of commercial motive; the immense amounts of cargo and passengers hauled by ship every year worldwide mean that even minute increases in efficiency would result in enormous savings in fuel. Increased maneuverability, moreover, could mean fewer accidents and greater safety for passengers, scientific instruments and the environment.

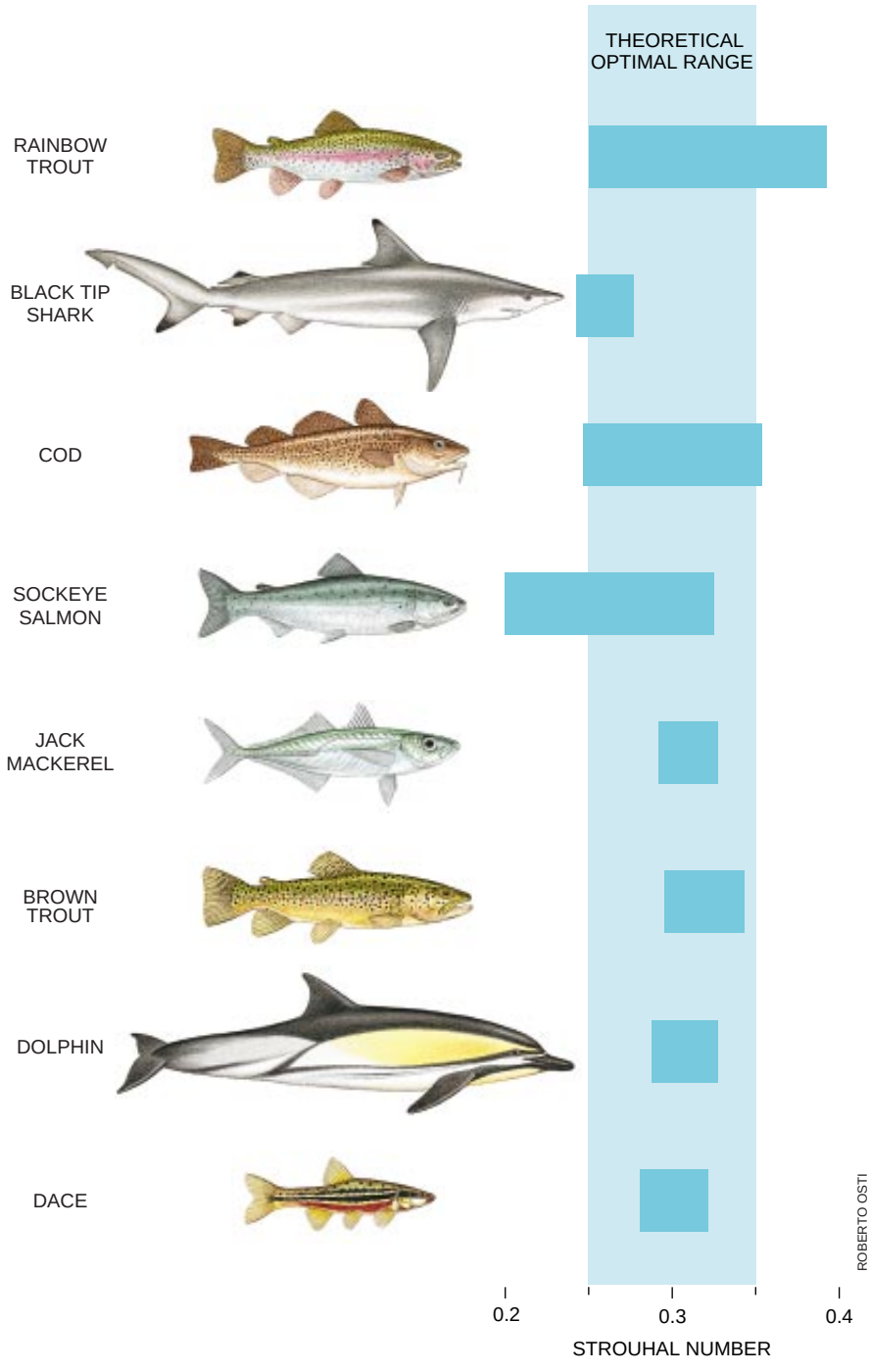
Such intriguing if distant possibilities were the topic of informal conversation with our colleagues at the Woods Hole Oceanographic Institution on Cape Cod, Mass., in the summer of 1989. In many places, the discussions would have been idle. But at Woods Hole, as at the Massachusetts Institute of Technology, the need for advanced and efficient propulsion systems is immediate. The two organizations are among several dozen around the world that are developing robotic, free-swimming craft that will one day explore ocean deeps, undertake military missions and help to maintain offshore oil installations. Extreme constraints on energy storage on board these so-called autonomous underwater vehicles demand propulsors more efficient than the propellers now used.

Developing them, however, would prove challenging. Replicating the performance of a fish by merely imitating its form and function would be impossible, because a smoothly and continuously flexing vehicle, with a fishlike body, is beyond the state of the art of today's robotics. Still, the prospective rewards of the effort were irresistible.

Besides the authors, our team consists of Mark Grosenbaugh of the Woods Hole Oceanographic Institution, Dick K. P. Yue of M.I.T. and a number of our students and postdoctoral associates, most notably Knut Streitlien of City College of New York and David S. Barrett of M.I.T. Our effort complements biological studies such as Lawrence C. Rome's study of power consumed by fish muscle, conducted at the University of Pennsylvania; Richard W. Blake's measurements of the fast-starting performance of pike, carried out at the University of British Columbia; and research into the stability of

MICHAEL S. TRIANTAFYLLOU and GEORGE S. TRIANTAFYLLOU are brothers who independently became interested in fluid dynamics during their undergraduate years at the National Technical University of Athens. Both went on to earn master's and doctorate degrees in ocean engineering from the Massachusetts Institute of Technology, where they collaborated on studies of wakes created by nonstreamlined objects in a flow. For the work on fish swimming described in this article, they gratefully acknowledge the support of the Advanced Research Projects Agency, the Office of Naval Research and the Sea Grant Program at M.I.T. Michael now teaches in M.I.T.'s department of ocean engineering and is director of the Ocean Engineering Testing Tank Facility. George is professor of mechanical engineering and member of the Benjamin Levich Institute for Physicochemical Hydrodynamics of City College of New York. Remarkably, both of them take breaks from their work by going swimming.

FISH ENCOUNTERING VORTICES senses the pressure variations of the spinning eddies as they move along its side. To capture energy from the vortices and boost its swimming efficiency, the fish instinctively times the flapping of its tail to create counterrotating whorls that meet and weaken the encountered ones.



FISH OF ALL KINDS flap their tails to create vortices that produce a jet of high propulsive efficiency. Key parameters describing the jet are related in a ratio known as the Strouhal number, defined as the product of the frequency of tail flapping (*yellow arrows*) times the jet's width (*purple*), divided by the fish's speed (*red*). A Strouhal number between 0.25 and 0.35 is a hallmark of efficient swimming.

fish swimming done by Paul W. Webb of the University of Michigan.

We started out by building simple foils that approximated the swish of a tail closely enough for us to reach new conclusions on the role of vortices in efficient swimming. Bolstered by these results, we built a fairly detailed replica of a bluefin tuna (*Thunnus thynnus*). The robotic, eight-link body and tail mechanism, which we called RoboTuna, let us further refine our findings and served as a prototype for the free-swimming model we are now fashioning.

Delphine Mystery

While planning our machines, we availed ourselves of the long trail of theoretical, experimental and biological studies of how fish swim. In 1936 the British zoologist James Gray created a stir by calculating the power that a dolphin would need to move at 20 knots, as some were reported to do. Gray assumed that the resistance of the moving dolphin was the same as that of a rigid model and estimated the power that the muscles of the dolphin could deliver. His conclusion, known as Gray's paradox, was that the dolphin was too weak, by a factor of about seven, to attain such speeds. The inescapable implication is that there are flow mechanisms at work around the body of the moving dolphin that lower its drag by a factor of seven.

Almost 60 years after its formulation, Gray's paradox has yet to be proved or disproved conclusively. (The biological and hydrodynamic tests that would be needed for scientific certainty require accuracies beyond the state of the art in both fields.) Nevertheless, it has spawned numerous studies and has led to the accumulation of a substantial body of theoretical and experimental results related to fish swimming. Despite all the studies and experiments, however, hardly any useful technologies can be traced even indirectly to the principles of fish swimming. In some earlier efforts, despite promising theoretical foundations, fish-inspired mechanisms performed poorly. Given the remarkable abilities of fish, this seemed to us the true paradox.

If the fish is as efficient a swimming machine as is generally thought, its primary thruster—its tail—must also be quite efficient. One of the puzzles we found from previous work, however, was that experiments conducted with fishlike tails achieved disappointingly low efficiency. Our first task, therefore, was to find out why this was so.

In a motor-driven craft, efficiency is the ratio of useful power (thrust times

forward velocity) divided by the power expended by the motor to drive the foil or propeller. Ideally, all the motor's power would be converted into propulsion, yielding a ratio of one. In practice, efficiency is always less than one, because some of the motor's power is wasted in wayward vortices and other undesirable turbulence as well as heat. For performance, the most important factor is the propulsor's efficiency at reasonably high levels of thrust; a device that is very efficient while producing only low levels of thrust is useless.

Any object in a flow, whether it is a wire in the wind or a swimming swordfish, creates a trail of spinning vortices. The wire obstructs the flow and leaves a wake, whereas the tail of a fish pushes water backward, establishing what is more properly known as a jet—a column of moving fluid that includes thrust-producing vortices. We became convinced that these jet vortices play the central role in the generation of thrust, and we argued that their optimal formation would increase efficiency tremendously.

From previous studies we had done on the vortices produced by a wire in a stream of air, we were well acquainted with a fluid-dynamic parameter known as the Strouhal number. It is the product of the frequency of vortex formation behind an object in a flow and the width of the wake, divided by the speed of the flow. What the number indicates, compactly, is how often vortices are created in the wake and how close they are. Interestingly, the ratio remains constant at about 0.2 for a variety of flow conditions and object shapes.

Although the Strouhal number was invented to describe the wakes behind flow obstructions, the similarities between wakes and jets are such that we realized we could use the number to describe jets. For a swimming fish, we defined the Strouhal number as the product of the frequency of tail swishing and the width of the jet, divided by the speed of the fish.

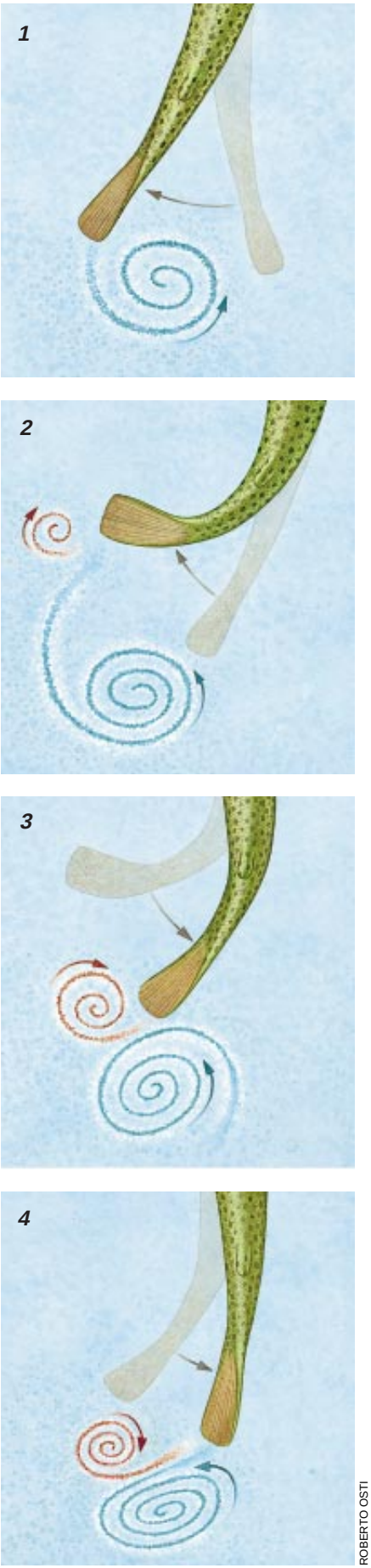
By analyzing data from flapping foils, we found that thrust-inducing vortices form optimally when the Strouhal number lies between 0.25 and 0.35. We anticipated that efficiency should be at a maximum for these values. Some preliminary experiments at the M.I.T. testing tank confirmed that the efficiency of a flapping foil does indeed peak when the Strouhal number is in this range.

With Grosenbaugh's collaboration, we subsequently analyzed a large amount of data collected about swimming fish. We found that fish of all sizes, from goldfish to sharks, swing their tail within the theoretically determined Strouhal number range of 0.25 to 0.35 [*see top illustration on opposite page*]. To show the formation of vortices and turbulence clearly, we conducted a separate trial in which we placed a small tropical fish in fluid that contained a suspension of tiny particles. By measuring the speed of the fish, as well as the frequency and the amplitude of its flapping tail, we calculated a Strouhal number of 0.30. Having satisfied ourselves of the number's importance in achieving high efficiency, we calculated its value for previous flapping-foil experiments that had reported disappointing efficiencies. None were even close to the 0.25 to 0.35 range. Returning to our laboratory with renewed zeal, we adjusted our foils to operate in this range—and measured efficiencies higher than 86 percent. In contrast, the small propellers used to drive underwater vehicles are typically no more than 40 percent effective.

Why Foils Are Efficient

What makes the high efficiency and high thrust of our foils possible is the manner in which the vortices are arranged behind the foil (or a fish's tail). The vortices become stronger as the load increases, but their rotational direction is always compatible with the desired direction of thrust, producing an efficient jet. A propeller, on the other hand, generates a long jet that rotates in the direction of propeller rotation, which is perpendicular to the direction of motion and needed thrust. All the power that goes into rotating this jet is wasted. The only way to minimize it and improve efficiency is to load the propeller very lightly, typically by giving it the largest possible diameter.

Another striking result from these experiments concerns the relation between efficiency and "angle of attack," the instantaneous angle between a foil's direction of motion and the plane formed by its leading and trailing edges. In our experiments, we found efficiency was at its peak when the largest angle of attack was between 15 and 25 degrees. This finding indicates the fun-

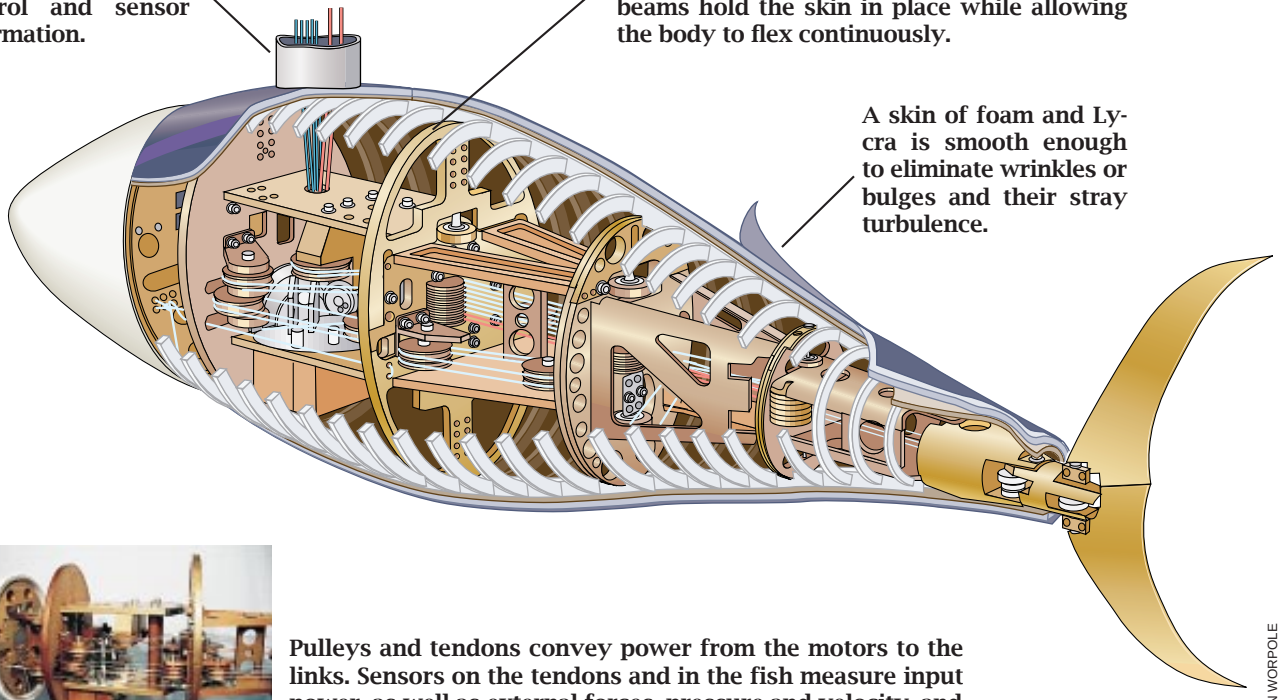


ROBERTO OSTI

FORCEFUL FLAP, followed in quick succession by another one in the reverse direction, produces a strong, sudden thrust well suited to pouncing on prey or a fast getaway. The initial flap makes a large vortex (*1*), and the second flap creates a different, counterrotating vortex (*2, 3*). Strong forward thrust and a stray but manageable lateral force result when the two vortices meet and combine to create a jet and are pushed away from the tail, weakening each other (*4*).

A Tuna of Aluminum and Lycra

The body of a robotic tuna consists of aluminum links connected by hinges. Six motors, external to the robot, supply the power to mimic the undulatory swimming of a real tuna. Separate systems of pulleys and tendons transfer torque from each motor, while isolating the motion of the links. As it swims in the Ocean Engineering Testing Tank Facility at the Massachusetts Institute of Technology, the robotic fish hangs from a carriage (*photograph, right*).



IAN WORPOLE

DAVE BARRETT

Pulleys and tendons convey power from the motors to the links. Sensors on the tendons and in the fish measure input power, as well as external forces, pressure and velocity, and track vortices as they move along the robot's side.

damental difference between steady airplane flight and flapping propulsion. The basic principles of fixed-wing flight require that, to avoid a stall, wings generally be kept to an angle of attack well below 15 degrees. Noticeable stall did not occur with the foil until the angle exceeded 30 degrees.

These results show that the criteria that indicate a stall for fixed wings do not apply to a flapping foil. True, what causes a stall in both cases is the sudden formation of uncontrolled vortices—above and behind the wings, in the case of aircraft, disrupting the normally smooth flow over them. With the flapping foil, however, vortices do not in and of themselves cause a stall. In fact, vortices—properly controlled and arranged—are essential to a foil's efficient operation, so it should be no surprise that they can be controlled over a wider range of angles to produce useful thrust.

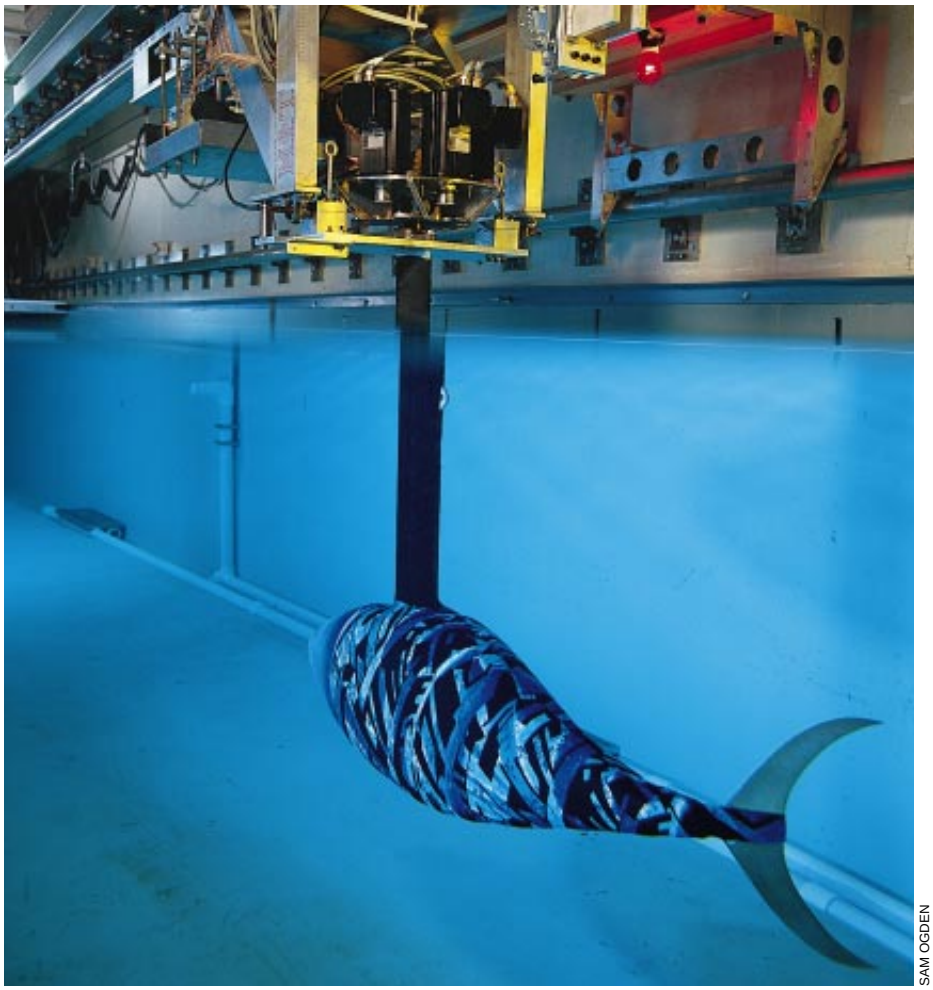
These findings, along with those concerning the Strouhal number, suggest that a properly designed foil could be a very attractive propulsor for ships, motor yachts and underwater vehicles. Given its natural advantages and the fact that development of the proper motors and gears is well within today's technological capabilities, this foil might be the first fish-inspired technological application. The use of an even number of countermoving foils, properly positioned, could minimize unpleasant swaying or vibration. Of course, future shipbuilders would also have to address structural reliability, the hydrodynamic shapes of sterns and other variables.

Trick of the Tail

The coincidence of high thrust and efficiency is not the only advantage of a flapping foil. It also offers the possibilities of more flexible operation, more maneuverability and, most intriguingly, tempting opportunities for recapturing kinetic energy from a wake.

Fish instinctively exert precise and effective control of the flow around their bodies to extract energy from waves, turbulence and even their own wakes. They have also evolved ways of controlling the flow so as to enhance their turning and starting. The underlying principles are not unique to fish or even to flapping propulsion. A propeller mounted on a ship is somewhat more efficient than one tested in a tank, because the moving propeller recovers some of the energy from the wake. The phenomenon is routinely exploited by ship designers.

Fish, marine animals and their mechanical imitators, however, are much better suited to this kind of control. Frolicking and leaping in the wakes of ships for miles on end, dolphins are clearly recovering energy by positioning their bodies and flapping their tails appropriately, as Gray noted decades ago and as Neil Bose of Memorial Uni-



SAM OGDEN

versity of Newfoundland has studied more recently. Fish can also recover energy from vortices in the ocean or even from vortices spun off from their own bodies.

The extraction of energy from unsteady flows using a stationary foil is called the Katzmayr effect, after the German engineer who first studied it, in 1927. In 1991 we explored a related phenomenon by placing a flapping foil some distance behind a cylinder in a stream. Rows of vortices generated by the cylinder moved toward the foil, which we could pitch and move sideways to encounter them in various positions. Systematic experiments confirmed that these adjustments could enhance or decrease efficiency.

Specifically, when the timing was right, vortices created by foil oscillations met incoming vortices spinning in the opposite direction. This effect weakened the vortices in the wake, resulting in the capture of energy by the foil and an increase in its efficiency. This mode, obviously the most desirable for acceleration or high-speed swimming, is only one of three possible with such a setup. With a shift in timing, we induced vortices spinning in the same direction to meet and reinforce one another, causing a strong jet flow with no immediately obvious practical use. In the third situation, we paired counterrotating vortices to create mushroom-shaped eddies; if generated by a fish's tail, they would slow the creature down. Overall we could vary the efficiency of the foil by a factor of at least two, depending on the mode.

This one set of findings cast new light on a diverse set of observations. Photographs taken by the American engineer Moe William Rosen in 1959 of the flow behind a small, fast fish clearly show vortices from the creature's tail interacting with oppositely spinning vortices from its body. It has been reported as well that fish such as salmon and trout exploit oncoming vortices, such as those created behind rocks, to boost their swimming efficiency during their arduous upstream voyages. Aided by a continuous parade of such vortices, it is even possible for a fish's swimming efficiency to exceed 100 percent.

Vorticity control is also fundamental to the astounding transient performance of some fish, whose fast starts, sudden accelerations and maneuverings are far superior to those of ships and submarines. What makes this agility possible, in essence, is the ability to produce sudden, very large forces. Ships and submarines, on the other hand, exert no control over the flow around their hulls and move at a slow pace, their very large wakes with uncontrolled vortices creating enormous drag forces.

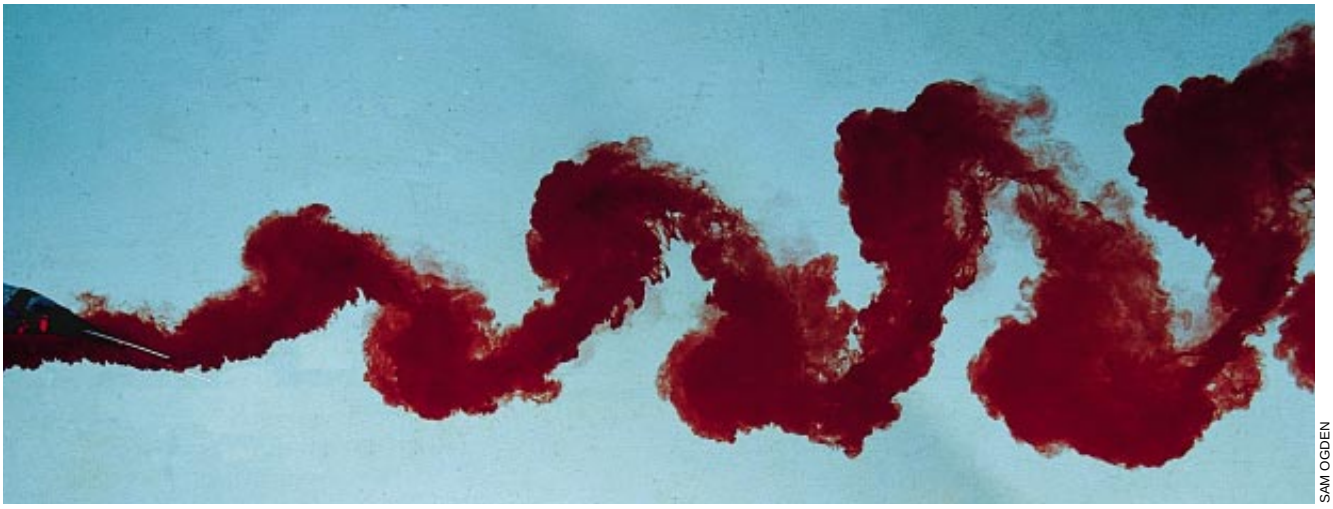
The control of vortices offers a novel solution. The idea is to produce favorable pressure gradients and then control them to optimize the response. Specifically, pitching and heaving a foil to a maximum angle and then back again produces a strong, sudden force, ideally suited to maneuvering and a fast start. The motion gives rise to a large initial vortex, followed quickly by another one spinning in the opposite direction. Sudden forward thrust, as well as a lateral force, results when the second vortex is briefly trapped between the first one and the surface of the foil.

This maneuver is exactly what a fast-starting, agile fish does with its tail. Just before shooting off in some direction, its body flexes sharply, with the forward half of the body oriented at 60 to 120 degrees with respect to the ultimate direction of motion. Such orientation is necessary for the fish to compensate for the lateral force that accompanies the thrust.

A Motorized Bluefin

As useful as the flapping foils were in elucidating the hydrodynamics of fish swimming, the real proof of the principles, as well as the first step toward transferring the technology, lies in constructing an artificial fish that uses them to swim. About two years ago we had become confident enough to begin doing just that. We selected the bluefin tuna as our model because of its well-known ability to cruise but also because of its size, which would fit nicely in M.I.T.'s testing tank.

The body and tail of the 49-inch RoboTuna are flexed by an eight-link mechanism of anodized aluminum, driven by six brushless motors and an assembly of strings and pulleys [*see box on these two pages*]. A set of densely packed "ribs" and a special skin of re-



SAM OGDEN

TAIL OF ROBOTIC TUNA spins off a trail of vortices of alternating orientation, made visible by dye: the tail swings to one side, creating a clockwise vortex, and then to the other, causing a counterclockwise one. Precise control of the timing and spacing of vortices is the main reason fish of all kinds swim as efficiently and skillfully as they do.

ticulated foam and conformal Lycra allow smooth flexing and keep stray turbulence to a minimum. We attached the entire assembly to a carriage, on which we mounted all the motors and control and communication equipment. A single strut encloses the cables for data and power.

Several sensors along the side of RoboTuna record flow pressure, just as fish use their "lateral line" sense organs to detect pressure variations. Along with force and motion transducers, they permit detailed evaluation of swimming forces and propulsive efficiency. Simultaneous measurement of forces lets us directly link flow features to swimming performance and also control the flow to enhance the model's propulsion and maneuvering. Soon the side-mounted pressure transducers will enable us to experiment with closed-loop control of vorticity, so that RoboTuna, like its natural counterpart, will be able to move its tail in response to oncoming vortices and the flow around its body. We made the flow around the robotic fish visible by using either dyes or a laser beam that causes microscopic particles in the water to phosphoresce.

In a few months, another generation of RoboTuna will take shape. We expect to begin building a free-swimming model, borrowing on the technology developed for the existing robot. This successor will be used to develop still more advanced technologies, based on our growing understanding of flow-control mechanisms, for possible application to commercial and naval vehicles.

More important, it will be a test bed for improving maneuvering and fast starting of vehicles using the fast generation and manipulation of large vortices. Such capabilities could prove invaluable even in oceanographic research, where underwater vehicles must sometimes operate in forbidding or confined environments. Near thermal vents, for example, temperatures can shoot up to hundreds of degrees Celsius in the space of a few feet or in a few seconds. In cluttered spaces, too, agility can sometimes stave off a collision or a catastrophic failure.

Nothing Like the Real Thing

The more sophisticated our robotic-tuna designs become, the more admiration we have for its flesh-and-blood model. Aware that we will never match the perfection of design of the living creature, we strive instead to uncover natural, useful mechanisms optimized by millions of years of evolution. Once identified, a kind of reverse engineering may enable us to devise novel ways of using these mechanisms. In time, these biologically inspired creations may even outperform their natural antecedents in useful ways—for instance, in surveying a stretch of seafloor.

