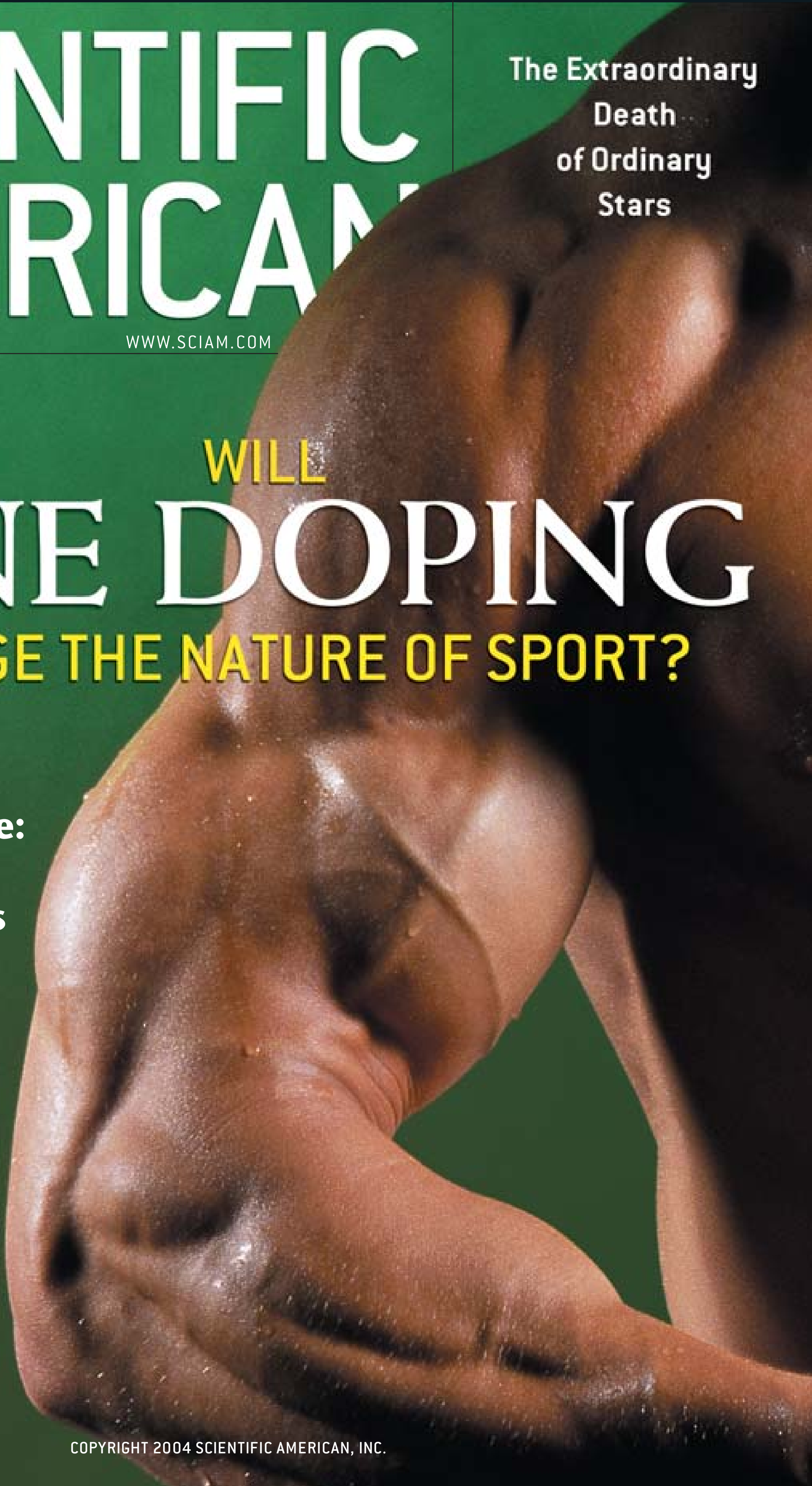
A 400-YEAR-OLD HOAX? • $1-MILLION PROOF FOR THE SHAPE OF SPACE

SCIENTIFIC AMERICAN

The Extraordinary Death of Ordinary Stars

JULY 2004 WWW.SCIAM.COM



WILL GENE DOPING CHANGE THE NATURE OF SPORT?

Mad Cow Disease: Faster Tests, Future Therapies

When Methane Ruled Climate

Nanosensors Based on Magnetic Effect

july 2004

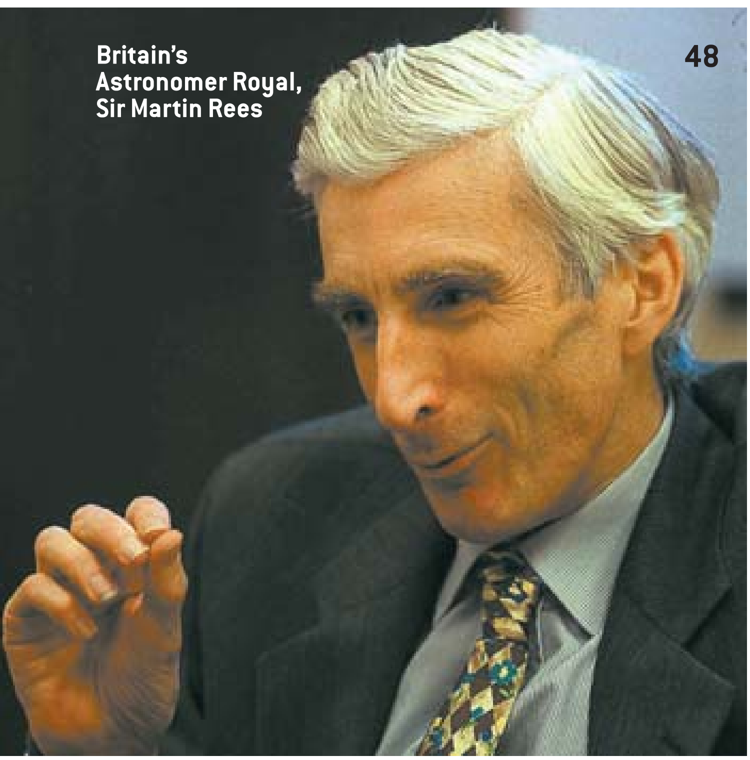
contents

SCIENTIFIC AMERICAN Volume 291 Number 1

features

SCIENTIFIC AMERICAN Volume 291 Number 1

departments



Britain's Astronomer Royal, Sir Martin Rees

48

columns

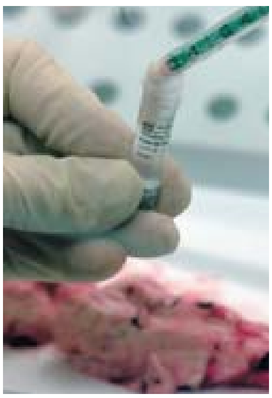
Cover photograph by Pete Saloutos, Corbis.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 2004 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40012504. Canadian BN No. 127387652RT; QST No. Q1015332537. Publication Mail Agreement #40012504. Return undeliverable mail to Scientific American, P.O. Box 819, Stn Main, Markham, ON L3P 8A2. Subscription rates: one year $34.97, Canada $49 USD, International $55 USD. 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; (212) 451-8877; fax: (212) 355-0408 or send e-mail to sacust@sciam.com **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.** Printed in U.S.A.

SA Perspectives

Testing Madness

Few ailments sound scarier than mad cow disease and its human counterparts. They incubate silently for years, slowly eating the brain away and leaving it full of holes. So it's not surprising that many people want the U.S. Department of Agriculture to test all cattle for the illness, formally called bovine spongiform encephalopathy (BSE). Certainly testing all 35 million cattle slaughtered annually would reopen trade with Japan, which has refused American beef since the discovery of a mad cow in Washington State last December. It might prevent BSE-free countries from dominating the export market. And consumers might simply feel better about their steaks, roasts and burgers. Too bad there's not much science to back up the proposal.



CATTLE BRAINS get tested for prions that cause BSE.

Commercial "rapid tests" are not designed to detect the disease reliably in most slaughtered bovines. They work best on those that have lived long enough to build up in their brains a detectable amount of prions, the proteins at the root of BSE. Typically those animals are older than 30 months or have symptoms, such as an inability to stand (called downer cattle).

Most U.S. bovines, however, reach slaughter weight before 24 months of age—before the tests can accurately detect incubating BSE. Most European countries recognize those limitations and target cattle 30 months and older. But using current kits on all slaughtered animals, at least 80 percent of which are younger than 30 months, may give misleading assurance about the safety of beef.

Do economic and emotional reasons justify that strategy? Testing costs about $25 to $35 per head, amounting to just a few extra pennies per pound. But in total, the "beef tax" would cost around $1 billion annually—for results that are equivocal.

When it comes to keeping consumers safe from prions, we can think of better uses for $1 billion. Like Europe, the U.S. should test cattle older than 30 months. Stricter and more complete enforcement of existing rules is even more critical. The USDA is supposed to check at least 200,000 cattle this year—what probably amounts to the bulk of U.S. downers, the category most likely to test positive. Yet reports of sloppiness have emerged. The most shocking occurred in Texas, where a downer somehow managed to avoid being tested after it was pulled by an inspector. The USDA's management, top-heavy with former beef officials, needs to take a more critical view of its relationship with the industry.

Also lost in the discussion is the surveillance of human prion diseases. Last year only about two thirds of all suspected human cases reached the national prion disease surveillance center at Case Western Reserve University, where brain postmortems are conducted. These examinations provide the evidence as to whether people are dying from prion infections—be they from mad cows or from deer and elk with chronic wasting disease. Additionally, they would help determine whether purported illness clusters, such as one tied to the now demolished Garden State Racetrack in New Jersey, have truly arisen from a common source.

Better assays are coming [see "Detecting Mad Cow Disease," by Stanley B. Prusiner, on page 86]. They hold promise for detecting prions in young cattle and in cow parts not previously found to be infectious. They may also prove effective in uncovering new prion maladies and in testing live humans. Only when such assays become validated will it make sense to target all cattle. Right now other measures rank higher.

THE EDITORS editors@sciam.com

REUTERS/CORBIS

How to Contact Us

EDITORIAL

For Letters to the Editors:
Letters to the Editors
Scientific American
415 Madison Ave.
New York, NY 10017-1111
or
editors@sciam.com
Please include your name and mailing address, and cite the article and the issue in which it appeared. Letters may be edited for length and clarity. We regret that we cannot answer all correspondence.

For general inquiries:
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-754-0550
fax: 212-755-1976
or
editors@sciam.com

SUBSCRIPTIONS

For new subscriptions, renewals, gifts, payments, and changes of address:
U.S. and Canada
800-333-1199
Outside North America
515-247-7631
or
www.sciam.com
or
Scientific American
Box 3187
Harlan, IA 51537

REPRINTS

To order reprints of articles:
Reprint Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8877
fax: 212-355-0408
reprints@sciam.com

PERMISSIONS

For permission to copy or reuse material from *SA*:
www.sciam.com/permissions
or
212-451-8546 for procedures
or
Permissions Department
Scientific American
415 Madison Ave.
New York, NY 10017-1111
Please allow three to six weeks for processing.

ADVERTISING

www.sciam.com has electronic contact information for sales representatives of Scientific American in all regions of the U.S. and in other countries.

New York
Scientific American
415 Madison Ave.
New York, NY 10017-1111
212-451-8893
fax: 212-754-1138

Los Angeles
310-234-2699
fax: 310-234-2670

San Francisco
415-403-9030
fax: 415-403-9033

Detroit
Karen Teegarden & Associates
248-642-1773
fax: 248-642-6138

Midwest
Derr Media Group
847-615-1921
fax: 847-735-1457

Southeast
Publicitas North America, Inc.
404-262-9218
fax: 404-262-3746

Southwest
Publicitas North America, Inc.
972-386-6186
fax: 972-233-9819

Direct Response
A&A Media Options, LLC
203-267-4251
fax: 203-267-1552

Belgium
Publicitas Media S.A.
+32-(0)2-639-8420
fax: +32-(0)2-639-8430

Canada
Derr Media Group
847-615-1921
fax: 847-735-1457

France and Switzerland
PEM-PEMA
+33-1-46-37-2117
fax: +33-1-47-38-6329

Germany
Publicitas Germany GmbH
+49-211-862-092-0
fax: +49-211-862-092-21

Hong Kong
Hutton Media Limited
+852-2528-9135
fax: +852-2528-9281

India
Convergence Media
+91-22-2414-4808
fax: +91-22-2414-5594

Japan
Pacific Business, Inc.
+813-3661-6138
fax: +813-3661-6139

Korea
Biscom, Inc.
+822-739-7840
fax: +822-732-3662

Middle East
Peter Smith Media & Marketing
+44-140-484-1321
fax: +44-140-484-1320

Sweden
Publicitas Nordic AB
+46-8-442-7050
fax: +46-8-442-7059

U.K.
The Powers Turner Group
+44-207-592-8331
fax: +44-207-630-9922

On the Web

WWW.SCIENTIFICAMERICAN.COM

FEATURED THIS MONTH

Visit **www.sciam.com/ontheweb**
to find these recent additions to the site:

The Boom in Bomb Detection

In the post-9/11 world and especially in the wake of the March 11 terrorist train bombings in Madrid, bomb detection has a higher priority than ever. Airport screening with x-ray machines is common; now other transportation modes are also being examined for their vulnerability. But there is no single technology that can be used to find all types of bombs. Future travelers, it seems, will be scanned, sniffed and zapped by an array of new high-tech devices.

Tourist Boats Force Killer Whales to "Shout" above the Din

Whale watching allows humans a glimpse of magnificent creatures in their natural habitat. But as the pastime becomes more popular, a new study suggests, noise from the boat traffic may be drowning out the animals' ability to hear one another's calls.

ASK THE EXPERTS

What causes hiccups?
William A. Whitelaw, professor in the department of medicine at the University of Calgary, explains.

NEW! GIVE THE GIFT OF Scientific American DIGITAL

The perfect present for any occasion.

EACH DIGITAL GIFT SUBSCRIPTION BRINGS:

- One year of access to 11+ years of *Scientific American*
- All current monthly issues before they reach the newsstands

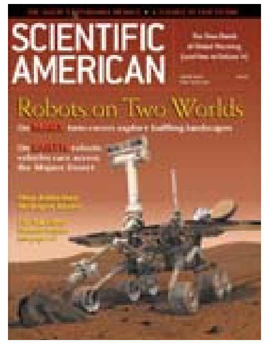
Give your gift today at www.sciamdigital.com/gift

HiENERGY TECHNOLOGIES, INC. *(top)*; FRED FELLEMAN *(bottom)*

Letters

EDITORS@SCIAM.COM

POLICY LEADERS OF THE WORLD, take note: readers of the March issue want you to pay attention to critical issues. One such concern mentioned by letter writers centered on the ways in which we elect candidates, in response to "The Fairest Vote of All," by Partha Dasgupta and Eric Maskin. Another priority—how we will avert environmental ills brought about by global warming—was raised by James Hansen's "Defusing the Global Warming Time Bomb." Details on reader reactions to those—and other articles in March—are on the pages that follow. But letter writers may also take note of this chestnut: "You can vote for whomever you want, but the government always gets in."



BRING OUT THE VOTE

Regarding "The Fairest Vote of All," by Partha Dasgupta and Eric Maskin: if the "true majority rule" system had been used in the last U.S. election, it is likely that some percentage of "Bush" voters would have selected the following ranking, to give the person generally perceived as the only other viable candidate as few points as possible:

Bush
Buchanan or Nader
Nader or Buchanan
Gore

Similarly, some percentage of "Gore" voters would have ranked Bush last to increase the impact of their vote. A "Nader" or "Buchanan" voter most likely would have ranked either Bush or Gore last for the same reason. The net outcome could have been a much stronger showing for Nader or Buchanan. It might even be more likely that a strong third-place candidate could win because of voters' attempts to keep an evident contender from beating their favored candidate.

Paul Sheneman
via e-mail

Dasgupta and Maskin apparently accept without discussion that a fair and desirable election is one that selects the candidate perceived by voters as best qualified. On the contrary, it is probably more important to the survival and stability of any organization that no minority faction feel powerless to affect the imposition of a candidate viewed as unacceptable. The fewest voters would be dissatisfied if they rated every candidate as "acceptable" or "unacceptable" and the candidate receiving the most acceptable votes was declared the winner.

William E. Tutt
Gainesville, Fla.

We question the authors' conclusion about the best replacement system. They use marketing hyperbole, adopting the term "true majority rule," for what political scientists call Condorcet voting.

Before discussing Condorcet, let's correct the authors' misrepresentations about instant runoff voting (IRV), another ranked-choice system, which simulates a series of runoff elections. We believe IRV is the best alternative for electing a single winner, such as president or mayor.

The authors dismiss IRV, using a distortion of the 2002 French presidential election. IRV, in fact, would have worked perfectly in that election. The top two candidates who advanced to the runoff were Chirac (19.8 percent) and Le Pen (17.4 percent). Nearly 63 percent of voters preferred other candidates. Under IRV, weak candidates would have been eliminated sequentially, and the majority of voters would have seen their votes coalesce behind the strongest candidates, Chirac and Prime Minister Jospin, in the final tally.

Now imagine a polarized election in which candidate A is favored by 55 percent of voters who all despise candidate B, who has 45 percent support. Now suppose candidate C joins the race and stresses C's likability and avoids any controversial issues. If 15 percent of the A supporters shift to candidate C, the result would be: A (40 percent), B (45 percent), C (15 percent). Even though 55 percent consider B the worst choice, B wins under plurality rules. Under IRV, C is eliminated in the runoff count and A wins with 55 percent. Under Condorcet, however, if A supporters rank C above B, whom they detest, and B supporters also rank C above A, because of disdain for A, candidate C can win. In fact, it is possible for the Condorcet winner to be someone nobody considers a particularly good candidate. Condorcet punishes candidates who take clear stands on controversial issues and rewards candidates who say little of substance.

IRV strikes a sensible balance between rewarding first-choice support and compromise appeal. Used in major national elections elsewhere, it also has an analogue within the American experience (traditional runoffs) that makes it a viable reform—one that has won the endorsement of Howard Dean and John McCain, been adopted by Utah Republicans and San Francisco voters, and been introduced in two dozen state legislatures.

Philip Macklin, professor of physics (emeritus)
Miami University, Oxford, Ohio
Terrill Bouricius, senior policy analyst
The Center for Voting and Democracy
Burlington, Vt.
Rob Richie, executive director
The Center for Voting and Democracy
Takoma Park, Md.

DASGUPTA AND MASKIN REPLY: Sheneman implies that we favor an electoral system in which candidates receive more points the higher they are ranked by voters, so that the one with the most points wins. But true majority rule, our proposed system, doesn't make use of "points" at all: the winner is simply the candidate preferred by a majority (more than 50 percent) of voters to any opponent. The system Sheneman is thinking of is called rank-order voting, which we take pains to criticize in our article. In contrast to rank-order voting, true majority rule is far less vulnerable to strategizing.

Tutt's proposal is called approval voting. In effect, it is a version of rank-order voting in which the voter is constrained to provide a ranking of candidates with just two tiers: "acceptable" and "unacceptable." But how is the voter to draw the line between the tiers? And even if the voter does have a clear sense of who is acceptable and who is not, he or she will have a strong incentive to vote strategically. Specifically, in our four-candidate example, Bush would be elected under approval voting if Gore backers included Bush as acceptable, whereas Gore would be elected if they did not. Thus, regardless of their true feelings about Bush's acceptability, they may be inclined, in Samuel Goldwyn's phrase, to include him out.

Contrary to Macklin, Bouricius and Richie, our article gives an accurate picture of the potential failings of IRV vis-à-vis the 2002 French election. If the six other candidates from that election were first eliminated in instant runoffs, the scenario in which the true majority winner, Jospin, is dropped next—leaving just Chirac and Le Pen—would be all too plausible. As for their A-B-C example, the writers argue that candidate A "should" win (and indeed does so, under IRV). But by their own assumptions, 60 percent of the electorate prefer C to A. How can it be democratic to elect A when C would beat him by a landslide in a head-to-head contest?

HOW WE VOTE is open to improvement.

TOBY TALBOT *AP Photo*; PHOTOILLUSTRATION BY JOHNNY JOHNSON

COUNTERING GLOBAL WARMING

In James Hansen's otherwise excellent article, "Defusing the Global Warming Time Bomb," I was disappointed to read his opinion that "there may be even better solutions, such as hydrogen fuel." While hydrogen clearly has an important role to play as a repository of energy, it is not likely to be a significant source of energy, because it requires at least as much energy to create molecular hydrogen as is recovered by its use as a fuel.

Fortunately, energy conservation measures available today not only could substantially decrease the production of greenhouse gases but also would be inexpensive—and might even pay a financial dividend. I believe it is the duty of the scientific community to keep this issue before the public and to press for general acceptance of energy conservation, with the goal of making it easier for those in leadership positions to support such initiatives.

Jack C. Childers, Jr.
Lutherville, Md.

ADDICTED TO CAFFEINE?

Regarding "The Addicted Brain," by Eric J. Nestler and Robert C. Malenka: Are caffeine and sugar also addictive substances?

Patricia Mathews
Albuquerque, N.M.

NESTLER AND MALENKA REPLY: Whether caffeine and sugar are addictive remains controversial. Caffeine unquestionably causes physical dependence. People who consume steady amounts on a daily basis exhibit a characteristic withdrawal syndrome (headache, fatigue, irritability) if they go without it for a day. This physical dependence is distinct from addiction, which can be defined as compulsive use of a drug despite horrendous consequences or as loss of control over drug use. By these latter definitions, very few people are truly addicted to caffeine.

Similarly, very few people eat sugar compulsively. An argument can be made, though, that the individuals who do display compulsive sugar consumption can be considered "addicted," and some work in laboratory animals shows that certain brain changes associated with compulsive drug use also occur with compulsive sugar consumption.

ERRATUM: The credit for the 1985 photograph of Curt Herzstark in "The Curious History of the First Pocket Calculator," by Cliff Stoll [Scientific American, January], was incomplete. The photograph was taken by Erhard Anthes and provided by Rick Furr.

50, 100 & 150 Years Ago
FROM SCIENTIFIC AMERICAN

Nobel Chemists ▪ Visionary Author ▪ Prescient (and Unlucky) Inventor

JULY 1954

PROTEIN CHEMIST—“In the study of the structure of a protein there are two questions to be answered. What is the sequence of amino acids in the polypeptide chain? What is the way in which the polypeptide chain is folded back and forth in the space occupied by the molecule? In this article we shall consider only the second question. The experimental technique of greatest value in the attack on this problem is that of X-ray diffraction. —Linus Pauling [et al.]” [*Editors’ note: Pauling was awarded the Nobel Prize in Chemistry several months after this article appeared.*]

DUST BOWLS—“In 1954 we have two dust bowls to shame us instead of one. The marginal soils of the southwestern plains, brought under the plow during the wartime agricultural boom, are now well on the way to complete breakdown. It is easy to blame this distressing situation on drought, but drought is a normal feature of climate on the southern Great Plains. The blame falls not on the elements, but on our refusal to adapt to them. The outbreaks of dust storms have closely followed the pattern of the original dust bowl in the 1930s. For two or three years the crops on lands of marginal fertility had failed. On the unprotected fields the exposed soil moved out with each wind of sufficient velocity to cause erosion. These areas expanded and coalesced into the two new dust bowls.”

JULY 1904

THE FUTURE—“We of the early twentieth century, and particularly that growing majority of us who have been born since the *Origin of Species* was written, perceive that man, and all the world of men, is no more than the present phase of a development so great and splendid that beside this vision all the exploits of humanity shrivel in the proportion of castles in the sand. We look back through countless millions of years and see the great will to live struggling out of the intertidal slime. We turn again toward the future, surely any thought of finality, any millennial settlement, has vanished from our minds. The question what is to come after man is the most persistently fascinating and the most insoluble question in the whole world.
—Herbert G. Wells”

ELEMENTAL CHEMIST—“The eminent English Scientist Sir William Ramsay, whose name is intimately associated with the new element radium, is one of the world’s youngest scientists, being only fifty-two years of age. Sir William Ramsay may be said to have first brought himself to the public notice by his brilliant discoveries of unknown and unsuspected constituents in the atmosphere (argon, helium, neon, krypton, and xenon)—discoveries made partly with the collaboration of Lord Rayleigh. The photograph of Sir William Ramsay was taken in his laboratory specially for the *Scientific American.*” [*Editors’ note: Ramsay was awarded the Nobel Prize in Chemistry several months after this article appeared.*]

SIR WILLIAM RAMSAY in his laboratory, 1904

SHIP STABILIZER—“The pitching of a ship in a rough sea is certainly a serious drawback both to the physical welfare of passengers and crew and to the expedition of any work made on board. Now Otto Schlick, a well-known naval engineer of Hamburg, Germany, has brought out an ingenious apparatus designed to diminish the amplitude of oscillation. This apparatus is based on the gyroscopic effect of a flywheel, mounted on board a steamer, and caused to rotate rapidly by a motor.”

JULY 1854

USELESS INVENTION?—“The Paris Correspondent for the ‘New York Times’ says: ‘An inventor, who considered himself on the point of final success, has just fallen victim to his own machine. This was a steam vehicle, running upon the ordinary post roads of France. M. Leroy was descending a hill, when the engine struck an obstacle, tipped over, and poured the contents of the boiler on to M. Leroy, who was badly scalded. He had spent ten years and all his money in perfecting his invention.’ He was a very foolish inventor to throw away his money on such an invention. To reproduce steam carriages for common roads, after the invention of railroads and locomotives, is like going to mill with corn in a bag, having a stone in one end to balance the grain in the other.”

news SCAN

TOXICOLOGY

Leading to Lead

CONFLICTING RULES MAY PUT LEAD IN TAP WATER **BY REBECCA RENNER**

The public reporting last year of high lead levels in the drinking water in Washington, D.C., has led to a congressional investigation, the firing of a D.C. health official, and calls for a review of the 1991 law that is supposed to keep the neurotoxic metal out of drinking water. That law, however, may not contribute to the problem as much as the changes made to disinfection procedures resulting from another water safety rule. The conflicting regulations mean that other municipalities may also soon find too much lead coming out of their faucets.

TASTE OF METAL: Modified disinfection methods may have changed the chemistry of drinking water in Washington, D.C., making it more likely to dissolve lead-encasing minerals in pipes.

PHILIP JAMES CORWIN *Corbis*

To date, at least 157 houses in D.C. have lead levels at the tap higher than 300 parts per billion (ppb), and thousands more have exceeded the Environmental Protection Agency's limit of 15 ppb. Residents have received contradictory advice about whether tap water is safe to drink and whether replacement of lead service lines will solve the problem.

Lead should not normally enter the flow, because layers of different lead-snaring minerals naturally build up inside the pipes. But these mineral scales act as a trap for lead only as long as they remain insoluble; a sudden shift in water chemistry can change that.

Such a change may have triggered the D.C. problems. In 2000 Washington Aqueduct, the area's water treatment plant, modified its procedures to comply with the 1998 Disinfection Byproducts Rule (DBR), which restricts the presence of so-called halogenated organic compounds in water. These compounds form when disinfectants, particularly chlorine, react with natural organic and inorganic matter in source water and in distribution systems. The DBR directs water companies to make sure that the by-products, which might cause cancer, stay below a certain level.

One of the most common ways to comply with the DBR is to use a mixture of chlorine and ammonia—called chloramines—instead of chlorine. Some 30 percent of major U.S. water companies currently take this route, and the proportion will probably grow as limits on disinfection by-products are tightened during the next few years. Because no one has investigated the effects of chloramines on corrosion in drinking-water systems, meeting DBR requirements may mean violating the 1991 lead-copper rule, which sets maximum limits on these metals (for lead, 15 ppb).

Evidence for chloramines' effect on Washington's pipes comes from EPA chemist Michael Schock. He discovered that different mineral scales—especially lead dioxide scales—are particularly vulnerable to changes in water chemistry. With chlorine, Washington's water was highly oxidizing. As a result, the mineral scales that formed consisted of lead dioxide, which Schock has found in every sample of Washington's lead service lines that he has examined. The switch to chloramines lowered the oxidizing potential of D.C.'s water, which probably dissolved the lead dioxide scale and thereby liberated the lead.

Corrosion scientists warned about potential conflicts between the two rules. "We were concerned that drastic changes in water treatment could disturb scales and mobilize metals," says one scientist involved in the investigation of the D.C. lead problem, who asked not to be named. Another researcher echoed the point: "There was essentially no research concerning interactions between the lead-copper rule and the DBR. There was zero consultation with corrosion scientists even though we screamed for it."

The EPA noted potential conflicts in a 1999 publication entitled *Microbial and Disinfection Byproducts Rules Simultaneous Compliance Guidance Manual*. But the document offers little in the form of specific procedural advice, scientists say.

Virginia Tech engineer Marc Edwards, a former EPA consultant who first called attention to the D.C. problem, has warned the agency and the water industry for years that changes in drinking-water treatment were liable to cause trouble for home plumbing systems. He believes that lead problems may lurk in other cities, too. Chemist Mark Benjamin of the University of Washington concurs, noting that the factors affecting corrosion—the pipe material, the mineral scales and the water quality—are universal in water systems. "It would be remarkable and unlikely to think that these factors just happened to combine in a unique way in Washington," he states.

Rebecca Renner covers environmental sciences from Williamsport, Pa.

GETTING THE LEAD OUT

Lead service pipes, the smaller pipes that branch out from the mains, are found in many U.S. cities in the Northeast and upper Midwest, according to the most recent national study, a 1990 American Water Works Association report. It tallied approximately 6.4 million lead connections and 3.3 million lead service lines. The report noted that 61,000 lead lines are replaced annually, but even so, millions are probably still in service. Chloramines most likely cause a problem in systems that have lead dioxide scales. Unfortunately, no one knows how many water systems have such scales.

BIOLOGY

Mickey Has Two Moms

NO SPERM NEEDED: MICE ARE BORN FROM TWO EGGS BY DIANE MARTINDALE

Being fatherless took on new meaning in April when a research team, led by Tomohiro Kono of the Tokyo University of Agriculture, created mice from two eggs. The group's achievement does not promise a new way to make babies; rather it helps to explain how egg and sperm work together and why males are vital in normal reproduction.

The process that created the mice is akin to parthenogenesis, in which an unfertilized egg develops on its own and produces viable offspring. It occurs in some lower creatures such as fleas, lizards and turkeys. The barrier to parthenogenesis in mammals is thought to be genetic imprinting, in which some genes needed for embryonic development are turned off in the female genome but switched on in the male genome, and vice versa. Thus, for an embryo to grow properly, it must have one set of chromosomes with a female imprint and the other with a male imprint. In past studies, mouse eggs have been induced to replicate without fertilization, but they survive only briefly.

Kono's team began with a genetically modified strain of mice in which the females produce eggs whose chromosomes have a malelike imprint. Specifically, the eggs lack the *H19* gene, which is normally imprinted, or turned off, in sperm. The mutation allows for the production of IGF-II, a growth factor that ordinarily comes only from the male genome and is crucial for embryonic development.

Moreover, Kono harvested immature eggs from the modified mice. This step is important because during egg formation all genetic imprints are erased before female imprints are established. Biologists believe that the chromosomes of very young eggs have not yet acquired female imprints and are in a state much closer to a male genome.

These immature, malelike gametes were then fused with mature, normal mouse eggs, chemically activated, and implanted into surrogate mice. Two mice were born: Kaguya, named after a Japanese fairy-tale character, grew to adulthood, mated and gave birth to a litter of pups with no apparent defects; the other was sacrificed at birth for genetic analysis.

KAGUYA, which was born from two eggs, bred and gave birth to normal pups.

COURTESY OF TOMOHIRO KONO

The experiment reveals the nature of imprinting and provides a useful tool for studying its role in development—faulty imprinting causes neurological disorders, abnormal growth and some cancers. The study also has implications in animal cloning and stem cell research, wherein defects in imprinting are often to blame for the high failure rates.

At this point, researchers do not think that the technique has implications in human fertility work. "The extreme genetic manipulations used by Kono's team are for now, at least, technically and ethically infeasible in humans," assures Azim Surani, a pioneer in imprinting studies at the Wellcome Trust/Cancer Research Institute at the University of Cambridge. Immature eggs would have to be plucked directly from the ovaries. But more troublesome, a woman would need to be genetically altered to produce eggs with the *H19* mutation so that the eggs can make IGF-II. An alternative might be to deliver IGF-II directly to the eggs, but the levels must be just right. Otherwise, the growth factor leads to abnormalities.

Kono's experiments also produced many dead and abnormal mice: only two mice resulted out of nearly 500 attempts. This low rate suggests that the risk of abnormalities could be very high. "The method is less efficient and riskier than cloning," Surani notes.

Most surprising to experts was how the subtle tweaking of just two genes removed the roadblock to producing live mice. What is more, genetic analysis revealed that the activity of many other genes in the mice had returned to normal, as though conventional fertilization had taken place. But simply altering *H19*, and thus IGF-II, is not enough to explain how these two mice made it to full term, argues Kevin Eggan, a developmental biologist at Harvard University. Some sort of random events "occurred in the two surviving mice, but no one knows what those are," he says. Kono's experiment may have shown that it is possible to do away with males in reproduction, but his findings have also reaffirmed their importance.

Diane Martindale is based in Toronto.

WHAT TO CALL AN EGGS-ONLY BIRTH?

The researchers who made Kaguya, the fatherless mouse, refer to the technique as parthenogenesis. But most experts challenge the accuracy of the term because they used two females. Kaguya did not develop from a single, unfertilized egg—a true parthenote—but from the union of two unfertilized eggs. The fusion of two eggs yields a gynogenote. But Kaguya is not even that, because the team used an immature, genetically modified egg in combination with a mature, normal one. Despite the complaint over the nomenclature, no one has offered up a better name.

MATERIALS SCIENCE

Magnetic Soot

CARBON NANOFOAM IS FOUND TO BE FERROMAGNETIC **BY GRAHAM P. COLLINS**

Recent decades have seen great interest in novel carbon structures such as buckyballs and nanotubes. In 1997 researchers in Australia discovered yet another form of carbon: a spidery, fractallike composition they dubbed nanofoam. At this year's March meeting of the American Physical Society, the group reported that this gossamer substance is ferromagnetic (like iron), the only type of pure carbon that has that property. The foam's magnetic behavior suggests that innovative uses might be possible, such as serving as a contrast-enhancing agent in magnetic resonance imaging.

Andrei V. Rode and his co-workers at the Australian National University in Canberra created carbon nanofoam when they blasted a glassy form of carbon with a series of short laser pulses in a container filled with inert argon gas. The pulses produced a plume of carbon vapor that settled as a thin layer on the vessel walls. To the naked eye, it looks like a conventional soot deposit.

The foam consists of clusters of about 4,000 atoms, with the clusters strung together to form a tenuous web. The clusters, which are each about six nanometers in diameter, seem to be made of graphite layers warped by the inclusion of seven-sided heptagons. That configuration gives the layers negative curvature—the "hyperbolic" saddle shape—the converse of what happens in buckyballs, in which pentagons replace some hexagons to form a soccer ball shape.