This goal is the guiding principle of the emerging science of biomimesis. By focusing research efforts and guiding the selection of parameters, details of the behavior and instincts of highly adapted, successful creatures can be a great asset in developing certain robots and other useful systems. Our project has required us to pose and answer fundamental questions about the mechanics of swimming.

For example, is fish swimming a simple perfection of hydrodynamic principles, constrained only by the mechanical limitations of muscle? Although the dolphin and the tuna are both fast and flex their bodies in similar ways while swimming, there are significant differences in the details of swimming as well. Are they both optimal solutions? If one is better than the other, is the superiority limited to certain situations? More important, as far as we are concerned, is there an even better design than either of them for swimming?

These are among the questions we hope to address with our next robotic tuna. The state of the art in mechanical systems suggests that it will take our best efforts to approach the breathtaking abilities of its living model, but we will be patient. After all, in the span of a few years we are learning processes that took eons to develop.

FURTHER READING

ANIMAL LOCOMOTION. James Gray. Weidenfeld & Nicolson, 1968.

SWIMMING AND FLYING IN NATURE, Vol. 2. Edited by Theodore Y. T. Wu, C. J. Brokaw and Christopher Brennen. Plenum Press, 1975.

LOCOMOTION BY SCOMBRID FISHES: HYDROMECHANICS MORPHOLOGY AND BEHAVIOR. J. J. Magnuson in *Fish Physiology*. Edited by W. S. Hoar and D. J. Randall. Academic Press, 1978.

OPTIMAL THRUST DEVELOPMENT IN OSCILLATING FOILS WITH APPLICATION TO FISH PROPULSION. G. S. Triantafyllou, M. S. Triantafyllou and M. A. Grosenbaugh in *Journal of Fluids and Structures,* Vol. 7, No. 2, pages 205–224; February 1993.

ACTIVE VORTICITY CONTROL IN A SHEAR FLOW USING A FLAPPING FOIL. R. Gopalkrishnan, Michael S. Triantafyllou, George S. Triantafyllou and David Barrett in *Journal of Fluid Mechanics,* Vol. 274, pages 1–21; September 10, 1994.

The Genetic Basis of Cancer

An accumulation of genetic defects can apparently cause normal cells to become cancerous and cancerous cells to become increasingly dangerous

by Webster K. Cavenee and Raymond L. White

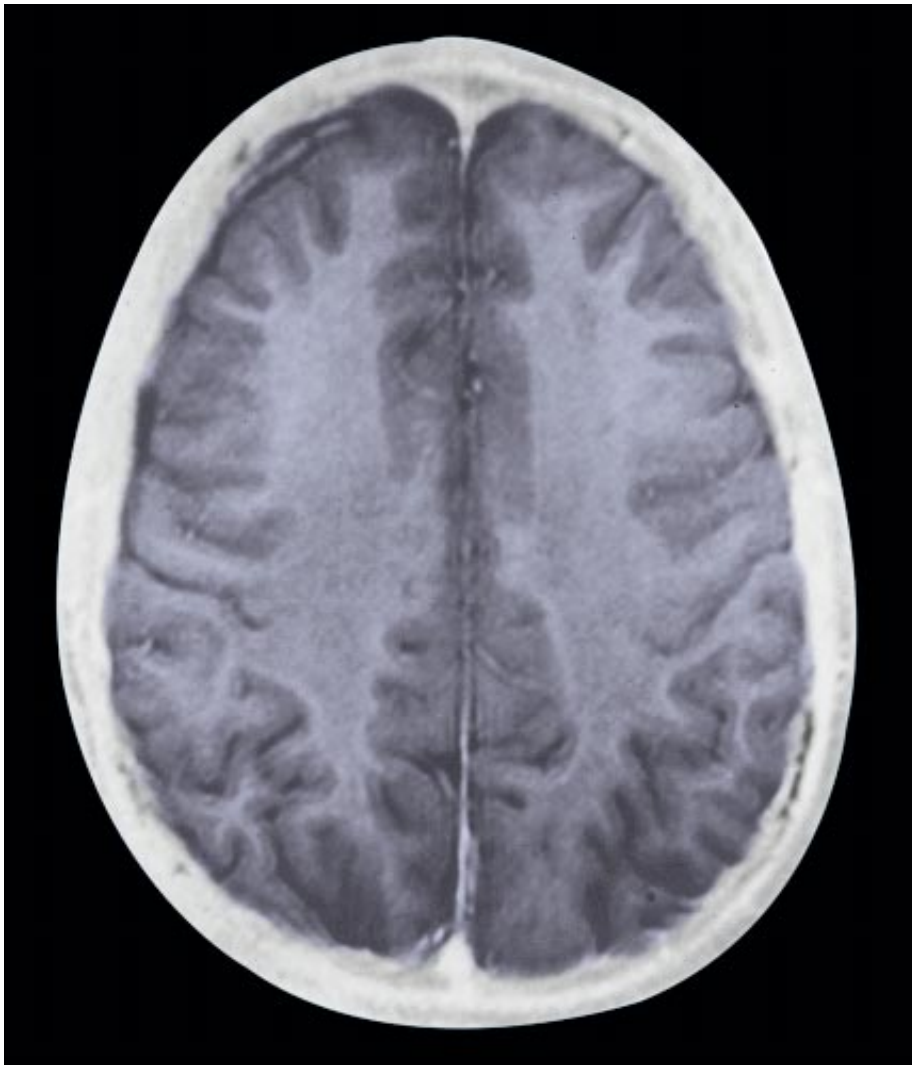
Patients stricken with cancer feel as if they have been invaded by an alien force. Yet malignancies arise from our own tissue. In fact, the weight of evidence today indicates that cancers generally derive from a single cell that is changed dramatically by a series of genetic alterations.

A healthy cell has a well-defined shape and fits neatly within the ordered array of cells surrounding it. It responds to the dictates of its environment, giving rise to daughter cells solely when the balance of stimulatory and inhibitory signals from the outside favors cell division. But the process of replication, or growth, carries the constant hazard of genetic mutations: random changes that can impair the regulatory circuits of a cell. If a single mutation occurs, the newly damaged cell, which may look normal and be slightly less responsive to external messages, may occasionally undergo unscheduled cell division.

Eventually, an accumulation of genetic damage can cause a daughter cell to become quite deaf to external messages and to display the signs of malignancy. In particular, it loses its distinctive shape and boundaries, ceases to respond to growth-inhibiting signals and gains the ability to replicate uncontrollably. The resulting mass, in turn, can compress and damage healthy tissue in its vicinity. What is worse, it can invade the barriers that separate one organ from another and can metastasize, establishing new colonies at distant sites.

Studies carried out over the past 20 years have begun to identify many of the genes that take part in this progression from normalcy to cancer. The ongoing research is confirming and extending early proposals that cancer develops primarily because cells suffer irreversible damage to particular classes of genes. It is also creating opportunities for improved diagnosis and therapy.

The emerging view of tumor progression reflects a convergence of several lines of research, the oldest of which still involves painstakingly looking at cells through a microscope. By 1914, for instance, the German cytologist Theodor Boveri had concluded from such



WEBSTER K. CAVENEE and RAYMOND L. WHITE collaborated at the University of Utah in the early 1980s. Cavenee is director of the Ludwig Institute for Cancer Research at the San Diego branch. He is also professor of medicine and member of the Center for Molecular Genetics at the University of California, San Diego. Before accepting his current posts, he had faculty appointments at the University of Cincinnati and McGill University. White, who has been on the faculty of the University of Utah since 1980, is director of the Huntsman Cancer Institute and chairman of the department of oncological sciences. This is his second article for *Scientific American.*

observations that malignant cells had abnormal chromosomes and that any event leading to such aberrancy would cause cancer.

Microscopic observations became considerably more specific after 1970, when new staining techniques, together with improved equipment, made it possible to distinguish each of the 23 pairs of chromosomes that collectively contain all the genes forming the blueprint for a human being. (All human cells, except for sperm and eggs, carry two sets of chromosomes—one inherited from the mother and one from the father.) Each chromosome takes up the stain in specific regions and thus becomes marked by a characteristic series of light and dark bands, a kind of bar code identifying the individual chromosome.

By comparing stained chromosomes from normal cells with those from tumors, investigators noted many different signs of genetic disarray in cancers. The chromosomes of tumors were often broken, with some of the pieces joined to other chromosomes. Individual chromosomes were present in multiple copies rather than the normal two. Whole chromosomes, or sometimes internal segments, seemed to have disappeared entirely. Unfortunately, until the 1980s researchers generally lacked the tools they needed to determine whether the chromosomal rearrangements were among the causes of cancer or were a by-product of its development.

Two Hits

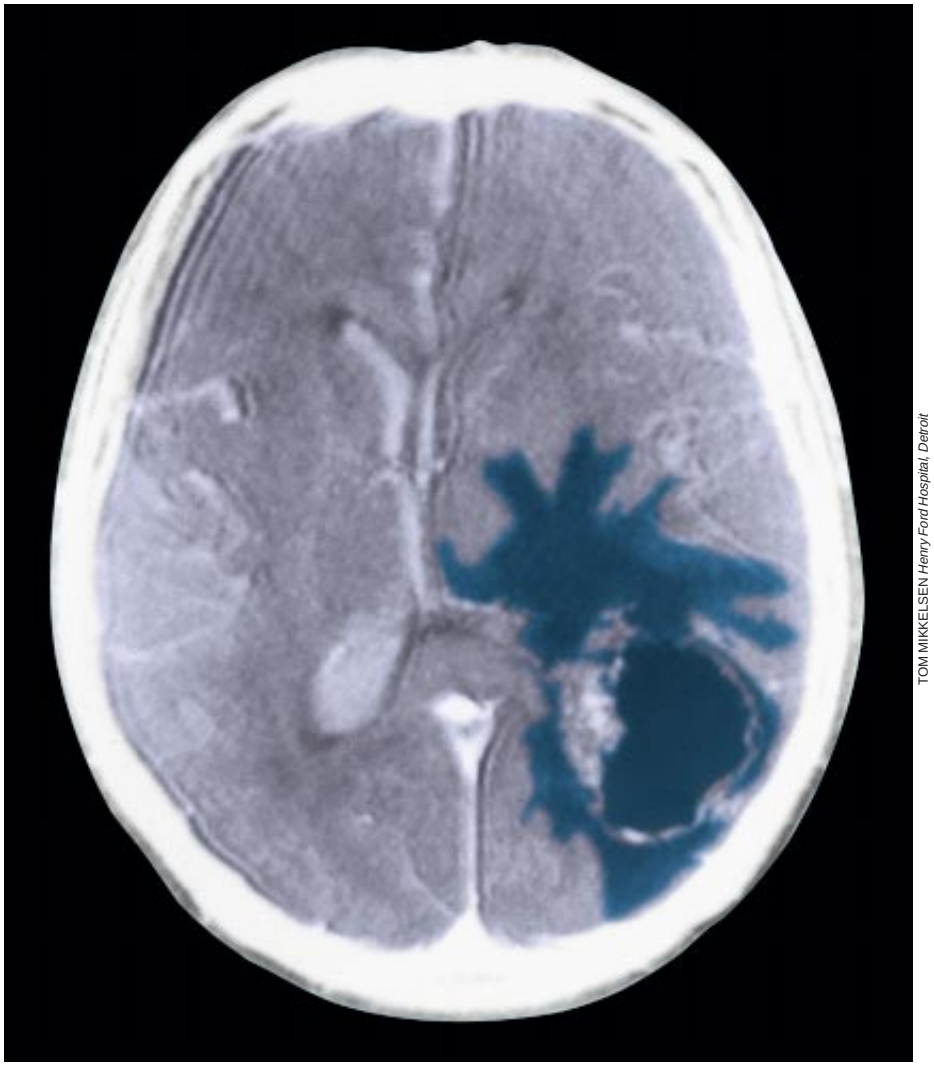
Quite different evidence that genes had a role to play came from observations that some extended families suffered an unusually high incidence of certain cancers. When particular diseases "run" in families in predictable patterns, an inherited defect is usually at fault.

Yet the discovery that some cancers could apparently be inherited also raised perplexing questions. A genetic defect passed to a child through the sperm or egg should appear in every cell of the body. Why, then, did people with inherited disease typically acquire only one or a few cancers and only at discrete sites? Further, did the existence of familial cancers necessarily mean that sporadic (nonfamilial) disease, which is much more common, also had a genetic basis? Or did sporadic cancers arise by completely different processes than inherited ones?

A proposal put forward in 1971 by Alfred G. Knudson, Jr., now at the Fox Chase Cancer Center in Philadelphia, seemed to offer an answer to both questions, although it took about a decade for his ideas to gain broad acceptance. Knudson had been puzzling over the cause of retinoblastoma, a rare childhood disorder in which malignant tumors develop in the retina before the age of six. He noted that sometimes the disease occurred in both eyes, but most of the time it affected only one eye. Moreover, children who were affected bilaterally often had close relatives afflicted with retinoblastoma.

A statistical analysis comparing the age at onset for each form of the disease showed that the bilateral type was usually diagnosed at an earlier age than was the unilateral type. Also, the shape of the age distribution curves suggested to Knudson that retinoblastoma resulted from two cellular defects arising at separate times. In bilateral disease the first defect was probably inherited and present in all cells of the body from the moment of conception. In unilateral disease the first defect probably arose during development or later and perhaps exclusively in retinal cells. In both cases, however, a tumor formed only if the first defect in a retinal cell was later accompanied by a second, independent one. Knudson's two-hit theory, as it is frequently called, turns out to be essentially correct for all cancers, not just retinoblastoma, although more than just two hits are often required.

The need for two hits—now known to constitute damage to genes—explains why patients in cancer-prone families are not riddled with tumors throughout their bodies: inheritance of just one genetic defect predisposes a person to



TOM MIKKELSEN *Henry Ford Hospital, Detroit*

CANCER OF THE BRAIN progressed in just three months from being invisible on a scan (*left*) to covering a large area of one hemisphere (*blue area, right*). The patient, whose initial complaint was an uncontrollable twitching in one eye, died two months after the second image was made. Recent evidence indicates that brain tumors and other malignancies arise when multiple genetic mutations combine to free a single cell from normal restraints on proliferation, invasiveness and movement.

cancer but does not cause it directly; a second event is required. Knudson's intuition that the causes of sporadic and familial cases can involve the same biochemical abnormalities has also been confirmed. But even back in the 1970s his insights provided justification for thinking that research aimed at discovering genetic and other cellular aberrations in rare familial cancers could shed light on the processes leading to sporadic malignancies.

Oncogenes Take Center Stage

As various researchers focused on the genetics of familial malignancies, other workers convinced that genes were at the root of cancer were taking a rather different approach to finding cancer-related genes. It had been known for many years that viruses can cause tumors in animals. That link had spurred a great deal of research aimed at identifying the cancer-causing genes carried by the viruses and at finding the host genes that were affected. Those efforts revealed, surprisingly, that the genes implicated in malignant diseases were often altered forms of human genes that the viruses had picked up during their travels. Other times the viruses activated host genes that were usually quiescent.

The normal versions of the pirated and activated genes—now called proto-oncogenes—carry codes specifying the composition of proteins that encourage cells to replicate. These growth-promoting genes come in many varieties. Some specify the amino acid sequences of receptors that protrude from the cell surface and bind to molecules known as growth factors. When bound by such factors, receptors issue an intracellular signal that ultimately causes cells to replicate. Others of the genes code for proteins that lie inside the cell and govern the propagation of the intracellular growth signal. Still others encode proteins that control cell division.

Discovery that the viral genes had human counterparts introduced the intriguing possibility that human cancers—including the majority not caused by viruses—might stem from mutations that convert useful proto-oncogenes into carcinogenic forms, or oncogenes. Consistent with this notion, studies indicated that alteration of just one copy, or allele, of these proto-oncogenes was enough to transform—render cancerous—some types of cells growing in culture. Such dominant mutations cause cells to overproduce a normal protein or to make an aberrant form that is overactive. In either case, the result is that stimulatory signals increase within the cell even when no such signals come from the outside.

Later studies supported a role for oncogenes—and also complicated matters. Notably, in 1982 and 1983, investigators in France and the U.S. conducted studies similar to the original cell-culture experiments, but with an important difference. Because normal cells would not grow indefinitely in a culture dish, those earlier studies had relied on rodent cells that were unusual in their ability to proliferate for a long time in culture. To eliminate this possibly confounding influence, François Cuzin of the University of Nice, Robert A. Weinberg of the Massachusetts Institute of Technology and H. Earl Ruley, then at Cold Spring Harbor Laboratory in New York State, asked whether single oncogenes could also transform normal rodent cells.

They found that mutations in at least two proto-oncogenes had to be present and that only certain combinations of mutations led to malignancy. These results suggested that individual oncogenes, though potentially quite powerful, were not able to cause tumors by themselves. A major effort was then launched to see whether human tumors carried oncogenic alterations of the types and combinations that were able to transform cells in culture.

For a while it seemed that oncogenes might explain most cases of cancer. This view was strengthened by discovery of more than a dozen of them in human tumors. The results were ultimately disappointing, however; a mere 20 percent of human tumors turned out to carry the expected alterations singly, and none of them had the pairs of cooperative alterations found in cultured cells. At the time, it also appeared that the inherited mutations responsible for predisposing people to familial cancers were not oncogenes. These were all strong hints that the full story was yet to be told.

Enter Tumor Suppressor Genes

Even before those hints attracted much attention, the two of us were beginning to suspect that damage to a different kind of gene might play a part in cancers. Such genes came to be known as tumor suppressors because many of them code for proteins that inhibit cell replication. In contrast to the mutations that activate oncogenes, mutations of these genes, we believed, would be recessive: they would affect cell function only when both alleles were damaged or lost. In testing this idea, we relied on new technology we had developed for the more general purpose of following the inheritance of genes and chromosomes through extended families [see "Chromosome Mapping with DNA Markers," by Ray White and Jean-Marc Lalouel; SCIENTIFIC AMERICAN, February 1988].

In the early 1980s, while collaborating at the University of Utah, we realized that our technique—which involved tracking genetic markers (identifiable segments of DNA) in tissues—could be used to determine whether segments of chromosomes carried by normal cells were missing in a tumor. For instance, if a selected region of a chromosome was deleted in a tumor, we could spot that loss by observing that a marker known to travel with that region was also missing.

Our experiments were focused by earlier studies of Jorge J. Yunis of the University of Minnesota and Uta Francke of Yale University. That research indicated a gene on chromosome 13 might be involved in retinoblastoma. With our DNA-marker technology, we were able to demonstrate in 1983 that large segments of chromosome 13 were missing in cells taken from sporadic as well as inherited retinoblastomas. This new evidence strongly supported the idea that the two hits hypothesized by Knudson could consist of the physical or functional loss of one allele of a gene followed by elimination of or damage to the normal copy. The missing DNA on chromosome 13, now known as the *RB* (retinoblastoma) gene, was isolated by Stephen H. Friend of Weinberg's laboratory in 1986 [see "Finding the Anti-Oncogene," by Robert A. Weinberg; SCIENTIFIC AMERICAN, September 1988].

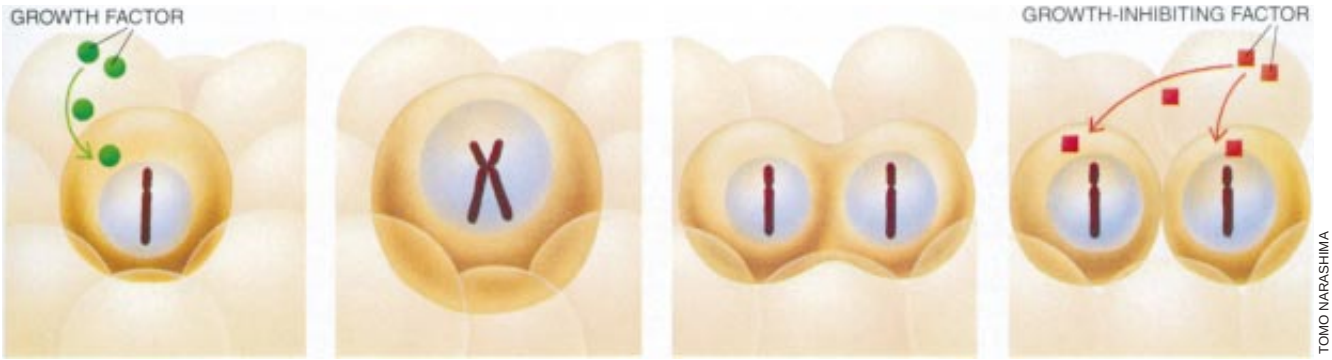
Subsequent studies have shown that recessive loss of the *RB* gene occurs in other cancers as well. What is more, inactivation or loss of DNA has now been shown to be a major feature in the genesis of every solid cancer examined so far. Breast cancer, prostate cancer, lung cancer, bladder cancer, pancreatic cancer and many others are marked by the disruption or elimination of multiple tumor suppressor genes.

By the late 1980s, then, there was good evidence that mutations in both proto-oncogenes and tumor suppressors could participate in causing cancer. It seemed reasonable to guess that some kinds of cancer resulted from a combination of such mutations. But did the mutations collect in the same cell or did some affect one cell, and others, different cells? A model of tumor progression proposed in the 1950s by Leslie Foulds of the Chester Beatty Research Institute in London and expanded in the 1970s by Peter C. Nowell of the University of Pennsylvania suggested that if both kinds of mutations were in-

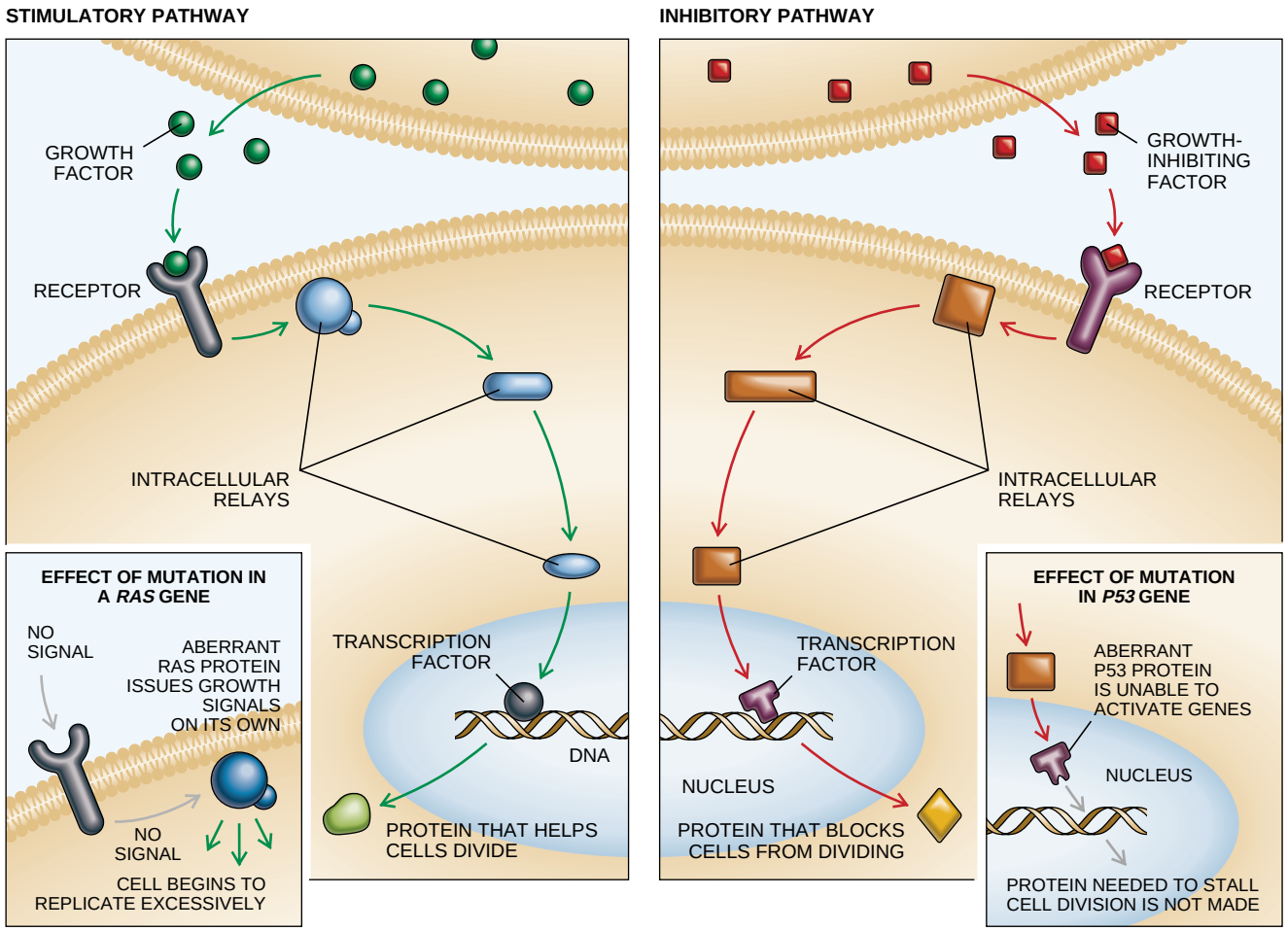
volved, they would accumulate in one cell and its direct descendants.

In this scheme, cancers are thought to arise and become more dangerous through a process known as clonal evolution. First, a single cell undergoes a genetic mutation that enables it to divide under conditions that cause normal cells to stop replicating. Because the inappropriately dividing cells copy their DNA and give identical sets to their offspring, the next generation of cells carries the same changes and shows the same inappropriate growth. Later, one of these cells or their descendants undergoes a mutation that further enhances its ability to escape normal regulation, perhaps allowing it to pass through surrounding tissue and enter the bloodstream. This mutation, too, is passed to daughter cells. Repetition of the process enables one cell to accumulate the mutations it needs to metastasize and colonize other organs.

If the theory were correct, it would mean the majority of cells in a tumor would carry the same defects. That being the case, therapy capable of counteracting one or more of those defects would be effective against all, or a great



TOMO NARASHIMA



JARED SCHNEIDMAN DESIGN

NORMAL CELL REPRODUCES ITSELF (*sequence at top*) in response to stimulation by external growth factors (*green*); it stops dividing in response to inhibitory factors (*red, far right*). For either reaction to occur, messages from the factors must be relayed deep into the target cell (*large panels*). Many cancer-causing genes are abnormal versions of ones that code for proteins in stimulatory pathways (*left panel*). The altered genes, called oncogenes, cause stimulatory proteins to be overproduced or overactive. In one example, mutation of a particular *ras* gene can lead to synthesis of a hyperactive ras protein (*inset at left*). Many other cancer-related genes code for proteins in inhibitory pathways (*right panel*) and are often called tumor suppressors. Damage to these genes can promote cancer if the defects prevent inhibitory proteins from being made or functioning properly—as often occurs when the *p53* gene is mutated (*inset at right*).

majority, of the cancer cells—a feature that is essential for eradicating any malignancy. For this reason and others, we set out to see if we could find evidence for the clonal evolution of tumors. One of us (White) focused primarily on colon cancer, and the other of us (Cavenee) on brain tumors. As part of this work, we had to identify many of the genes involved in these cancers.

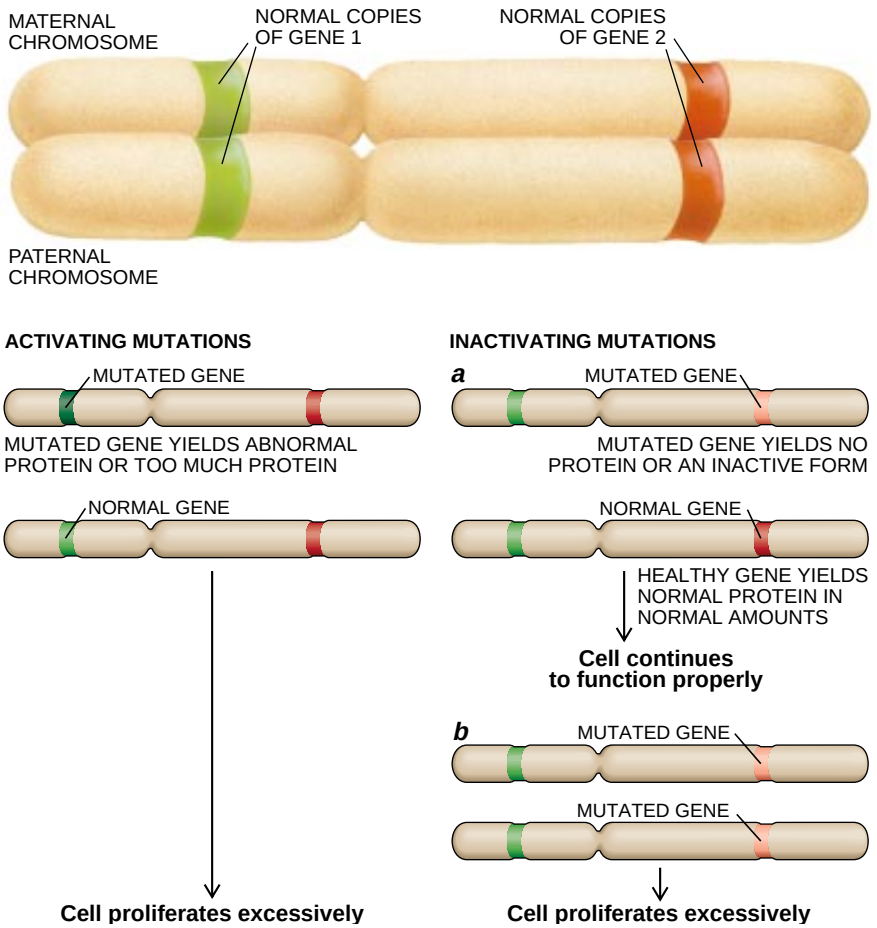
The Genetics of Colon Cancer

White turned to colon cancer in part because it usually emerges from a well-defined precursor—the colon polyp. If a cancer developed in a clonal fashion, mutations arising in an early stage of tumor development would be expected to be present in later stages, and each successive stage would be marked by additional mutations. To test this expectation experimentally, it is necessary to collect samples from the successive stages and compare their genes. In colon disease, samples are fairly easy to obtain. As a polyp, which is initially microscopic, becomes larger and more irregular, it becomes readily accessible to the gastroenterologist (who removes it for therapeutic purposes) and thus to the experimentalist.

Colon cancer also held appeal for our purpose because families that were genetically prone to a rare disease called familial adenomatous polyposis had been identified and were available for study. In affected individuals the colon becomes carpeted with hundreds or thousands of polyps, one or more of which is likely to become cancerous in midlife. Clearly, an inherited defect in some gene—called *APC* (for adenomatous polyposis coli)—was necessary for polyp formation and, in turn, for the development of colon cancer in such patients. It also seemed possible that appearance of a defect in the *APC* gene was one of the earliest steps, if not the first step, leading to many cases of sporadic colon cancer. If that gene could be isolated, these ideas could be tested, and investigators would have at least one of the genes needed for evaluating whether colon cancer developed in a clonal manner.