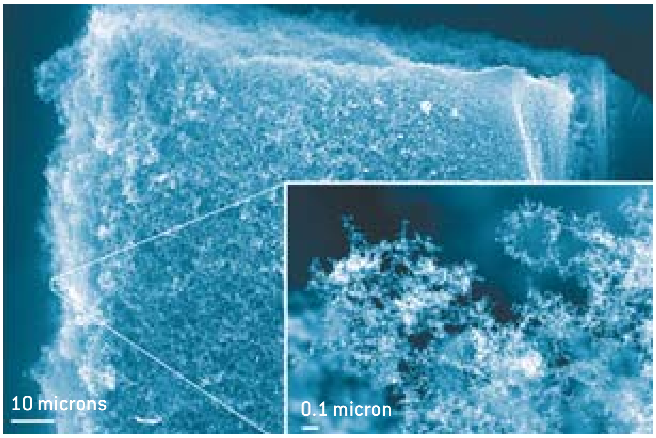
So much empty space permeates the web that the nanofoam's density is only a few times that of air at sea level, comparable with the lowest-density solid known, porous materials called silica aerogels. The foam is similar to carbon aerogels but with 1⁄100 their density. Unusual properties of the foam were apparent from the beginning. It held a charge so well that it clung electrostatically to the production vessel, making it difficult to extract. This trait indicated that the foam was a poor electrical conductor, unlike carbon aerogels.

The researchers also found that the nanofoams had numerous unpaired electrons, which require carbon atoms with fewer than four bonds. Rode and his collaborators propose that these unpaired electrons occur at "topological and bonding defects" associated with the saddle-curved sheets of atoms. The convoluted sheets would stabilize these unpaired electrons by protecting them from reacting with one another.

The presence of unpaired electrons suggested that the foam should have magnetic properties as well. A simple test bore out this suspicion: "Freshly produced foam samples were attracted to a permanent magnet," Rode states. The group investigated the nanofoam's magnetic properties in collaboration with researchers from the Foundation for Research and Technology-Hellas and the University of Crete, both in Heraklion, Greece, and from the Ioffe Physical-Technical Institute in St. Petersburg, Russia. Not only was the foam attracted to a magnet, but below –183 degrees Celsius it could acquire a permanent magnetization, like a piece of iron. In other words, it is ferromagnetic, a property shared by no other instance of carbon's many structures (called allotropes).

"It took us a long time to investigate the magnetic properties of the foam, to confirm the results, and to exclude the possibility of impurities in the foam," Rode says—contaminants such as iron or nickel particles from the stainless-steel container used could have confounded the magnetic results.

The group's current focus is on finding how to control the properties of the foam by adjusting the conditions under which it is produced. "The major challenge in our research," Rode explains, "is understanding how the laser beam intensity and repetition rate and the gas pressure influence the foam's properties." A nanofoam with the right characteristics could someday be used to enhance magnetic resonance imaging: granules of it injected into the bloodstream would cause the blood to show up very clearly on the scan. It might also have applications in spintronic devices, in which the spin or magnetism of electrons is utilized.



FRACTAL STRUCTURE of carbon nanofoam is apparent in these electron micrograph images. Tiny clusters, each containing a few thousand carbon atoms, are strung together to form the tenuous web. The recently discovered magnetic properties of the foam could lead to novel applications.

ANDREI V. RODE *Australian National University*

TRAIPSING THROUGH ALLOTROPES

Pure carbon comes in numerous molecular structures, or allotropes, in addition to the familiar graphite, diamond, buckyballs and nanotubes:

- **Lonsdaleite: Has the same bond type as diamond, but the atoms are arranged in a hexagonal pattern; also called hexagonal diamond.**
- **Chaoite: Produced when graphite is shocked by a meteorite impact. Also has hexagonal arrangement of atoms.**
- **Schwarzite: Hypothetical structure in which hexagonal layers are warped into "negatively curved" saddle shapes by the presence of heptagons.**
- **Filamentous carbon: Fibers made of small plates stacked in long chains.**
- **Carbon aerogels: Very low density, porous structures analogous to the better-known silica aerogels.**
- **Amorphous: Any form of carbon in which there is disorder instead of an extensive regular lattice structure. Can be classified by the relative proportions of diamondlike and graphitelike bonds.**

LINGUISTICS

Baby Talk Beginnings

INFANT PACIFICATION MAY HAVE LED TO THE ORIGIN OF LANGUAGE BY KATE WONG

When a staff member brings a baby into the offices of *Scientific American,* a small crowd inevitably forms around the infant, and although the onlookers all have rather different personalities and mannerisms, they tend to talk to the baby in the same singsong way. Vowels are lingered over, phrases are repeated in high-pitched voices, and questions carry exaggerated inflections. Sound familiar? This is motherese, the distinctive speech that human adults across the globe instinctively use when addressing babies. And according to a new theory, it holds a key to the emergence of language.

In a paper slated for the August *Behavioral and Brain Sciences,* Florida State University physical anthropologist Dean Falk proposes that just as motherese forms the scaffold for language acquisition during child development, so, too, did it underpin the evolution of language. Such baby talk itself originated, she posits, as a response to two other hallmarks of human evolution: upright walking and big brains.

In contrast to other primates, humans give birth to babies that are relatively undeveloped. Thus, whereas a chimpanzee infant can cling to its quadrupedal mother and ride along on her belly or back shortly after birth, helpless human babies must be carried everywhere by their two-legged caregivers. Assuming, as many anthropologists do, that early humans had chimplike social structures, moms did most of the child rearing. But having to hold on to an infant constantly would have significantly diminished their foraging efficiency, Falk says.

She argues that hominid mothers therefore began putting their babies down beside them while gathering and processing food. To placate an infant distressed by this separation, mom would offer vocal, rather than physical, reassurance and continue her search for sustenance. This remote comforting, derived from more primitive primate communication systems, marked the start of motherese, Falk contends. And moms genetically blessed with a keen ability to read and control their children, so the theory goes, would successfully raise more offspring than those who were not. As mothers increasingly relied on vocalization to control the emotions of their babies—and, later, the actions of their mobile juveniles—words precipitated out of the babble and became conventionalized across hominid communities, ultimately giving rise to language.

GOO GOO: Baby talk may have enabled early moms to keep their infants calm while foraging for food.

JENNIE WOODCOCK *Reflections Photolibrary/Corbis*

Falk's report has generated a number of objections. Paleoanthropologist Karen R. Rosenberg of the University of Delaware and her colleagues, for example, balk at the suggestion that early hominid mothers set their children down in the first place, observing that, Westerners aside, modern caregivers rarely do this, preferring to carry them in their arms or in slings. Protohuman moms probably fashioned baby slings, too, they suggest, both for ease of transportation and to keep the young warm by holding them close to their bodies. If so, they need not have developed a way to control their babies from a distance.

Linguists likewise demur. Falk's account sheds considerable light on the origins of speech, writes Derek Bickerton of the University of Hawaii at Honolulu in an accompanying commentary. Unfortunately, he continues, it reveals nothing about the origins of language. He charges that the hypothesis fails to address how the two fundamental features of language—namely, referential symbols and syntactic structure—arose, noting that speech is merely a language modality, as are Morse code and smoke signals. Falk's scenario does not explain how mother's melodic utterances acquired meaning in the first place, Bickerton insists.

Popular wisdom holds that language is a relatively modern invention, one that appeared in the past 100,000 years or so. But if Falk is right, baby talk—and perhaps full-blown language—evolved far earlier than that: the fossil record indicates that by 1.6 million years ago, early members of our genus, *Homo,* were fully bipedal and probably giving birth to undeveloped infants. And future discoveries of even older modern-looking human fossils could root our yen for yakking deeper still. It seems certain only that we have not heard the last word on the first words.

ASTROPHYSICS

Growing Pains

OLD, MASSIVE GALAXIES WHEN NONE SHOULD BE **BY GEORGE MUSSER**

Bow down before dark matter: that is one of the messages of 20th-century astronomy. Some unseeable material rules the cosmos, and ordinary matter is just along for the ride. It sounds like the culmination of the Copernican revolution, the ultimate displacement of humanity from a central role in the grand scheme of things. But lately researchers have been thinking that the lesson in humility has gone too far. What dark matter demands, ordinary matter doesn't always obey meekly.

TOO BIG, TOO SOON? Giant galaxy J02399 packed in 300 billion solar masses when the universe was only 2.4 billion years old. Mass was inferred from the motion of carbon monoxide (*the density of which is shown by the color intensity*); dust (*contour lines*) traces the galaxy's shape.

Inklings of the spunkiness of ordinary matter have emerged over the past decade as observers have peered deeper into space and therefore farther back in time. According to the standard dark matter scenario, galaxies should have formed hierarchically: subgalactic scraps came first and slowly consolidated into full-fledged galaxies. Yet many galaxies seem to have jumped the gun: they were too big too early. "The mass assembly of massive galaxies is extremely and remarkably rapid—and much more rapid than is seen in the models," says Reinhard Genzel of the Max Planck Institute for Extraterrestrial Physics in Garching, Germany.

For instance, sensitive infrared observations have spotted giant galaxies just a couple of billion years after the big bang, which is early by cosmological standards. Many of these systems contain mature stars and so must have arisen even earlier. Moreover, the mix of elements in galaxies indicates that bigger ones are older than their smaller brethren, another blow to the hierarchical paradigm.

Genzel and his colleagues' latest work tightens the screws even further. They focus on submillimeter galaxies, so called because astronomers see them in light with a wavelength a bit shorter than one millimeter. Because such light is hard to detect, these galaxies were discovered only in 1997, even though they are some of the brightest objects in the universe. Genzel's team has measured the orbital speed of gas clouds within 11 of these systems, giving the first unambiguous measurement of the mass of galaxies in the early universe: greater than 100 billion solar masses, as hefty as the biggest galaxies in the present-day universe. Extrapolated to the whole sky, the team's work implies 50 million of

REINHARD GENZEL ET AL. IN *ASTROPHYSICAL JOURNAL*, VOL. 584; 2003. REPRODUCED BY PERMISSION OF AAS these heavyweights, 100 times as many as models predict.

"They're absolutely right," admits theorist Carlos Frenk of the University of Durham in England. "The models we put out three years ago did not produce enough big galaxies" in the distant past. Some claim that the findings cast a pall over the very concept of dark matter, but Genzel, Frenk and others say that the dark matter is behaving as it should; it is the ordinary matter mixed in with it that is causing the trouble.

Dark matter may seem exotic, but cosmologists regard it as the essence of simplicity. It is "cold," endowed with little energy, and it responds only to the force of gravity. Ordinary matter, in contrast, is a cauldron of nuclear reactions, shock waves, magnetism, turbulence—a mess that cosmologists whimsically call gastrophysics.

Models used to assume that as dark matter clumps, ordinary matter just follows along. But gastrophysics stirs the pot. As gas pools, it turns into stars, whose outflows and explosions push material back out into intergalactic space—a process of negative feedback. This rebellion is most effective in small clumps of dark matter, where gravity is too weak to contain the stellar spatter. So building a small galaxy is harder than it looks. Conversely, the same processes can actually amplify star formation in large galaxies.

Theorists nowadays include this feedback in their simulations, but the observations by Genzel's team suggest that they haven't gone far enough. Frenk and his colleagues suspect that big, powerful stars are more common than astronomers have been assuming. Another group, led by Gian Luigi Granato of the Astronomical Observatory of Padua, postulates that large black holes act as a kind of galactic thermostat: as stars funnel matter into them, they spew material that chokes off star formation. In both cases, the extra feedback causes galaxy formation to occur abruptly rather than progressively over time, as the hierarchical paradigm would suggest.

So even if dark matter ultimately dictates the overall course of cosmic events, ordinary matter has the consolation of being the life of the universe, softening the brute forces of nature like a flower box in the city.

TOO OLD IN A YOUNG UNIVERSE

Even by the exacting standards of astronomers, the galaxy that Alan Stockton of the University of Hawaii at Manoa and his colleagues discovered last year is a real head-scratcher. It is massive (300 billion stars or thereabouts) and mature (its reddish hue implies that the stars are two billion years old) in a comparatively youthful period of the universe (some 2.6 billion years after the big bang). Stranger still, it seems to have a disk shape, something like the Milky Way. For a disk to endure, the galaxy could not have collided with another sizable galaxy. Because such collisions are the usual triggers for rapid star formation, astronomers are left wondering how the galaxy managed to create so many stars so quickly.

BIOTECH

Bugging for Guidance

NO ONE IS SURE WHO REGULATES GENETICALLY MODIFIED INSECTS **BY JR MINKEL**

MOSQUITO could be genetically manipulated so that it does not harbor the malaria-causing parasite.

For all the media attention genetically modified crops have received, the plants are relatively easy to control compared with what lies down the road—genetically modified insects (GMIs). Although most field trials of such insects are years away, experts say that the science has advanced rapidly and that regulators need to begin establishing rules now for assessing their potential effects on the environment and public health.

Modified insects are meant to combat a variety of pests and diseases that afflict humans, plants and beneficial insects such as the honeybee. Researchers expect the risks to be small, but they still have not been studied. "We're not talking about the [Flavr Savr] tomato," comments Thomas Scott, entomologist at the University of California at Davis. In some cases, "we're talking about human-blood-sucking, free-ranging, pathogen-transmitting organisms."

Investigators first want to know which regulatory agencies will grant approval for interstate transport and permanent release of GMIs. Given the range of possible applications of the creatures, the Food and Drug Administration, the Environmental Protection Agency and the U.S. Department of Agriculture all potentially have the authority to regulate them. How their oversight will be divided and coordinated, though, is not clear, according to a January report by the nonpartisan Pew Initiative on Food and Biotechnology. Overlapping jurisdictional boundaries could burn researchers who get approval from one authority only to find that another also claims jurisdiction.

Scientists also want clear risk-assessment guidelines for preparing an application for permanent release, explains Marjorie Hoy, a University of Florida entomologist who conducted a short field trial of a transgenic predatory mite carrying a marker gene but is uncertain about the legal procedures to release such mites permanently. "After going through the process, I realized that even if I had some really good genes and a really good system, I wasn't sure what I'd have to do to get it into the field," Hoy says.

The agency that seems poised to take charge is the USDA Animal and Plant Health Inspection Service (APHIS). It already inspects agriculturally important modified insects coming into the country or moving between states and has rules in place covering field trials of engineered plant pests such as the cotton-attacking pink bollworm.

The service is also drafting rules governing the transportation and release of livestock pests, APHIS scientist Bob Rose says. Almost all insects that carry human disease bite farm animals as well, he observes, adding that between APHIS and the EPA, most applications of genetically modified insects are covered fairly clearly. The EPA regulates research on "paratransgenic" organisms—those containing symbiotic microbes engineered to counteract pathogens—by viewing them as microbial pesticides, which fall under the Federal Insecticide, Fungicide, and Rodenticide Act. "GMI regulation is not really the mess some folks would like others to believe it is," Rose insists. Mosquito researchers, however, remain skeptical that the livestock rules will cover everything. "There is a policy issue of whether you want an agency with expertise for livestock health taking the lead on human health issues," argues Michael Rodemeyer, executive director of the Pew Initiative.

The regulatory history of modified crops suggests that a strong, central authority could help prevent agencies being played against one another, notes entomologist Mark Winston of Simon Fraser University in British Columbia. He adds that such an authority might also ask whether we need engineered insects where integrated pest management could work just as well. "Nobody's asking that question in a regulatory way," Winston remarks.

Regulatory uncertainties are not necessarily a cause for alarm, points out Mark Benedict, an entomologist at the Centers for Disease Control and Prevention. Insects modified for human disease control haven't been moneymakers so far, and researchers say they aren't eager to release them for short trials, much less permanently, without oversight.

JR Minkel is based in New York City.

ACHIEVING PEST CONTROL

Scientists have two major ways of altering an insect to control disease:

- Dwindle the population by interfering with reproduction. A gene inserted into insects would be passed down to their young, killing them before they could mate. (The released insects would be immune.) The foreign gene would gradually disappear as the insects carrying it died. But this method may be hard to implement against mosquitoes, for example, because the number of modified insects needed to swamp the wild population could be too high to be practical.
- Introduce a gene that makes the insect less likely to harbor or transmit a pathogen. The risks are potentially high because it would require spreading a foreign gene throughout a population and maintaining it permanently.

TONY BRAIN *Science Photo Library*

BY THE NUMBERS

Undercutting Fairness

STATES AND LOCALITIES UNDERMINE TAX PROGRESSIVENESS **BY RODGER DOYLE**

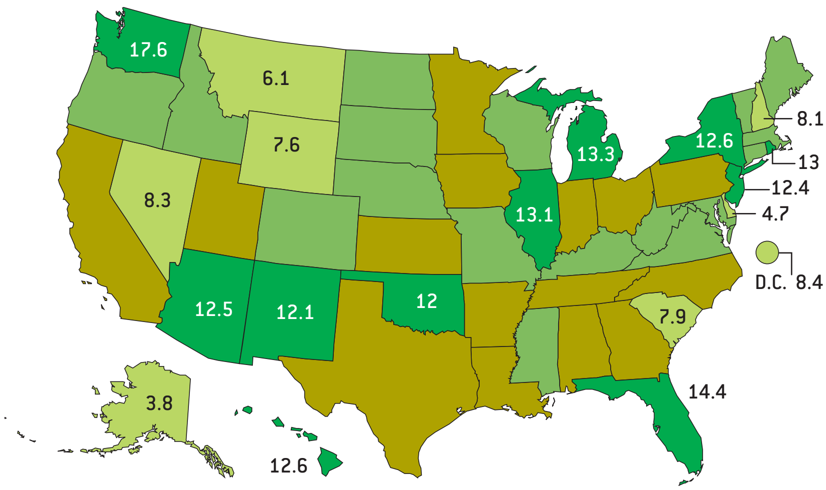
For most, the federal income tax probably does not rank as a great achievement of the 20th century. But it is efficient and mostly fair, thanks to its progressive rates. Few realize, however, that state and local taxes are so strongly regressive that they cancel out much of the progressivity of the federal tax.

This conclusion comes from the Institute for Taxation and Economic Policy, a Washington, D.C.–based research group. It conducted a state-by-state analysis of the tax burden on families headed by those younger than 65 years of age—namely, their total tax as a percentage of their income. (Elderly families were excluded because state tax systems often treat them differently.) It found that those in the lowest 20 percent income bracket paid at a rate 2.2 times that of the top 1 percent, whereas the middle 20 percent paid at a rate 1.8 greater.

As illustrated on the map, the tax burden for the bottom 20 percent varies widely by state. At one extreme is the state of Washington, where this group pays at a rate 5.7 times that of the top 1 percent. At the other extreme is Delaware, where the tax burden is virtually the same for all income groups. A map for the middle 20 percent income group would display a fairly similar pattern.

The type of tax levied explains the differences among the states. Washington, for example, relies primarily on sales and excise taxes, whereas Delaware relies mostly on a progressive income tax. Sales taxes, which are levied on a percentage basis, and excise taxes, which are levied as fixed fees, take a larger share of income from low- and middle-income families than from the rich and nullify the progressive effect of income taxes.

Most state and local tax systems are archaic and not merely because of regressivity. Property taxes, the dominant source of local income, have traditionally financed schools, a custom that results in inadequate funding in lower-income districts. Most states still rely heavily on local property taxes for schools, but a few, such as New Hampshire and Vermont, have implemented statewide property taxes to raise the equity of school funding.



Percent of Income Paid in State and Local Taxes by Bottom 20 Percent, 2002

Less than 9 | 9 to 10.4 | 10.5 to 11.9 | 12 or more

Most sales taxes were enacted in the 1930s and did not apply to services, which were a relatively small part of consumer spending then. Today services account for 60 percent of spending. Only New Hampshire, Hawaii and South Dakota tax services comprehensively. A broader tax, particularly one that exempted necessities such as utilities, would be less regressive because services are consumed disproportionately by the wealthy.

State and local governments are shifting away from progressivity. Their revenues from income taxes fell 10 percent from 2000 to 2003; during the same period, sales taxes rose 6 percent and property taxes, 20 percent.

Rodger Doyle can be reached at rdoyle2@adelphia.net

FURTHER READING

The Progressivity of State Tax Systems. Andrew Reschovsky in *The Future of State Taxation.* Urban Institute Press, 1998.

State Sales and Income Taxes: An Economic Analysis. George R. Zodrow. Texas A&M University Press, 1999.

Who Pays? A Distributional Analysis of the Tax System in All 50 States. Second edition. Robert S. McIntyre et al. Institute on Taxation and Economic Policy, January 2003. Available at www.ctj.org

State and local taxes as a percentage of income, 2002:

INCOME GROUP	SALES AND EXCISE	PROPERTY	INCOME	TOTAL TAX*
Lowest 20 percent	7.8	3.1	0.6	11.4
Second 20 percent	6.4	2.3	1.6	10.3
Middle 20 percent	5.1	2.5	2.3	9.6
Fourth 20 percent	4.1	2.6	2.7	8.8
Next 15 percent	3.1	2.6	3.2	7.7
Next 4 percent	2.0	2.3	3.8	6.5
Top 1 percent	1.1	1.4	4.8	5.2

* Reflects deduction for federal taxes

RODGER DOYLE; SOURCE: *Robert S. McIntyre et al.* (map and table)

DATA POINTS: SEED MONEY

Enough nations ratified the International Treaty on Plant Genetic Resources for Food and Agriculture for the "seed treaty" to take force on June 29. Signatories have pledged to commit funds to conserve the genetic diversity of the world's food crops. The law also prohibits the patenting of seeds, although some ambiguity in the wording of the relevant section—article 12.3.d—leaves open that possibility. The U.S., which has not ratified the treaty, and nine others voted to delete that article.

World's crop varieties lost in the past century: 95 percent

Number of crops that feed most people: 150

Number of crops that provide 80 percent of food energy: 12

Top four crops:
Rice
Maize
Potatoes
Wheat

Food energy supplied by these four: At least 50 percent

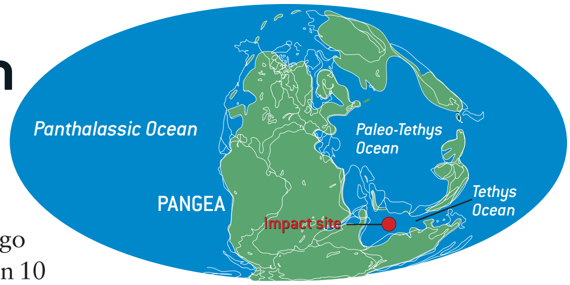
Number of governments that negotiated the treaty: 164

Number of ratifications needed: 40

SOURCES: Intermediate Technology Development Group (www.itdg.org); United Nations Food and Agriculture Organization

EXTINCTIONS

Permian Percussion



IMPACT SITE is shown in red, among the continents as they appeared 250 million years ago.

The impact that may have triggered the largest extinction in the earth's history may have struck near Australia. The end-Permian mass extinction 250 million years ago wiped out seven in 10 land species and nine in 10 marine species, far worse than the K-T extinction that claimed the dinosaurs 65 million years ago. A team of scientists contends that a buried 125-mile-wide crater called Bedout, off northwestern Australia, resulted from a collision with a Mount Everest–size meteor, not from volcanism as previously thought. Seafloor rock samples dating roughly to the end Permian reveal glass inside crystal, a feature the researchers say occurred because of melting induced by shock waves from the impact. The team also found quartz fractured in multiple directions, a potential sign of cosmic strike; volcanic activity fractures quartz along one direction. The report, appearing online May 13 from *Science*, noted that the putative end-Permian and K-T impacts both might have initiated large-scale volcanism. *—Charles Choi*

EARTH SCIENCE

Hot Stuff Coming Through

One debate in the global warming issue involves past discrepancies in data. Satellite readings of the troposphere—the atmospheric layer closest to Earth—showed a warming trend of less than 0.1 degree Celsius per decade, far smaller than surface temperatures suggested. Evidently, the satellite data were off because the stratosphere above the troposphere disguised warming trends. Scientists at the University of Washington analyzed measurements from U.S. National Atmospheric and Oceanic Administration satellites from 1979 to 2001. The probes measured microwaves emitted by atmospheric oxygen to determine its temperature. About one fifth of the troposphere signals actually came from the stratosphere, which is cooling about five times as fast as the troposphere is warming. After the researchers compensated for this stratospheric effect, they reported in the May 6 *Nature* that satellite readings closely resembled surface temperature measurements: they both predict an overall global warming of about 0.17 degree C per decade. *—Charles Choi*

ENVIRONMENT

When Air Quality Hits "Mutant"

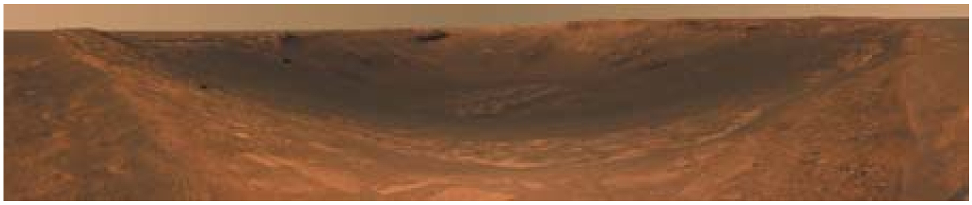
Air pollution can trigger heritable changes, according to studies in birds and rodents. To find out which components of air pollution are mutagens, scientists at McMaster University and their colleagues exposed two groups of lab mice to air at a location near two steel mills and a major highway. One group, however, breathed air passed through a HEPA filter. This experiment was repeated in a rural area with two other groups of mice. After 10 weeks of exposure, the mice were bred. The offspring of mice that breathed unfiltered, polluted air inherited mutations twice as often from their fathers as the offspring from any of the other three groups. The researchers suggest in the May 14 *Science* that the culprits are microscopic airborne particles of soot and dust that frequently have toxins such as polycyclic aromatic hydrocarbons attached. *—Charles Choi*

SMOG—seen here hovering over Los Angeles—can affect genes.

LUCY READING, AFTER MAP BY CHRISTOPHER R. SCOTESE (*top*); ROBERT LANDAU *Corbis* (*bottom*); ILLUSTRATION BY MATT COLLINS

MARS

Rover and Over



ENDURANCE CRATER on Mars is rendered in "true" color—the way the human eye would perceive it.

Having already recovered spectacular signs of dried-up Martian water, NASA has extended the mission of the twin rovers Spirit and Opportunity. Opportunity trundled for six weeks from Eagle Crater to a deeper crater called Endurance, whose layers of exposed bedrock may give clues to the lifetime and size of the surrounding ancient sea. It reached the crater rim in early May and will circle the 130-meter-wide, stadium-size divot, taking panoramic pictures of its walls and bottom. The rover may also drive down inside to study the crater's mineralogy and chemistry. Half a world away, Spirit has set off across the plains of Gusev Crater and was slated to have arrived at Columbia Hills by mid-June. The missions could last for hundreds of Martian days before enough dust settles onto the rovers' solar panels and thereby chokes off electrical power, says Matt Golombek, member of the rover science team at the Jet Propulsion Laboratory in Pasadena, Calif. *—JR Minkel*

BIOLOGY

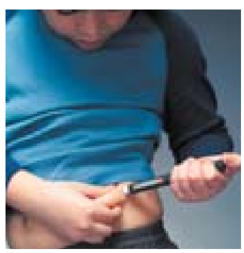
Mixing in Mitochondria

The energy-producing mitochondria mimic the nucleus in that they have DNA as well. The parallel with nuclear DNA has now gotten stronger. Konstantin Khrapko of Harvard Medical School and his colleagues analyzed mitochondrial DNA (mtDNA) from an individual whose muscle cells contained 10 percent maternal mtDNA and 90 percent paternal—unusual because mtDNA in sperm is ordinarily destroyed during embryonic development, meaning that mtDNA normally passes only from mother to child. The researchers detected mixing, or recombination, between the two lineages in 0.7 percent of the DNA molecules. "It has exciting implications for mitochondrial DNA repair and replication," Khrapko says, noting that cells use one type of recombination to help repair damage to nuclear DNA. It is unclear whether such mixing affects "molecular clocks" based on mtDNA and used to track ancestral human movements; the mixing could be unique to this case. See the May 14 *Science*. *—JR Minkel*

DIABETES

Beta from Beta

In type 1, or juvenile-onset, diabetes, the immune system destroys insulin-producing beta cells, causing a lifelong dependency on insulin therapy. Replacing those beta cells, thereby curing diabetes, seemed possible: previous studies hinted that the body has stem cells in the pancreas that give rise to the cells. Investigators at Harvard University now suggest that beta cells can proliferate by duplicating themselves. They engineered mice to have beta cells possessed of a genetic marker they could switch on with the drug tamoxifen. After the mice were given the drug, all the new beta cells had activated marker genes, indicating that they came from preexisting beta cells. Pancreatic stem cells may still exist but perhaps give rise to only a small fraction of beta cells. If human beta cells originate like their mouse counterparts, treating diabetes might mean boosting the growth of remaining beta cells. The study appears in the May 6 *Nature*. *—Charles Choi*



INSULIN doses would be unneeded if beta cells could be replenished.

BRIEF POINTS

- **Microscopic bits of nylon, polyester and other plastics are spread throughout marine habitats, such as beaches and ocean floors. The environmental consequences are still unknown.**

  Science, May 7, 2004

- **A new kind of lunar mineral, an iron-silicon substance named hapkeite, was made when micrometeorite impacts on the moon vaporized bits of metal that redeposited on rocks.**

  Proceedings of the National Academy of Sciences USA, May 4, 2004

- **Food allergy cause? Dendritic cells, which recognize foreign proteins, normally die once they activate attacking T cells. In food allergy, dendritic cells remain alive, suggesting they keep T cells revved up.**

  Immunology, May 2004

- **Dolphins swim fast thanks to the soft, flaky skin they shed every two hours. Both softness and shedding reduce drag resulting from turbulent vortices that form next to their bodies.**

  Journal of Turbulence, May 2004

NASA/JPL/CORNELL (*top*); CONEYL JAY *Science Photo Library* (*bottom*)

Overcoming Self

A company tries to turn the immune system against cancer   By GARY STIX

The long projections that stretch out from the dendritic cell give it its name, one derived from the Greek word for "treelike." The job of the dendritic cell is that of an educator. This elite member of the human immune system grabs a piece of a foreign invader (an antigen)—whether from a virus, bacterium or another organism—and sends out an alarm. It waves a piece of antigen that acts as a signal so that T cells can rush in and dispatch the interloper.

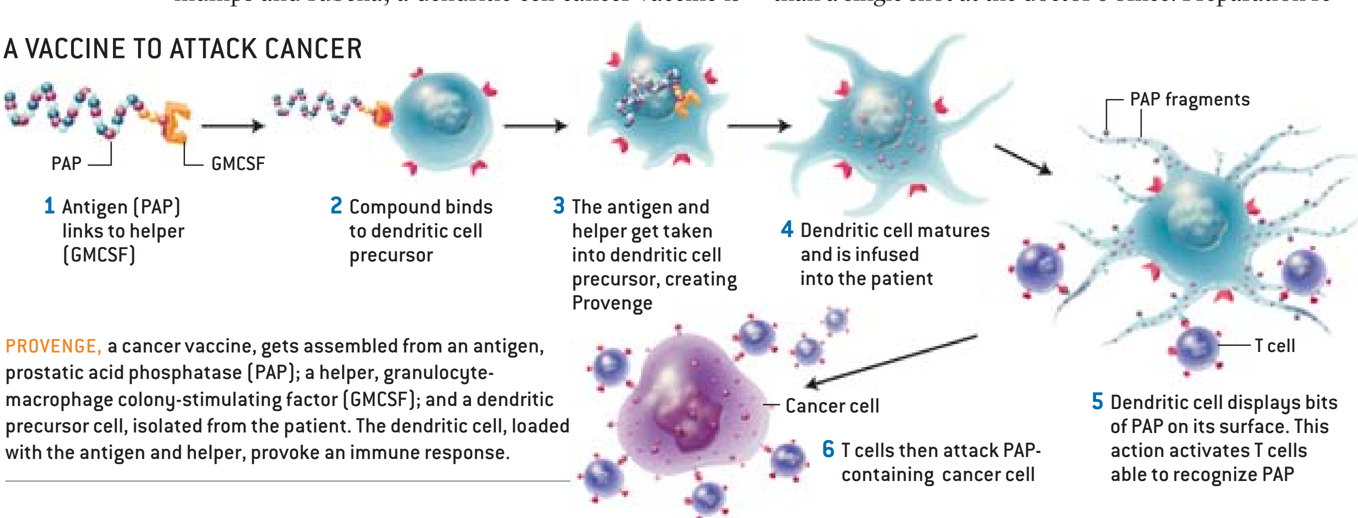
In principle, the actions of the dendritic cell suggest a wholly new approach to cancer therapeutics, except for one hitch. In the terminology of immunologists, cancer cells are "self"—not encroaching outsiders. A late-stage clinical trial of a cancer vaccine using dendritic cells may be completed by next year and may prove whether or not such a drug can overcome self.

Dendreon—which occupies the former West Coast research building of Bristol-Myers Squibb, a glass-and-steel structure a few blocks from the Seattle waterfront—has been put on fast-track status by the U.S. Food and Drug Administration to test a vaccine against prostate cancer in a Phase III study (testing in large numbers of humans). Unlike a vaccine for measles, mumps and rubella, a dendritic cell cancer vaccine is not used as a preventative but as a therapy for patients who have already acquired cancer. The dendritic cell is the only one capable of triggering an initial T cell immune response. So, used in therapy, it is potentially more effective than other approaches to cancer vaccines that just inject antigens directly.

Dendreon's experience illustrates both the ups and downs of such sophisticated immunological approaches against malignancies. Research began by ferreting out an antigen that could be displayed by a dendritic cell. It turns out that an enzyme, prostatic acid phosphatase (PAP), is found almost exclusively in prostate cancer cells in men with an advanced state of the disease. Cancer cells' innate ability to evade the immune system means that an antigen alone does not suffice to call in the troops. PAP needs a helper. A type of molecule called a cytokine—in this case, granulocyte-macrophage colony-stimulating factor (GMCSF)—is fused to PAP. GMCSF appears to ensure that dendritic cells take up the fusion protein—made up of PAP and GMCSF—and that they send out a danger signal to recruit T cells, thus creating an autoimmune response to the patient's cancer cells.

Cancer vaccination with dendritic cells involves more than a single shot at the doctor's office. Preparation re-

A VACCINE TO ATTACK CANCER



PROVENGE, a cancer vaccine, gets assembled from an antigen, prostatic acid phosphatase (PAP); a helper, granulocyte-macrophage colony-stimulating factor (GMCSF); and a dendritic precursor cell, isolated from the patient. The dendritic cell, loaded with the antigen and helper, provoke an immune response.