In 1987 Mark Leppert in White's laboratory at Utah and Walter F. Bodmer and his colleagues at the Imperial Cancer Research Fund in London separately demonstrated, through use of the marker technology described earlier, that the *APC* gene resided near the middle of the long arm of chromosome 5. Intensive work, often collaborative, by White's laboratory and those of two other investigators—Yusuke Nakamura of the Cancer Institute in Tokyo and Bert Vogelstein of Johns Hopkins University—eventually revealed the precise location of the gene. The research also identified several inherited *APC* mutations that appeared in sporadic as well as familial colon tumors. This work thus defined a first step in the evolution of colon cancer. It also provided additional confirmation of the speculation that the same genes are often mutated in both inherited and sporadic tumors.

The groups found, too, that all the cancer-related mutations in the *APC* gene led to production of an incomplete protein. Evidently, cells could operate relatively normally if they retained one normal *APC* allele and thus made some amount of the full APC protein. But if both alleles became damaged, a needed brake on replication disappeared. The precise function of the *APC* gene is unclear, but now that the gene is in hand, its normal responsibilities and its role in cancer should soon be defined.

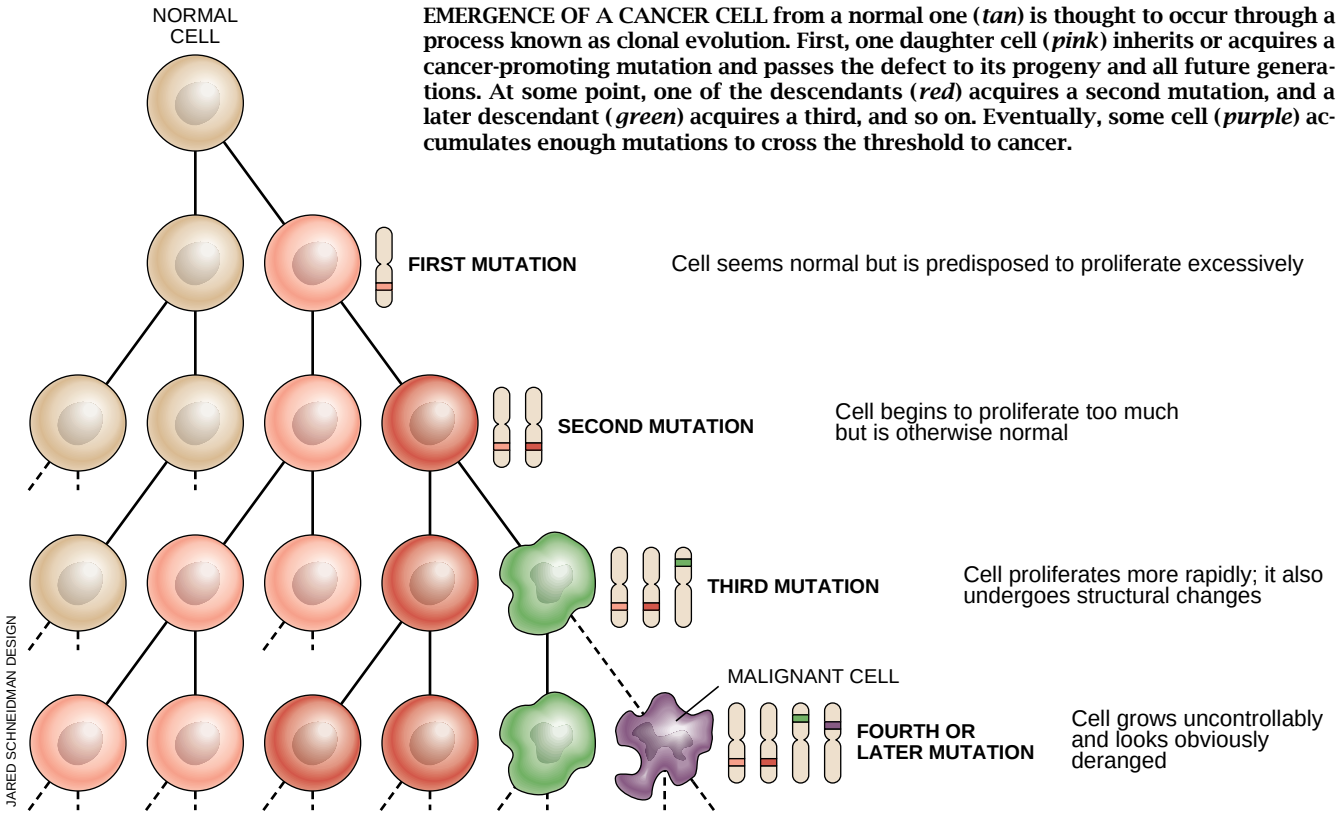


GENES ARE INHERITED IN MATCHING PAIRS—one from the mother and one from the father (*top*). Sometimes mutation of a single copy pushes a cell toward cancer (*left*)—such as when it leads to production of a protein that activates excessive cell division. (Oncogenic mutations fall into that category.) Other times both copies must be altered—such as when a gene coding for a protein that stalls cell division is inactivated (*right*). If only one copy of such a gene is affected (*a*), the other copy can still generate the needed protein. But if both copies are hobbled (*b*), an important brake on tumor development is lost.

Multiple Defects

The steps that follow immediately after the *APC* gene is inactivated are still obscure. In many cases, however, later mutation in a single allele of a particular proto-oncogene seems to push a polyp toward malignancy. This gene, as Manuel Perucho observed when he was at Cold Spring Harbor Laboratory, is one of several *ras* genes. The protein normally made under the direction of this gene sits under the cell membrane and relays stimulatory messages from growth factor receptors to other molecules in the cytoplasm. The mutant version does not wait for signals from the outside but issues its own autonomous growth signals.

Vogelstein and his group have shown that large polyps and colon cancers often carry only mutated copies of two additional tumor suppressor genes. One



EMERGENCE OF A CANCER CELL from a normal one (*tan*) is thought to occur through a process known as clonal evolution. First, one daughter cell (*pink*) inherits or acquires a cancer-promoting mutation and passes the defect to its progeny and all future generations. At some point, one of the descendants (*red*) acquires a second mutation, and a later descendant (*green*) acquires a third, and so on. Eventually, some cell (*purple*) accumulates enough mutations to cross the threshold to cancer.

is *p53,* which resides on chromosome 17 and is now known to be involved in many different cancers. The normal protein product of this gene functions in several biochemical pathways, including those enabling a cell to repair damage to DNA. The other is a gene—probably *DCC* (for deleted in colorectal cancer)—that resides on chromosome 18. *DCC* codes for a protein that appears on the cell surface and helps colon cells stick to one another.

The discovery that genetic changes in the *APC* gene occur early and persist, whereas other changes appear only in later stages, fits well with the theory of clonal evolution. But that conclusion was initially statistical and based on examining tissues removed from many different patients. That approach could not demonstrate conclusively that mutations appearing in one generation of cells are passed to later generations of those same cells. Another strategy, however, provided more convincing results.

Sometimes the polyp from which a cancer has emerged can be identified at the edge of a cancer. By comparing the DNA in a polyp with that in its adjacent cancer, Vogelstein showed that every mutational hit found in a polyp also appeared in the corresponding cancer, as would be expected if the tumor formed by clonal evolution. Further, the cancer invariably included mutations that were not found in the polyp, as would also be expected if the added mutations accounted for the increased aggressiveness of a cancer. For instance, some polyps carried a *ras* mutation without a *p53* defect, but the cancers growing from the polyps had both mutations. As yet, there is no strong evidence that mutation of *ras, p53* and *DCC* genes must happen in any particular order for a polyp to become cancerous, although the *ras* mutation seems to come first fairly often.

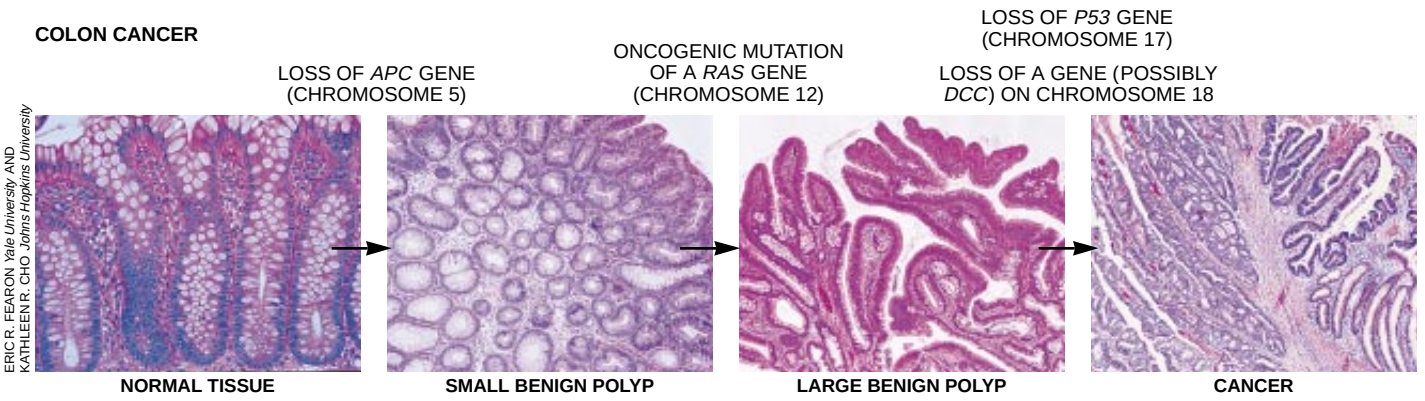
Brain Tumors Reveal Their Secrets

In spite of these encouraging findings, study of colon cancer has a major analytical limitation. To truly demonstrate that a given clone of cells is undergoing progressive changes in its genes, one needs to examine the same tumor over time. In the case of colon cancer, tumors are almost always removed at the earliest stage of detection. Such practice makes good clinical sense, but it prevents sequential observations. This consideration led Cavenee to seek out a disease in which removal of a tumor is sometimes followed by the reappearance of the tumor in a more aggressive form at the same site. In 1987, while he was at the Ludwig Institute for Cancer Research at McGill University, he and his co-workers settled on cancers known as astrocytomas—the most common tumors that originate in the brain.

Cancer of the brain is defined somewhat differently than it is in other tissues. In that organ, cells do not need to invade connective tissue or metastasize in order to be lethal; sadly, proliferation at a site critical to survival can sometimes be enough to kill a patient. Hence, most masses in the brain are called cancers. Cavenee's group examined progression of astrocytomas from their less malignant to more malignant stages, as determined by the size and shape of the tumors and by the structure of their constituent cells.

When the investigators began this work in 1987, they did not have the blueprint of genetic change that was emerging for colon cancer. They therefore began by laying the groundwork for future studies of individual patients. They obtained tumors from many different patients, grouped them according to stages, or grades, of advancing disease, and compared the genetic derangements found in each stage.

Over the next four years they made good headway. They learned, for instance, that tumors of every grade had inactivating alterations in chromosome 17, in a gene they had not yet identified. Moreover, the proportion of tumors displaying the mutation in the lowest stage was equal to that in all other stages; this pattern is a sign that the mutation came early and was retained. If a mutation generally occurred later in disease,

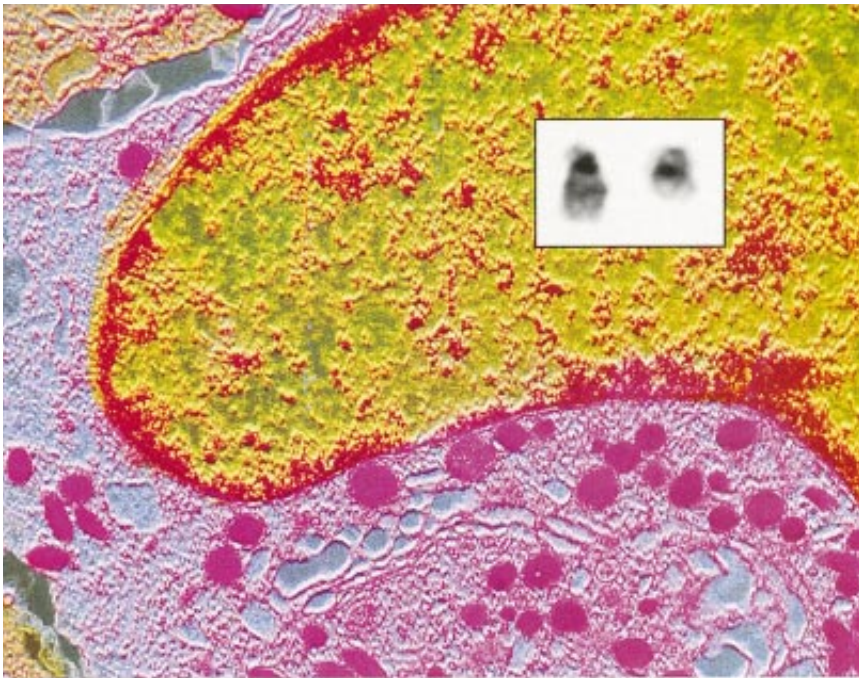


ERIC R. FEARON *Yale University* AND KATHLEEN R. CHO *Johns Hopkins University*

GENETIC CHANGES indicated at the top are among those thought to participate frequently in the development of colon cancer (*left*) or in the progression of a common brain cancer (astrocytoma) from its mildest to its most aggressive

the frequency would rise in the later stages. By the end of the 1980s Vogelstein's laboratory established that mutations in the *p53* gene, on chromosome 17, were among the most common alterations in human cancer. Subsequent analysis of Cavenee's tissue samples confirmed his growing suspicion that the chromosome 17 mutation was actually a defect in the *p53* gene.

Aware that a particular region of chromosome 9 was deleted in other kinds of brain tumors, C. David James on Cavenee's team, in conjunction with V. Peter Collins of the Ludwig Institute in Stockholm, examined this chromosome as well. Middle- and late-stage astrocytomas, but not early ones, often showed a loss in both copies of this chromosome. Thus, the deletion probably encouraged progression to middle-stage tumors from a lesser stage. The lost region contains a cluster of genes that code for proteins known as interferons. Such proteins can draw the attention of the immune system to diseased cells, and so elimination of their genes presumably helps cancer cells evade immune destruction. The missing region may additionally include two newly discovered genes, called *multiple tumor suppressors 1* and *2,* whose protein products are involved in regulating cell division. Disappearance of any of these genes could potentially contribute to a variety of cancers.



PETER C. NOWELL *University of Pennsylvania*

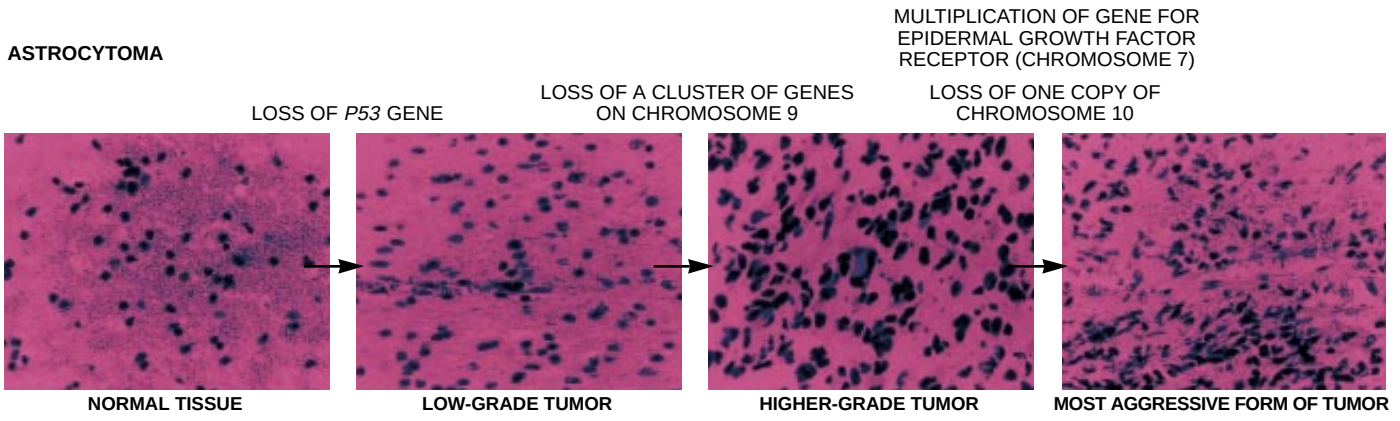
PHILADELPHIA CHROMOSOME (*at right in inset*) was the first chromosomal abnormality ever linked to a specific cancer. In the 1960s Peter C. Nowell of the University of Pennsylvania observed that the appearance of an unusually small chromosome in white blood cells was a hallmark of leukemia. It is now known that the aberrant structure forms when a normal version of chromosome 22 (*at left in inset*) swaps genetic material with another chromosome, in the process giving up more than it receives. Unfortunately, the DNA gained by chromosome 22 combines with a preexisting gene to form a hybrid oncogene.

The tissue studies also extended reports by Axel Ullrich of Genentech, Michael D. Waterfield of the Ludwig Institute in London and Joseph Schlessinger of the Weizmann Institute of Science in Israel that chromosomes in astrocytomas often carry more than one copy of the gene specifying the receptor for epidermal growth factor. Because each copy can be used to make the protein, cells will carry extra receptors on their surface. That abundance, in turn, can cause cells to overreact to the presence of the growth factor. This alteration seems to participate in bringing tumors from a middle to a late stage of disease.

Finally, Cavenee's group found that virtually all the end-stage tumors examined were missing one copy of chromosome 10 and that the loss was rare in earlier stages. This pattern says the loss is probably involved in advancement to the most virulent stage. Regrettably, though, we do not yet know which gene or genes on the lost chromosome are most important to the progression.

These results suggested by 1991 that formation of brain tumors involves, at a minimum, inactivation of the *p53* gene, loss of a gene on chromosome 9, oncogenic amplification of the gene for the epidermal growth factor receptor and, at a very late stage, loss of at least one copy of chromosome 10. But stronger proof that astrocytomas are caused by the accumulation of these, and possibly other, defects in cells required examining genetic changes in the cancer of single individuals over time.

At about that time Tom Mikkelsen joined Cavenee's laboratory and took on the challenge of comparing the genetic makeup of original astrocytomas with that of later recurrences arising at the same sites. This task was impossible earlier not only because the genes involved were not known but also be-



form (*right*). Other genes not listed here play roles as well. Unless otherwise indicated, the term "gene loss" indicates that both copies of a tumor suppressor gene have been damaged or deleted. The images show magnified slices of tissue.

cause matched pairs of tumors are hard to obtain. A patient seen initially at one institution may be cared for elsewhere when the cancer returns. Also, physicians do not remove tumors that reappear if it is thought that surgery is unlikely to extend survival. Luckily, however, two distinguished clinicians—Mark L. Rosenblum of the University of California at San Francisco and Karl Schwechheimer of Albert Ludwigs University in Freiberg, Germany—had come forward with collections of frozen tissue that included a few matched sets.

To Cavenee's satisfaction and delight, the genetic analysis of these tissues—done in collaboration with David Sidransky in Vogelstein's group—fulfilled the predictions of the theory of clonal evolution. The initial tumors possessed fewer mutations than did the recurrences. These alterations included one or more of the genetic hits (such as damage to chromosome 17) that had been identified in the low-grade tumors analyzed previously. And, most significant, the corresponding high-grade versions possessed each alteration found in the primary tumor as well as additional defects (of the kinds identified in the earlier studies). For reasons that are not obvious, progression of astrocytomas seems to follow a more defined sequence of genetic changes than is apparent in colon cancer.

Next on the Agenda

The collected results we have described offer strong support for the idea that cancer develops and becomes more dangerous primarily because cells in a single lineage accumulate defects in genes that normally regulate cell proliferation. Changes in other kinds of genes, many of which have not yet been identified, presumably facilitate the ability of tumors to grow, invade local tissue and establish distant metastases. Hormones and other factors in the environment around the genetically altered cells almost certainly enhance their genetically defined deregulation.

Questions remain. Why do cell types differ in the mix of mutations they require in order to become cancerous? And how is it possible for five or more mutations to accumulate in cells? After all, the probability is actually quite small that any given cell bearing a permanent mutation in a cancer-related gene will independently gain another mutation in such a gene.

Newly discovered genetic aberrations found in a second form of inherited colon tumors (hereditary nonpolyposis colon cancer) may offer a partial answer to the last question. The affected genes specify proteins responsible for identifying and repairing mistakes made when DNA in a replicating cell is copied. If these repair genes themselves are damaged, the number of mutations passed to daughter cells will go up dramatically. The daughter cells may then deliver DNA carrying still more mutations to their progeny. Defects in repair genes may thus play a role in making late-stage tumors highly aggressive. They may even account for the astonishingly fast rate at which some tumors arise and become killers.

Mutations in certain genes can also be especially devastating if the mutations have multiple effects. As a case in point, damage to the *p53* gene can apparently do more than release a brake on proliferation. Certain mutations seem to reduce the ability of cells to limit blood vessel formation. As extra vessels grow in a tumor, they help to nourish the mass and to serve as conduits through which malignant cells can spread to distant sites. In parallel, the abnormal proteins yielded by the altered gene may aid tumor cells in resisting the destructive effects of radiation.

As investigators gain clarity on the specific groups of genetic changes that lead to and exacerbate particular forms of cancer, their insights should point the way to practical benefits for patients. When the mutations follow in a fairly set sequence, their identification in a patient's tumor should be of value for clarifying the stage of disease and thus for tailoring therapy to the individual's needs. In addition, knowledge of the genes that are mutated in a primary tumor may make it possible to detect recurrences of some cancers earlier than is now possible—by spotting mutations that have occurred in tissues not yet displaying detectable masses.

Expanded understanding of the genetic bases of cancer can also be expected to lead to the introduction of drugs that will counteract the effects of selected mutations and thereby slow tumor development or halt it altogether. Some evidence suggests it may not be necessary to correct the effects of every mutation; doing so for one or two genes may well prove to be sufficient for taming renegade cells.

The process by which normal cells become cancerous and grow ever more dangerous is undoubtedly even more complicated than has been discovered so far. But continued investigation of the genetic changes underlying specific cancers seems a rational way to tease apart many of those complexities—and to gain new leads for treatment.

FURTHER READING

THE CLONAL EVOLUTION OF TUMOR CELL POPULATIONS. Peter C. Nowell in *Science,* Vol. 194, pages 23–28; October 1, 1976.

GENETIC AND EPIGENETIC LOSSES OF HETEROZYGOSITY IN CANCER PREDISPOSITION AND PROGRESSION. Heidi J. Scrable, Carmen Sapienza and Webster K. Cavenee in *Advances in Cancer Research,* Vol. 54, pages 25–62; 1990.

A GENETIC MODEL FOR COLORECTAL TUMORIGENESIS. Eric R. Fearon and Bert Vogelstein in *Cell,* Vol. 61, No. 5, pages 759–767; June 1, 1990.

TUMOR SUPPRESSOR GENES. Robert A. Weinberg in *Science,* Vol. 254, pages 1138–1146; November 22, 1991.

Bonobo Sex and Society

The behavior of a close relative challenges assumptions about male supremacy in human evolution

by Frans B. M. de Waal

At a juncture in history during which women are seeking equality with men, science arrives with a belated gift to the feminist movement. Male-biased evolutionary scenarios—Man the Hunter, Man the Toolmaker and so on—are being challenged by the discovery that females play a central, perhaps even dominant, role in the social life of one of our nearest relatives. In the past few years many strands of knowledge have come together concerning a relatively unknown ape with an unorthodox repertoire of behavior: the bonobo.

The bonobo is one of the last large mammals to be found by science. The creature was discovered in 1929 in a Belgian colonial museum, far from its lush African habitat. A German anatomist, Ernst Schwarz, was scrutinizing a skull that had been ascribed to a juvenile chimpanzee because of its small size, when he realized that it belonged to an adult. Schwarz declared that he had stumbled on a new subspecies of chimpanzee. But soon the animal was assigned the status of an entirely distinct species within the same genus as the chimpanzee, *Pan.*

The bonobo was officially classified as *Pan paniscus,* or the diminutive *Pan.* But I believe a different label might have been selected had the discoverers known then what we know now. The old taxonomic name of the chimpanzee, *P. satyrus*—which refers to the myth of apes as lustful satyrs—would have been perfect for the bonobo.

FRANS B. M. DE WAAL was trained as an ethologist in the European tradition, receiving his Ph.D. from the University of Utrecht in 1977. After a six-year study of the chimpanzee colony at the Arnhem Zoo, he moved to the U.S. in 1981 to work on other primate species, including bonobos. He is now a research professor at the Yerkes Regional Primate Research Center in Atlanta and professor of psychology at Emory University.

The species is best characterized as female-centered and egalitarian and as one that substitutes sex for aggression. Whereas in most other species sexual behavior is a fairly distinct category, in the bonobo it is part and parcel of social relations—and not just between males and females. Bonobos engage in sex in virtually every partner combination (although such contact among close family members may be suppressed). And sexual interactions occur more often among bonobos than among other primates. Despite the frequency of sex, the bonobo's rate of reproduction in the wild is about the same as that of the chimpanzee. A female gives birth to a single infant at intervals of between five and six years. So bonobos share at least one very important characteristic with our own species, namely, a partial separation between sex and reproduction.

A Near Relative

This finding commands attention because the bonobo shares more than 98 percent of our genetic profile, making it as close to a human as, say, a fox is to a dog. The split between the human line of ancestry and the line of the chimpanzee and the bonobo is believed to have occurred a mere eight million years ago. The subsequent divergence of the chimpanzee and the bonobo lines came much later, perhaps prompted by the chimpanzee's need to adapt to relatively open, dry habitats [see "East Side Story: The Origin of Humankind," by Yves Coppens; SCIENTIFIC AMERICAN, May 1994].

In contrast, bonobos probably never left the protection of the trees. Their present range lies in humid forests south of the Zaire River, where perhaps fewer than 10,000 bonobos survive. (Given the species' slow rate of reproduction, the rapid destruction of its tropical habitat and the political instability of central Africa, there is reason for much concern about its future.)

If this evolutionary scenario of ecological continuity is true, the bonobo may have undergone less transformation than either humans or chimpanzees. It could most closely resemble the common ancestor of all three modern species. Indeed, in the 1930s Harold J. Coolidge—the American anatomist who gave the bonobo its eventual taxonomic status—suggested that the animal might be most similar to the primogenitor, since its anatomy is less specialized than is the chimpanzee's. Bonobo body proportions have been compared with those of the australopithecines, a form of prehuman. When the apes stand or walk upright, they look as if they stepped straight out of an artist's impression of early hominids.

Not too long ago the savanna baboon was regarded as the best living model of the human ancestor. That primate is adapted to the kinds of ecological conditions that prehumans may have faced after descending from the trees. But in the late 1970s, chimpanzees, which are much more closely related to humans, became the model of choice. Traits that are observed in chimpanzees—including cooperative hunting, food sharing, tool use, power politics and primitive warfare—were absent or not as developed in baboons. In the laboratory the apes have been able to learn sign language and to recognize themselves in a mirror, a sign of self-awareness not yet demonstrated in monkeys.

Although selecting the chimpanzee as the touchstone of hominid evolution represented a great improvement, at least one aspect of the former model did not need to be revised: male superiority remained the natural state of affairs. In both baboons and chimpanzees, males are conspicuously dominant over females; they reign supremely and often brutally. It is highly unusual for a fully grown male chimpanzee to be dominated by any female.

Enter the bonobo. Despite their common name—the pygmy chimpanzee—bonobos cannot be distinguished from the chimpanzee by size. Adult males of

FRANS LANTING *Minden Pictures*

BONOBO FEMALE interacts with an infant. Juvenile bonobos depend on their mothers for milk and transport for up to five years. They are extremely well tolerated by adults, who have rarely been seen to attack or threaten them.

the smallest subspecies of chimpanzee weigh some 43 kilograms (95 pounds) and females 33 kilograms (73 pounds), about the same as bonobos. Although female bonobos are much smaller than the males, they seem to rule.

Graceful Apes

In physique, a bonobo is as different from a chimpanzee as a Concorde is from a Boeing 747. I do not wish to offend any chimpanzees, but bonobos have more style. The bonobo, with its long legs and small head atop narrow shoulders, has a more gracile build than does a chimpanzee. Bonobo lips are reddish in a black face, the ears small and the nostrils almost as wide as a gorilla's. These primates also have a flatter, more open face with a higher forehead than the chimpanzee's and—to top it all off—an attractive coiffure with long, fine, black hair neatly parted in the middle.

Like chimpanzees, female bonobos nurse and carry around their young for up to five years. By the age of seven the offspring reach adolescence. Wild females give birth for the first time at 13 or 14 years of age, becoming full grown by about 15. A bonobo's longevity is unknown, but judging by the chimpanzee it may be older than 40 in the wild and close to 60 in captivity.

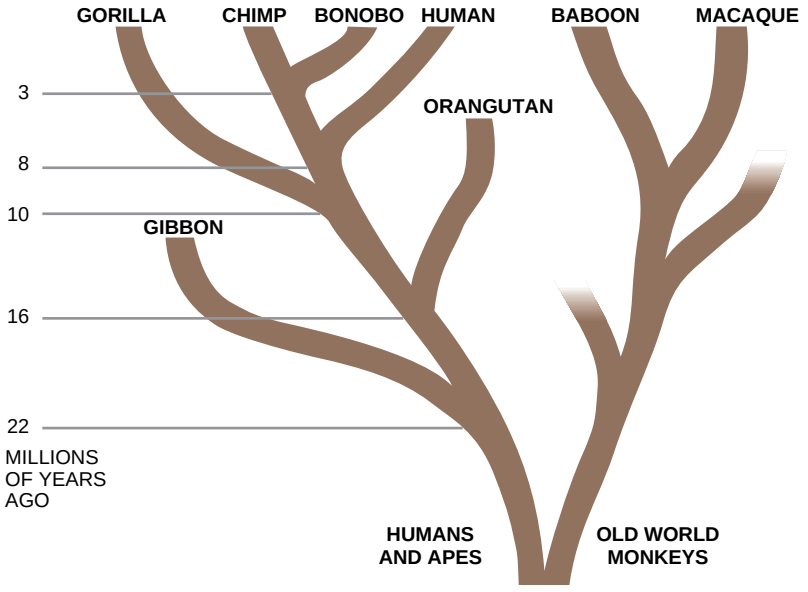
Fruit is central to the diets of both wild bonobos and chimpanzees. The former supplement with more pith from herbaceous plants, and the latter add meat. Although bonobos do eat invertebrates and occasionally capture and eat small vertebrates, including mammals, their diet seems to contain relatively little animal protein. Unlike chimpanzees, they have not been observed to hunt monkeys.

Whereas chimpanzees use a rich array of strategies to obtain foods—from cracking nuts with stone tools to fishing for ants and termites with sticks—tool use in wild bonobos seems undeveloped. (Captive bonobos use tools skillfully.) Apparently as intelligent as chimpanzees, bonobos have, however, a far more sensitive temperament. During World War II bombing of Hellabrun, Germany, the bonobos in a nearby zoo all died of fright from the noise; the chimpanzees were unaffected.