ALICE Y. CHEN

quires isolating the patient's white blood cells and then separating out a mix of dendritic cells and precursor cells still undergoing maturation (both types may produce an immune reaction). This melange is cultured with fusion proteins for 40 hours and afterward is injected back into the patient, a process that is repeated on three occasions. "Some people look at this and say that they couldn't imagine a more complicated way to do things," says David Urdal, Dendreon's chief scientific officer. But getting the antigen to the dendritic cells, which are normally widely dispersed in the body, by injecting it directly has proved an ineffective form of administration.

Clinical trials with the drug, called Provenge, did not at first go according to plan. Dendreon's stock plunged in early 2002 after an analysis by a company statistician estimated that the drug might not meet the trial's goal of delaying disease progression for patients for whom hormone and other therapies had already failed. After the 127-patient trial ended—this first-ever Phase III trial of a dendritic cell cancer vaccine just having missed achieving statistical significance—researchers took a closer look at the data.

A subset of the patients, those with a less aggressive form of the disease (seven or less on the Gleason scale), had improved significantly on the therapy with only relatively minor side effects. The additional analysis suggested that this population of patients, which accounts for 75 percent of the 75,000 men with late-stage prostate cancer (patients who failed hormone therapy), could benefit. But the FDA prefers not to use after-the-fact analyses as the basis for approving a drug. Investigators can analyze and reanalyze the data until they find a group of patients that seems to have gotten better. To have merit, a solid clinical trial must meet the goal, or end point, that it establishes at its outset. So Dendreon went back to the FDA and got approval to enroll an additional 275 patients whose Gleason scores registered seven or less as part of a second trial of Provenge that was already under way.

Dendreon also continued to follow patients from its first trial. In January of this year, it reported at an industry meeting on projections that patients with the lower Gleason scores would experience a longer survival time—8.4 months more than patients receiving a placebo. Moreover, 53 percent of the Provenge patients were still alive at 30 months, compared with 14 percent taking a placebo. The presentation sent the company's stock soaring and facilitated Dendreon's raising of $150 million in the public markets in January.

Cancer vaccinologists often compare their field with that of monoclonal antibodies a decade ago, when those molecules that can, say, block a specific receptor on a cell had fallen into disrepute. In recent years, monoclonals such as Rituxan (for lymphoma) and Herceptin (for breast cancer) have staged a rousing comeback, and they now represent the vanguard of cancer therapies.

Not everyone concurs with this comparison. Monoclonals are often described as weapons that attack a particular target. "With a monoclonal, you have a smart bullet against cancer cells," says Matthew Geller, senior biotechnology analyst for CIBC World Markets. "With a vaccine, you're trying to teach the immune system to go after cancer cells, something it hasn't been able to figure out in millions and millions of years of evolution."

Cancer vaccinologists hope their nascent discipline follows the trajectory of monoclonal antibodies, which triumphed after a series of early setbacks.

Another cancer vaccine researcher also had qualms. Pramod K. Srivastava, a professor at the University of Connecticut School of Medicine and a founder of Antigenics, a cancer vaccine company, questions whether there is any evidence in the scientific literature that the antigen PAP induces a protective immune response, although Dendreon emphasizes that the antigen must be fused with the cytokine GMCSF and loaded into a dendritic cell to create autoimmunity. Srivastava contends that the cytokine or the dendritic cell itself might be producing some immune response, not the PAP. "The antigen might simply be there for the ride," he says.

The results of Dendreon's new trial, which are scheduled to be available next year, will show whether or not the skeptics are right. But so far the findings are good enough that the company has been in discussions with both a pharmaceutical and a biotechnology firm for a marketing collaboration for Provenge. Cancer vaccine researchers sometimes cautiously use the "C" word. The possibility of a curative effect stems from the idea that a vaccine might be able to raise a lasting response by the immune system to cancer cells, unlike a short-lived chemical therapy.

Urdal of Dendreon makes a slightly different pitch for the future of cancer vaccines. He speculates that long-lasting immunity might turn cancer from a terminal disease into a chronic one that stabilizes and allows patients to live out their lives, even though traces of malignancy might remain. Cancer vaccines could then move to the forefront of immunotherapies that would treat the disease by using the body's own defenses rather than by plying patients with toxic chemicals. SA

If It's Broke, Fix It

Two economists propose solutions for patent system reform By GARY STIX

In the 19th century the Netherlands issued patent after patent that was neither novel nor practical, a situation that would be familiar to anyone in the contemporary U.S. patent community. The Dutch parliament back then adopted an unusual approach to the problem: in 1869 it voted 49 to eight to abolish patenting, a decision that was not rescinded until 1910, under heavy pressure from the country's trading partners. The U.S. has never gone ahead with such a radical step. But reformists continue to debate a multitude of ideas on how patent quality can be improved.

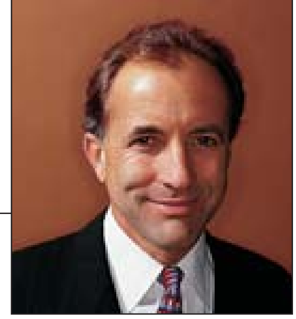
Later this year a book by Adam B. Jaffe of Brandeis University and Josh Lerner of Harvard Business School will describe what is wrong with the current system and then outline how it might be revamped. The book—*Innovation and Its Discontents: How Our Broken Patent System Is Endangering Innovation and Progress, and What to Do about It* (Princeton University Press)—is intended as an antidote to structural changes in the patent system made during the past two decades that have dramatically increased the rights of patent holders [see "The Silent Revolution," by Gary Stix; Staking Claims, June].

Under the system envisaged by Jaffe and Lerner, most patents would get only a cursory examination, because they would raise few objections given that they are "economically unimportant." For instance, patent 6,701,872—"a method and apparatus for automatically exercising a curious animal"—is unlikely to be contested by another inventor. Excepting the large number of such patents will let examiners devote more time to a few critical cases. When an examiner decided that a patent should be issued, a "pre-grant opposition" process would begin that allowed others to point to previous technology—"prior art"—that shows a patent is not new or inventive (obvious) and therefore should not be granted. If such a procedure had been in place in 1999, Vergil L. Daughtery III of Americus, Ga., would probably not have been able to get his first patent for a certain type of financial instrument—"an expirationless option"—that had been anticipated during the 1960s in papers written by economist Paul Samuelson of the Massachusetts Institute of Technology.

Even if no prior art can be found and a patent is granted, it may still not meet the requisite obviousness standard. A case in point is Amazon.com's "One Click" patent, issued in 1999, for making online purchases. Although there was no preexisting patent or technical paper, it gave Amazon an exclusive right to a practice that was widely used in the software industry at the time. To avoid such blunders, a procedure must be put in place to allow for reexamination once a patent has already been issued. Existing review procedures—whether before or after issuance—are simply inadequate, Jaffe and Lerner assert.

Even if these new procedures are instituted, bad patents will still be granted—and suits will be brought to invalidate them. Current law gives the patent holder an advantage because it presumes that a patent is valid and places the onus on the plaintiff to present "clear and convincing" evidence that an error has been made by the examiners. The requirement for a trial by jury complicates these cases, because juries often have difficulty grasping the intricacies of both the technology and the subtleties of patent law. Uncertain about whether the burden of proof has been met, juries are as likely as not to side with the patent holder, making it difficult to mount effective court challenges. Jaffe and Lerner suggest that judges, not juries, rule in these cases, increasing the likelihood that plaintiffs will get a fair hearing. All these changes, the authors contend, will go a long way toward ensuring that the system is not biased against those who question bad patents. SA

JENNIFER KANE



God's Number Is Up

Among a heap of books claiming that science proves God's existence emerges one that computes a probability of 67 percent By MICHAEL SHERMER

In his 1916 poem "A Coat," William Butler Yeats rhymed: "I made my song a coat/Covered with embroideries/Out of old mythologies/From heel to throat."

Read "religion" for "song," and "science" for "coat," and we have a close approximation of the deepest flaw in the science and religion movement, as revealed in Yeats's denouement: "But the fools caught it,/Wore it in the world's eyes/As though they'd wrought it./Song, let them take it/For there's more enterprise/In walking naked."

Naked faith is what religious enterprise was always about, until science became the preeminent system of natural verisimilitude, tempting the faithful to employ its wares in the practice of preternatural belief. Although most efforts in this genre offer little more than scientistic cant and religious blather, a few require a response from the magisterium of science, if for no other reason than to protect that of religion; if faith is tethered to science, what happens when the science changes? One of the most innovative works in this genre is *The Probability of God* (Crown Forum, 2003), by Stephen D. Unwin, a risk management consultant in Ohio, whose early physics work on quantum gravity showed him that the universe is probabilistic and whose later research in risk analysis led him to this ultimate computation.

If faith is tethered to science, what happens when the science changes?

Unwin rejects most scientific attempts to prove the divine—such as the anthropic principle and intelligent design—concluding that this "is not the sort of evidence that points in either direction, for or against." Instead he employs Bayesian probabilities, a statistical method devised by 18th-century Presbyterian minister and mathematician Reverend Thomas Bayes. Unwin begins with a 50 percent probability that God exists (because 50–50 represents "maximum ignorance"), then applies a modified Bayesian theorem:

$$P_{after} = \frac{P_{before} \times D}{P_{before} \times D + 100\% - P_{before}}$$

The probability of God's existence after the evidence is considered is a function of the probability before times D ("Divine Indicator Scale"): 10 indicates the evidence is 10 times as likely to be produced if God exists, 2 is two times as likely if God exists, 1 is neutral, 0.5 is moderately more likely if God does not exist, and 0.1 is much more likely if God does not exist. Unwin offers the following figures for six lines of evidence: recognition of goodness (D = 10), existence of moral evil (D = 0.5), existence of natural evil (D = 0.1), intranatural miracles (prayers) (D = 2), extranatural miracles (resurrection) (D = 1), and religious experiences (D = 2).

Plugging these figures into the above formula (in sequence, where the P_{after} figure for the first computation is used for the P_{before} figure in the second computation, and so on for all six Ds), Unwin concludes: "The probability that God exists is 67%." Remarkably, he then confesses: "This number has a subjective element since it reflects *my* assessment of the evidence. It isn't as if we have calculated the value of pi for the first time."

Indeed, based on my own theory of the evolutionary origins of morality and the sociocultural foundation of religious beliefs and faith, I would begin (as Unwin does) with a 50 percent probability of God's existence and plug in these figures: recognition of goodness (D = 0.5), existence of moral evil (D = 0.1), existence of natural evil (D = 0.1), intranatural miracles (D = 1), extranatural miracles (D = 0.5), and religious experiences (D = 0.1). I estimate the probability that God exists is 0.02, or 2 percent.

Regardless, the subjective component in the formula relegates its use to an entertaining exercise in thinking—on par with mathematical puzzles—but little more. In my opinion, the question of God's existence is a scientifically insoluble one. Thus, all such scientistic theologies are compelling only to those who already believe. Religious faith depends on a host of social, psychological and emotional factors that have little or nothing to do with probabilities, evidence and logic. This is faith's inescapable weakness. It is also, undeniably, its greatest power. SA

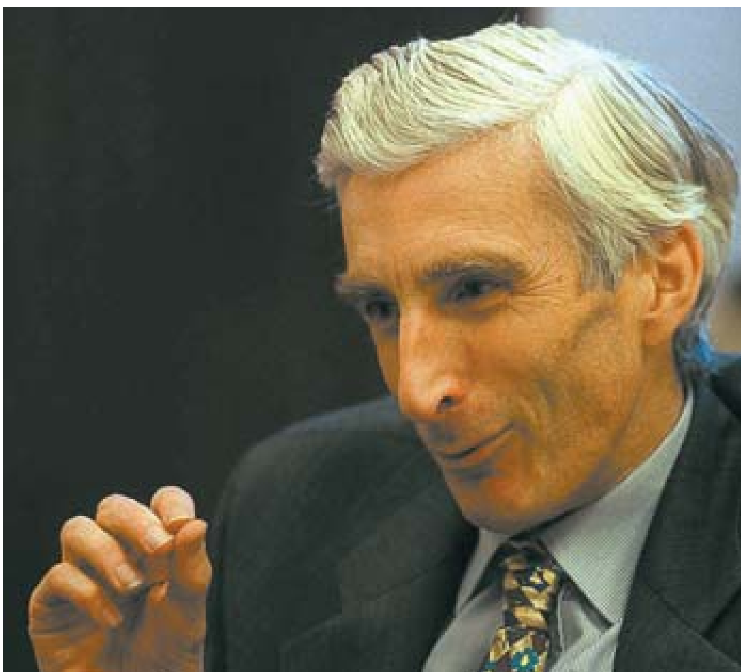
Michael Shermer is publisher of Skeptic *(www.skeptic.com) and author of* The Science of Good and Evil.

BRAD HINES

Doom and Gloom by 2100

Unleashed viruses, environmental disaster, gray goo—astronomer Sir Martin Rees calculates that civilization has only a 50–50 chance of making it to the 22nd century By JULIE WAKEFIELD

Death and destruction are not exactly foreign themes in cosmology. Black holes can rip apart stars; unseen dark energy hurtles galaxies away from one another. So maybe it's not surprising that Sir Martin Rees, Britain's Astronomer Royal, sees mayhem down on Earth. He warns that civilization has only an even chance of making it to the end of this century. The 62-year-old University of Cambridge astrophysicist and cosmologist feels so strongly about his grim prognostication that last year he published a popular book about it called *Our Final Hour*.



ANGELA ROWLINGS *AP Photo*

SIR MARTIN REES: LIFE AMONG STARS

- Knighted in 1992; became Astronomer Royal in 1995.
- Career choice in an alternative universe: music composer.
- Has bet $1,000 that a bioterror or "bioerror" incident will claim one million lives by 2020 (see www.longbets.org/9).
- "We can't enjoy the benefits of science without confronting the risks."

The book (entitled *Our Final Century* in the U.K.) represents a distillation of his 20 years of thinking about cosmology, humankind and the pressures that have put the future at risk. In addition to considering familiar potential disasters such as an asteroid impact, environmental degradation, global warming, nuclear war and unstoppable pandemics, Rees thinks science and technology are creating not only new opportunities but also new threats. He felt compelled to write *Our Final Hour* to raise awareness about both the hazards and the special responsibilities of scientists.

As one himself, Rees was among the first to posit that giant black holes power quasars, and his work on quasar distribution helped to refute the theory that the cosmos exists in a steady state. Rees directed Cambridge's Institute of Astronomy until 1992; he then served for a decade as a Royal Society Research Professor before assuming the mastership of Cambridge's Trinity College. Since 1995 Rees has also held the honorary title of U.K. Astronomer Royal, once an active post based at Greenwich Observatory and first held by John Flamsteed and then Edmond Halley.

Astronomers are well positioned to ponder the fate of humanity, Rees insists, because they have a unique vantage point in terms of the vast timescales of the future. "Astronomers have a special perspective to see ourselves as just a part of a process that is just beginning rather than having achieved its end," he says. "And perhaps this gives an extra motive to be concerned about what happens here on Earth in this century."

Innovation is changing things faster than ever before, and such increasing unpredictability leaves civilization more vulnerable to misadventure as well as to disaster by design. Advances in biotechnology, in terms of both increasing sophistication and decreasing costs, means that weaponized germs pose a huge risk. In a wager he hopes to lose, Rees has bet $1,000 that a biological incident will claim one million lives by 2020. "In this increasingly interconnected world where individuals have more power than ever before at their fingertips, society should worry more about some kind of massive calamity, however improbable," Rees states.

In calculating the coin-flip odds for humanity at 2100, Rees adds together those improbabilities, including those posed by self-replicating, nanometer-size robots. These nanobots might chew through organic matter and turn the biosphere into a lifeless "gray goo," a term coined by nanotech pioneer K. Eric Drexler in the 1980s. Gray goo achieved more prominence last year after Prince Charles expressed concern about it and Michael Crichton used it as the basis for his novel *Prey*.

It's not just out-of-control technology that has Rees worried. Basic science can present a threat. In July 1999 *Scientific American* ran a letter by Princeton University physicist Frank Wilczek, who pointed to "a speculative but quite respectable possibility" that the Brookhaven National Laboratory's Relativistic Heavy Ion Collider (RHIC) could produce particles called strangelets. These subatomic oddities could grow by consuming nearby ordinary matter. Soon after, a British newspaper posited that a "big bang machine"—that is, RHIC—could destroy the planet.

The ensuing media flurry led then Brookhaven director John H. Marburger to pull together an outside panel of physicists, who concluded that the strangelet scenario was remote, about a one-in-50-million chance of killing six billion people. (Another panel, convened by CERN near Geneva, drew a similar conclusion.) In *Our Final Hour,* Rees noted that the chances can be expressed differently—namely, that 120 people might die from the RHIC experiments. He thinks experts should debate in public the merits and risks of such work.

Some researchers were not pleased with Rees's position. Subir Sarkar, a University of Oxford cosmologist who considers Rees a true "guru" for his wide-ranging perspective and contributions to astrophysics and cosmology, contends nonetheless that Rees was "irresponsible in making a big deal of the negligible probability" connected with the particle collisions at RHIC. Rees acknowledges that other doomsday scenarios rank much higher in terms of a "risk calculus." Yet he maintains that if the safety criteria used for nuclear reactors are applied—in terms of maximum acceptable probability of deaths multiplied by number at risk—the probability of global catastrophe from any particle acceleration experiment would need to be below about one in a trillion.

Perhaps more important than his *Our Final Hour* arguments is Rees's ability to popularize technical subjects. "He is, by any account, one of the clearest and most readable expositors of current science to the general public," asserts friend and colleague Peter Meszaros, a Pennsylvania State University astrophysicist. Rees has written six books for the lay reader (as well as several *Scientific American* articles).

It's possible to tip the balance to civilization's advantage, Rees concludes, believing that environmental and biomedical issues should be higher on the political agenda. To raise the debate above the level of rhetoric, however, the public must be better informed. He looks to the U.S. to take a leadership role. But so far he finds its handling of the controversies over stem cell research and global warming to be wanting: the U.S. "has been rather remiss in tackling issues that are taken more seriously elsewhere in the world, especially environmental problems."

If humanity loses, would it really matter to the rest of the universe? Life exists thanks to a happy combination of physical constants. Tweak a few, and life as we know it becomes impossible. Those who ponder whether we were *meant* to be here or whether our universe is part of a multiverse, consisting of universes with different physical parameters, sometimes invoke the anthropic principle. It basically states that the universe must be able to

APOCALYPSE SOON? A 2003 bioterror drill in Cambridge, Mass.

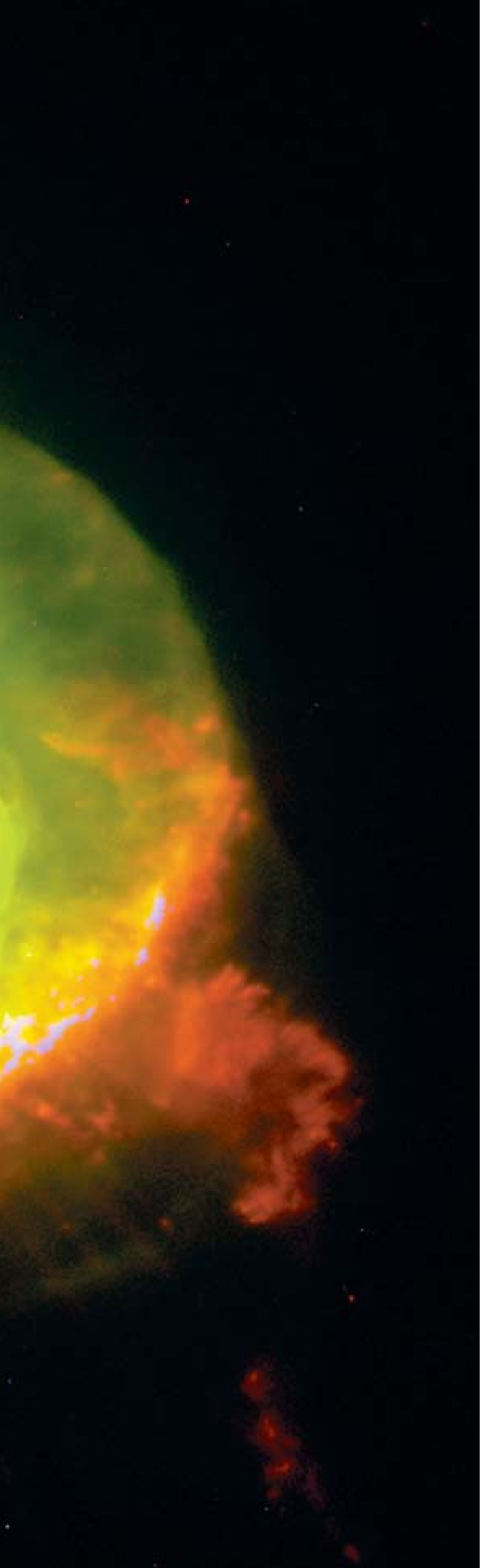
KRT/CHICAGO TRIBUNE

spawn intelligent life because we are here to observe it. "Anthropic reasoning will be irrelevant if the 'final theory' defines all the constants of physics uniquely, but unavoidable if it doesn't," Rees states. "The latter option is favored by an increasing proportion of theorists"—in other words, science may be able to explain the numbers only with an anthropic argument.

Anthropic reasoning would seem to cast a supernatural pall over science. But Rees doubts that revelations from cosmology will ever resolve the controversy between science and religion. For a start, he sees no qualitative change in the debate since Newton's time: scientific explanations remain perpetually incomplete. "If we learn anything from the pursuit of science, it is that even something as basic as an atom is quite difficult to understand," Rees declares. "This alone should induce skepticism about any dogma or any claim to have achieved more than a very incomplete and metaphorical insight into any profound aspect of our existence." Or nonexistence, depending on the coin flip. SA

Julie Wakefield is based in Washington, D.C.

CAT'S EYE NEBULA (NGC 6543) is one of the galaxy's most bizarre planetary nebulae—a multilayered, multicolored gas cloud some 3,000 light-years from the sun. Such nebulae have nothing to do with planets; the term is a historical vestige. Instead they are the slowly unfolding death of modest-size stars. Our own sun will end its life much like this. The intricacy of the Cat's Eye, seen by the Hubble Space Telescope in 1994, sent astronomers scrambling for an explanation.



J. P. HARRINGTON *University of Maryland*, K. J. BORKOWSKI *North Carolina State University* AND NASA

THE EXTRAORDINARY Deaths OF ORDINARY Stars

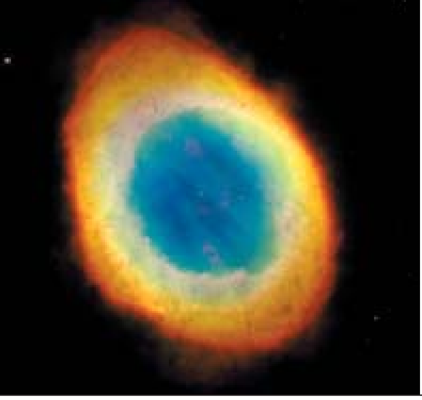
The demise of the sun in five billion years will be a spectacular sight. Like other stars of its ilk, the sun will unfurl into nature's premier work of art: a planetary nebula

By Bruce Balick and Adam Frank

Within easy sight of the astronomy building at the University of Washington sits the foundry of glassblower Dale Chihuly. Chihuly is famous for glass sculptures whose brilliant flowing forms conjure up active undersea creatures. When they are illuminated strongly in a dark room, the play of light dancing through the stiff glass forms commands them to life. Yellow jellyfish and red octopuses jet through cobalt waters. A forest of deep-sea kelp sways with the tides. A pair of iridescent pink scallops embrace each other like lovers.

For astronomers, Chihuly's works have another resonance: few other human creations so convincingly evoke the glories of celestial structures called planetary nebulae. Lit from the inside by depleted stars, fluorescently colored by glowing atoms and ions, and set against the cosmic blackness, these gaseous shapes seem to come alive. Researchers have given them such names as the Ant, the Starfish Twins and the Cat's Eye. Hubble Space Telescope observations of these objects are some of the most mesmerizing space images ever obtained.

Planetary nebulae were named, or rather misnamed, two cen-



RING NEBULA (M 57)

Like all great art, planetary nebulae do more than captivate us. They cause us to question our perception of the world.

turies ago by English astronomer William Herschel. He was a prodigious discoverer of nebulae—fuzzy, cloudlike objects visible only through a telescope. Many had a vaguely round shape that reminded Herschel of the greenish planet Uranus (which he had discovered), and he speculated they might be planetary systems taking shape around young stars. The name stuck even though the opposite turned out to be true: this type of nebula consists of gas molted from dying stars. It represents not our past but our future and our fate. In five billion years or so the sun will end its cosmic tenure in the elegant violence of a planetary nebula.

Like all great art, planetary nebulae do more than captivate us. They cause us to question our perception of the world. In particular, they pose challenges to stellar evolution theory, the physics that describes the life story of stars. This theory is a mature, supposedly well developed branch of science, one of the foundations on which all our understanding of the cosmos is based. Yet it has trouble accounting for the complex figures evident in the Hubble images. If stars are born round, live round and die round, how do they create such elaborate patterns as ants, starfish and cat's-eyes?

Death Becomes Them

OVER THE PAST CENTURY, astronomers have come to realize that stars cleanly separate into two distinct classes as they die. The elite massive stars—those with a birth weight exceeding eight solar masses—explode suddenly as supernovae. More modest stars, such as the sun, have a drawn-out death. Instead of detonating, they spend their last years burning their fuel spasmodically, like an automobile engine running out of gas.

Nuclear reactions in such a star's core, the source of power for nearly its entire life, deplete the available hydrogen, then the helium. As the nuclear burning moves outward to the fresh material in a shell surrounding the core, the star bloats into a so-called red giant. When the hydrogen in the shell, too, is exhausted, the star takes to fusing helium there. In the process, it becomes unstable. Deep convulsions, combined with the pressure of radiation and other forces, heave the distended and loosely bound surface layers outward into space, creating a planetary nebula.

Since the 18th century, astronomers have imaged and catalogued about 1,500 planetary nebulae, and another 10,000 may lurk out there, hidden behind dense clouds of dust in our galaxy. Whereas a supernova goes off in the Milky Way every few centuries, a new planetary nebula forms every year, as hundreds of older ones fade into obscurity. Supernovae may be flashier, but their debris is roiling and chaotic, lacking the symmetry and intricacy of these nebulae.

Planetary nebulae are not as airy and tranquil as their images suggest. Au contraire, they are massive and tempestuous. Each contains the equivalent of about a third the mass of the sun, including nearly all of the star's remaining unburned nuclear fuel. Initially the loosely bound outer layers stream off the star at 10 to 20 kilometers per second—a relatively slow outflowing wind that will carry the bulk of the nebula's eventual mass. As the star strips down to its still hot core, it evolves from orange to yellow, then white, and finally blue. When its surface temperature exceeds about 25,000 kelvins, it bathes the surrounding gas in harsh ultraviolet light, which has enough punch to dismember molecules and strip atoms of their electrons.

The stellar wind carries ever less mass at ever increasing speed. After 100,000 to one million years, depending on the original mass of the star, it ceases altogether, and the remnant star settles down as an extremely dense and hot white dwarf—a stellar ember crushed by gravity into a nearly crystalline orb about the size of Earth.

Because the forces that are supposed to drive off mass from dying stars are spherically symmetrical, astronomers before the 1980s thought of planetary nebulae as expanding spherical bubbles [see

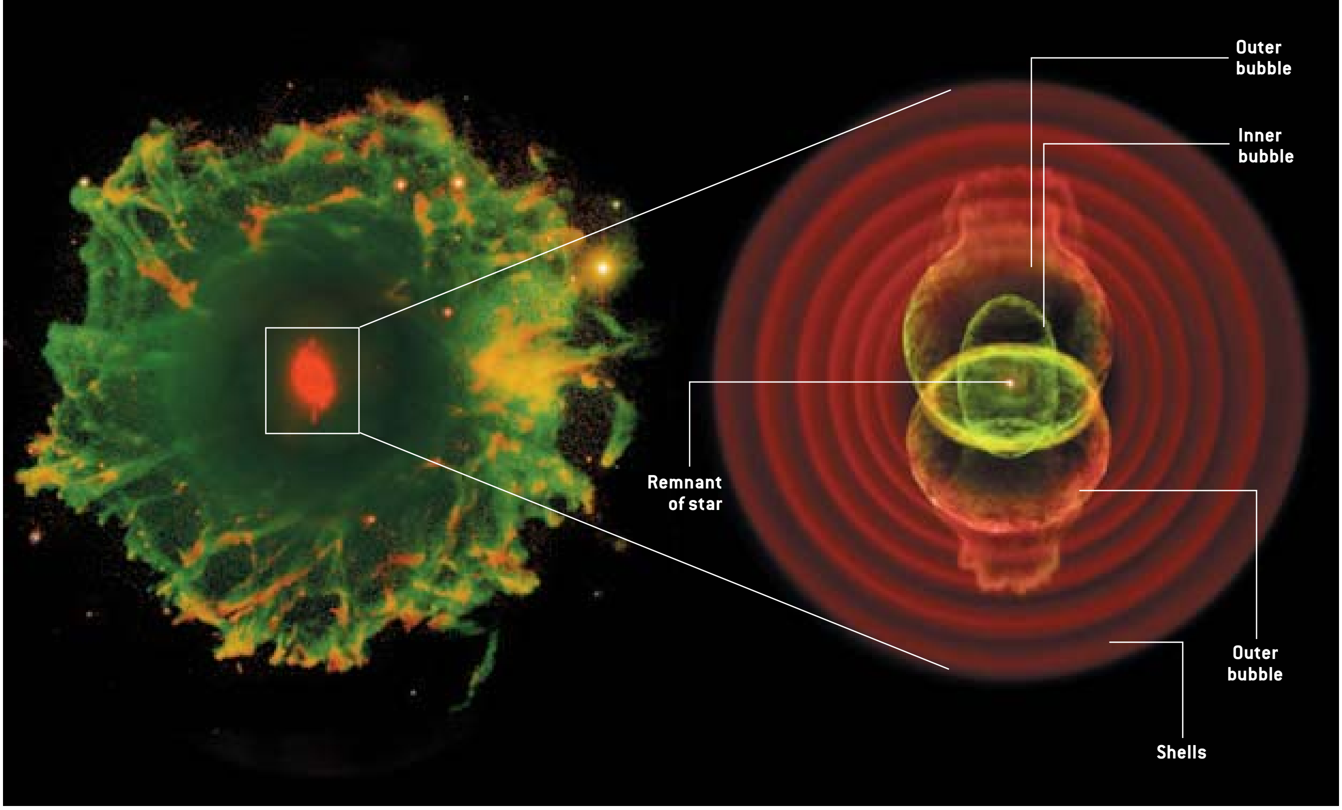
Overview/*Planetary Nebulae*

- Adorning the entire Milky Way like so many Christmas tree ornaments, planetary nebulae are the colorful remnants of modest stars—those less than eight solar masses. As these stars sputter toward death, they molt their outer layers in the form of a "wind" that blows outward at up to 1,000 kilometers per second. The stars gradually strip down to their deeper, hotter layers, the ultraviolet light from which ionizes the wind and causes it to fluoresce.
- Hubble Space Telescope images have revealed nebulae with surprisingly complex shapes, which are still only vaguely understood. Magnetic fields trapped in the core and released into the wind may play a role. So may close companion stars or large planets, whose tidal forces shepherd gas into giant rings that, in turn, funnel the wind into an hourglasslike shape.

NASA/ESA AND J. HESTER *Arizona State University*; NASA AND HUBBLE HERITAGE TEAM (STScI/AURA)

THE CAT'S EYE DISSECTED

The image on the preceding pages shows just part of the full glory that is the Cat's Eye. A ground-based telescope image (*left*) reveals the "eyelashes"—a ragged outer band of gas. The inner region, or "pupil," which an artist has reconstructed here (*right*), consists of a remnant star encased in an egg-shaped layer of gas, which in turn is surrounded by two off-center bubbles, which in turn are surrounded by concentric gas shells. Evidently the star ejected material in distinct stages over the course of millennia. The upper part of the nebula is tilted toward the viewer.



R. CORRADI *Institute of Astrophysics of the Canary Islands* (*left*); DON DIXON (*right*)

"Planetary Nebulae," by Martha and William Liller; SCIENTIFIC AMERICAN, April 1963]. Since then, the picture has steadily gotten far more complicated—and far more interesting.

Whistling in the Dark

THE FIRST SIGN THAT planetary nebulae are more than just stellar burps came in 1978, when ultraviolet observations showed that dying stars continue to blow winds long after they eject their outer gaseous layers. Though tenuous, these later winds top out at about 1,000 kilometers per second, 100 times as fast as the denser winds that preceded them.

To account for their effects, Sun Kwok of the University of Calgary, Christopher R. Purton of the Dominion Radio Astrophysical Observatory in Canada and M. Pim Fitzgerald of the University of Waterloo borrowed a stellar winds model that had been developed for other astrophysical phenomena. The idea is that when the fast winds ram into the slower ones upstream, a dense rim of compressed gas forms at the interface, much like the rim of snow at the front of a plow. The rim of gas surrounds a nearly empty (but very hot) cavity, and over time the fast wind clears out an ever larger volume.

This model, now called the interacting stellar winds hypothesis, works well for round or nearly round planetary nebulae. Observers in the 1980s, however, began to realize that round nebulae are the exception, perhaps just 10 percent of the total population. Many of the others have a prolate, or egglike, shape. The most spectacular, though rare, nebulae comprise two bubbles on opposite sides of the dying star. Astronomers call them "bipolar." "Butterfly" or "hourglass" would be a more vivid description.

To explain these shapes, the two of us, along with Vincent Icke and Garrelt Mellema, then both at Leiden University in the Netherlands, extended the interacting winds concept. Suppose that the slow winds first manage to form a dense torus orbiting the equator of the star. Later, this torus gently deflects the outflowing stellar winds in a polar direction. An elliptical nebula results. Hourglasslike nebulae are those with a very tight, very dense torus. The torus serves as a nozzle, as your lips do when you whistle, collimating your exhaled breath into a narrow jet of air. Similarly, the torus strongly deflects the fast winds, producing a mirror-image pair of jets or hourglass-shaped streams of gas.

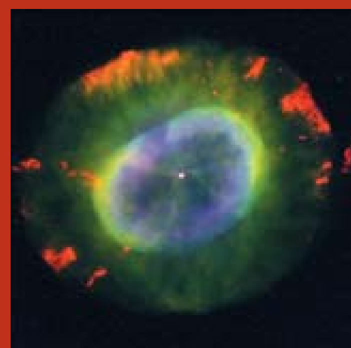
The model was simple, and it nicely fit

The Art of Planetary Nebulae

Hubble Space Telescope images have revealed planetary nebulae to be far more intricate and diverse than theorists ever expected.