Bonobos are also imaginative in play. I have watched captive bonobos engage in "blindman's buff." A bonobo covers her eyes with a banana leaf or an arm or by sticking two fingers in her eyes. Thus handicapped, she stumbles around on a climbing frame, bumping into others or almost falling. She seems to be imposing a rule on herself: "I cannot look until I lose my balance." Other apes and monkeys also indulge in this game, but I have never seen it performed with such dedication and concentration as by bonobos.

Juvenile bonobos are incurably playful and like to make funny faces, sometimes in long solitary pantomimes and at other times while tickling one another. Bonobos are, however, more controlled in expressing their emotions—whether it be joy, sorrow, excitement or anger—than are the extroverted chimpanzees. Male chimpanzees often engage in spectacular charging displays in which they show off their strength: throwing rocks, breaking branches and uprooting small trees in the process. They keep up these noisy performances for many minutes, during which most other members of the group wisely stay out of their way. Male bonobos, on the other hand, usually limit displays to a brief run while dragging a few branches behind them.

Both primates signal emotions and intentions through facial expressions and hand gestures, many of which are also present in the nonverbal communication of humans. For example, bonobos will beg by stretching out an open



LAURIE GRACE

EVOLUTIONARY TREE of primates, based on DNA analysis, shows that humans diverged from bonobos and chimpanzees a mere eight million years ago. The three species share more than 98 percent of their genetic makeup.

hand (or, sometimes, a foot) to a possessor of food and will pout their lips and make whimpering sounds if the effort is unsuccessful. But bonobos make different sounds than chimpanzees do. The renowned low-pitched, extended "huuu-huuu" pant-hooting of the latter contrasts with the rather sharp, high-pitched barking sounds of the bonobo.

Love, Not War

My own interest in bonobos came not from an inherent fascination with their charms but from research on aggressive behavior in primates. I was particularly intrigued with the aftermath of conflict. After two chimpanzees have fought, for instance, they may come together for a hug and mouth-to-mouth kiss. Assuming that such reunions serve to restore peace and harmony, I labeled them reconciliations.

Any species that combines close bonds with a potential for conflict needs such conciliatory mechanisms. Thinking how much faster marriages would break up if people had no way of compensating for hurting each other, I set out to investigate such mechanisms in several primates, including bonobos. Although I expected to see peacemaking in these apes, too, I was little prepared for the form it would take.

For my study, which began in 1983, I chose the San Diego Zoo. At the time, it housed the world's largest captive bonobo colony—10 members divided into three groups. I spent entire days in front of the enclosure with a video camera, which was switched on at feeding time. As soon as a caretaker approached the enclosure with food, the males would develop erections. Even before the food was thrown into the area, the bonobos would be inviting each other for sex: males would invite females, and females would invite males and other females.

Sex, it turned out, is the key to the social life of the bonobo. The first suggestion that the sexual behavior of bonobos is different had come from observations at European zoos. Wrapping their findings in Latin, primatologists Eduard Tratz and Heinz Heck reported in 1954 that the chimpanzees at Hellabrun mated *more canum* (like dogs) and bonobos *more hominum* (like people). In those days, face-to-face copulation was considered uniquely human, a cultural innovation that needed to be taught to preliterate people (hence the term "missionary position"). These early studies, written in German, were ignored by the international scientific establishment. The bonobo's humanlike sexuality needed to be rediscovered in the 1970s before it became accepted as characteristic of the species.

Bonobos become sexually aroused remarkably easily, and they express this excitement in a variety of mounting positions and genital contacts. Although chimpanzees virtually never adopt face-to-face positions, bonobos do so in one out of three copulations in the wild. Furthermore, the frontal orientation of the bonobo vulva and clitoris strongly suggest that the female genitalia are adapted for this position.

Another similarity with humans is increased female sexual receptivity. The tumescent phase of the female's genitals, resulting in a pink swelling that signals willingness to mate, covers a much longer part of estrus in bonobos than in chimpanzees. Instead of a few days out of her cycle, the female bonobo is almost continuously sexually attractive and active [*see illustration on page 86*].

Perhaps the bonobo's most typical sexual pattern, undocumented in any other primate, is genito-genital rubbing (or GG rubbing) between adult females. One female facing another clings with arms and legs to a partner that, standing on both hands and feet, lifts her off the ground. The two females then rub their genital swellings laterally together, emitting grins and squeals that probably reflect orgasmic experiences. (Laboratory experiments on stump-tailed macaques have demonstrated that women are not the only female primates capable of physiological orgasm.)

Male bonobos, too, may engage in pseudocopulation but generally perform a variation. Standing back to back, one male briefly rubs his scrotum against the buttocks of another. They also practice so-called penis-fencing, in which two males hang face to face from a branch while rubbing their erect penises together.

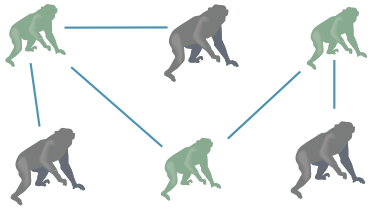
The diversity of erotic contacts in bonobos includes sporadic oral sex, massage of another individual's genitals and intense tongue-kissing. Lest this leave the impression of a pathologically oversexed species, I must add, based on hundreds of hours of watching bonobos, that their sexual activity is rather casual and relaxed. It appears to be a completely natural part of their group life. Like people, bonobos engage in sex only occasionally, not continuously. Furthermore, with the average copulation lasting 13 seconds, sexual contact in bonobos is rather quick by human standards.

That sex is connected to feeding, and even appears to make food sharing possible, has been observed not only in zoos but also in the wild. Nancy Thompson-Handler, then at the State University of New York at Stony Brook, saw bonobos in Zaire's Lomako Forest engage in sex after they had entered trees loaded with ripe figs or when one among them had captured a prey animal, such as a small forest duiker. The flurry of sexual contacts would last for five to 10 minutes, after which the apes would settle down to consume the food.

One explanation for the sexual activity at feeding time could be that excitement over food translates into sexual arousal. This idea may be partly true.

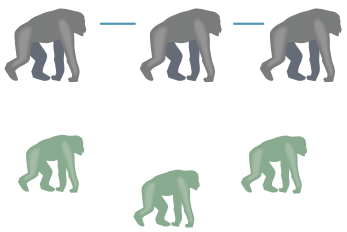
Social Organization among Various Primates

BONOBO



Bonobo communities are peace-loving and generally egalitarian. The strongest social bonds (*blue*) are those among females (*green*), although females also bond with males. The status of a male (*purple*) depends on the position of his mother, to whom he remains closely bonded for her entire life.

CHIMPANZEE

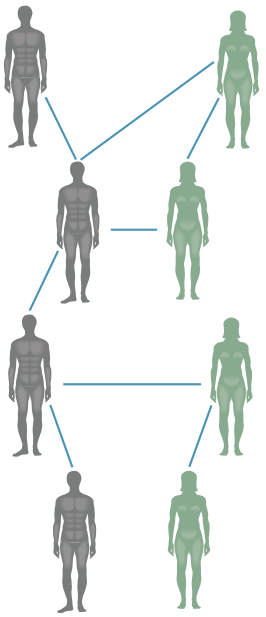


In chimpanzee groups the strongest bonds are established between the males in order to hunt and to protect their shared territory. The females live in overlapping home ranges within this territory but are not strongly bonded to other females or to any one male.

GIBBON

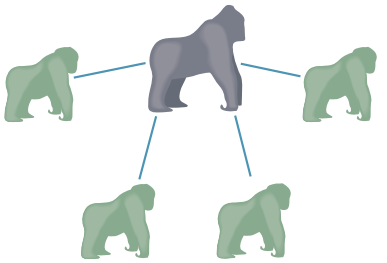
Gibbons establish monogamous, egalitarian relations, and one couple will maintain a territory to the exclusion of other pairs.

HUMAN



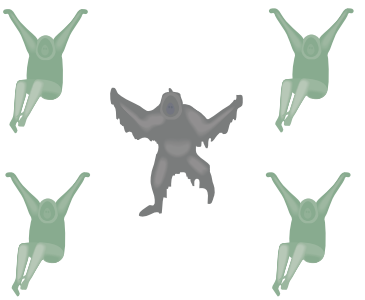
Human society is the most diverse among the primates. Males unite for cooperative ventures, whereas females also bond with those of their own sex. Monogamy, polygamy and polyandry are all in evidence.

GORILLA



The social organization of gorillas provides a clear example of polygamy. Usually a single male maintains a range for his family unit, which contains several females. The strongest bonds are those between the male and his females.

ORANGUTAN



Orangutans live solitary lives with little bonding in evidence. Male orangutans are intolerant of one another. In his prime, a single male establishes a large territory, within which live several females. Each female has her own, separate home range.

LAURIE GRACE

Yet another motivation is probably the real cause: competition. There are two reasons to believe sexual activity is the bonobo's answer to avoiding conflict.

First, anything, not just food, that arouses the interest of more than one bonobo at a time tends to result in sexual contact. If two bonobos approach a cardboard box thrown into their enclosure, they will briefly mount each other before playing with the box. Such situations lead to squabbles in most other species. But bonobos are quite tolerant, perhaps because they use sex to divert attention and to diffuse tension.

Second, bonobo sex often occurs in aggressive contexts totally unrelated to food. A jealous male might chase another away from a female, after which the two males reunite and engage in scrotal rubbing. Or after a female hits a juvenile, the latter's mother may lunge at the aggressor, an action that is immediately followed by genital rubbing between the two adults.

I once observed a young male, Kako, inadvertently blocking an older, female juvenile, Leslie, from moving along a branch. First, Leslie pushed him; Kako, who was not very confident in trees, tightened his grip, grinning nervously. Next Leslie gnawed on one of his hands, presumably to loosen his grasp. Kako uttered a sharp peep and stayed put. Then Leslie rubbed her vulva against his shoulder. This gesture calmed Kako, and he moved along the branch. It seemed that Leslie had been very close to using force but instead had reassured both herself and Kako with sexual contact.

During reconciliations, bonobos use the same sexual repertoire as they do during feeding time. Based on an analysis of many such incidents, my study yielded the first solid evidence for sexual behavior as a mechanism to overcome aggression. Not that this function is absent in other animals—or in humans, for that matter—but the art of sexual reconciliation may well have reached its evolutionary peak in the

bonobo. For these animals, sexual behavior is indistinguishable from social behavior. Given its peacemaking and appeasement functions, it is not surprising that sex among bonobos occurs in so many different partner combinations, including between juveniles and adults. The need for peaceful coexistence is obviously not restricted to adult heterosexual pairs.

Female Alliance

Apart from maintaining harmony, sex is also involved in creating the singular social structure of the bonobo. This use of sex becomes clear when studying bonobos in the wild. Field research on bonobos started only in the mid-1970s, more than a decade after the most important studies on wild chimpanzees had been initiated. In terms of continuity and invested (wo)manpower, the chimpanzee projects of Jane Goodall and Toshisada Nishida, both in Tanzania, are unparalleled. But bonobo research by Takayoshi Kano and others of Kyoto University is now two decades under way at Wamba in Zaire and is beginning to show the same payoffs.

Both bonobos and chimpanzees live in so-called fission-fusion societies. The apes move alone or in small parties of a few individuals at a time, the composition of which changes constantly. Several bonobos traveling together in the morning might meet another group in the forest, whereupon one individual from the first group wanders off with others from the second group, while those left behind forage together. All associations, except the one between mother and dependent offspring, are of a temporary character.

Initially this flexibility baffled investigators, making them wonder if these apes formed any social groups with stable membership. After years of documenting the travels of chimpanzees in the Mahale Mountains, Nishida first reported that they form large communities: all members of one community mix freely in ever changing parties, but members of different communities never gather. Later, Goodall added territoriality to this picture. That is, not only do communities not mix, but males of different chimpanzee communities engage in lethal battles.

In both bonobos and chimpanzees, males stay in their natal group, whereas females tend to migrate during adolescence. As a result, the senior males of a chimpanzee or bonobo group have known all junior males since birth, and all junior males have grown up together. Females, on the other hand, transfer to an unfamiliar and often hostile group where they may know no one. A chief difference between chimpanzee and bonobo societies is the way in which young females integrate into their new community.

On arrival in another community, young bonobo females at Wamba single out one or two senior resident females for special attention, using frequent GG rubbing and grooming to establish a relation. If the residents reciprocate, close associations are set up, and the younger female gradually becomes accepted into the group. After producing her first offspring, the young female's position becomes more stable and central. Eventually the cycle repeats with younger immigrants, in turn, seeking a good relation with the now established female. Sex thus smooths the migrant's entrance into the community of females, which is much more close-knit in the bonobo than in the chimpanzee.

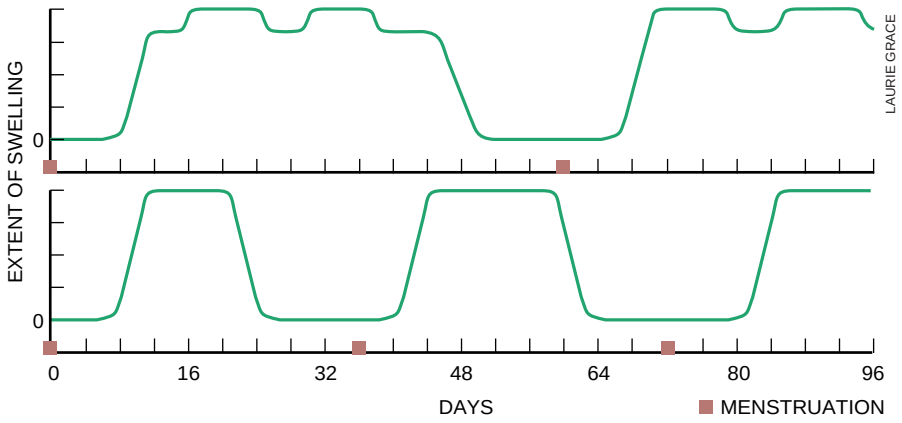
Bonobo males remain attached to their mothers all their lives, following them through the forest and being dependent on them for protection in aggressive encounters with other males. As a result, the highest-ranking males of a bonobo community tend to be sons of important females.

What a contrast with chimpanzees! Male chimpanzees fight their own battles, often relying on the support of other males. Furthermore, adult male chimpanzees travel together in same-sex parties, grooming each other frequently. Males form a distinct social hierarchy with high levels of both competition and association. Given the need to stick together against males of neighboring communities, their bonding is not surprising: failure to form a united front might result in the loss of lives and territory. The danger of being male is reflected in the adult sex ratio of chimpanzee populations, with considerably fewer males than females.

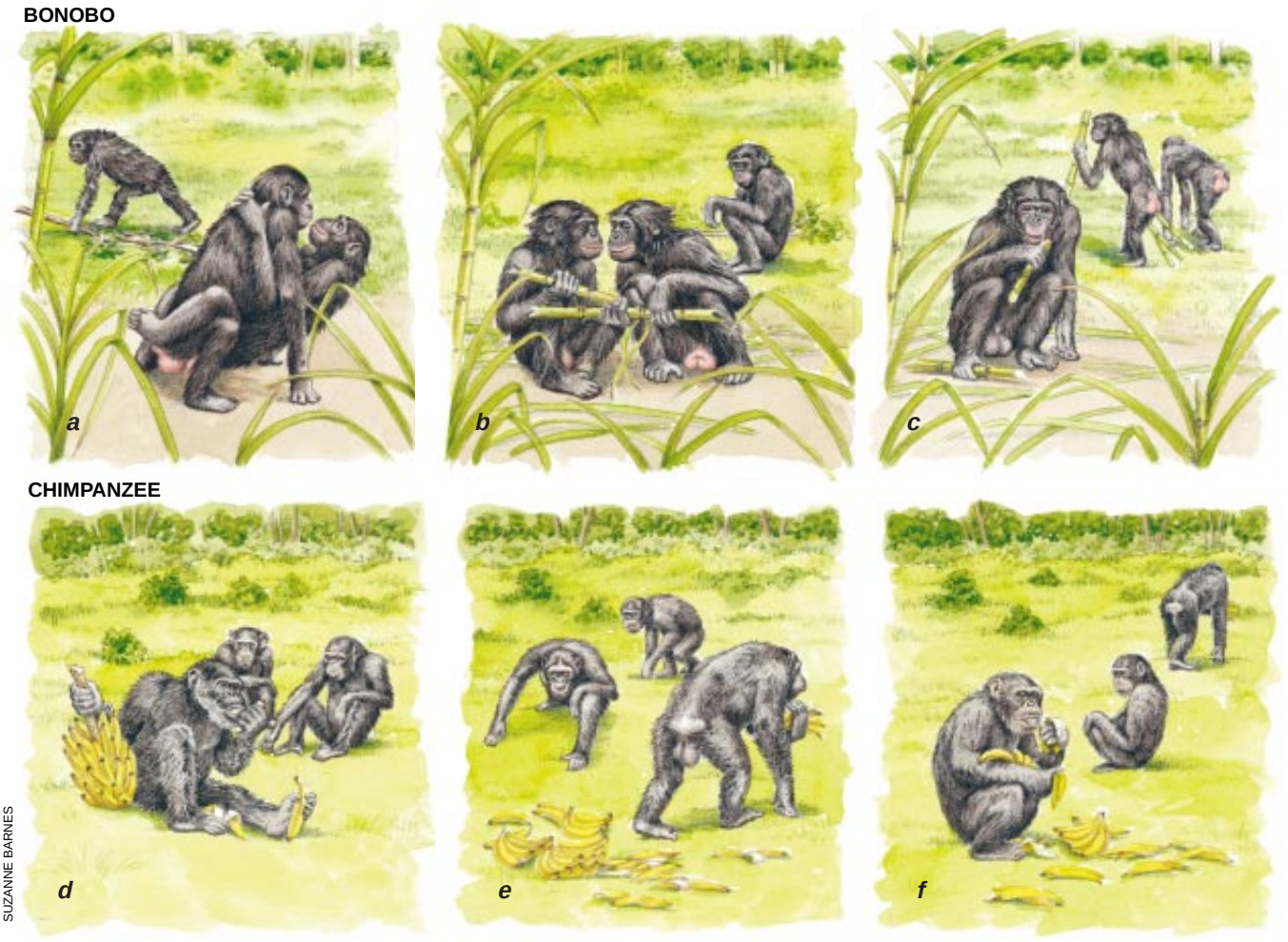
Serious conflict between bonobo groups has been witnessed in the field, but it seems quite rare. On the contrary, reports exist of peaceable mingling, including mutual sex and grooming, between what appear to be different communities. If intergroup combat is indeed unusual, it may explain the lower rate of all-male associations. Rather than being male-bonded, bonobo society gives the impression of being female-bonded, with even adult males relying on their mothers instead of on other males. No wonder Kano calls mothers the "core" of bonobo society.

The bonding among female bonobos violates a fairly general rule, outlined by Harvard University anthropologist Richard W. Wrangham, that the sex that stays in the natal group develops the strongest mutual bonds. Bonding among male chimpanzees follows naturally because they remain in the community of their birth. The same is true for female kinship bonding in Old World monkeys, such as macaques and baboons, where males are the migratory sex.

Bonobos are unique in that the migratory sex, females, strongly bond with same-sex strangers later in life. In setting up an artificial sisterhood, bonobos can be said to be secondarily bonded. (Kinship bonds are said to be primary.) Although we now know *how* this happens—through the use of sexual contact and grooming—we do not yet know *why* bonobos and chimpanzees differ in this respect. The answer may lie in the different ecological environments of bonobos and chimpanzees—such as the abundance and quality of food in the forest. But it is uncertain if such explanations will suffice.



FEMALE RECEPTIVITY for sex, manifested by swollen genitals, occupies a much larger proportion of the estrus cycle of bonobos (*top*) than of chimpanzees (*bottom*). The receptivity of bonobos continues through lactation. (In chimpanzees, it disappears.) This circumstance allows sex to play a large part in the social relations of bonobos. The graph was provided by Jeremy Dahl of the Yerkes Primate Center.



SUZANNE BARNES

DOMINANCE BY BONDING is evinced by female bonobos, who engage in genito-genital (GG) rubbing before eating sugarcane (*a*), while a bigger male displays to no avail. The females then share the food without competition (*b*). Only when they leave can the male get to the sugarcane (*c*). In male-dominated chimpanzee society the male eats first (*d*), while the females wait at a safe distance. After he leaves (*e*), carrying as many bananas as he can, the dominant female gets what is left (*f*). (Small amounts of sugarcane and bananas are provided at some research sites in Zaire.)

Bonobo society is, however, not only female-centered but also appears to be female-dominated. Bonobo specialists, while long suspecting such a reality, have been reluctant to make the controversial claim. But in 1992, at the 14th Congress of the International Primatological Society in Strasbourg, investigators of both captive and wild bonobos presented data that left little doubt about the issue.

Amy R. Parish of the University of California at Davis reported on food competition in identical groups (one adult male and two adult females) of chimpanzees and bonobos at the Stuttgart Zoo. Honey was provided in a "termite hill" from which it could be extracted by dipping sticks into a small hole. As soon as honey was made available, the male chimpanzee would make a charging display through the enclosure and claim everything for himself. Only when his appetite was satisfied would he let the females fish for honey.

In the bonobo group, it was the females that approached the honey first. After having engaged in some GG rubbing, they would feed together, taking turns with virtually no competition between them. The male might make as many charging displays as he wanted; the females were not intimidated and ignored the commotion.

Observers at the Belgian animal park of Planckendael, which currently has the most naturalistic bonobo colony, reported similar findings. If a male bonobo tried to harass a female, all females would band together to chase him off. Because females appeared more successful in dominating males when they were together than on their own, their close association and frequent genital rubbing may represent an alliance. Females may bond so as to outcompete members of the individually stronger sex.

The fact that they manage to do so not only in captivity is evident from zoologist Takeshi Furuichi's summary of the relation between the sexes at Wamba, where bonobos are enticed out of the forest with sugarcane. "Males usually appeared at the feeding site first, but they surrendered preferred positions when the females appeared. It seemed that males appeared first not because they were dominant, but because they had to feed before the arrival of females," Furuichi reported at Strasbourg.

Sex for Food

Occasionally, the role of sex in relation to food is taken one step further, bringing bonobos very close to humans in their behavior. It has been speculated by anthropologists—including C. Owen Lovejoy of Kent State University and Helen Fisher of Rutgers University—that sex is partially separated from reproduction in our species because it serves to cement mutually profitable relationships between men and


FRANS LANTING *Minden Pictures*


FRANS LANTING *Minden Pictures*


FRANS DE WAAL


FRANS LANTING *Minden Pictures*

BEHAVIOR among bonobos is often reminiscent of that among humans. A female and an infant play (*left*); two juveniles practice sex without penetration (*top center*); a bonobo walks upright, using his hands to carry food (*right*); and a male and female have sex (*bottom center*), after which the female leaves with one of the male's two oranges.

women. The human female's capacity to mate throughout her cycle and her strong sex drive allow her to exchange sex for male commitment and paternal care, thus giving rise to the nuclear family.

This arrangement is thought to be favored by natural selection because it allows women to raise more offspring than they could if they were on their own. Although bonobos clearly do not establish the exclusive heterosexual bonds characteristic of our species, their behavior does fit important elements of this model. A female bonobo shows extended receptivity and uses sex to obtain a male's favors when—usually because of youth—she is too low in social status to dominate him.

At the San Diego Zoo, I observed that if Loretta was in a sexually attractive state, she would not hesitate to approach the adult male, Vernon, if he had food. Presenting herself to Vernon, she would mate with him and make high-pitched food calls while taking over his entire bundle of branches and leaves. When Loretta had no genital swelling, she would wait until Vernon was ready to share. Primatologist Suehisa Kuroda reports similar exchanges at Wamba: "A young female approached a male, who was eating sugarcane. They copulated in short order, whereupon she took one of the two canes held by him and left."

Despite such quid pro quo between the sexes, there are no indications that bonobos form humanlike nuclear families. The burden of raising offspring appears to rest entirely on the female's shoulders. In fact, nuclear families are probably incompatible with the diverse use of sex found in bonobos. If our ancestors started out with a sex life similar to that of bonobos, the evolution of the family would have required dramatic change.

Human family life implies paternal investment, which is unlikely to develop unless males can be reasonably certain that they are caring for their own, not someone else's, offspring. Bonobo society lacks any such guarantee, but humans protect the integrity of their family units through all kinds of moral restrictions and taboos. Thus, although our species is characterized by an extraordinary interest in sex, there are no societies in which people engage in it at the drop of a hat (or a cardboard box, as the case may be). A sense of shame and a desire for domestic privacy are typical human concepts related to the evolution and cultural bolstering of the family.

Yet no degree of moralizing can make sex disappear from every realm of human life that does not relate to the nuclear family. The bonobo's behavioral peculiarities may help us understand the role of sex and may have serious implications for models of human society.

Just imagine that we had never heard of chimpanzees or baboons and had known bonobos first. We would at present most likely believe that early hominids lived in female-centered societies, in which sex served important social functions and in which warfare was rare or absent. In the end, perhaps the most successful reconstruction of our past will be based not on chimpanzees or even on bonobos but on a three-way comparison of chimpanzees, bonobos and humans.

FURTHER READING

THE PYGMY CHIMPANZEE: EVOLUTIONARY BIOLOGY AND BEHAVIOR. Edited by Randall L. Susman. Plenum Press, 1984.

THE COMMUNICATIVE REPERTOIRE OF CAPTIVE BONOBOS (*PAN PANISCUS*) COMPARED TO THAT OF CHIMPANZEES. F.B.M. de Waal in *Behaviour,* Vol. 106, Nos. 3–4, pages 183–251; September 1988.

PEACEMAKING AMONG PRIMATES. F.B.M. de Waal. Harvard University Press, 1989.

UNDERSTANDING CHIMPANZEES. Edited by Paul Heltne and Linda A. Marquardt. Harvard University Press, 1989.

THE LAST APE: PYGMY CHIMPANZEE BEHAVIOR AND ECOLOGY. Takayoshi Kano. Stanford University Press, 1992.

CHIMPANZEE CULTURES. R. Wrangham, W. C. McGrew, F.B.M. de Waal and P. Heltne. Harvard University Press, 1994.

Protein-Based Computers

Devices fabricated from biological molecules promise compact size and faster data storage. They lend themselves to use in parallel-processing computers, three-dimensional memories and neural networks

by Robert R. Birge

The world's most advanced supercomputer does not require a single semiconductor chip. The human brain consists of organic molecules that combine to form a highly sophisticated network able to calculate, perceive, manipulate, self-repair, think and feel. Digital computers can certainly perform calculations much faster and more precisely than humans can, but even simple organisms are superior to computers in the other five domains. Computer designers may never be able to make machines having all the faculties of a natural brain, but many of us think we can exploit some special properties of biological molecules—particularly proteins—to build computer components that are smaller, faster and more powerful than any electronic devices on the drawing boards thus far.

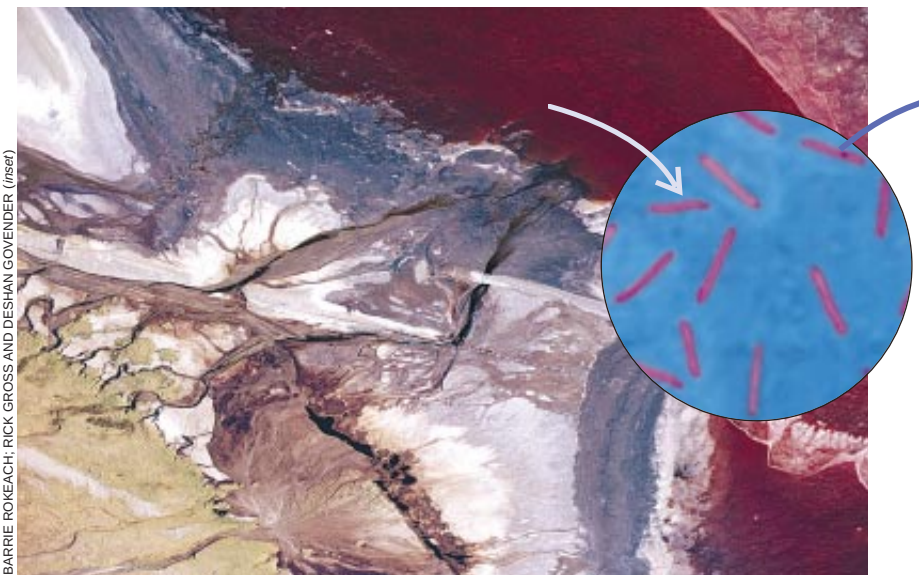
The size issue is especially pressing. Since the 1960s the computer industry has been compelled to make the individual components on semiconductor chips smaller and smaller in order to manufacture larger memories and more powerful processors economically. These chips essentially consist of arrays of switches, usually of the kind known as logic gates, that flip between two states—designated as 0 and 1—in response to changes in the electric current passing through them. (Computers typically represent all information in terms of such binary digits, or bits.) If the trend toward miniaturization continues, the size of a single logic gate will approach the size of molecules by about the year 2030.

But there is a serious roadblock. Each factor of two in miniaturization increases the cost of manufacturing a chip by a factor of five. At some point the search for ever smaller electronic devices may be limited by economics rather than physics [see "The Wall," by Gary Stix, "Science and Business," SCIENTIFIC AMERICAN, July 1994]. On the other hand, the use of biological molecules as the active components in computer circuitry may offer an alternative approach that is more economical.

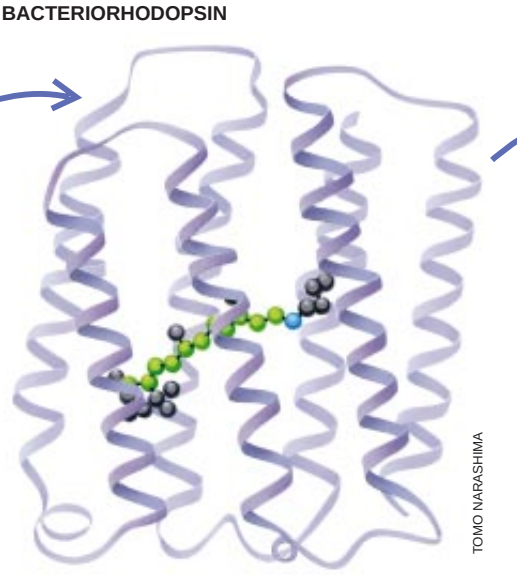
Molecules can potentially serve as computer switches because their atoms are mobile and change position in a predictable way. If we can direct that atomic motion and thereby consistently generate at least two discrete states in a molecule, we can use each state to represent either 0 or 1. Such switches offer reductions in the size of hardware because they are themselves small—about one thousandth the size of the semiconductor transistors used today as gates (which measure about one micron, or a millionth of a meter, across). Indeed, a biomolecular computer could in principle be one fiftieth the size of a present-day semiconductor computer composed of a similar number of logic elements. In the computer business, smaller gate size generally makes for a faster device, and protein-based computers could theoretically operate 1,000 times faster than modern computers.