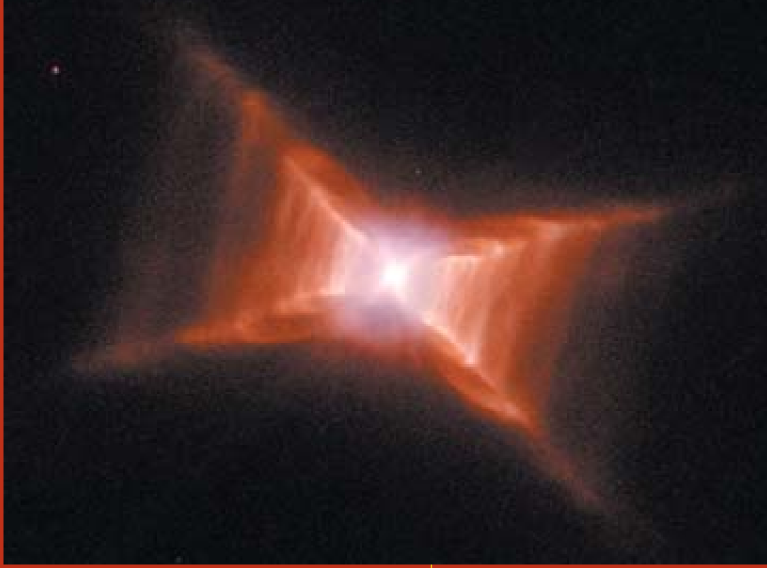
The Stingray nebula (Hen 3-1357), the youngest known planetary, started to glow just 20 years ago. A companion star and a torus of gas may account for its shape.

Encased in a dense, dusty, carbon-rich torus (*upper right*), the central star of the Bug nebula (NGC 6302) is one of the hottest known.



The Blue Snowball nebula (NGC 7662) contains so-called FLIERS (*red splotches*), fast-moving knots of gas of uncertain origin.

At the center of the Twin Jet nebula (M 2-9) are a binary star system and a gaseous disk 10 times the diameter of Pluto's orbit. Blue shows hydrogen ions; red, oxygen atoms; and green, nitrogen ions.



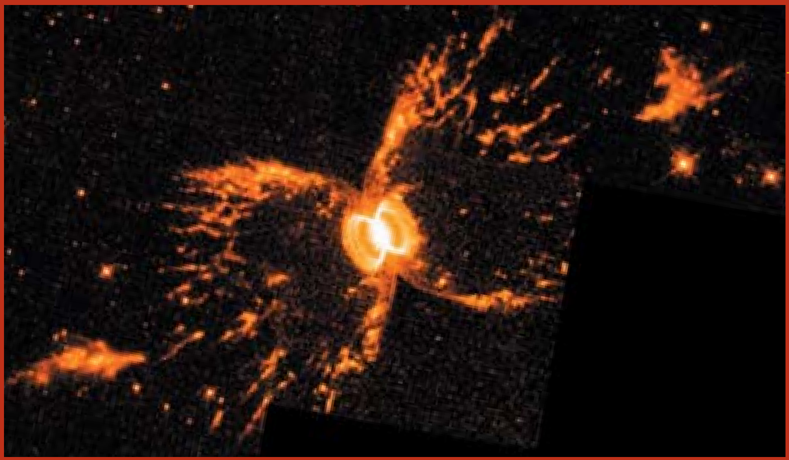
Gas streams out from the central star of the Ant nebula (Menzel 3) at 1,000 kilometers per second.

The Red Rectangle nebula (HD 44179) has a boxy shape because we are seeing nested cones of gas from the side. For an interactive image, visit www.spacetelescope.org/images/html/zoomable/heic0408a.html

Like a searchlight, the central star of the Egg nebula (CRL 2688) illuminates concentric shells of dust, which extend over a tenth of a light-year from the star. The colors represent light polarized in different directions.

The Dandelion Puff Ball nebula (NGC 6751) is an example of an elliptical planetary nebula. Red, green and blue correspond to weakly, moderately and strongly ionized gas, respectively.

This image of the Southern Crab nebula (He2-104), which captures the glow of nitrogen gas, reveals a small, bright nebula embedded in a larger one. The red giant that created the nebula is orbited by a white dwarf star.



ESA/NASA AND A. ZIJLSTRA *University of Manchester* (*Bug*); M. BOBROWSKY *Orbital Sciences Corp.* AND NASA (*Stingray*); B. BALICK AND J. ALEXANDER *University of Washington*, A.HAJIAN *U.S. Naval Observatory*, Y. TERZIAN *Cornell University*, M. PERINOTTO *University of Florence*, P. PATRIARCHI *Arcetri Observatory* AND NASA (*Blue Snowball*); B. BALICK, V. ICKE *Leiden University*, G. MELLEMA *Stockholm University* AND NASA (*Twin Jet*); NASA/ESA AND HUBBLE HERITAGE TEAM (STScI/AURA) (*Ant*); NASA/ESA, H. VAN WINCKEL *Catholic University of Leuven* AND M. COHEN *University of California, Berkeley* (*Red Rectangle*); NASA AND HUBBLE HERITAGE TEAM (STScI/AURA) (*Egg*); NASA AND HUBBLE HERITAGE TEAM (STScI/AURA) (*Dandelion Puff Ball*); R. CORRADI *Institute of Astrophysics of the Canary Islands*, M. LIVIO *STScI*, B. BALICK, U. MUNARI *Astronomical Observatory of Padua/Asiago*, H. SCHWARZ *Nordic Optical Telescope* AND NASA (*Southern Crab*)

Terra-Cotta

TOASTED BY THE RED GIANT SUN, the future Earth will at least be a good spot to watch the unfolding of a planetary nebula.

Planetary nebulae are a glimpse into the future of our own solar system. When the sun reaches the eleventh hour of its life, it will swell to the size of Earth's present orbit, causing Mercury and Venus to burn up like giant meteors. Earth will escape this fate because the sun will have blown out some of its material, weakening its gravity so that our planet slips into a new, larger orbit. The ochre-red sun will fill the noon sky. As one edge sets in the west, the other will begin to rise in the east. Though cooler than today (2,000 kelvins rather than 5,800 kelvins), it will still bake the planet's surface to a nice hard finish.

In these reduced circumstances, Earth will witness the formation of a planetary nebula from the inside. The sun will eject its outer layers in an extreme version of the present-day solar wind. Eventually the red behemoth will be stripped to its core, which will quickly settle down as a white dwarf star. Lit by this blue-hot pinprick, objects on Earth will cast sharp-edged, pitch-black shadows; sunrise and sunset will take no longer than an eyeblink. Exposed rock will turn to plasma as ultraviolet radiation from the dwarf destroys all molecular bonds, coating the surface with an eerie iridescent fog, constantly lifting and swirling. As the dwarf radiates away its energy, it will fade into a cold, dark cinder. Thus, our world will end first in fire, then in ice. *—B.B. and A.F.*

all the images available by 1993. Supercomputer simulations confirmed the viability of the basic idea, and new observations verified that the slow wind really did appear denser near the equator. We did not attempt to explain why the slow wind would be ejected as a torus, hoping that particular detail would be filled in later.

Our confidence in the model was quickly deflated. In 1994 Hubble took its first clear image of a planetary nebula, the Cat's Eye (designated NGC 6543), first seen by Herschel. That fateful picture blew us off our chairs. One of its two crossed ellipses, a thin rim surrounding an ellipsoidal cavity, matched the model. But what were all those other structures? No one had predicted that clumpy red-colored regions would lace the nebula; jetlike streaks immediately outside it were stranger still. At best, the interacting winds model could be just partly correct.

Sweating the Theoreticians

A POPULAR SCIENTIFIC IDEA is not easy to overturn, even when faced with images like those from Hubble. We went into professional denial, hoping the Cat's Eye was an anomaly. It was not. Other Hubble images soon established beyond doubt that some fundamental piece was missing from our picture of how stars die. Egos aside, this was the best place for scientists to find themselves. When cherished ideas are in ruins at your feet, nature is challenging you to look at the world anew: What have you missed? What have you not thought of before?

In such situations, it helps to focus on the most extreme cases, because they are where the unknown shaping forces may be operating most distinctly. Among planetary nebulae, the most extreme cases are the bipolar objects. The Hubble images of these objects look as if they had been taken from Georgia O'Keeffe's exquisite flower series. The small-scale features that dapple the nebulae come in mirror-image pairs, one on each side of the nebula. This reflection symmetry implies that the entire structure was assembled coherently by organized processes operating near the stellar surface, something like the making of a snowflake or sunflower.

For these objects, the interacting winds model makes a readily testable prediction: once gas leaves the torus, it

RON MILLER

Planetary nebulae are not as airy and tranquil as their images suggest. They are **massive and tempestuous**.

RETINA NEBULA (IC 4406)

NASA AND HUBBLE HERITAGE TEAM (STScI/AURA)

flows outward at a steady speed, which in turn produces a distinctive Doppler shift in the light emitted by the gas. Unfortunately, the model fails this test. In 1999 one of us (Balick) and Romano Corradi (now at the Institute of Astrophysics of the Canary Islands) and their collaborators used Hubble to study the Southern Crab nebula (designated He2-104). They found that its expansion velocity increased in proportion with distance from the star. The gas farthest away got there simply because it was moving the fastest. Extrapolating back in time, the lovely hourglasslike nebula seemed to have formed in a single eruption from the star about 5,700 years ago. That made the interacting winds model, which presumes that a continuous wind shapes the nebula, irrelevant.

Even stranger, Corradi and his colleagues found that the Southern Crab nebula was really two nebulae, one nested inside the other like Russian *matryoshka* dolls. We had guessed that the inner nebula was simply the younger of the two, but observations clearly showed that both nebulae had exactly the same pattern of increasing speed with distance. Thus, all of the complex structure must have formed during just one lavishly choreographed event six millennia ago. To this day, we puzzle over these findings.

The coffin lid of the interacting winds model was hammered shut in the late 1990s, when Kwok, Raghvendra Sahai and John Trauger of the Jet Propulsion Laboratory in Pasadena, Calif., Margaret Meixner of the University of Illinois and their co-workers published a new class of Hubble images. Their targets were very young planetary nebulae, caught before or shortly after the star ionized and heated them. Astronomers had expected that these objects would be smaller but otherwise similar versions of the more mature variety. Once again we were wrong: Embryonic and juvenile planetary nebulae have far more playful shapes. Their multiple axes of symmetry simply cannot be explained by the nozzle we had hypothesized. As Sahai and Trauger intimated in their 1998 paper on these objects, the time had come to find a different paradigm.

Stirring the Pot

THE OUTLINES OF fruitful theories for the shaping of planetary nebulae continue to emerge. The trick is to develop models that embrace the entire vexing array of observations. Researchers now agree that one of the principal players is the gravitational influence of companion stars. At least 50 percent of all the "stars" you see at night are really pairs of stars orbiting each other. In most of these systems, the stars are so far apart that they develop independently. But in a small fraction, the gravity of one star can deflect or even control the material flowing out of another. This fraction matches the fraction of planetary nebulae that are bipolar.

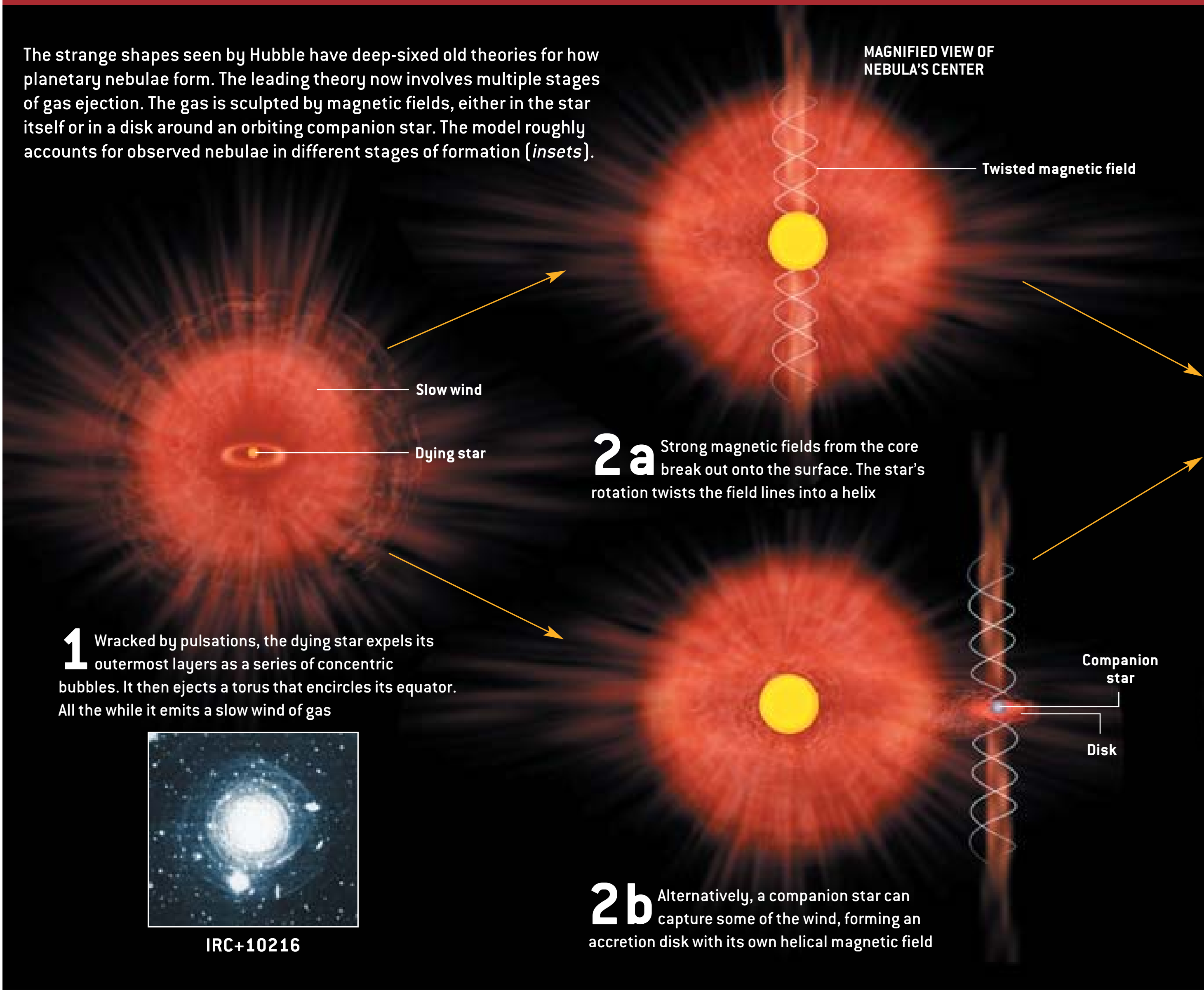
Mario Livio of the Space Telescope Science Institute and his former student Noam Soker of the Technion-Israel Institute of Technology championed this idea many years before it became fashionable [see "Planetary Nebulae," by Noam Soker; SCIENTIFIC AMERICAN, May 1992]. According to their scenario, the companion captures the material flowing from the dying star. In a system where the orbits are smaller than Mercury's and an orbital "year" is measured in Earth days, this transfer is cumbersome. By the time that material from the dying star reaches the companion, the latter has scooted well ahead in its orbit. The material drawn tidally from the large dying star thus forms a tail that chases the denser companion star from behind. This tail eventually settles into a dense, thick disk that swirls around the companion. Later simulations show that even a companion with an orbit as wide as Neptune's could scoop up an accretion disk.

The saga can take an interesting twist. As the dying star swells in size, it can swallow up its companion and disk. The result is a case of cosmic indigestion. The companion and disk enter a spiral orbit inside the body of the larger star, reshaping and flattening it from within. The outflows can thrash about, forming curved jets. Gradually the companion sinks deeper into the star until it merges with the core, at which point the outflow is cut off. This process could explain why some nebulae appear to result from an outflow that came to an abrupt end.

THE AUTHORS

BRUCE BALICK and *ADAM FRANK* have published dozen of papers, both observational and theoretical, on planetary nebulae and their precursor stars. Balick remembers deciding to become an astronomer at age five when his father read him a book about the planets. He has worked in fields ranging from star formation to active galactic nuclei and is now chair of the astronomy department at the University of Washington. Frank fell in love with astronomy around the same age, inspired by the covers of the science-fiction magazine *Amazing Stories* in his father's library. Growing up in the New York area, he soon discovered he could see only four or five stars in the night sky, so his attention turned to theory. Now a professor at the University of Rochester, Frank is interested in many topics in astrophysical fluid dynamics, from the death of stars to the birth of planets.