At this stage no one is seriously proposing a purely biomolecular computer. Far more likely, at least for the near future, is the use of hybrid technology in which molecules and semiconductors are used in combination. Such an approach should provide computers that are one fiftieth the size and as much as 100 times faster than current ones.

Biological molecules also appeal because they can be designed one atom



BARRIE ROKEACH; RICK GROSS AND DESHAN GOVENDER (*inset*)



TOMO NARASHIMA

at a time—giving engineers the control they need to manufacture gates able to perform exactly as an application requires. Further, bioelectronic computers should help in the ongoing pursuit of more adaptable computers. Computer scientists are already enhancing the versatility of electronic devices by developing new configurations of computer hardware known as architectures.

Researchers have introduced parallel-processing architectures, which allow multiple sets of data to be manipulated simultaneously. In order to expand memory capacities, they are devising hardware that stores data in three dimensions instead of the usual two. And scientists have built neural networks that mimic the learning-by-association capabilities of the brain, an ability necessary for significant progress toward artificial intelligence. The ability of certain proteins to change their properties in response to light should simplify the hardware required for implementation of these architectures.

Although no computer components made entirely or partly from proteins are on the market yet, ongoing international research efforts are making exciting headway. It seems reasonable to predict that hybrid technology combining semiconductor chips and biological molecules will move from the realm of science fiction to commercial application fairly soon. Liquid-crystal-display technology offers a prime example of a hybrid system that has achieved commercial success. Most laptop computers today depend on liquid-crystal displays, which combine semiconductor devices and organic molecules to control the intensity of the image on screen.

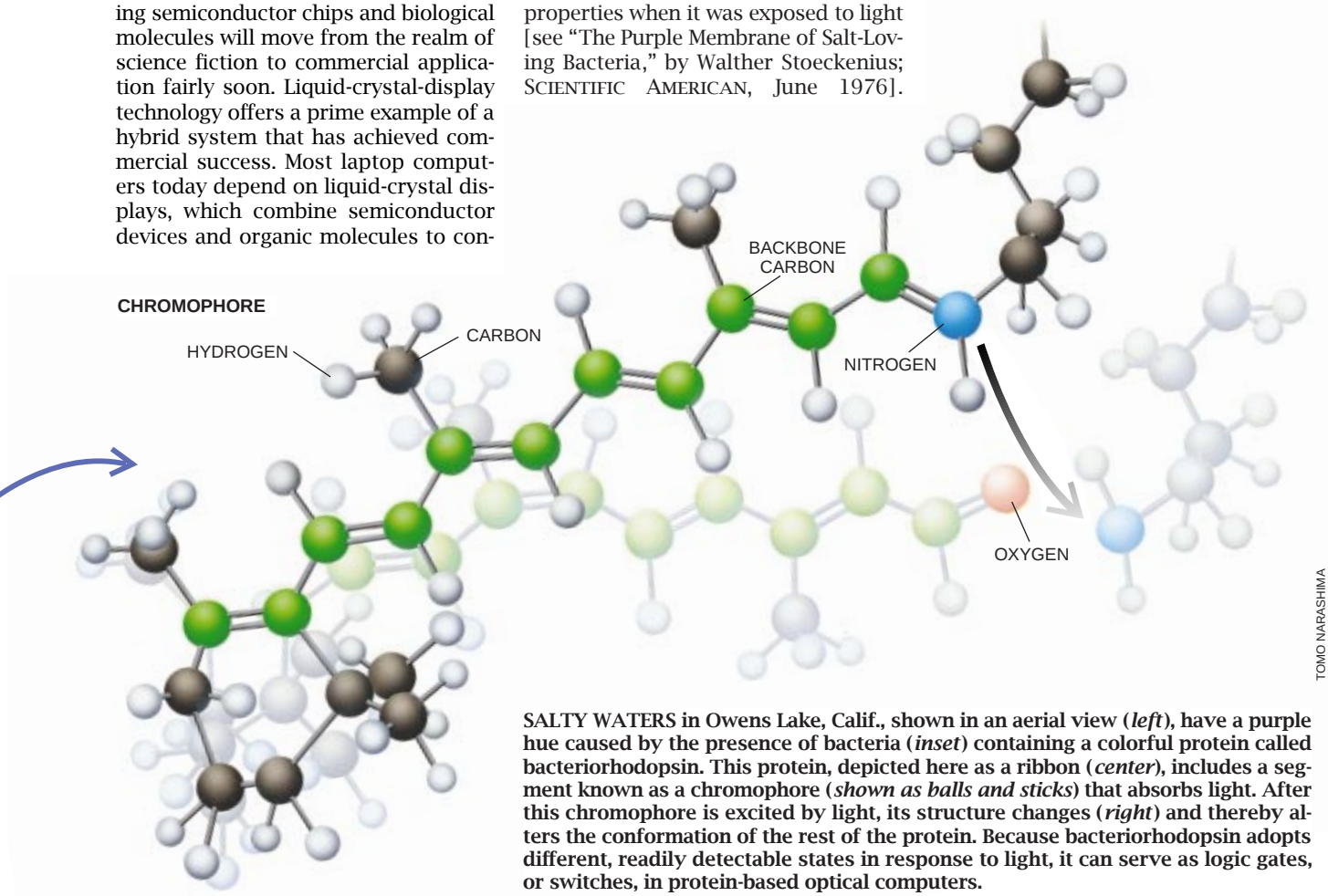
ROBERT R. BIRGE is professor of chemistry, director of the W. M. Keck Center for Molecular Electronics and research director of the New York State Center for Advanced Technology in Computer Applications and Software Engineering at Syracuse University. In addition to working on the development of protein-based electronic devices and hybrid computers, he investigates laser spectroscopy and quantum theory of proteins. He received his B.S. in chemistry from Yale University in 1968, his Ph.D. in chemical physics from Wesleyan University in 1972 and was a National Institutes of Health postdoctoral fellow at Harvard University from 1973 to 1975.

Several biological molecules are under consideration for use in computer hardware, but the bacterial protein bacteriorhodopsin has generated the most interest. During the past 10 years, my laboratory and others in North America, Europe and Japan have built prototype parallel-processing devices, three-dimensional data storage hardware and neural networks based on this protein.

Origins in the Salt Marsh

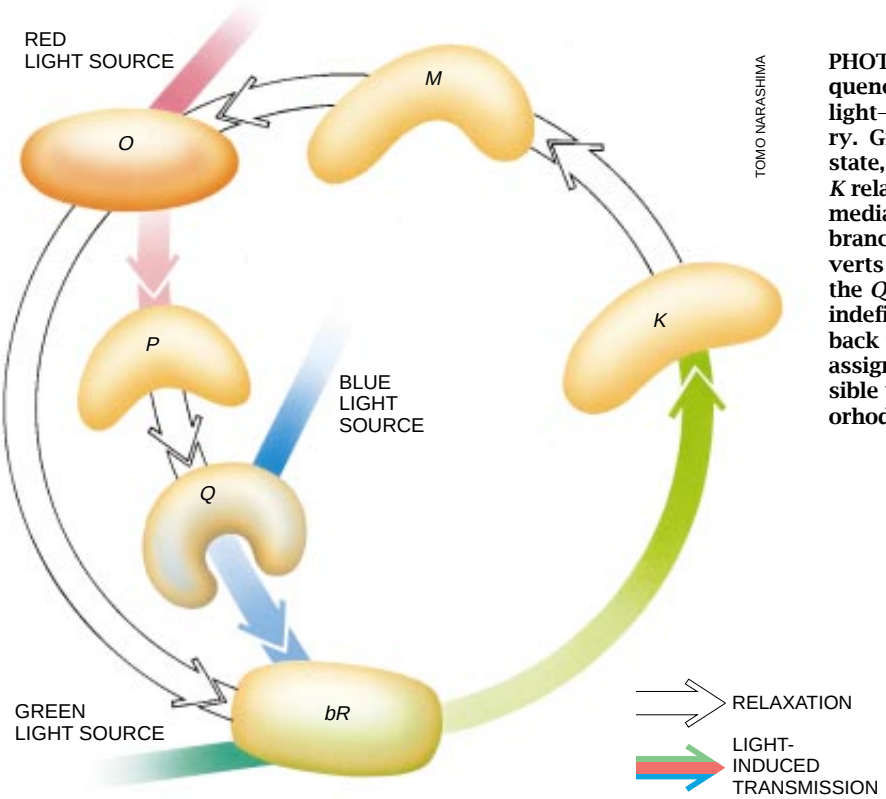
Interest in bacteriorhodopsin dates back to the early 1970s, when Walther Stoeckenius of the University of California at San Francisco and Dieter Oesterhelt, now at the Max Planck Institute for Biochemistry in Martinsried, discovered that the protein exhibited unusual properties when it was exposed to light [see "The Purple Membrane of Salt-Loving Bacteria," by Walther Stoeckenius; SCIENTIFIC AMERICAN, June 1976]. Found in the membrane of *Halobacterium salinarium,* bacteriorhodopsin enables the bacterium to grow when the concentration of oxygen is insufficient to otherwise sustain the organism. When struck by light, the protein changes its structure and transports a proton across the membrane, thereby supplying energy to maintain cell metabolism.

Soviet scientists were the first to recognize and develop the potential of bacteriorhodopsin for computing. Soon after it was discovered, the late Yuri A. Ovchinnikov of the Shemyakin Institute of Bioorganic Chemistry in Moscow assembled a team of scientists from five Soviet institutions to work on biomolecular electronics as part of what came to be called Project Rhodopsin. Ovchinnikov obtained a good deal of funding



TOMO NARASHIMA

SALTY WATERS in Owens Lake, Calif., shown in an aerial view (*left*), have a purple hue caused by the presence of bacteria (*inset*) containing a colorful protein called bacteriorhodopsin. This protein, depicted here as a ribbon (*center*), includes a segment known as a chromophore (*shown as balls and sticks*) that absorbs light. After this chromophore is excited by light, its structure changes (*right*) and thereby alters the conformation of the rest of the protein. Because bacteriorhodopsin adopts different, readily detectable states in response to light, it can serve as logic gates, or switches, in protein-based optical computers.



TOMO NARASHIMA

PHOTOCYCLE of bacteriorhodopsin—the sequence of structural changes induced by light—allows for the storage of data in memory. Green light transforms the initial resting state, known as *bR*, to the intermediate *K*. Next *K* relaxes, forming *M* and then *O*. If the *O* intermediate is exposed to red light, a so-called branching reaction occurs. Structure *O* converts to the *P* state, which quickly relaxes to the *Q* state—a form that remains stable almost indefinitely. Blue light, however, will convert *Q* back to *bR*. Any two long-lasting states can be assigned the binary value 0 or 1, making it possible to store information as a series of bacteriorhodopsin molecules in one state or the other.

for such research because he had the ear of Soviet military leaders and was able to convince them that by exploring bioelectronics, Soviet science could leapfrog the West in computer technology.

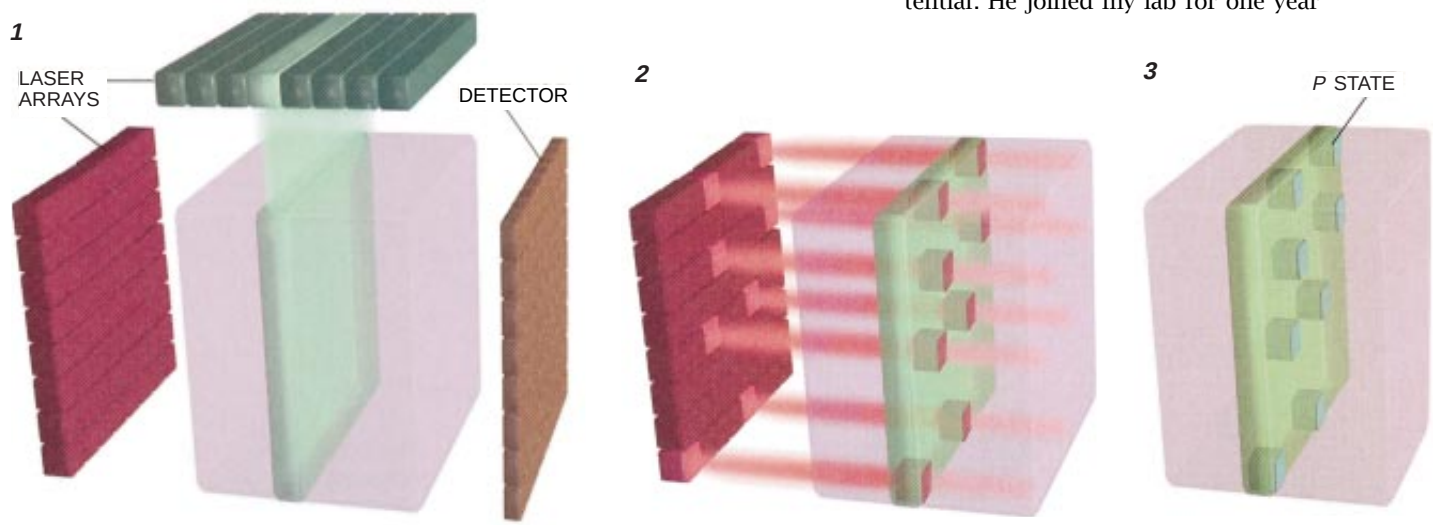
Many aspects of this ambitious project are still considered military secrets and may never be revealed. We do know that the Soviet military made microfiche films, called Biochrome, out of bacteriorhodopsin. Informal reports from former Soviet scientists now in the U.S. indicate that researchers there also made optical data processors using protein technology. The details of their most impressive accomplishment, a processor for military radar, remain obscure.

I became interested in bacteriorhodopsin in the 1970s, while I was studying the biochemical basis of vision at the University of California at Riverside. My work had initially focused on a related protein, rhodopsin, present in the retina of mammals. Both rhodopsin and bacteriorhodopsin are complex proteins that include a light-absorbing component known as a chromophore. The chromophore absorbs energy from light, triggering a complex series of internal motions that result in dramatic changes in the structure of the larger protein. These changes alter the protein's optical and electrical characteristics. For example, when rhodopsin absorbs light in the human eye, the change in structure releases energy that serves as an electrical signal able to convey visual information to the brain.

Computer Applications

At first I was concerned purely with understanding how such light-activated changes to rhodopsin occurred. During the late 1970s, however, I became interested in bacteriorhodopsin as well. I had also decided to apply my knowledge of its properties to the design of computer memories and processors based on the protein. Albert F. Lawrence, then at Hughes Aircraft Company, played an important role in convincing me that bioelectronics had potential. He joined my lab for one year

WRITING DATA



WRITING OF INFORMATION into cubes of bacteriorhodopsin (*purple*), and reading out of that information, is accomplished with laser beams. The writing process is begun by firing green laser beams through a plane of the cube (*1*); this step begins the protein's photocycle. Then, red lasers are fired (*2*) at the particular set of molecules in the plane (*green*) to be converted to the binary 1 state; the remaining molecules represent binary 0. The targeted molecules first form the *P* state (*3*),

to explore the use of biological materials in optical memories.

We focused on bacteriorhodopsin instead of rhodopsin because of the former's greater stability and better optical properties. Also, it can be prepared in large quantities. The components of computers must be able to withstand changes in their environment without breaking apart. Bacteriorhodopsin naturally functions in salt marshes where temperatures can exceed 150 degrees Fahrenheit and where the molecule is often exposed to intense light.

The applications under study for computer processors and the memories on which they operate exploit what is called the photocycle—the series of structural changes bacteriorhodopsin undergoes in response to light. (In its resting state the molecule is known as *bR,* and each intermediate in the series is identified by a letter of the alphabet.) The various intermediates can be used to represent bits of data.

Moreover, the intermediates absorb light in different regions of the spectrum. As a consequence, we can read the data by shining laser beams on the molecules and noting the wavelengths that do not pass through to the detector. Because we can alter the structure of bacteriorhodopsin with one laser and then, with another laser, determine which intermediates have formed, we have the needed basis for writing to and then reading from memory.

Most devices under study make use of the resting state and one intermediate of bacteriorhodopsin. One state is designated as 0 and the other as 1, and switching between the states is controlled by a laser beam. Many early memory devices based on bacteriorhodopsin could operate only at the extremely cold temperature of liquid nitrogen, at which the light-induced switching between the initial *bR* structure and an intermediate known as the *K* state could be controlled. These devices were very fast compared with semiconductor switches (the *bR* to *K* conversion takes place in a few trillionths of a second, compared with the few billionths of a second that common semiconductor devices require). But the need for such low temperatures precluded general application.

Today most bacteriorhodopsin-based devices function at or near room temperature, a condition under which another intermediate, *M,* is stable. Although most bacteriorhodopsin-based memory devices incorporate the *bR*-to-*M* switch, other structures may actually prove more useful in protein-based computer systems.

Parallel Processing

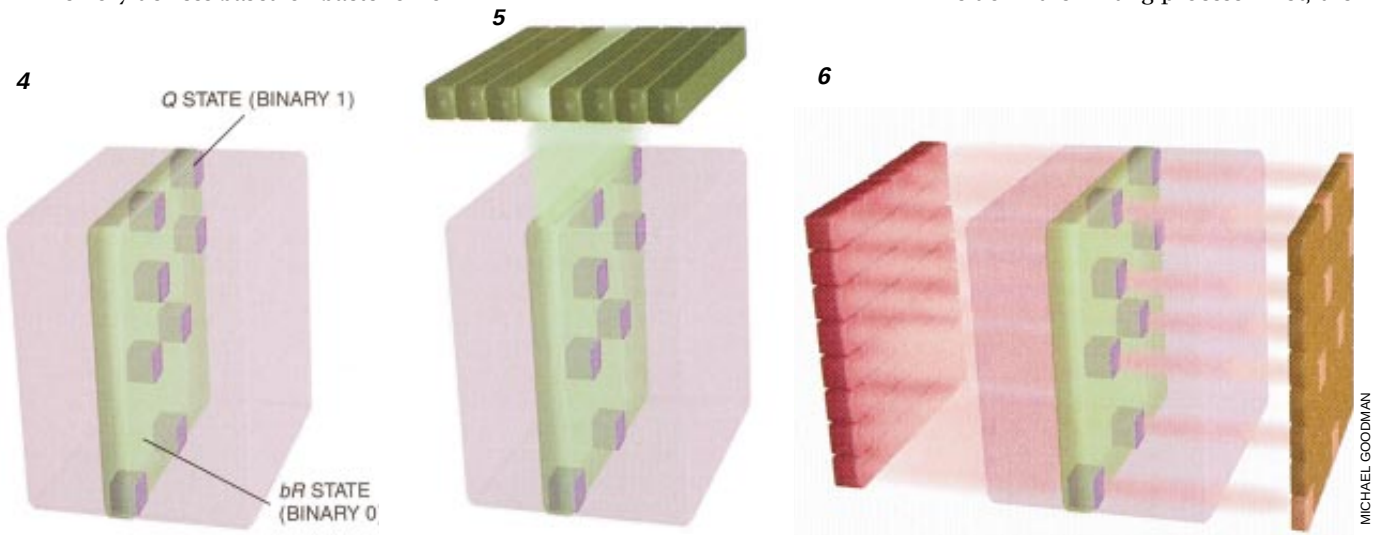
Certain of the intermediates produced after bacteriorhodopsin is initially exposed to light will change to unusual structures when they absorb energy from a second laser beam, in a process known as a sequential one-photon architecture. For example, such a branching reaction occurs from the *O* intermediate to form *P* and *Q*. These structures are generated by two consecutive pulses of laser light—first green light, then red. Although *P* is fairly short-lived, it relaxes into a form known as *Q,* which is stable for extended periods, even up to several years. Because of its extended stability, the *Q* state has great significance in the search for long-term, high-density memory.

The intermediates *P* and *Q,* formed in the sequential one-photon process, are particularly useful for parallel processing. For writing data in parallel, our approach incorporates another innovation: three-dimensional data storage. A cube of bacteriorhodopsin is surrounded by two arrays of laser beams placed 90 degrees from each other. One array of lasers, all set to green and called paging beams, activates the photocycle of the protein in any selected square plane, or page, within the cube. After a few milliseconds, when the number of *O* intermediates reaches near maximum, the other laser array—this time of red beams—is fired.

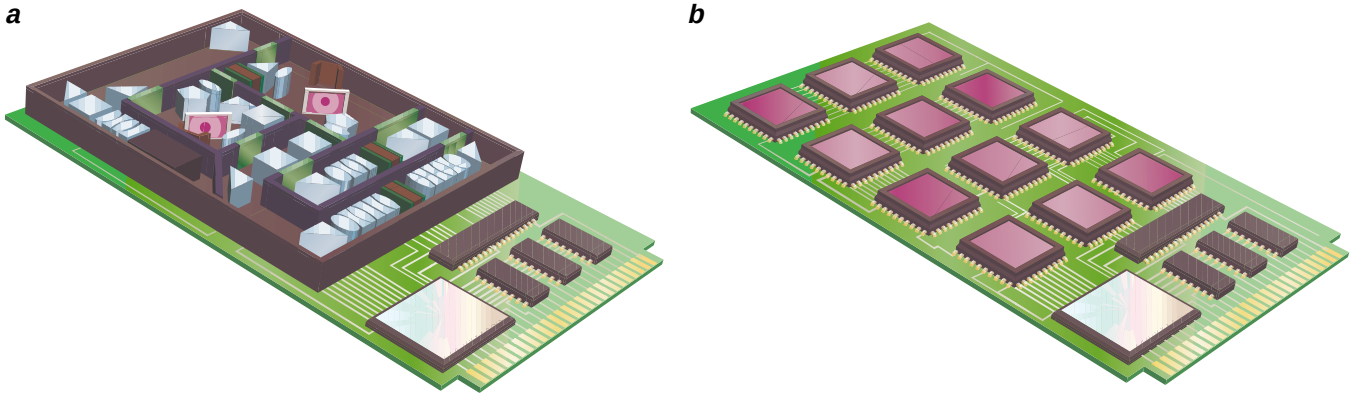
This second array is programmed to illuminate only the region of the activated square where data bits are to be written, switching the molecules there to the *P* structure. The *P* intermediate then relaxes its structure to form the highly stable *Q* state. If we assign the *bR* structure to binary state 0 and both *P* and *Q* to binary state 1, the process is analogous to the binary switching that takes place in semiconductor or magnetic memory. Because the laser array can activate molecules in various places throughout the chosen illuminated page, multiple data locations, known as addresses, can be written to simultaneously—in other words, in parallel.

Our system for reading stored memory—whether during processing or during the extraction of a result—relies on the selective absorption of red light by the *O* intermediate. To read multiple bits of data in parallel, we start just as we do in the writing process. First, the

READING DATA



then relax to the *Q* structure (*4*). Reading from the protein-based memory is begun by again activating the plane with green light (*5*). Then, red lasers of low intensity are fired. Molecules that were originally in the *bR* state absorb the red light, and molecules in the *P* or *Q* state allow the low light levels to pass through. Hence, the resulting pattern of dark and light—that is, 0's and 1's—can be picked up by a detector placed directly opposite from the red laser array (*6*).



MICHAEL GOODMAN

COMPUTERS OF THE FUTURE might be hybrids, consisting of cards with both proteins (*purple*) and semiconductors. The cards shown here, which have not yet been built, could provide associative memory (*a*) and three-dimensional memo-

green paging beam fires at the square of protein to be read, starting the normal photocycle of the molecules in the *bR* state. After two milliseconds, the entire laser array is turned on at a very low intensity of red light. The molecules that are in the binary 1 state (*P* or *Q* intermediates) do not absorb these red beams or change their state.

But the molecules that started out in the original binary 0 state (*bR*) do absorb the beams (but do not change their structure), because they have cycled to the red-absorbing *O* intermediate. A detector images the light passing through the cube of memory and records the location of *O* and of *P* or *Q* structures—or in terms of binary code, the detector reads 0's and 1's. The process is complete in approximately 10 milliseconds, a rate of 10 megabytes per second for each page of memory.

Three-Dimensional Memories

In addition to facilitating parallel processing, three-dimensional cubes of bacteriorhodopsin provide much more memory space than do two-dimensional optical memories. For example, a relatively recent, nonbiological memory system incorporates a thin film of magnetic material that is written on by a laser beam and erased by a magnetic field. These memories are two-dimensional because data are stored on the surface of the disk. Such two-dimensional memories have a storage capacity that is limited to about 100 million bits per square centimeter.

In contrast, three-dimensional optical memories can theoretically approach storage densities of one trillion bits per cubic centimeter. In practice, optical and hardware limitations lower possible densities for volumetric memories. Nevertheless, most investigators believe a 300-fold improvement in storage capacity over two-dimensional devices should be possible. Indeed, I anticipate that the major near-term impact of bioelectronics on computer hardware will be in the area of volumetric memory.

Speed is also an important benefit of volumetric memories. The combination of three-dimensional storage with the use of parallel architectures enhances the speed of such memories, just as parallel processing in the human brain overcomes relatively slow neural processes and allows the brain to be a thinking machine with fast reflexes and rapid decision-making capability. The entire writing process described above takes place in about 10 milliseconds. If we illuminate a square measuring 1,024 bits by 1,024 bits within a larger cube of protein, we can write 1,048,576 bits of data, or about 105 kilobytes, into memory in a 10-millisecond cycle. These values represent an overall write speed of 10 million characters per second, comparable to slow semiconductor memory. Yet each memory device can access more than one data cube, and the speed of the memory is proportional to the number of cubes operating in parallel. Thus, an eight-cube memory would operate much faster, at 80 million characters per second.

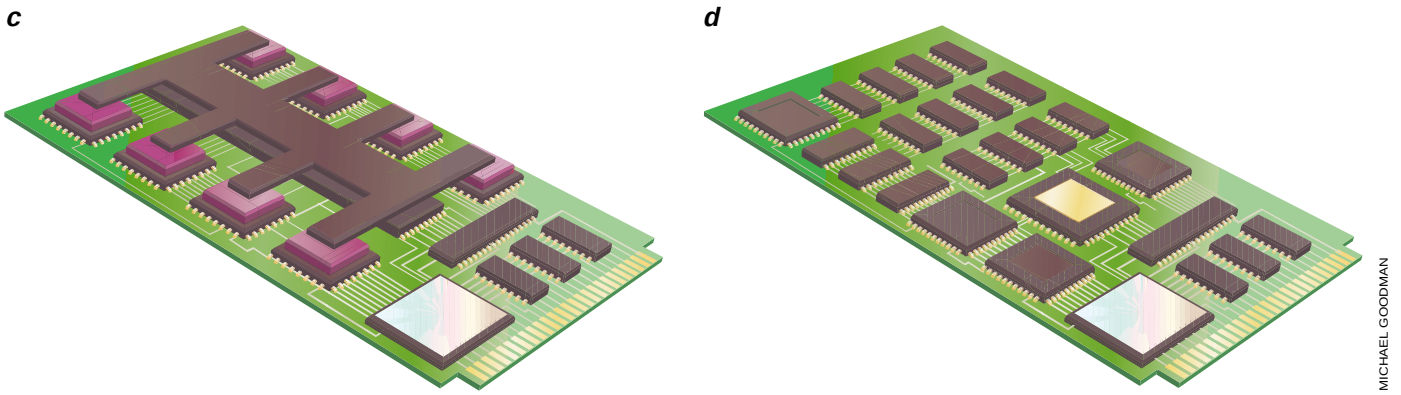
Cubes of memory must be extremely uniform in their composition to ensure accurate reading and writing, because too many or too few molecules in one region will distort information stored there. Manufacturing the cubes in low gravity can produce the needed homogeneity for memory devices. Two space shuttle flights investigating this possibility were sponsored by the W. M. Keck Center for Molecular Electronics at Syracuse University in collaboration with BioServe Space Technologies, the U.S. Air Force Rome Laboratory and the National Aeronautics and Space Administration. The results have been encouraging, and more flights are planned.

Several other types of computer systems based on bacteriorhodopsin are being investigated. For example, biological molecules seem to hold promise as components of the associative memories needed for neural networks and, eventually, for artificial intelligence.

Neural Networks

Associative memories operate rather differently from the memories that dominate current computer architectures. This type of architecture takes a set of data, often in the form of an image, and scans the entire memory bank until it finds a data set that matches it. In some cases, the computer will find the closest match if it cannot find a perfect match, in a sense taking an educated guess at an answer. Because the human brain operates in a neural, associative mode, many computer scientists believe large-capacity associative memories will be required if we are to achieve artificial intelligence.

My laboratory has developed an associative-memory device that relies on the holographic properties of thin films of bacteriorhodopsin. Holograms allow multiple images to be stored in the same segment of memory, permitting large data sets to be analyzed simultaneously. The memory system is based on the classic design of Eung G. Paek and Demetri Psaltis of the California Institute of Technology [see "Optical Neural Computers," by Yaser S. Abu-Mostafa and Demetri Psaltis; SCIENTIFIC AMERICAN, March 1987]. We find that bacteriorhodopsin offers distinct advantages over the photorefractive crystals used to fabricate these memories. Because the protein is more sensitive to light than are inorganic crystals, lower light levels can be employed. In consequence, less energy is needed for writing to and reading from memory, and the speed of these processes improves. Further, bacteriorhodopsin can be written to and read



ry—32 gigabytes of permanent memory (*b*) and eight gigabytes of removable memory (*c*). When combined with a semiconductor central processing unit (*d*), these cards form a complete computer system with enhanced capabilities.

from many more times than can crystals, which suffer from fatigue after repeated read-write cycles.

As studies of natural bacteriorhodopsin continue, many laboratories are also exploring the value of modified forms of the protein in computer devices. Specifically, they are studying genetically engineered versions of the protein, in which one amino acid replaces another in order to enhance the properties needed for particular applications. For example, the lifetime of the *M* state in the photocycle can be lengthened by removal of an internal amino acid from the protein, as shown by Norbert Hampp and Christoph Bräuchle of the University of Munich, in collaboration with Oesterhelt.

Of course, biomolecular computers represent the ultimate goal. As I mentioned earlier, however, most scientists believe the first step in the development of protein-based computers will be the generation of hybrid systems that combine the best features of semiconductor and molecular architectures. In particular, hybrid technology, composed in part of high-density, protein-based memory, may help solve the lingering problem of memory capacity.

During the past decade, the speed of computer processors increased almost 1,000 times, whereas external data storage capacities increased by only a factor of 50. Also, the transfer of data within the computer remains the principal bottleneck that limits performance. Parallel processing and light-based interconnections, both made faster with hybrid computers that exploit the efficient switching of biological molecules, allow for the storage, transfer and manipulation of massive amounts of data.

To explore the possible value of hybrid computers, my laboratory is currently designing one that contains four types of memory units or processors, known as cards. The card with the central processing unit of this computer will consist of traditional semiconductor technology. Two cards will contain protein-based volumetric memory with a total capacity of roughly 40 gigabytes. One of these cards will be a fast, permanent, random-access memory using no moving parts; the other will offer less expensive, removable, long-term data storage. The fourth card will contain an associative memory based on films of bacteriorhodopsin.

The Future of Computers

The hybrid computer we envision would be highly flexible. By taking advantage of particular combinations of the memory cards described above, the computer should be able to handle large pools of data, carry out complex scientific simulations or serve as a unique platform for investigations of artificial intelligence. With close to a terabyte (10^{12} bytes) of memory in cubes of bacteriorhodopsin, this machine would handle large databases with alacrity. Associative memory processing coupled with volumetric memory would make database searches many orders of magnitude faster than is currently possible. Because this hybrid computer can be designed to function as a neural associative computer capable of learning and of analyzing data and images in much the same way as the human brain, the likely importance of hybrid computers to studies in artificial intelligence cannot be underestimated.

Although my group and others have had remarkable success developing volumetric memories and associative processors, more work is needed before a fully operational hybrid computer can be built. Along the way toward developing a powerful yet reasonably priced design, other competing architectures may replace many of the hardware components we have described. Nevertheless, we are confident that hybrid computers of some type will be available within the next eight years.

We further expect that during the next two decades, they will evolve into the dominant architectures for certain types of computing, such as for scientific calculations and multimedia applications. Personal computer users will benefit by having large and inexpensive memory boards that have many gigabytes of data storage and removable memory components that contain a few gigabytes of data storage in a small cube. Imagine the advantage of carrying in your pocket a small cube storing the equivalent of a comprehensive encyclopedia and all the words you have written in the past 10 years.

But the most dramatic application may well be found in yet another realm. With terabytes of data storage, neural associative capabilities and a high capacity for parallel processing, hybrid computers will, for the first time, incorporate the three crucial hardware requirements for artificial intelligence. We are indeed at the threshold of an exciting new era in computing.

FURTHER READING

THREE-DIMENSIONAL OPTICAL STORAGE MEMORY. Dimitri A. Parthenopoulos and Peter M. Rentzepis in *Science,* Vol. 245, pages 843–845; August 25, 1989.

BACTERIORHODOPSIN: A BIOLOGICAL MATERIAL FOR INFORMATION PROCESSING. Dieter Oesterhelt, Christoph Bräuchle and Norbert Hampp in *Quarterly Reviews of Biophysics,* Vol. 24, No. 4, pages 425–478; November 1991.

PROTEIN-BASED OPTICAL COMPUTING AND MEMORIES. Robert R. Birge in *Computer,* Vol. 25, No. 11, pages 56–67; November 1992.

PROTEIN-BASED THREE-DIMENSIONAL MEMORY. Robert R. Birge in *American Scientist,* Vol. 82, No. 4, pages 348–355; July–August 1994.

Environmental Degradation in Ancient Greece

Contrary to the view that the ancients lived in harmony with their environment, archaeological and geologic evidence shows that they often abused the land

by Curtis N. Runnels

The stark Greek landscape charms everyone who sees it. For thousands of years, Greeks and visitors alike have sung the praises of this small country, famous for its Bronze Age civilizations and the cultural achievements of its people in classical times. But is the countryside one sees today the result of climate acting alone, or have humans played a part by clearing forests and causing soil erosion?

Recent archaeological work is changing a long-standing view of the impact of agriculture on the land in Greece. The evidence mounts for episodes of deforestation and catastrophic soil erosion over the past 8,000 years. Many scholars believe they resulted from a long history of human land use and abuse. This new perspective on human settlement and its impact on the natural environment stands in sharp contrast to the views of the 19th-century Romantics, who saw the ancient Greeks as careful stewards of a land they held to be filled with gods.

Indeed, it remains a widely espoused opinion today that the destructive ecological practices of modern civilization are a new development. The popular press frequently carries reports of people who advocate returning to the balanced and reverential regard they suppose our ancestors had for the natural world. The Garden of Eden is a primal myth of Western civilization, and it was preceded in classical antiquity by the belief in the Golden Age—a time, alas now lost, when human beings were said to have lived in innocent harmony with their natural environment.

Sir Peter B. Medawar (in *The Limits of Science,* a book of his essays published in 1984) described this kind of thinking as "Arcadian." He compared the concepts of Utopia and Arcadia and concluded that Arcadia is closer to the ideal of a Golden Age than are the Utopias of Thomas More and Francis Bacon. Their Utopias were places where science-based technology was employed for the melioration of society, whereas Arcadia "is the conception farthest removed from Utopia, for one of its principal virtues is to be pastoral, prescientific and pretechnological. In Arcadia, mankind lives in happiness, ignorance and innocence, free from the diseases and psychic disquiet that civilization brings with it—living indeed in that state of inner spiritual tranquillity which comes today only from having a substantial private income derived from trustee securities."

The effects of Arcadian thinking can be seen in the debate about the degree to which the ancient Greeks were responsible for the deforestation and erosion that have reduced much of Greece to a barren, stony—if picturesque—wasteland. To be sure, environmental degradation (if not its causes) was noted in ancient times; many references to it appear in the writings of the ancients, particularly Plato and Aristotle. They give accurate and apparently eyewitness accounts of deforestation and soil erosion in the fourth century B.C. Archaeology can now confirm that despoilment of the natural environment took place in antiquity, but the evidence allows one also to place the responsibility for this destruction primarily on the inhabitants of that time.

CURTIS N. RUNNELS earned degrees in archaeology from the University of Kansas and Indiana University and has taught at Stanford University. He is associate professor of archaeology at Boston University. Runnels has worked for 20 years in Greece, specializing in the study of early prehistory and the relation between human settlement and landscape through time. His most recent book is *A Greek Countryside: The Southern Argolid from Prehistory to the Present Day,* written with Michael H. Jameson and Tjeerd H. van Andel.

A New Kind of Archaeology

Over the past two decades, the introduction of multidisciplinary and intensive regional survey techniques in Greece has revolutionized Greek archaeology. One aspect of this revolution is a shift away from the investigation of single sites to studies of the natural and cultural history of entire regions. This approach is achieved in part by using new methods of fieldwork, particularly searching large areas (typically more than 100 square kilometers) with teams of people who record every artifact, feature and site that can be detected. Another innovation is the large-scale use of remote sensing, ranging from ground-penetrating radar to satellite images, to assist the archaeologist in detecting past human activity. Further aid has come from the new discipline of geoarchaeology, which combines the techniques of geology and archaeology to improve the interpretation of the natural contexts of ancient cultures.

An example of the collaboration between archaeologists and geoarchaeologists can be seen in the study of landscape change in southern Greece. Two major projects in which I have participated since 1979 in the Argolid (the northeastern part of the Peloponnese peninsula) combined walking tours designed to identify archaeological sites with geologic surveys that attempted to reconstruct the history of the landscape during the Pleistocene and Holocene epochs of the past 50,000 years.

One of these projects targeted the Argive plain. According to Aristotle, it had undergone considerable alteration in the Bronze Age, from about 3000 to

COURTESY OF THE NATIONAL GALLERY OF ART, AILSA MELLON BRUCE FUND

GOLDEN AGE of Greece is represented in Claude Lorrain's painting *The Judgment of Paris* (1646), in which the goddess of strife throws down a golden apple inscribed "to the fairest" and says that Paris must award it to one of three goddesses. The persistent view of the ancient Greeks as careful stewards of a land that they saw as guided by gods and goddesses is belied by recent archaeological evidence of soil erosion caused by human land abuse.

1000 B.C. To investigate ancient land use, I joined archaeologist Berit Wells, director of the Swedish Institute in Athens, and Eberhard A. W. Zangger, a geoarchaeologist now at the University of Heidelberg. Our goal was to survey two valleys, Berbati and Limnes, which lie on the northern edge of the Argive plain and have a combined area of 60 square kilometers.

This project, sponsored by the Swedish Institute and headed by Wells, had two parts. One was a geoarchaeological study of the two valleys by Zangger, who had done a similar study of the Argive plain while working on his doctoral dissertation at Stanford University. The other was an intensive archaeological survey.

The survey brought to light evidence of human activity from the Middle Paleolithic period, some 50,000 years ago, and showed that agriculturists entered the area about 7,000 years ago, settling on the edges of the Berbati Valley near the best soils and supplies of water. After this initial colonization, the size of the original Neolithic settlements began to grow, and new ones were founded. By the Late Neolithic period (between 4000 and 3000 B.C.), more than 20 settlements and smaller sites were scattered over the region. This expansion continued for the first few centuries of the third millennium—the Early Bronze Age in cultural terms.

Ancient Land Use

From the larger settlements came evidence of agriculture and many artifacts such as pottery, querns for grinding grain and simple stone tools of flint and obsidian. Some of the smaller sites doubtless served special purposes (as huts for shepherds and storehouses for tools, for example) and were not necessarily inhabited permanently. Nevertheless, they demonstrate the more intensive use of the landscape at this time because they are located on steep slopes, at high elevations and in areas where only soils of marginal productivity ever existed.

In addition to sites where past human activity left concentrations of artifacts, our searching technique revealed tens of thousands of individual Neolithic items scattered over the countryside. These thin scatters of artifacts are found in areas now uninhabited and in some cases uninhabitable because they are devoid of soil and vegetation.

The pattern of settlements and artifacts reveals the structure of what archaeologists call fossil cultural landscapes. Particularly interesting are the numerous finds of ground and polished hard-stone axes and flint blades, which have edges coated with silica deposited when the blades were used to cut grasses and other plants. These common agri-

cultural tools belong to the Neolithic period and the Early Bronze Age and mark regions where agricultural fields once existed. We believe grazing and farming were the principal activities on the slopes of the Berbati and Limnes valleys from 4000 to 3000 B.C. and that they resulted in one or more episodes of catastrophic soil erosion, which left alluvial fans in the valley bottoms and deep layers of sediments to mantle the Argive plain.

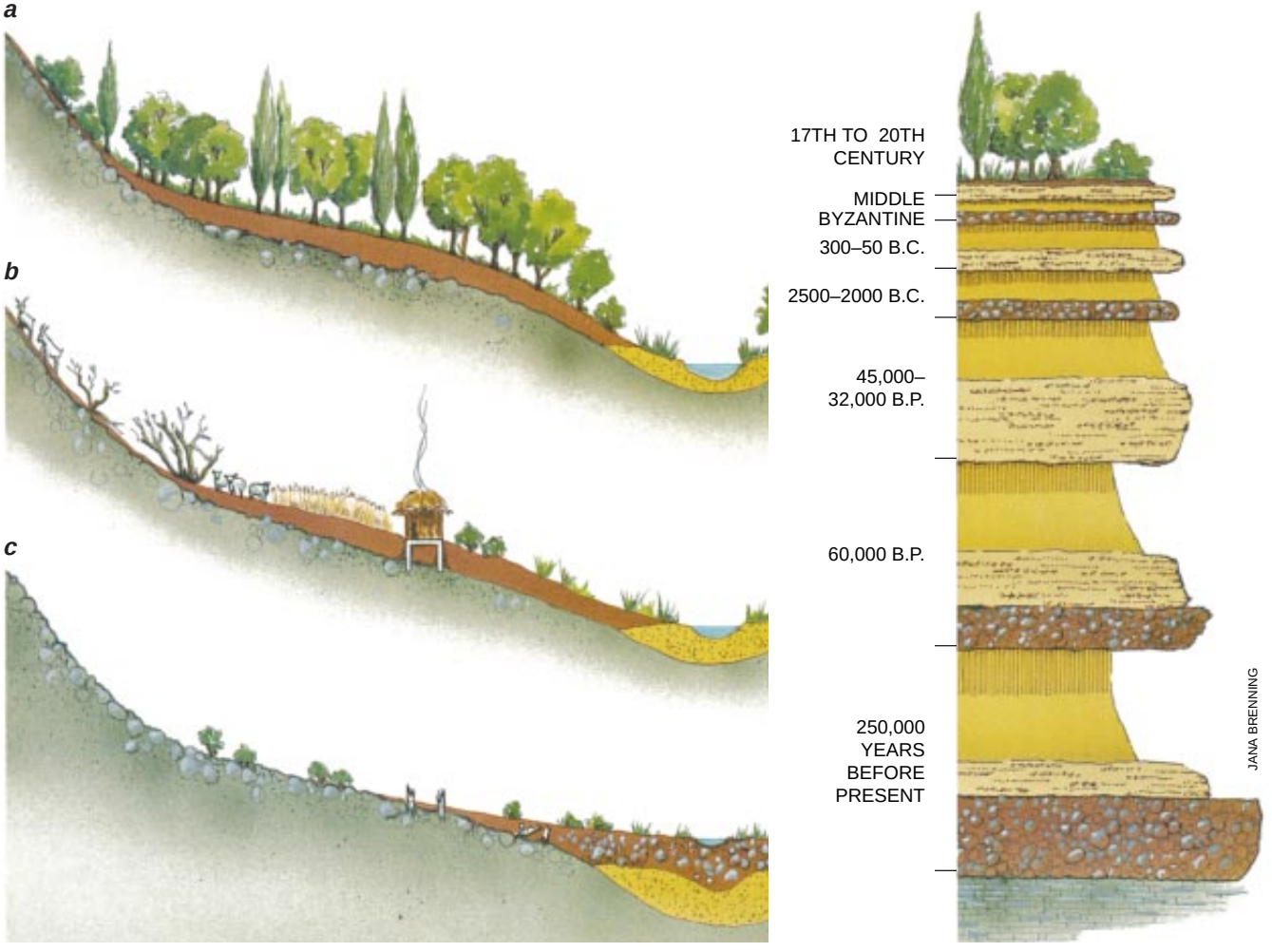
One episode occurred at the end of the Neolithic period, well after the expansion of settlement in the Berbati and Limnes valleys. This erosion buried at least one Neolithic site in the Argive plain, which was found in one of Zangger's cores that was drilled through the thick alluvium. What appears to have been a second similar event came in the third millennium, at the end of the Early Bronze Age. The erosional deposits of this period spread across the Argive plain to make up most of its present surface. Nor were these the last events to be recorded. Flooding eclipsed parts of the town of Tiryns at the end of the Late Bronze Age (around 1200 B.C.). After this natural disaster, the slopes of Berbati were abandoned and for the most part never inhabited again.

Confirmation of the picture offered by the archaeological record comes from pollen taken from a seven-meter core obtained from the now vanished Lake Lerna in the Argive plain and analyzed by Susanne Jahns of the University of Göttingen. It shows that deciduous oak trees were common in the Argolid in the Holocene epoch. By the middle of the fourth millennium B.C., however, the oak pollen had dropped off sharply and was replaced by pollen of hornbeam, pine, scrub oak and heather. These plants are endemic to cleared and disturbed land. Today the slopes surrounding the Argive plain consist of bare, rocky land covered with a thin scrub vegetation of low bushes and herbaceous plants, with only isolated pockets of trees—chiefly pines.

Other projects in Greece have found signs of soil erosion. The episodes do not correlate in time with the periods of erosion in the Argolid. If climate were the sole cause, most of Greece would have simultaneously been affected. The variation in timing therefore points to a human role. Clearing natural vegetation from the slopes, which was necessary to bring them under cultivation and to open them for grazing by sheep and goats, made the soil unstable and thereby triggered extensive and permanent erosion.

The Soil Tells a Story

Our second project produced further evidence of regional land abuse. This study was carried out in the southern Argolid, a remote part of the Peloponnese peninsula. With historian Michael H. Jameson of Stanford and geo-



LAND ABUSE by ancient Greeks followed a consistent pattern (*left*). At first (*a*), a mixed forest held the soil in place on a slope. When the forest was cleared for farming (*b*), the soil stayed in place for a time, but eventually erosion carried it off the slope, depositing alluvium in the valley bottom (*c*). As a result, a typical geologic profile (*right*) from the southern Argolid region of Greece shows a sequence of erosional deposits and intervening soils (with their approximate ages). In each period the soil that gradually forms above the deposits is thinner and less developed and is less able to support vegetation.

archaeologist Tjeerd H. van Andel of the University of Cambridge, I conducted an archaeological reconnaissance of some 250 square kilometers. The survey identified more than 350 sites of human activity spanning a period of some 50,000 years.

An interesting pattern of settlement emerged when we plotted the numbers of sites by archaeological period. It turned out that the sites are not distributed evenly through time but instead exhibit a series of abrupt peaks and valleys, suggesting that periods of expansion of sites over the land were followed by periods of abandonment. Somewhat to our surprise, we found that this unexpected pattern of settlement could be correlated with the geologic history of the region.

This correlation resulted from an analysis of the erosional history of the area made by van Andel and Kevin O. Pope, also at Stanford. An innovative feature of their work was the study of soils. Soils can be distinguished from the deposits of erosion because they form on the deposits as a result of the chemical weathering and transformation of underlying sediments—a process that can take thousands of years in the semiarid parts of Greece. Van Andel and Pope identified at least seven cycles of soil erosion that were followed by times of stability when thick profiles of soil had the opportunity to accumulate. The age-related features of soil profiles make it possible to correlate the different erosional deposits and soil profiles with archaeological findings.

Three of the erosional events that left marks in the geologic record took place during the last ice age—at about 272,000, 52,000 and 33,000 years ago. They can be attributed to global climate changes. But four episodes belong to the past 5,000 years. Each episode—at about 2500 B.C., 350 to 50 B.C., A.D. 950 to 1450 and in recent times—was followed by a period of stability when substantial soil profiles formed. Although small-scale climate changes may partly explain this pattern, we place the chief blame on the activities of the local inhabitants. Our evidence for this claim is twofold: the correlation between the periods of erosion and the periods of intense human settlement, and the formation of soil during the periods when the human impact was minimal.

Our dating technique began with a search for archaeological sites and artifacts that are covered by alluvium, buried in it or sitting on it. By combining archaeological and geologic data, we were able to date the times of erosion closely. An important element in this work was the recognition that layers of soils could be used to determine periods of surface stability that contrasted with the erosional periods.

From these data we concluded that the clearing of land during times of intensive human settlement gave rise to soil erosion, which in turn caused the people to abandon their settlements or at least to scale back their activities. And the reduced human activity permitted erosional deposits to stabilize and soils to form.

Soil erosion on a similar scale has been reported from other parts of Greece—the northern provinces of Macedonia and Thessaly and the islands of Euboea in the center of the country and of Crete in the south. The episodes date from as early as the sixth millennium B.C. and continue through virtually every historical era to the present day.

The archaeological study of ancient environmental catastrophes has only just begun, and it is perhaps too early to say that such analyses can always distinguish the effects of human habitation from those of climate. Nevertheless, it is clear from the variety and different timing of these events in neighboring regions that a climatic explanation is unlikely to account for all the data. We have learned enough to maintain that the changes in the natural landscape in antiquity are at least partly the result of shortsighted human activity. The lesson to be drawn from our work is that people have a long history of misusing the land: environmental catastrophes are not an innovation of the modern world.

Paradise Lost

"In the time of the Trojan wars the Argive land was marshy and could only support a small population, whereas the land of Mycenae was in good condition (and for this reason Mycenae was the superior). But now the opposite is the case... the land of Mycenae has become... dry and barren, while the Argive land that was formerly barren owing to the water has now become fruitful. Now the same process that has taken place in this small district must be supposed to be going on over whole countries and on a large scale."

—Aristotle, from *Meteorologica,* Book 1, Chapter 14

EUGENE G. SCHULZ

A typical landscape in the Argolid region of Greece today

FURTHER READING

FIVE THOUSAND YEARS OF LAND USE AND ABUSE IN THE SOUTHERN ARGOLID, GREECE. Tjeerd H. van Andel, Curtis N. Runnels and Kevin O. Pope in *Hesperia,* Vol. 55, No. 1, pages 103–128; January–March 1986.

LAND USE AND SOIL EROSION IN PREHISTORIC AND HISTORICAL GREECE. Tjeerd H. van Andel, Eberhard Zangger and Anne Demitrack in *Journal of Field Archaeology,* Vol. 17, No. 4, pages 379–396; Winter 1990.

THE BERBATI-LIMNES ARCHAEOLOGICAL SURVEY: THE 1988 SEASON. Berit Wells, Curtis Runnels and Eberhard Zangger in *Opuscula Atheniensia,* Vol. 18, No. 15, pages 207–238; 1990.

NEOLITHIC TO PRESENT SOIL EROSION IN GREECE. Eberhard Zangger in *Past and Present Soil Erosion.* Edited by Martin Bell and John Boardman. Oxbow Books, 1992.

THE GEOARCHAEOLOGY OF THE ARGOLID. Eberhard Zangger. Mann Verlag, 1993.

Seeking the Criminal Element

by W. Wayt Gibbs, *staff writer*

Scientists are homing in on social and biological risk factors that they believe predispose individuals to criminal behavior. The knowledge could be ripe with promise—or rife with danger

BRUCE DAVIDSON *Magnum Photos*

TEXAN TEENS playing with gang signs and loaded guns are acting their age—most adolescents dabble in delinquency for several years. But a small fraction grow into the chronic felons that commit the majority of violent crimes. Can scientists identify the dangerous few before they attack—and if so, what then?

Imagine you are the father of an eight-year-old boy," says psychologist Adrian Raine, explaining where he believes his 17 years of research on the biological basis of crime is leading. "The ethical dilemma is this: I could say to you, 'Well, we have taken a wide variety of measurements, and we can predict with 80 percent accuracy that your son is going to become seriously violent within 20 years. We can offer you a series of biological, social and cognitive intervention programs that will greatly reduce the chance of his becoming a violent offender.'

"What do you do? Do you place your boy in those programs and risk stigmatizing him as a violent criminal even though there is a real possibility that he is innocent? Or do you say no to the treatment and run an 80 percent chance that your child will grow up to (a) destroy his life, (b) destroy your life, (c) destroy the lives of his brothers and sisters and, most important, (d) destroy the lives of the innocent victims who suffer at his hands?"

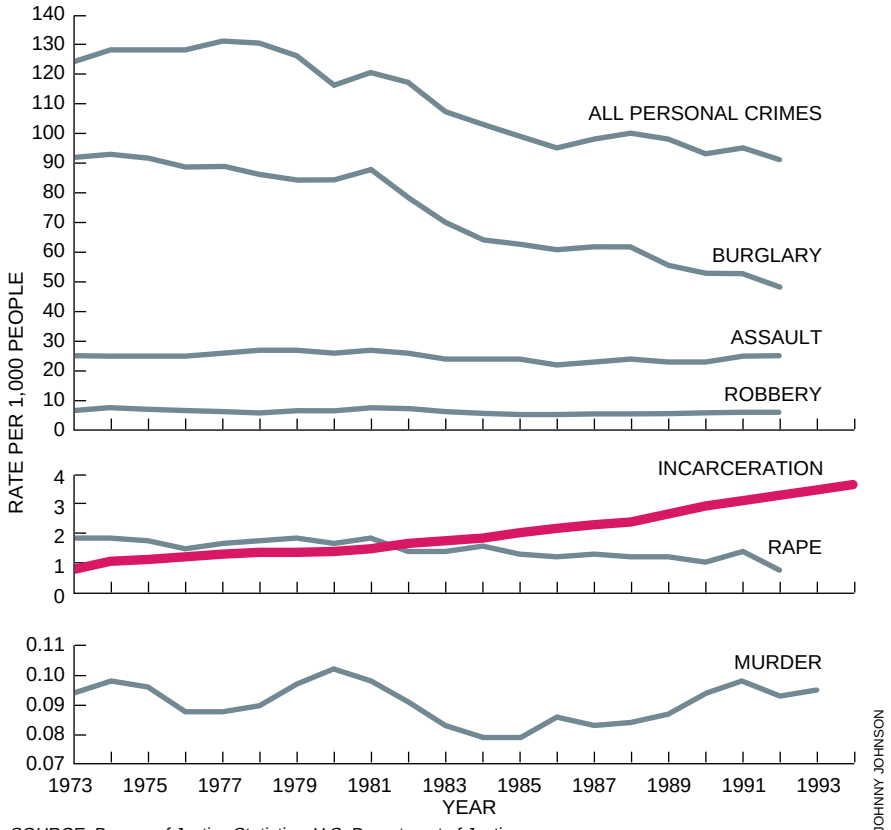
For now, such a Hobson's choice is purely hypothetical. Scientists cannot yet predict which children will become dangerously aggressive with anything like 80 percent accuracy. But increasingly, those who study the causes of criminal and violent behavior are looking beyond broad demographic characteristics such as age, race and income level to factors in individuals' personality, history, environment and physiology that seem to put them—and society—at risk. As sociologists reap the benefits of rigorous long-term studies and neuroscientists tug at the tangled web of relations between behavior and brain chemistry, many are optimistic that science will identify markers of maleficence. "This research might not pay off for 10 years, but in 10 years it might revolutionize our criminal justice system," asserts Roger D. Masters, a political scientist at Dartmouth College.

"With the expected advances, we're going to be able to diagnose many people who are biologically brain-prone to violence," claims Stuart C. Yudofsky, chair of the psychiatry department at Baylor College of Medicine and editor of the *Journal of Neuropsychiatry and Clinical Neurosciences.* "I'm not worried about the downside as much as I am encouraged by the opportunity to prevent tragedies—to screen people who might have high risk and to prevent them from harming someone else." Raine, Yudofsky and others argue that in order to control violence, Americans should trade their traditional concept of justice based on guilt and punishment for a "medical model" based on prevention, diagnosis and treatment.

But many scientists and observers do worry about a downside. They are concerned that some researchers underplay the enormous complexity of individual behavior and overstate scientists' ability to understand and predict it. They also fear that a society desperate to reduce crime might find the temptation to make premature or inappropriate use of such knowledge irresistible.

Indeed, the history of science's assault on crime is blemished by instances in which incorrect conclusions were used to justify cruel and unusual punishments. In the early 1930s, when the homicide rate was even higher than it is today, eugenics was in fashion. "The eugenics movement was based on the idea that certain mental illness and criminal traits were all inherited," says Ronald L. Akers, director of the Center for Studies in Criminology and Law at the University of Florida. "It was based on bad science, but they thought it was good science at the time." By 1931, 27 states had passed laws allowing compulsory sterilization of "the feeble-minded," the insane and the habitually criminal.

Studies in the late 1960s—when crime was again high and rising—revealed that many violent criminals had an extra Y chromosome and thus an extra set of "male" genes. "It was a dark day for science in Boston when they started screening babies for it," recalls Xandra



SOURCE: Bureau of Justice Statistics, U.S. Department of Justice

CRIME RATES have not responded consistently to "get tough" approaches to incarceration. Since the early 1970s the proportion of Americans behind bars has more than tripled. Property crime (including burglary, robbery and personal larceny) has dropped about 30 percent, but violent crime remains high.

O. Breakefield, a geneticist at Massachusetts General Hospital. Subsequent studies revealed that although XYY men tend to score lower on IQ tests, they are not unusually aggressive.

False Positive ID

Social science studies on the causes of crime have been less controversial, in part because they have focused more on populations than on individuals. But as consensus builds among criminologists on a few key facts, researchers are assembling these into prediction models that try to identify the juveniles most likely to lapse into delinquency and then into violent crime.

Perhaps their most consistent finding is that a very small number of criminals are responsible for most of the violence. One study tracked 10,000 males born in Philadelphia in 1945 for 27 years; it found that just 6 percent of them committed 71 percent of the homicides, 73 percent of the rapes and 69 percent of the aggravated assaults attributed to the group.

Preventing just a small fraction of adolescent males from degenerating into chronic violent criminals could thus make a sizable impact on the violent crime rate, which has remained persistently high since 1973 despite a substantial decline in property crime. (Females accounted for only 12.5 percent of violent crime in 1992.) "For every 1 percent that we reduce violence, we save the country $1.2 billion," Raine asserts.

The problem, says Terrie E. Moffitt, a psychologist at the University of Wisconsin who is conducting long-term delinquency prediction studies, is that "a lot of adolescents participate in antisocial behavior"—87 percent, according to a survey of U.S. teens. "The vast majority desist by age 21," she says. The dangerous few "are buried within that population of males trying out delinquency. How do you pick them out? Our hypothesis is that those who start earliest are at highest risk."

Marion S. Forgatch of the Oregon Social Learning Center tested that hypothesis on 319 boys from high-crime neighborhoods in Eugene. Last November at the American Society of Criminology meeting, she reported her findings: boys who had been arrested by age 14 were 17.9 times more likely to become chronic offenders than those who had not, and chronic offenders were 14.3 times more likely to commit violent offenses. "This is a good way of predicting," she says.

Good is a relative term. For if one were to predict that every boy in her study who was arrested early would go on to commit violent crimes, one would be wrong more than 65 percent of the time. To statisticians, those so misidentified are known as false positives. "All of these predictors have a lot of false positives—about 50 percent on average," says Akers, who recently completed a survey of delinquency prediction models. Their total accuracy is even lower, because the models also fail to identify some future criminals.

The risk factors that Akers says researchers have found to be most closely associated with delinquency are hardly surprising. Drug use tops the list, followed by family dysfunction, childhood behavior problems, deviant peers, poor school performance, inconsistent parental supervision and discipline, separation from parents, and poverty. Numerous other controlled studies have found that alcoholism, childhood abuse, low verbal IQ and witnessing violent acts are also significant risk factors. Compared with violent behavior, however, all these experiences are exceedingly common. The disparity makes it very difficult to determine which factors are causes and which merely correlates.

Preventive Intervention

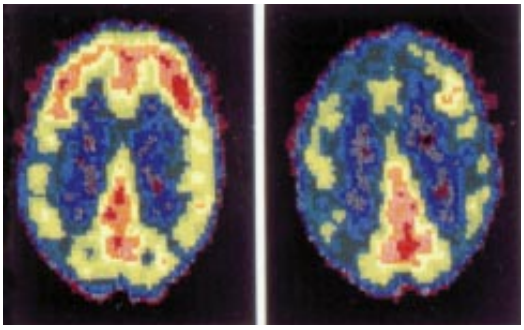
The difference is important, notes Mark W. Lipsey of Vanderbilt University, because "changing a risk factor if it is not causal may have no impact," and the ultimate goal of prediction is to stop violence by intervening before it begins. Unfortunately, improvements in predictive models do not necessarily translate into effective intervention strategies. Lipsey recently analyzed how well some 500 delinquency treatment programs reduced recidivism. "The conventional wisdom that nothing works is just wrong," he concludes. But he concedes that "the net effect is modest"—on average, 45 percent of program participants were rearrested, versus 50 percent of those left to their own devices. Half of that small apparent improvement, he adds, may be the result of inconsistency in the methods used to evaluate the programs.

Some strategies do work better than others, Lipsey discovered. Behavioral programs that concentrated on teaching job skills and rewarding prosocial attitudes cut rearrest rates to about 35 percent. "Scared straight" and boot camp programs, on the other hand, tended to increase recidivism slightly.

Patrick H. Tolan of the University of Illinois at Chicago has also recently published an empirical review of delinquency programs. To Lipsey's findings he adds that "family interventions have repeatedly shown efficacy for reducing antisocial behavior and appear to be among the most promising interventions to date." According to Forgatch, two experiments in Eugene, Ore., showed that teaching parents better monitoring and more consistent, less coercive discipline techniques reduces their kids' misbehavior. "We should make parenting skills classes compulsory for high school students," argues Raine of the University of Southern California.

Unfortunately, Tolan observes, family intervention is difficult and rarely attempted. The most common kinds of programs—counseling by social workers, peer mediation and neighborhood antiviolence initiatives—are hardly ever examined to see whether they produce lasting benefits. "It usually is hard to imagine that a good idea put into action by well-meaning and enlightened people cannot help," he noted in the paper. "It may seem that any effort is better than nothing. Yet our review and several of the more long-term and sophisticated analyses suggest that both of these assumptions can be dangerously wrong. Not only have programs that have been earnestly launched been ineffective, but some of our seemingly best ideas have led to worsening of the behavior of those subjected to the intervention."

Many researchers are thus frustrated that the Violent Crime Control and Law Enforcement Act of 1994 puts most of its $6.1 billion for crime prevention in untested and controversial programs, such as "midnight basketball" and other after-school activities. "Maybe these programs will help; maybe they won't," Tolan says. "No one has done a careful evaluation." The Crime Act does not insist that grant applicants demonstrate or even measure the effectiveness of their approach. For these and other reasons, Republicans vowed in their "Contract with America" to repeal all prevention programs in the Crime Act and to increase funding for prison construction. But that strategy also ignores research. "We do know," Tolan asserts, "that locking kids up will not reduce crime and may eventually make the problem worse."

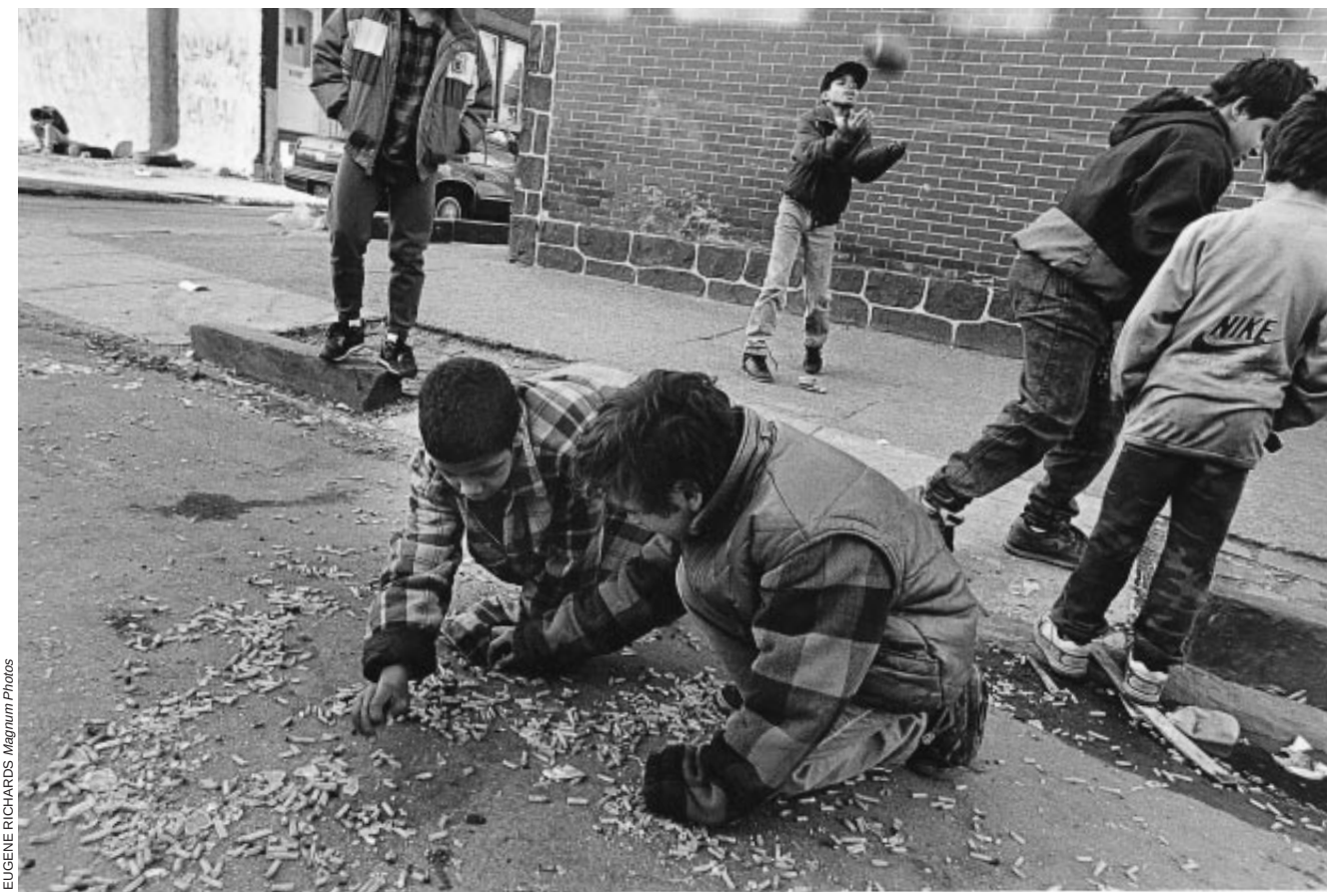


ADRIAN RAINE AND MONTE S. BUCHSBAUM

BRAIN OF MURDERER (*right*) shows less activity in the frontal cortex (*top third of image*) than the brain of a nonviolent subject of the same age and sex. In one study of 22 murderers, about 75 percent had low frontal activity, which is believed to indirectly regulate aggressive impulses.

All in Our Heads?

The failure of sociology to demonstrate conclusively effective means of controlling violent crime has made some impatient. "There is a growing recognition that we're not going to solve



EUGENE RICHARDS *Magnum Photos*

POOR PARENTAL SUPERVISION is a major risk factor for later delinquency. These children in Philadelphia play with empty crack cocaine vials. Parent training programs have been among the most successful in reducing kids' antisocial behavior.

The Tangled Roots of Violence

The failure of expensive prison booms and welfare programs to beat back the historically high violent crime rates of the past 20 years has prepared fertile ground for new approaches to crime control. Encouraged by research that tentatively links a few instances of antisocial aggression with biological abnormalities, some politicians and activists are turning to science, perhaps too hastily, to identify and treat those who are likely to become dangerous.

Take the case of Everett L. "Red" Hodges, a California oilman who has spent more than $1 million to support research that implicates the trace metal manganese as a marker for violent criminal behavior. Hodges was struggling to tame a delinquent son in 1984 when he came across a *Science News* story on a study that had found high levels of lead, cadmium and copper in the head hair of violent felons.

Intrigued, Hodges offered funding to Louis A. Gottschalk, a psychiatrist at the University of California at Irvine, to conduct a controlled study to replicate the results. Analysis of hair clipped from convicted and accused felons at a prison and two county jails in southern California revealed no unusual levels of lead, cadmium or copper. But Gottschalk did find that average levels of manganese were about 3.6 times higher in the alleged felons than in men of similar age and race at local barbershops. "A new paradigm is opening in criminal justice," Hodges says, beaming. "It's a marker."

That judgment may be premature. Critics of Gottschalk's research, published in 1991 in a psychiatric (rather than a nutrition) journal, point out that average manganese levels varied from 2.2 parts per million in the prisoners to just 0.71 in one of the groups of jail inmates. Previous studies had found *lower* manganese levels in inmates than in control subjects. Skeptics also note that Gottschalk threw a wide net, measuring levels of 23 trace metals. "If you look at enough variables, you're bound to find a statistically significant association," comments Curtiss D. Hunt of the Grand Forks Human Nutrition Research Center in North Dakota. "But it may be meaningless." Hunt adds that the concentration of a metal in the hair does not tell one how much is

KELLY NEWS AND ENTERTAINMENT *KCRA, Sacramento*

any problem in society using just one discipline," says Diana Fishbein, a professor of criminology at the University of Baltimore. "Sociological factors play a role. But they have not been able to explain why one person becomes violent and another doesn't."

Some social scientists are looking to psychiatrists, neurologists and geneticists to provide answers to that question, ready or not. "Science must tell us what individuals will or will not become criminals, what individuals will or will not become victims, and what law enforcement strategies will or will not work," wrote C. Ray Jeffery, a criminologist at Florida State University, last year in the *Journal of Research in Crime and Delinquency*.

As medical researchers have teased out a few tantalizing links between brain chemistry, heredity, hormones, physiology and assaultive behavior, some have become emboldened. "Research in the past 10 years conclusively demonstrates that biological factors play some role in the etiology of violence. That is scientifically beyond doubt," Raine holds forth. The importance of that role is still very much in doubt, however.

As with social risk factors, no biological abnormality has been shown to *cause* violent aggression—nor is that likely except in cases of extreme psychiatric disorder. But researchers have spotted several unusual features, too subtle even to be considered medical problems, that tend to appear in the bodies and brains of physically aggressive men. On average, for example, they have higher levels of testosterone, a sex hormone important for building muscle mass and strength, among other functions. James M. Dabbs, Jr., of Georgia State University has found in his experiments with prison inmates that men with the highest testosterone concentrations are more likely to have committed violent crimes. But Dabbs emphasizes that the link is indirect and "mediated by numerous social factors," such as higher rates of divorce and substance abuse.

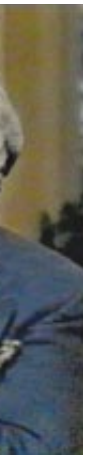
"Low resting heart rate probably rep-

EUGENE RICHARDS *Magnum Photos*

SINS OF THE PARENT are often visited on the child. Delinquents are more likely to have parents who abuse drugs or alcohol, commit crimes or beat them. But risk factors are generally poor predictors: most children of such parents do not become chronic criminals.

in the blood or the brain. "We know so little about manganese's role in the body that we haven't even set an RDA [recommended daily allowance] for it."

Hodges remains convinced he is on the right track. "Violence can be detected and treated," he argues. In 1987 a mugger fractured the skull of another of Hodges's sons. That year Hodges founded the Violence Research Foundation (VRF) to lobby public officials to experiment with treatment programs that use what he calls "the power of nutrition" to pacify violent criminals.



The VRF found an ally in Senator Robert Presley of California, who pushed through a bill in 1989 authorizing a study of male prisoners by Stephen Schoenthaler of California State University at Stanislaus. In the first part of the study, 402 offenders were divided randomly into three groups and given vitamin-mineral supplements equivalent to the RDA, three times the RDA or a placebo. Preliminary results showed that rule violations among the first group dropped 38 percent during the study. Strangely, the behavior of inmates getting the higher dose did not improve significantly, and violations rose 20 percent among the placebo group.

Although encouraging, the equivocal results are so inconclusive that Schoenthaler has decided not to publish them until he completes further studies with more controls. Hodges, however, has publicized the results widely at conferences and on television talk shows (*left*), much to the scientist's annoyance. "We have asked that all reports on the study be embargoed until the final paper goes through the peer-review process," Schoenthaler says, "but he [Hodges] continues to make an example of it."

With two studies in hand, Hodges has redirected his crusade to Washington. "What we need is a leap of faith from the Justice Department," he says. So he has begun searching for converts. Hodges claims to have the support of former attorney general Edwin Meese, with whom he has met several times, and Senator Tom Harkin of Iowa. "I have an invitation from [Utah senator] Orrin G. Hatch to come up to Washington as soon as he becomes chairman of the Judiciary Committee," Hodges stated in December. A member of Hatch's staff says that Hodges's "material is under review, but no agreements have been made yet."

Trace element deficiencies are just one of many frequently cited but poorly demonstrated claims that nutritional problems can cause criminal and violent behavior. A 1992 report by the Federal Bureau of Prisons stated that correctional facilities in 46 states have incorporated a wide array of dietary intervention and testing programs, even though "such programs are perceived by many physicians, scientific researchers, registered dietitians, and other health care professionals as an incorporation of food faddism into public policy."

—Steven Vames and W. Wayt Gibbs

resents the best replicated biological correlate of antisocial behavior," Raine observes, pointing to 14 studies that have found that problem children and petty criminals tend to have significantly lower pulses than do well-behaved counterparts. A slower heartbeat "probably reflects fearlessness and underarousal," Raine theorizes. "If we lack the fear of getting hurt, it may lead to a predisposition to engage in violence." But that hypothesis fails to explain why at least 15 studies have failed to find abnormal heart rates in psychopaths.

Jerome Kagan, a Harvard University psychologist, has suggested that an inhibited "temperament" may explain why the great majority of children from high-risk homes grow up to become law-abiding citizens. One study tested pulse, pupil dilation, vocal tension and blood levels of the neurotransmitter norepinephrine and the stress-regulating hormone cortisol to distinguish inhibited from uninhibited, underaroused two-year-olds. An expert panel on "Understanding and Preventing Violence" convened by the National Research Council suggested in its 1993 report that inhibited children may be protected by their fearfulness from becoming aggressive, whereas uninhibited children may be prone to later violence. The panel concluded that "although such factors in isolation may not be expected to be strong predictors of violence, in conjunction with other early family and cognitive measures, the degree of prediction may be considerable."

Perhaps the most frequently cited biological correlate of violent behavior is a low level of serotonin, a chemical that in the body inhibits the secretion of stomach acid and stimulates smooth muscle and in the brain functions as a neurotransmitter. A large body of animal evidence links low levels of serotonin to impulsive aggression. Its role in humans is often oversimplified, however. "Serotonin has a calming effect on behavior by reducing the level of violence," Jeffery wrote in 1993 in the *Journal of Criminal Justice Education.* "Thus, by increasing the level of serotonin in the brain, we can reduce the level of vi-

STEPHEN SHAMES *Matrix*

CHILDHOOD AGGRESSIVENESS, seen in this boy threatening his brother with a broom, is one of the strongest known predictors of later violence. Yet in 1990 a 17-year study found that of 209 hyperaggressive preschoolers predicted to develop antisocial behavior, 177 did not.

KENNETH JARECKE *Contact Press Images*

ANDREW LICHTENSTEIN *Impact Visuals*

SOCIAL RISK FACTORS that seemingly push youths toward violent behavior include repeatedly witnessing assaults, heavy drinking or drug use (implicated in about 60 percent of all offenses), association with deviant peers and gun possession.

olence." A front-page article in the *Chicago Tribune* in December 1993 explained that "when serotonin declines… impulsive aggression is unleashed."

Such explanations do violence to the science. In human experiments, researchers do not generally have access to the serotonin inside their subject's braincase. Instead they tap cerebrospinal fluid from the spinal column and measure the concentration of 5-hydroxyindoleacetic acid (5-HIAA), which is produced when serotonin is used up and broken down by the enzyme monoamine oxidase (MAO). Serotonin does its job by binding to any of more than a dozen different neural receptors, each of which seems to perform a distinct function. The low levels of 5-HIAA repeatedly seen in violent offenders may indicate a shortage of serotonin in the brain or simply a dearth of MAO—in which case their serotonin levels may actually be high. Moreover, serotonin can rise or drop in different regions of the brain at different times, with markedly different effects.

Environment, too, plays a role: nonhuman primate studies show that serotonin often fluctuates with pecking order, dropping in animals when they are threatened and rising when they assume a dominant status. The numerous pathways through which serotonin can influence mood and behavior confound attempts to simply "reduce the level of violence" by administering serotonin boosters such as Prozac, a widely prescribed antidepressant.

Nevertheless, the link between 5-HIAA and impulsive aggression has led to a concerted hunt for the genes that control the production and activity of serotonin and several other neurotransmitters. "Right now we have in our hand many of the genes that affect brain function," says David Goldman, chief of neurogenetics at the National Institute

For Biological Studies, Minorities Need Not Apply

Scientists pursuing the role of biology in violent behavior have been twice shy since 1992, when shrill public criticism forced the National Institutes of Health to withdraw financial support of a conference on the ethical implications of "Genetic Factors in Crime" and compelled former health secretary Louis Sullivan to abort his proposed "Violence Initiative." Led by firebrand psychiatrist Peter Breggin, critics charged that in a society where blacks account for 12.4 percent of the population but 44.8 percent of arrests for violent crimes, such research plays into the hands of racists.

The controversy did little to dissuade scientists from their studies, which continue to grow in number. The NIH has reinstated funding for the genetics conference and increased its budget for violence-related research to $58 million. Most Violence Initiative projects have found support in other programs. And last December the National Science Foundation began soliciting proposals for a $12-million, five-year violence research consortium.

But the political wrangling seems to have intimidated investigators from including minorities in any violence studies with a biological tinge—and from collecting medical data in multiracial studies. Designers of an 11,000-subject, eight-year study of the causes of crime in Chicago, for example, were forced last summer to ditch plans to collect blood and urine samples when Breggin organized rallies to block the project, says Felton Earls, a Harvard University professor and co-director of the study. As a result of such pressure, asserts Adrian Raine of the University of Southern California, "all the biological and genetic studies conducted to date have been done on whites. Scientifically, we can make no statements on the biological basis of violence and crime in blacks or Hispanics or Asians."

There is no reason to suspect that any genetic connection links race to antisocial behavior. But there is reason to be concerned that ostensibly objective biological studies, blindly ignoring social and cultural differences, could misguidedly reinforce racial stereotypes. Still, Earls, Raine and other researchers emphasize that biological factors, if they exist, are important only insofar as they protect individuals from—or make them vulnerable to—bad influences in their family, school and neighborhood. Research that excludes those who are most burdened by such pressures may be most expedient, but is it most useful?

STEPHEN SHAMES *Matrix*

Shown here (*left to right*) are boys in Omaha detained after a drive-by shoot-out, underage drinking in New York City and the aftermath of a shooting in Houston.

on Alcohol Abuse and Alcoholism. Although none has yet been shown to presage violence, "I believe the markers are there," he says. But he warns that "we're going to have to understand a whole lot more about the genetic, environmental and developmental origins of personality and psychiatric disease" before making use of the knowledge.

Yudofsky is less circumspect. "We are now on the verge of a revolution in genetic medicine," he asserts. "The future will be to understand the genetics of aggressive disorders and to identify those who have greater tendencies to become violent."

A Compelling Option

Few researchers believe genetics alone will ever yield reliable predictors of behavior as complex and multifarious as harmful aggression. Still, the notion that biologists and sociologists might together be able to assemble a complicated model that can scientifically pick out those who pose the greatest threat of vicious attack seems to be gaining currency. Already some well-respected behavioral scientists are advocating a medical approach to crime control based on screening, diagnostic prediction and treatment. "A future generation *will* reconceptualize nontrivial recidivistic crime as a disorder," Raine predicted in his 1993 book, *The Psychopathology of Crime.*

But the medical model of crime may be fraught with peril. When the "disease" is intolerable behavior that threatens society, will "treatment" necessarily be compulsory and indefinite? If, to reexamine Raine's hypothetical example, prediction models are judged reliable but "biological, social and cognitive intervention programs" are not, might eight-year-old boys be judged incorrigible before they have broken any law?

Calls for screening are now heard more often. "There are areas where we can begin to incorporate biological approaches," Fishbein argues. "Delinquents need to be individually assessed." Masters claims that "we now know enough about the serotonergic system so that if we see a kid doing poorly in school, we ought to look at his serotonin levels."

In his 1993 article Jeffery emphasized that "attention must focus on the 5 percent of the delinquent population who commit 50 percent of the offenses.... This effort must identify high-risk persons at an early age and place them in treatment programs *before* they have committed the 10 to 20 major felonies characteristic of the career criminal."

Yudofsky suggests a concrete method to do this: "You could ask parents whether they consider their infant high-strung or hyperactive. Then screen more closely by challenging the infants with provocative situations." When kids respond too aggressively, he suggests "you could do careful neurologic testing and train the family how not to goad and fight them. Teach the children nonviolent ways to reduce frustration. And when these things don't work, consider medical interventions, such as beta blockers, anticonvulsants or lithium.

"We haven't done this research, but I have no doubt that it would make an enormous impact and would be immediately cost-effective," Yudofsky continues. While he bemoans a lack of drugs designed specifically to treat aggression, he sees a tremendous "opportunity for the pharmaceutical industry," which he maintains is "finally getting interested."

But some worry that voluntary screening for the good of the child might lead to mandatory screening for the protection of society. "It is one thing to convict someone of an offense and compel them to do something. It is another thing to go to someone who has not done anything wrong and say, 'You look like a high risk, so you have to do this,' " Akers observes. "There is a very clear ethical difference, but that is a very thin line that people, especially politicians, might cross over."

Even compelling convicted criminals to undergo treatment raises thorny ethical issues. Today the standards for proving that an offender is so mentally ill that he poses a danger to himself or others and thus can be incarcerated indefinitely are quite high. The medical model of violent crime threatens to lower those standards substantially. Indeed, Jeffery argues that "if we are to follow the medical model, we must use neurological examinations in place of the insanity defense and the concept of guilt. Criminals must be placed in medical clinics, not prisons." Fishbein says she is "beginning to think that treatment should be mandatory. We don't ask offenders whether they want to be incarcerated or executed. They should remain in a secure facility until they can show without a doubt that they are self-controlled." And if no effective treatments are available? "They should be held indefinitely," she says.

Unraveling the mystery of human behavior, just like untangling the genetic code of human physiology, creates a moral imperative to use that knowledge. To ignore it—to imprison without treatment those whom we define as sick for the behavioral symptoms of their illness—is morally indefensible. But to replace a fixed term of punishment set by the conscience of a society with forced therapy based on the judgment of scientific experts is to invite even greater injustice.

FURTHER READING

THE PSYCHOPATHOLOGY OF CRIME. Adrian Raine. Academic Press, 1993.

UNDERSTANDING AND PREVENTING VIOLENCE. Edited by A. J. Reiss, Jr., and J. A. Roth. National Academy Press, 1993.

WHAT WORKS IN REDUCING ADOLESCENT VIOLENCE. Patrick Tolan and Nancy Guerra. Available from the Center for the Study and Prevention of Violence, University of Colorado, 1994.

Crime statistics and violence prevention program information are available at gopher://justice2.usdoj.gov:70/1/ojp on the World Wide Web.

Turning the Tables Around

Up on the 67th floor of Ruff Towers, two employees from We-Haulit-4U Moving dragged the last of nine tables into a storeroom. The door clicked shut.

"Done," said Dan, breathing heavily. "Final check, then I'll buy us lunch at the Plushy Pink Pizza Palace. Okay, two oak square tables, six rectangular tables—four pine and two formica—and one giant antique mahogany square."

"Right," said Max, crossing off each item on his clipboard. "Say, it's a bit crowded in here."

"Jam-packed. Wall-to-wall tables, except where we're standing."

"We were lucky to fit them in so well."

"Sure were," said Dan, glancing around the room. "Oh, no. Look! There's a leak in the ceiling—dripping right onto the antique. We'll have to move it. It will be ruined if we don't."

"This is not going to be easy," Max observed.

"Can't we pile it on top of the other tables?"

"Not a chance. The ceiling is too low."

"We'd be okay provided we could move the two formica tables under the leak and the antique table to the opposite corner, just in case the leak spreads [*see diagram on this page*]. We can slide the tables into the space that's left, one by one, thereby making new spaces to slide more of them into."

"Won't we get trapped?"

"No. We can crawl underneath the tables," Max said.

Dan stooped to peer under a table. "You're right—there's plenty of room."

Thirty minutes later they had successfully moved the antique table to the middle of the right-hand wall, more or less out of harm's way, but now one of the pine tables was getting wet, and the door was blocked [*see top illustration on page 109*].

"What we need," Max mused, "is a map."

"Max, we can see where the tables are."

"Not a map of the room. A map of the puzzle."

Dan stared at him. "Have you gone crazy? Puzzles don't have maps."

"I hate to contradict you, but puzzles do have conceptual maps, imaginary maps in the brain. Maps that show you all the positions in the puzzle and how to get from one to another."

Dan nodded. "It's going to be a pretty complicated map, Max. There are an awful lot of positions—and moves."

"True. So we'd better find some way to break the problem into simpler pieces. Hey! That's it. First of all, let's find out what we can do easily. Then we can string those together somehow."

"Well, if you've got a square hole with just the two smallest tables in it, you can move them around pretty freely," Dan said.

"That's the idea. A sort of subpuzzle, where you move only a few tables inside some well-defined area [*see bottom illustration on page 109*]. Hmm. There's one, a bit more complicated, where you have a rectangular region containing only two rectangular tables and two square ones."

"So you could assume that the positions that differ from one another by shuffling tables around inside one of these subpuzzles are effectively the same," Dan said. "That must cut the list of positions down quite a bit."

"Yeah. And sometimes there's only one sensible way to continue moving the tables, if you don't want to undo what you've already done."

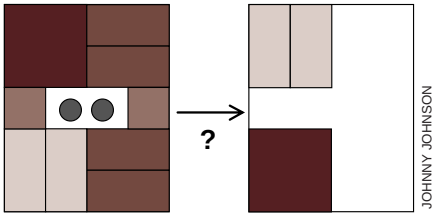
"So provided you know where you're starting from and where you're trying to go, sequences like that can be left off the map?"

"Precisely. Hand me that clipboard." In no time, Max had drawn a map showing some of the possible positions and moves [*see upper illustration on page 110*].

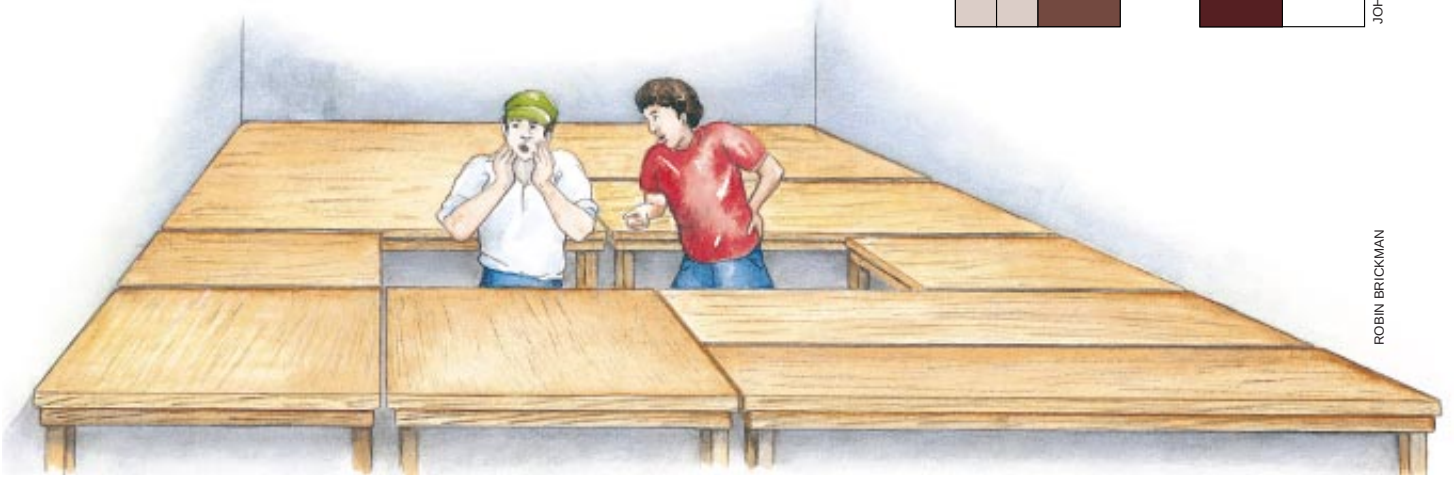
"I've marked the start and the finish positions," Max said. "Then there are six different ways to place key tables, labeled A, B, C, D, E and F."

"I'd have expected more than six."

"There are more. This is just part of the map. But six are enough to solve the puzzle. Now shut up and listen.

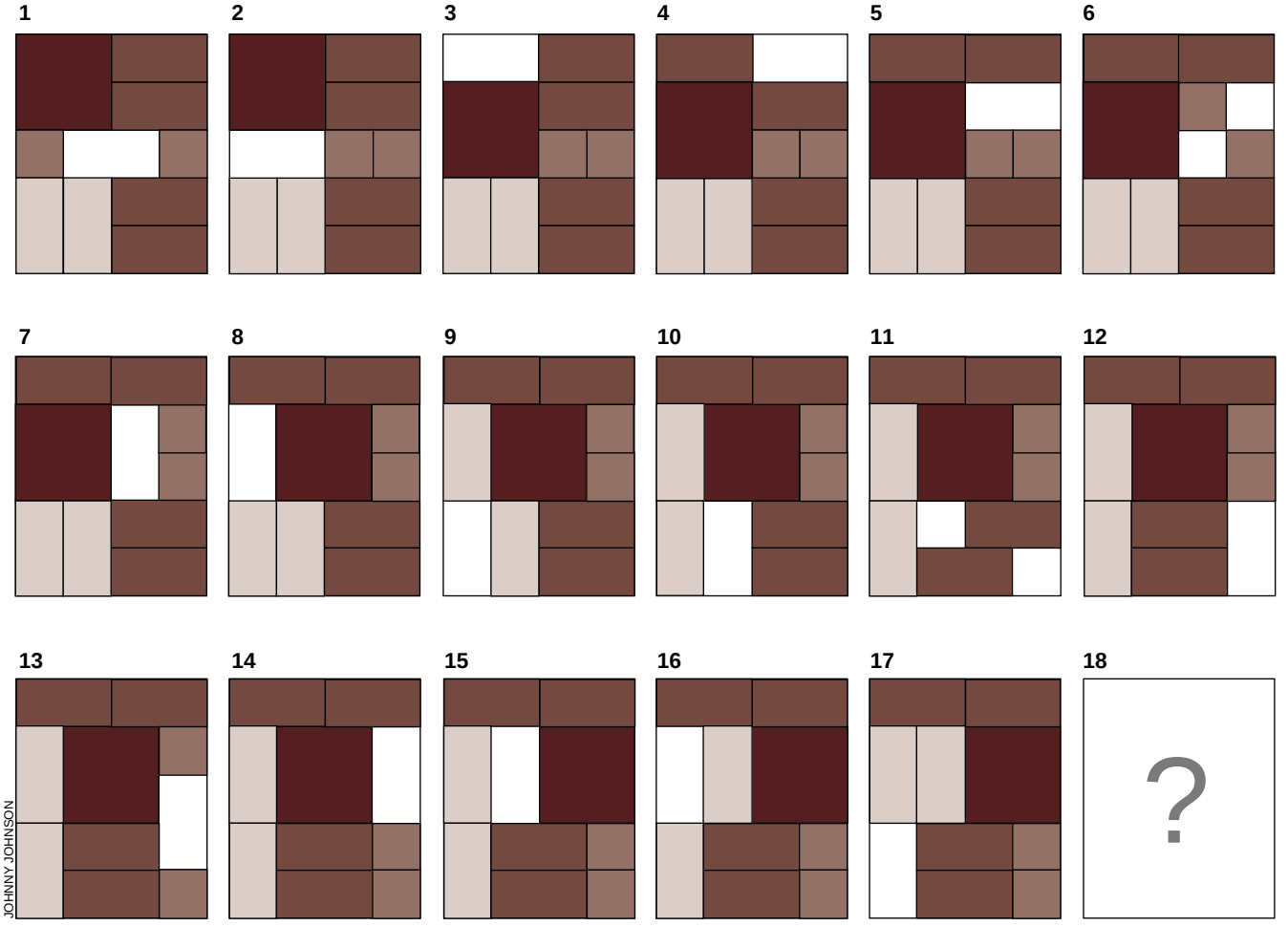


JOHNNY JOHNSON



ROBIN BRICKMAN

***MOVERS** Dan and Max, shown as two circles on the diagram, are surrounded by nine tables. To safeguard the seven wood tables against water damage, they must move the tables until the two formica ones fit in the top left corner under the leak.*



The lines show sequences of forced moves—in the sense that if you know where to start and where to finish, the moves in between are fairly obvious because there's really only one choice you can make at each step, right?"

"Okay, I see that."

"Good. I've shaded in rectangular regions where there's a subpuzzle to solve. To show which one, I've drawn little pictures of the start and finish positions within the rectangle, at the appropriate ends of the connecting lines."

Dan's mouth opened like a goldfish's. "Sorry, I don't quite follow."

"Well, suppose you want to know how to move from C to E. Look at the horizontal line that joins them; it passes through two little diagrams. If you replace the shaded area in C with the left diagram and the shaded area in E with the right one, that will give you the start and finish positions. Because the moves in between are 'forced,' it doesn't take very long to work them out. If you make a copy of the puzzle using bits of card, you can move them around and check."

"What does 'DEAD END' mean?"

"What do you think? Now, what does the map tell us?"

"Where things are and how to go from one to another. Well, clues to those things."

"It tells us even more than that. It tells us that one way to solve the puzzle is to go along the route START-C-A-B-FINISH."

Dan's face lit up in admiration. "You could go START-C-D-B-FINISH instead?"

"Sure. Or even START-C-E-F-D-B-FINISH—but that would be an unnecessarily complicated route."

Dan was getting into it now. "Or START-C-D-F-E-C-D-B-A-B-D-C-E—"

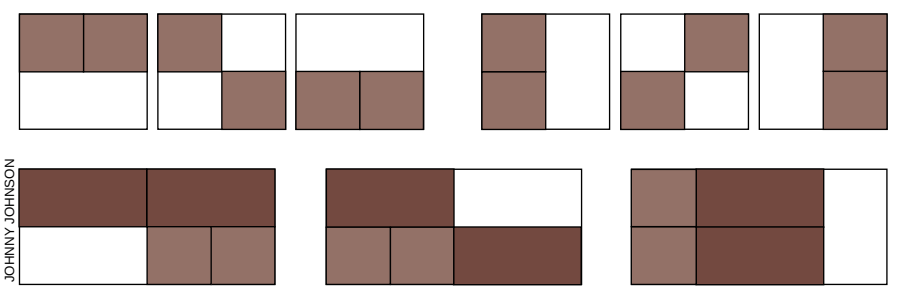
"Yes," Max interrupted, "but that would be an even more unnecessarily complicated route."

***SEQUENCE OF MOVES** illustrates how Dan and Max began rearranging the tables and managed to slide the large square antique over to the right wall.*

"I'll settle for the simplest one."

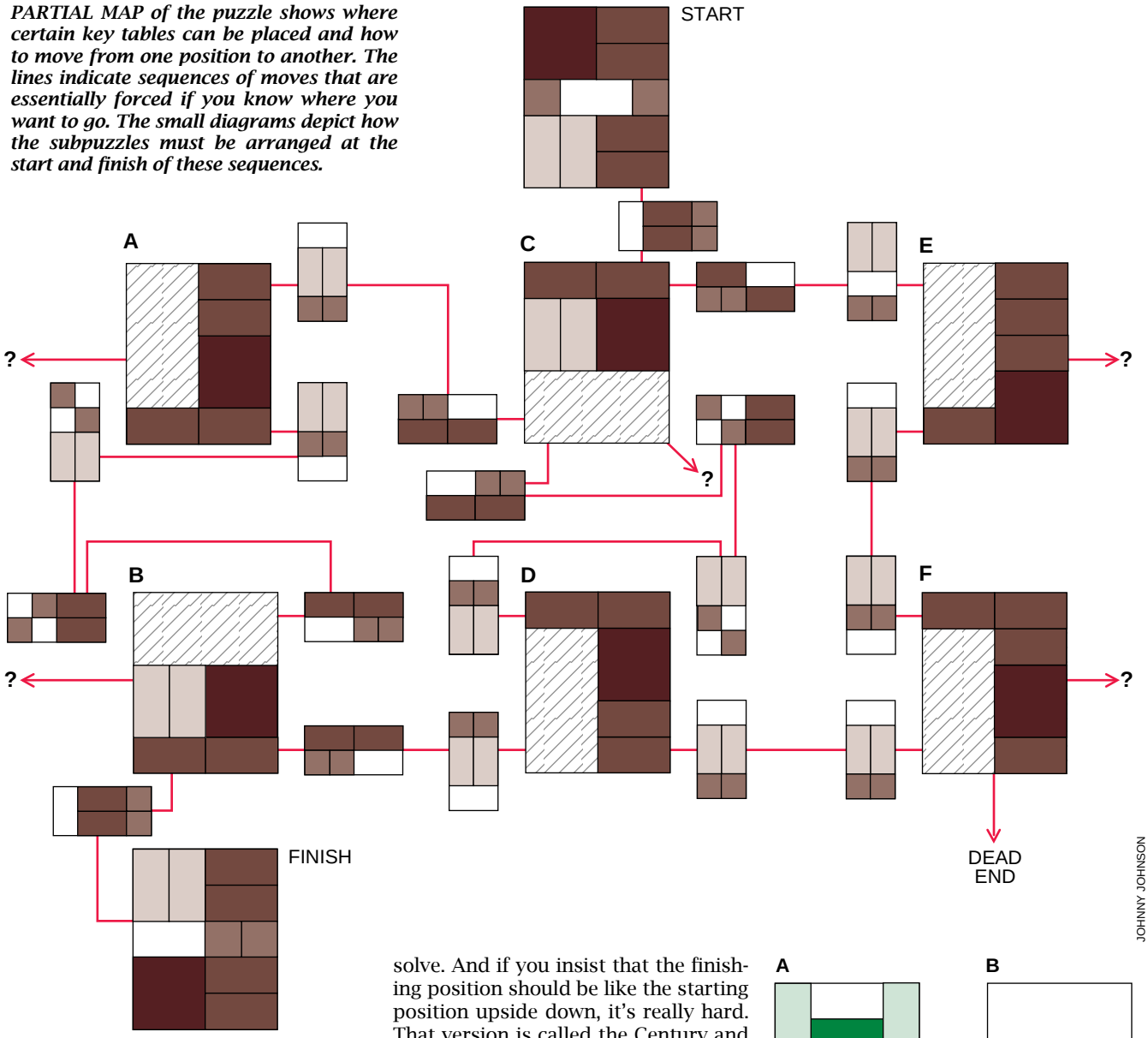
"Fine by me. Let's get these tables moving!"

In a little while they had moved the tables so that the antique sat far from the leak, the two formica tables were under the leak, and the door stood free and clear. Having missed both lunch and dinner, Dan and Max were starved. They raced to the first floor and set off for the Plushy Pink Pizza Palace, which



***SUBPUZZLES** provide some useful maneuvers. In each, the tables can easily be rearranged without going outside the marked boundaries.*

PARTIAL MAP of the puzzle shows where certain key tables can be placed and how to move from one position to another. The lines indicate sequences of moves that are essentially forced if you know where you want to go. The small diagrams depict how the subpuzzles must be arranged at the start and finish of these sequences.



was open all night.

"You know," Dan said, "that wasn't so hard."

"Not once we worked out that map. But we were lucky. It was a simple one."

"That's because you used some tricks to simplify it."

"The tricks help, but there are plenty of sliding block puzzles with far more complicated maps, even when you use every trick you can think of." [For additional puzzles, see *Winning Ways,* Vol. 2: *Games in Particular,* by Elwyn R. Berlekamp, John H. Conway and Richard K. Guy. Academic Press, 1982.]

"Like what?"

"Well, there's one called the Donkey Puzzle, probably from 19th-century France. That's a good bit harder. The Century Puzzle, invented around 1980, is harder still. It takes 100 moves to solve. And if you insist that the finishing position should be like the starting position upside down, it's really hard. That version is called the Century and a Half Puzzle because it requires 151 moves."

Max ordered a deep-pan pizza with extra cheese. Dan ordered a special with a lot of extra toppings—pepperoni, tuna, capers, pineapple, hot tamales, a whole banana, chewing gum, licorice and a lighted sparkler.

When the pizzas arrived, Dan's didn't look quite right. Most of the ingredients were upside down, including the crust. The waitress had included a whole tuna and set the licorice on fire.

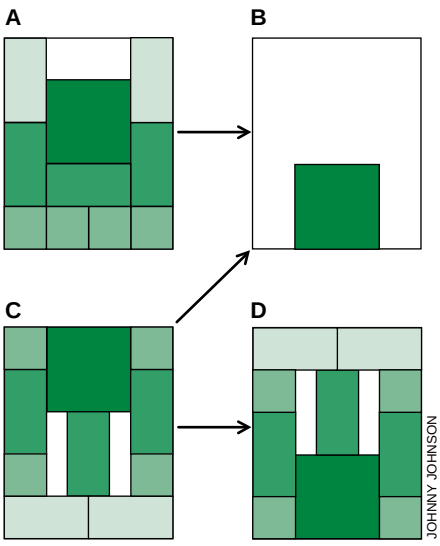
"Enjoy your puzzle, sir," she said over her shoulder.

"Send it back," Max suggested.

"No, no, you heard what she said. I can't resist a challenge." He straightened his back, squared his shoulders and reached for the clipboard.

"What are you doing?" Max asked.

"Just wait until I've made a map of this pizza."



DONKEY PUZZLE involves moving the blocks from the position shown in diagram A to that in diagram B. The Century Puzzle requires 100 moves to reconfigure the blocks from C to B. The Century and a Half Puzzle, from C to D, takes 151 moves.

BOOK REVIEWS by Philip Morrison

The Big Picture

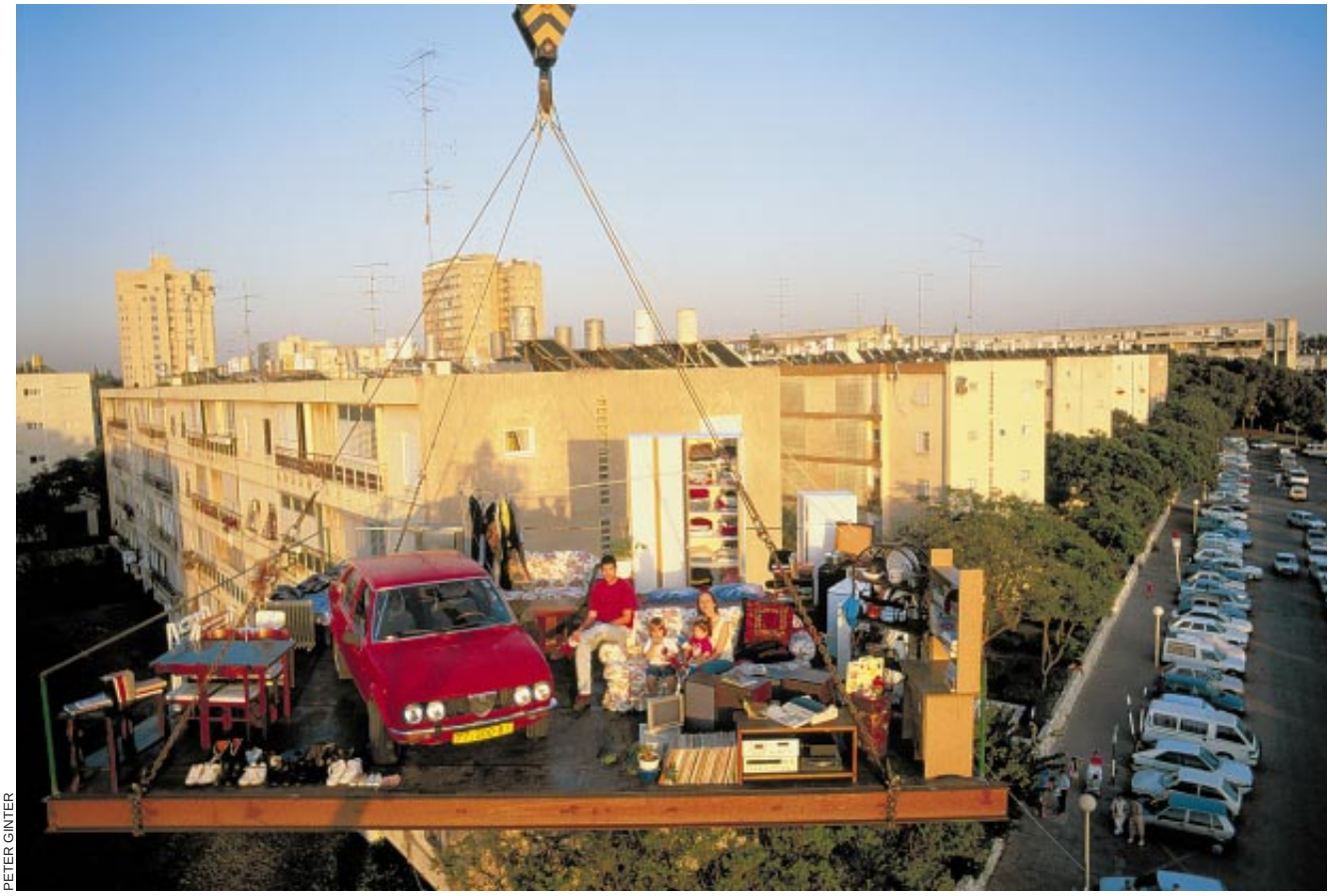
MATERIAL WORLD: A GLOBAL FAMILY PORTRAIT, by Peter Menzel. Text by Charles C. Mann. Sierra Club Books, distributed by Random House ($30). CD-ROM from StarPress Multimedia (1-800-782-7944; $59.95).

Our world now counts about a billion households. This remarkable book, well supported by the United Nations, vividly presents its own cross section of families today, with the energy and intimacy of a work of art for our times, choosing just one family from each of 30 lands. It carries a quiet challenge: it is the prosperous minority in this world who can most easily move toward an equitable and stable future.

Californian Peter Menzel, creator of the work, is a globe-trotting photojournalist of wide experience and acclaim. He selected the final list of countries by modifying the experts' cool economic and social criteria to emphasize interesting and timely visits—those "we can learn from and that I wanted to see." Finally, he and a respected local companion knocked on doors of typical houses in typical neighborhoods to seek one cooperative family. Each family had to be willing and able to take part in the Big Picture, one that would display the whole family outside their home surrounded by their material goods, moved out for the day to allow the camera visual access. In crowded Tel Aviv the enterprising photographer hired a construction crane to lift high a platform that bore the entire contents of the small apartment with its tenants, their red Alfa Romeo, computer, menorah and all, held in front of their large apartment house for picture taking.

Not everyone wants to put such a show before the neighbors. But it worked out wonderfully. We even get a full list of those holdings that were not moved into camera range: many pigs, chickens, fish ponds, rice storage bins, laden freezers, groves of orange and banana trees, and "two acrobatic airplanes." The Big Picture, a documented day of hard work for all, was preceded by a few days of candid snapshots—2,000 rolls total—and lots of video as well. A long list of questions about daily living and cherished values was put to each family as the visitors and hosts came to know one another. These friendly visits took place between late 1992 and early 1994.

Whom will you meet? Rice farming is the single occupation most represented, just as the world works today. In northern India the young parents and four kids are seen on the family string bed next to three heavy bags from the last rice harvest. All they had was in the picture, and that is not much. Their windowless house of stucco or plaster measures 12 feet by 30. They named their most valued possessions: four or



PETER GINTER

ISRAELI FAMILY poses outside their apartment (end unit on top floor), surrounded by their worldly possessions.

five prints and figurines of gods and goddesses; their fondest hopes were for a cow or two. One of the most endearing spreads here shows the wife's feet among those of her friends, all being painted with the traditional red dyes and newer nail polish that would proudly set off their silver anklets and toe rings.

In Haiti the peasant family we meet is poorer still, growing some sugarcane and millet to sell. Their breakfast consisted of potatoes, smoked fish, coffee; potatoes alone make a typical lunch; for a typical dinner—nothing at all. A toy VW Bug, its four wheels missing, adds a touching note. In January 1994 fear of armed authority was real in Haiti; the photographer, too, "got an adrenaline rush" from the militia search at rifle point on his way back to the city. A Thai rice-farming household is by comparison rich, its appliances many in their two-story frame house with its four windowed rooms, two water buffalo in front of the paddy, and the motor scooter most valued among their goods.

We meet about 10 diverse families from the developed world, among them those of an Icelandic airline pilot, a cable splicer near Houston, a salaried Japanese warehouseman and an affluent and worried professor of political science at Kuwait University, whose four cars and a sofa 45 feet long appear in his family's Big Picture.

Three instructive color spreads offer a dozen or two color photographs each, to sample the current televisions, meals and toilets of the world. Most of these families are in daily touch with outside events; only four out of the 30 own neither a radio nor a television. The work week is a long one everywhere outside the developed world; even in most developed countries the toil of women remains long.

This is a record of striking value, in particular for the schools. It should appear anew every 20 years. No time traveler could bring back a more valuable souvenir!

Grace under Pressure

THE NEW ALCHEMISTS: BREAKING THROUGH THE BARRIERS OF HIGH PRESSURE, by Robert M. Hazen. Times Books, 1993 ($23).

Wherever you live on the earth there is a region nearby where diamonds abound—but it is out of sight 100 miles underground. The geologists have inferred this from indirect evidence, although we have never seen diamonds form in nature. They sit stably in the hot depths for several billion years, until some local crustal weakness admits a sudden explosion.

The high-pressure material cools as it expands violently upward, to punch out a narrow shaft or pipe and to exit with fearful force, spalling a funnel-shaped cavity in the surface rocks. The overburden is flung high, to fall back as a jumbled mass of debris across the diamond-bearing plug. Such is the celebrated site at Kimberley, South Africa, the first understood among diamond pipes.

The rocky outrush must be explosively rapid, for hot diamonds cannot survive long at reduced pressure; they turn within a minute to stable graphite. The geologists have found within some pipes the telltale crystal shapes of diamond in "incredibly rich" masses, now gone to graphite. The natural diamonds we know are the rare survivors that reached the surface before first reverting. Only cooling quenches the atomic rearrangement.

Diamonds in your hand are metastable; they may endure for a long time in human terms, but not forever. At white-hot temperatures under pressures of 50,000 or 75,000 atmospheres, it is the graphite that is metastable and soon turns to small diamond crystals, most easily if dissolved in iron or nickel. "Kimberlite taught scientists what they had to do to make diamonds:... squeeze carbon and heat it unmercifully.... Those were tricks worth learning."

Not all the tricks were taught by nature. This book is a chronicle, often at firsthand, of the livelier tricks of diamond making, now become a multibillion-dollar industry, indispensable in metalworking and in drilling for oil, not to mention one-hour eyeglass production or thermonuclear bombs. Here is a lively tangle of the physics of materials, its experimental ingenuity and theoretical insight, with lots of intrigue—personal, corporate and governmental.

Even the circumstances of the first true synthesis of diamond are still not beyond doubt: it might or might not have been done and recognized in Stockholm in 1953, although the claim made for Schenectady late in 1954 has its strengths. The first patents for diamond synthesis were certainly issued to the General Electric Company, although delayed from publication for a few years by U.S. government classification as strategic. Once classification was lifted in 1959, GE lawyers filed patent papers across the globe.

One week later De Beers, Pretoria sovereign over the diamond trade, submitted full diamond-making patents of its own. How much was leakage, how much convergence, is still clouded and still fascinating. The efforts at patent

breaking led to a decade of subsequent combat between the well-documented GE defenders and the ingenious assailants mustered by De Beers. Would you believe the audacious witness George Kennedy, a U.C.L.A. diamond physicist known both for his originality and his precision? He was versatile enough to have lectured thrice in one single day at Oxford—on high-pressure physics, on new exotic orchids and on the dating of pre-Columbian pottery. On his own off-campus time, the nonpareil Kennedy married a Hollywood heiress and scandalously burned down her house while smoking in bed, but he soon made amends, finding them a magnificent joint residence in the Santa Monica Mountains.

By now patent licenses abound, and diamonds grow for GE, De Beers and many others. GE operates its scores of massive 1,000-ton presses around the clock in its plant in the suburbs of Columbus, Ohio, where the well-run "pressure cookers" synthesize some 30 tons of tiny diamonds a year. Smaller, newer presses, in which six cleverly aligned tungsten carbide anvils press hard on the faces of a cube-shaped sample chamber, make a number of larger diamonds, fully competitive in smaller niches of the trade.

Synthetic diamonds appear now at about 100 tons a year, most for industry. The big mines in three continents yield about 10 tons a year, half for the gem trade. Meanwhile Du Pont makes diamond abrasives in noisy caverns by passing high-explosive shocks through graphite-copper mixes. Whole filmy layers of diamond can be deposited from vapor, as the winds from carbon stars made tiny diamond plates that drifted into the early solar system. More recently diamonds have been made from the new molecular buckyballs (C_{60}) well squeezed at room temperature.

But the most unexpected innovation is not diamond synthesis. It is the diamond-anvil pressure cell developed in 1959 by Alvin Van Valkenburg and Charles Weir of the National Bureau of Standards. (An independent Chicago team shares credit.) "An elaborate nutcracker" of two gem diamonds, the little device reproduces the pressures of planetary depths right there on the desktop under the experimenter's hand. Turn a small crank to wind a sturdy screw that presses together two flat diamond faces (nothing is stronger or more transparent) that squeeze between them a soft-metal gasket ring. High pressure appears on whatever is placed in the ring. Ten or 20 pounds of push can bring a sample volume that is only one tenth of a millimeter across up to pressures as high as 100,000 atmospheres. There is still ample room for pinhead samples and probing photons in all variety. It is the high *density* of energy that makes high pressure, even when the total energy is minor.

Faraday could have done this; our author thinks even Isaac Newton might have. Heating is available now by careful design, and x-rays, laser light and meticulous spectroscopy are welcome. Special diamond shaping distributes stress over curved faces and can push the pressure up 10-fold: such desktop diamond vises now outperform all the heavy steel and carbide presses in the world, be they five stories high, as is one in Moscow.

Of course, the little cells cannot put out much new material; their product is knowledge. Many minerals expected in the earth's interior are now opened to pressure probing. It looks as if the earth's mantle might converge to one optimal crystal form, whatever its mix of materials. That old grail of the high-pressure lab, hydrogen atoms squeezed to merge into an electron-sharing metal, is also strongly hinted at but still eludes. Plenty of novelty is ahead at atomic energy densities, even though the far grander energy densities between colliding protons will not be studied in the echoing tunnels of Texas.

Local Fluff and Other Stuff

THE GUIDE TO THE GALAXY, by Nigel Henbest and Heather Couper. Cambridge University Press, 1994 ($49.95; paperbound, $24.95).

It is galaxies that shine out from the gravitational potholes in our cosmos. We dwell right inside one showy specimen so that surely we are obliged to map our own backyard. This book is a comprehensive tour guide, with many square, colorful pages (although often the color is computer-contrived, as conventional as the red that once marked maps of the British Empire). Dust is both pierced and disclosed by powerful infrared arrays, x-rays single out dilute pools of fiercely hot gas and the radio dishes analyze both for cold molecules and for magnetized relativistic plasma. Even the coded colors are packed with meaning.

The chapters open with a useful history of what we know. The black river of the Milky Way that dominates a medieval Korean map of the visible sky is a splendid icon. Like Mount Everest, the Milky Way is there. (Its current loss to the eye within the wasteful skyglow of our cities is a major deprivation, com-

parable with and more widespread than the scarceness of the blue whale.)

Our knowing guides begin the tour in the suburbs of the Milky Way. We gain perspective from the galaxies that comprise its near neighbors, called the Local Group. Our own view of the star forest is lost in trees. By such comparisons we came to expect and to find how the stars circle the carousel of the disk (turning clockwise as viewed from the north), locally bobbing and even weaving under complicated Newtonian pulls.

Then we map the sights in three arms of this, our spiral. The Local Arm (sometimes called after Orion, where a nearby starburst glows) lies between the Perseus Arm, just outward, and the arm next inward, named for Sagittarius. It is in the Perseus Arm, for instance, that we find the Crab Nebula and the brightest radio source, Cassiopeia A, two relics of very recent supernova explosions. Our own Arm might as well have been named for any bright constellation, since all the bright nearby stars are within it. Hundreds of important landmarks—skymarks?—that bear familiar names (the Hyades, the Jewel Box, Kepler's supernova) are colorfully mapped and indexed at differing scales. It becomes pretty clear how easily we oversimplify the first maps we make. Ages and motions allow inferences that no static map can make clear, especially because without a deeply physical study we view the remote sky only in two and not in all four dimensions of space-time.

Magnify two small regions of special interest. First, inspect our precincts on the small scale of 50 light-years. We reside within a gas cloud of modest hydrogen density, dubbed the Local Fluff, not extending out as far as Sirius or Arcturus. Radio astronomy maps its physical bounds. But that fluff is itself within a hotter, more dilute gas bubble a couple of hundred light-years in diameter. The implied sequence is quite general; newly formed hot stars do blow bubbles in the same hydrogenous medium from which they formed, and those blowing winds, sometimes the shock waves of full supernova explosions, then compress the gas they reach, often to initiate star formation anew.

The many infrared images here are essential, but they so lack the crispness of the optical photographs that they impelled this viewer over and over again to wipe his spectacles. This is no fault in the wonderful observations. We badly miss a page or so to explain how these infrared images are made and their current limitations. Still, any readers who like to see and to know just where they are will enjoy this textually and visually nourishing volume by two deft writers.

ESSAY by Steven Shackley

Relics, Rights and Regulations

What a wonderful idea it seems, especially in these politically correct days: force museums and universities to return sacred relics and excavated burials that scientists stole from Native Americans and Hawaiians years ago. In November 1990 that logic led the U.S. Congress and President George Bush to enact a new law, the Native Americans Graves Protection and Repatriation Act (NAGPRA).

Most institutions openly embrace the need to repatriate Native American artifacts, but NAGPRA represents a quintessential example of good intentions gone badly awry. In typical fashion, the lawmakers gave little forethought to the social and economic consequences of the act, not only for museums but also for tribal organizations, which must now sift through reams of inventory lists describing objects and skeletal remains that may or may not be associated with their ancestors. The law is straining museums' meager resources and may foster legal battles that will do little for scientists or for native groups trying to rebuild their ethnic heritage.

Under NAGPRA, institutions have had to prepare complete, lengthy lists of holdings that might properly belong to Native American groups. The deadline for the summaries of this material has already passed; full inventories of archaeological artifacts are due in November 1995. But the Department of the Interior, which oversees the enforcement of NAGPRA, has yet to generate the federally mandated guidelines to aid the short-staffed museums in their tasks. Recently Congress allocated $2.3 million to assist both museums and native groups in implementing parts of the law. That sum could not fund the work at my museum, let alone the hundreds of affected institutions around the country.

In the meantime some members of the NAGPRA Review Committee, an organization of Native Americans and nonnative "professionals," have been interpreting the law to suit their ends. Several sections of the original text of the law state that it applies only to those "tribes" that have received federal recognition through due process. The review committee now suggests that groups that are *not* federally recognized should be consulted, too. The effect is to compel museums and universities to serve as intermediaries between the federal government and unofficial native groups that are attempting to use claims filed under NAGPRA as a way to gain federal recognition. The goal, of course, is to get the attendant rights and funding. NAGPRA has turned into an unintended piece of civil rights legislation, one that may require Congress to grant new entitlements even as the old ones become prime targets for cutting.

I do not mean to say that all the law's effects have been bad. Museums and universities are now communicating with native groups to an extent unheard of a few years ago. In the course of negotiations, many of those groups are realizing that they have no proper way to catalogue, store and preserve the sacred artifacts and burial remains covered by NAGPRA. In addition, such groups often no longer practice the original rituals in which the items were used, or if they do, they may produce objects more useful than the old, fragile relics. Native Americans and scientists are thus finding that their interests often dovetail.

One example of the new spirit of collaboration is taking place at the Santa Barbara campus of the University of California. Phillip L. Walker, a professor there and a member of the NAGPRA Review Committee, entered into an agreement with the Chumash tribe to create an on-campus ossuary that is controlled by the Chumash but open to Walker and to the university. This particular solution was possible in part because of the unambiguous history of the Chumash.

There is good archaeological and historic evidence that the modern Chumash are direct descendants of the natives whose remains have been excavated on the California mainland and nearby Channel Islands. It therefore seemed sensible to grant the Chumash control over most of the artifacts found in this region. As long as the current Chumash group continues in power, the agreement worked out by Walker and the university will stand. If, as is often the case in politics, the next Chumash leadership disagrees with its predecessor's decisions, Walker could lose access to the collection.

In much of North America, working out such curatorial compromises is complicated by solid evidence that modern tribes are not directly descended from the prehistoric inhabitants of the same area. Many groups, particularly early hunter-gatherers, moved frequently among many regions. The federally recognized Native Americans living in an area today may have no more lineal connection to the makers of the local archaeological remains than do the excavators themselves. One solution favored by certain native groups is to give the remains to the local tribe regardless of any empirical connection between modern and ancient dwellers. Some Native Americans—and some scientists, including me—oppose such action on ritual and rational grounds. Allowing one set of people to dictate the treatment of relics that belong to another runs directly contrary to the ideal that motivated NAGPRA in the first place.

I have deliberately not argued for the scientific value of keeping collections by universities; this topic has been argued to death. It is only fair that native groups should be able to regain possession of artifacts that legitimately belong to them. But it is worth noting that genetic research related to skeletal remains has reached a critical threshold; future work would have great import for science and for native groups alike. Archaeological research into the geography of tribal territories, patterns of migration and interactions between groups could actually aid Native Americans in their land claims.

The greater problem here is the bureaucracy in charge of the process. At a 1992 meeting of the American Anthropological Association, C. Timothy McKeown of the Department of the Interior stated that he would feel the department had done its job if all parties were dissatisfied. No wonder the electorate is so angry at government. Universities, museums and native groups appear, despite the many obstacles, to be finding resolutions to the problem of repatriation. In the end, I am sure the NAGPRA Review Committee and the Department of the Interior will take credit for the beneficial solutions that result.

STEVEN SHACKLEY is an assistant research archaeologist at the P. A. Hearst Museum of Anthropology at the University of California, Berkeley.