AS A STAR DIES, A NEBULA IS BORN



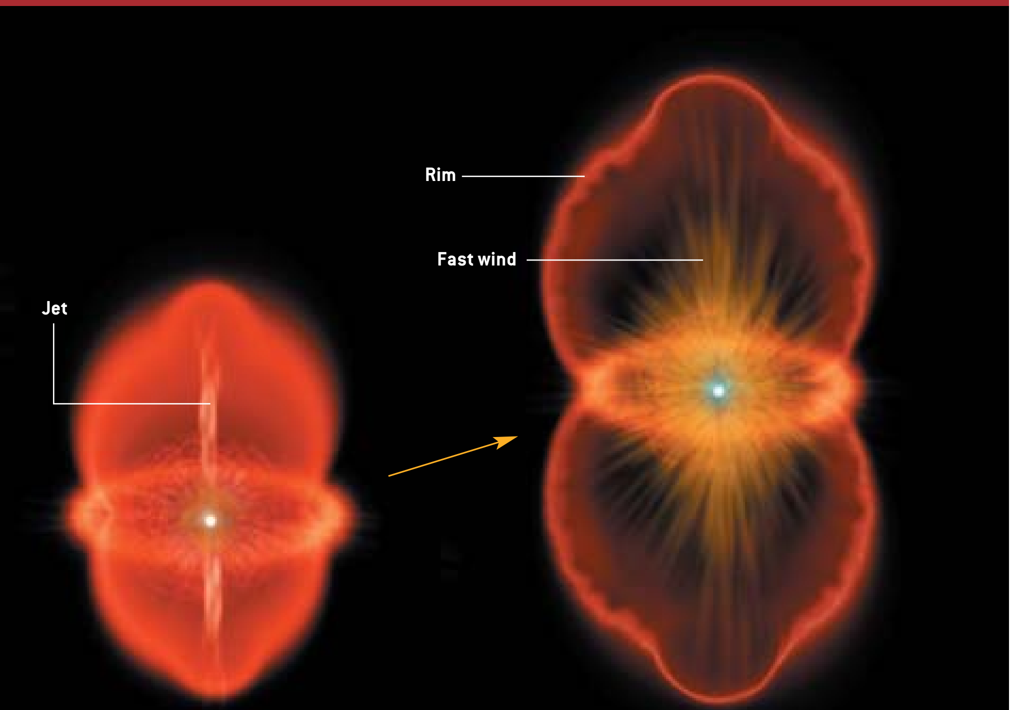
Magnetic Guidance

COMPANION STARS in binary systems are not the only plausible sculptors of planetary nebulae. Another player may be powerful magnetic fields embedded in either the star or the disk that forms around its companion. Because much of the gas in space is ionized, magnetic fields can guide its motion. Strong fields act like stiff rubber bands that shape the gas flow, much as Earth's magnetic field snares particles from the solar wind and guides them into the polar regions to trigger auroras. Conversely, strong winds can stretch, bend or entangle the fields.

In the mid-1990s Roger A. Chevalier and Ding Luo of the University of Virginia proposed that outflowing stellar winds carry hoops of magnetic field. The tug of war between the gas and the field can collimate the outflow into exotic shapes. Unfortunately for the model, it predicts that the field must begin in a weak state and play no role in generating the wind. That is a problem, because active magnetic fields on the surfaces of stars do seem to be instrumental in launching winds.

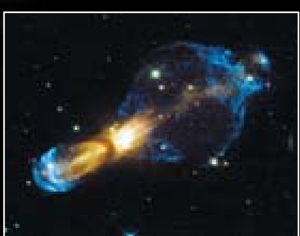
Another route has been to explore how strong magnetic fields can fling matter into space. As convection roils a dying star, fields anchored to the core rise with buoyant gas to the surface and, if the core is rotating rapidly, get wound up like a spring. As they break out at the surface, they snap and shoot material outward. A similar process can occur in a magnetized accretion disk. In fact, the star and accretion disk can each power a set of winds. A misalignment of their axes might produce some of the strange multipolar shapes seen in young planetary nebulae. Along with Eric G. Blackman of the University

DON DIXON (*illustrations*); COURTESY OF N. MAURON AND P. J. HUGGINS IN *ASTRONOMY AND ASTROPHYSICS*, VOL. 359; 2000 (*IRC+10216*)

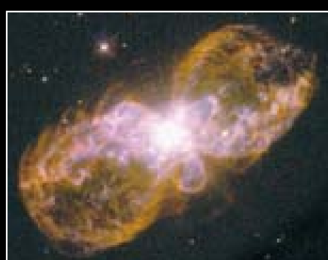


3 Whatever its origin and location, the magnetic field funnels gas into a short-lived jet, which plows into the slow-moving wind. Meanwhile the torus causes the wind to take on an hourglass shape

4 The star emits a fast-moving wind, which hits the slow-moving wind from behind and builds up a rim of gas



CALABASH (OH 231.8+4.2)



HUBBLE'S DOUBLE BUBBLE

of Rochester, Sean Matt of McMaster University and their colleagues, one of us (Frank) is studying these effects. The key is that magnetic fields, like binary stars, provide extra forces that can generate a far greater range of shapes than the interacting winds model can.

Our understanding of how dismembered stars are sculpted into planetary nebulae has made some progress but is still immature. The overall description of stellar death is well accepted. Stars evolve in such a way that their engines sputter as they shut down and shed their outer layers into space. In fact, the theory of stellar structure and evolution is one of the most successful scientific theories of the 20th century. It exquisitely explains observations of most stars—their light output, their colors, even most of their quirks. But large gaps clearly remain, especially at the very beginning and very end of stars' lives.

Not far from the University of Rochester is the Eastman School of Music. There some of the world's best young musicians and composers struggle every day to develop ways to express their creative visions. Those of us who study the death of sunlike stars find ourselves in a similar position. We believe that we have identified the instruments of how dying stars shape their outflows. What we do not yet understand is how these laws are orchestrated to create something as harmoniously structured as a planetary nebula. What powers the stellar winds? When are companion stars important? What role do magnetic fields play? What creates multiple-lobed nebulae?

We are hardly the only astrophysicists to be awed, puzzled and challenged by enigmatic images from Hubble and other instruments over the past decade. Nearly every field of astronomical research has a similar tale to tell. New information ultimately upends the best of theories in every field of research. That is the nature of progress. Discovery is often disruptive. It clears out old niches and prepares the way for big (and often disorienting) leaps forward. Scientific theories are built to be used, but they must be mistrusted, tested and improved. SA

MORE TO EXPLORE

The Shapes of Planetary Nebulae. Bruce Balick in *American Scientist,* Vol. 84, No. 4, pages 342–351; July 1996.

Cosmic Butterflies: The Colorful Mysteries of Planetary Nebulae. Sun Kwok. Cambridge University Press, 2001.

Shapes and Shaping of Planetary Nebulae. Bruce Balick and Adam Frank in *Annual Review of Astronomy and Astrophysics,* Vol. 40, pages 439–486; 2002.

A variety of Web sites have images of planetary nebulae:

www.astro.washington.edu/balick/WFPC2
www.blackskies.com/intro.html#NEBULAE
hubblesite.org/newscenter/archive/category/nebula/planetary
ad.usno.navy.mil/pne

For more on stellar evolution, see:
www.astronomynotes.com/evolutn/s1.htm
www.blackskies.com/neb101.htm
observe.arc.nasa.gov/nasa/space/stellardeath/stellardeath_intro.html

NASA/ESA AND V. BUJARRABAL *National Astronomical Observatory of Spain* (*Calabash*); B. BALICK, V. ICKE, G. MELLEMA AND WFPC2/HST/NASA (*Double Bubble*)

Gene therapy for restoring muscle lost to age or disease is poised to enter the clinic, but elite athletes are eyeing it to enhance performance.

Can it be long before gene doping changes the nature of sport?

By H. Lee Sweeney

Athletes will be going to Athens next month to take part in a tradition begun in Greece more than 2,000 years ago. As the world's finest specimens of fitness test the extreme limits of human strength, speed and agility, some of them will probably also engage in a more recent, less inspiring Olympic tradition: using performance-enhancing drugs. Despite repeated scandals, doping has become irresistible to many athletes, if only to keep pace with competitors who are doing it. Where winning is paramount, athletes will seize any opportunity to gain an extra few split seconds of speed or a small boost in endurance.

Sports authorities fear that a new form of doping will be undetectable and thus much less preventable. Treatments that regenerate muscle, increase its strength, and protect it from degradation will soon be entering human clinical trials for muscle-wasting disorders. Among these are therapies that give patients a synthetic gene, which can last for years, producing high amounts of naturally occurring muscle-building chemicals.

This kind of gene therapy could transform the lives of the elderly and people with muscular dystrophy. Unfortunately, it is also a dream come true for an athlete bent on doping. The chemicals are indistinguishable from their natural counterparts and are only generated locally in the muscle tissue. Nothing enters the bloodstream, so officials will have nothing to detect in a blood or urine test. The World Anti-Doping Agency (WADA) has already asked scientists to help find ways to prevent gene therapy from becoming the newest means of doping. But as these treatments enter clinical trials and, eventually, widespread use, preventing athletes from gaining access to them could become impossible.

Is gene therapy going to form the basis of high-tech cheating in athletics? It is certainly possible. Will there be a time

ATHLETES BUILD MUSCLE through intensive training. This Olympic-class rower's back displays the result of his hard work. But gene therapy could allow athletes to build more muscle, faster, and to stay strong longer without further effort.

RANDY HARRIS (*opposite page*); JEN CHRISTIANSEN (*this page*) when gene therapy becomes so commonplace for disease that manipulating genes to enhance performance will become universally accepted? Perhaps. Either way, the world may be about to watch one of its last Olympic Games without genetically enhanced athletes.

Loss Leads to Gain

RESEARCH TOWARD genetically enhancing muscle size and strength did not start out to serve the elite athlete. My own work began with observing members of my family, many of whom lived well into their 80s and 90s. Although they enjoyed generally good health, their quality of life suffered because of the weakness associated with aging. Both muscle strength and mass can decrease by as much as a third between the ages of 30 and 80.

There are actually three types of muscle in the body: smooth muscle, lining internal cavities such as the digestive tract; cardiac muscle in the heart; and skeletal muscle, the type most of us think of when we think of muscle. Skeletal muscle constitutes the largest organ of the body, and it is this type—particularly the strongest so-called fast fibers—that declines with age. With this loss of strength, losing one's balance is more likely and catching oneself before falling becomes more difficult. Once a fall causes a hip fracture or other serious injury, mobility is gone completely.

Skeletal muscle loss occurs with age in all mammals and probably results from a cumulative failure to repair damage caused by normal use. Intriguingly, aging-related changes in skeletal muscle resemble the functional and physical changes seen in a suite of diseases collectively known as muscular dystrophy, albeit at a much slower rate.

In the most common and most severe version of MD—Duchenne muscular dystrophy—an inherited gene mutation results in the absence of a protein called dystrophin that protects muscle fibers from injury by the force they exert during regular movement. Muscles are good at repairing themselves, although their normal regenerative mechanisms cannot keep up with the excessive rate of damage in MD. In aging muscles the rate of damage may be normal, but the repair mechanisms become less responsive. As a result, in both aging and Duchenne MD, muscle fibers die and are replaced by infiltrating fibrous tissue and fat.

In contrast, the severe skeletal muscle loss experienced by astronauts in microgravity and by patients immobilized by disability appears to be caused by a total shutdown of muscles' repair and growth mechanism at the same time apoptosis, or programmed cell death, speeds up. This phenomenon, known as disuse atrophy, is still not fully understood but makes sense from an evolutionary perspective. Skeletal muscle is metabolically expensive to maintain, so keeping a tight relation between muscle size and its activity saves energy. Skeletal muscle is exquisitely tuned to changing functional demands. Just as it withers with disuse, it grows in size, or hypertrophies, in response to repeated exertions. The increased load triggers a number of signaling pathways that lead to the addition of new cellular components within individual muscle fibers, changes in fiber type and, in extreme conditions, addition of new muscle fibers.

Raising IGF-I allows us to break the connection between muscle use and its size.

To be able to influence muscle growth, scientists are piecing together the molecular details of how muscle is naturally built and lost. Unlike the typical cell whose membrane contains liquid cytoplasm and a single nucleus, muscle cells are actually long cylinders, with multiple nuclei, and cytoplasm consisting of still more long tiny fibers called myofibrils [*see box on opposite page*]. These myofibrils, in turn, are made of stacks of contractile units called sarcomeres. Collectively, their shortening produces muscle contractions, but the force they generate can damage the muscle fiber unless it is channeled outward. Dystrophin, the protein missing in Duchenne muscular dystrophy patients, conducts this energy across the muscle cell's membrane, protecting the fiber.

Yet even with dystrophin's buffering, muscle fibers are still injured by normal use. In fact, that is believed to be one way that exercise builds muscle mass and strength. Microscopic tears in the fibers caused by the exertion set off a chemical alarm that triggers tissue regeneration, which in muscle does not mean production of new muscle fibers but rather repairing the outer membrane of existing fibers and plumping their interior with new myofibrils. Manufacturing this new protein requires activation of the relevant genes within the muscle cell's nuclei, and when the demand for myofibrils is great, additional nuclei are needed to bolster the muscle cell's manufacturing capacity.

Local satellite cells residing outside the muscle fibers answer this call. First these muscle-specific stem cells proliferate by normal cell division, then some of their progeny fuse with the muscle fiber,

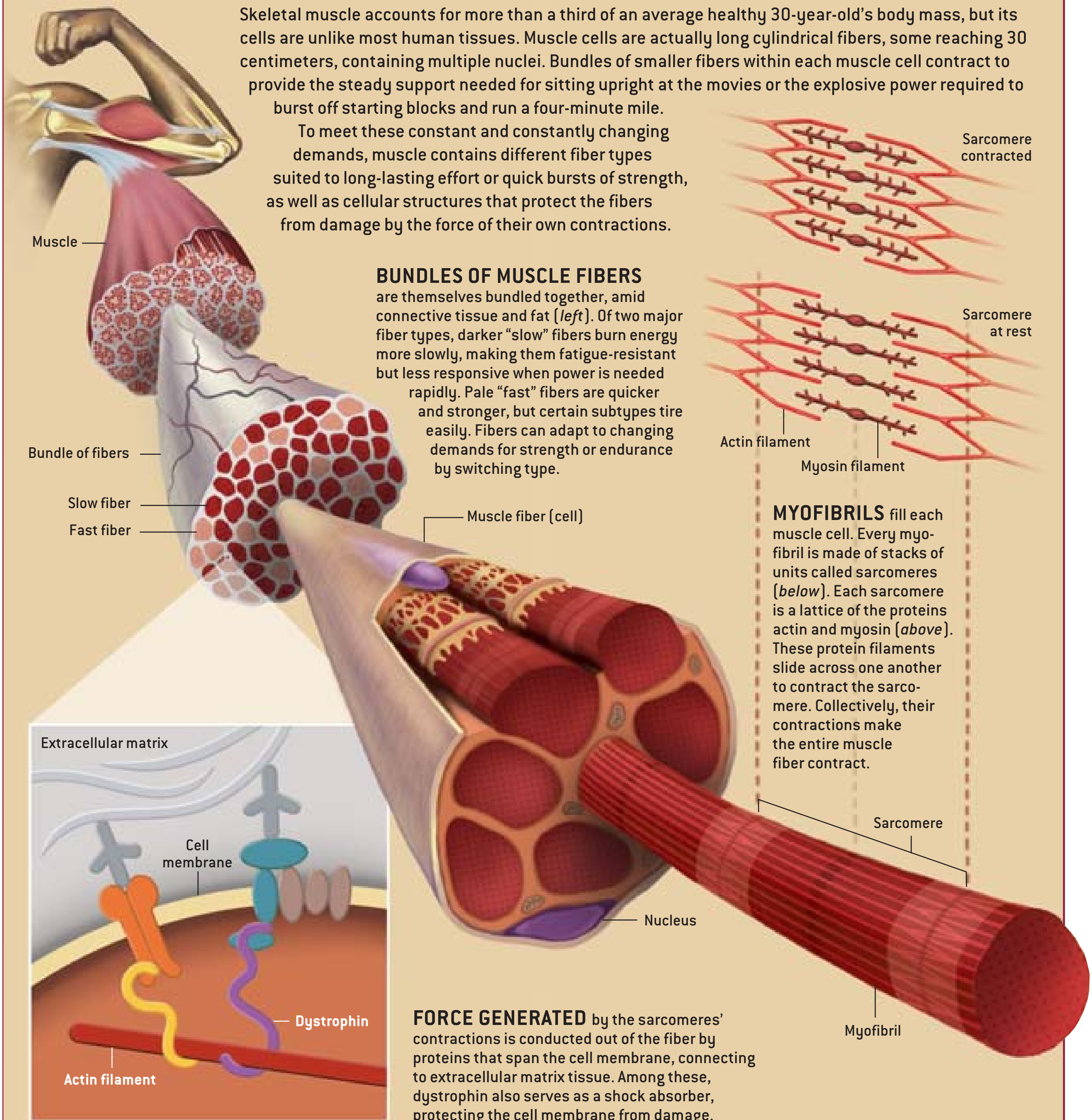
Overview/*Molecular Muscle Building*

- Muscle growth and repair are controlled by chemical signals, which are in turn controlled by genes. Muscle lost to age or disease can be replaced by boosting or blocking these signals with the addition of a synthetic gene.
- Athletes could use the same technique to enhance muscle size, strength and resilience, and the treatment might be undetectable.
- When gene therapy enters the medical mainstream, preventing its abuse will be difficult, but attitudes toward genetic enhancement may also change.

THE BODY'S POWERHOUSE

Skeletal muscle accounts for more than a third of an average healthy 30-year-old's body mass, but its cells are unlike most human tissues. Muscle cells are actually long cylindrical fibers, some reaching 30 centimeters, containing multiple nuclei. Bundles of smaller fibers within each muscle cell contract to provide the steady support needed for sitting upright at the movies or the explosive power required to burst off starting blocks and run a four-minute mile.

To meet these constant and constantly changing demands, muscle contains different fiber types suited to long-lasting effort or quick bursts of strength, as well as cellular structures that protect the fibers from damage by the force of their own contractions.



BUNDLES OF MUSCLE FIBERS are themselves bundled together, amid connective tissue and fat (*left*). Of two major fiber types, darker "slow" fibers burn energy more slowly, making them fatigue-resistant but less responsive when power is needed rapidly. Pale "fast" fibers are quicker and stronger, but certain subtypes tire easily. Fibers can adapt to changing demands for strength or endurance by switching type.

MYOFIBRILS fill each muscle cell. Every myofibril is made of stacks of units called sarcomeres (*below*). Each sarcomere is a lattice of the proteins actin and myosin (*above*). These protein filaments slide across one another to contract the sarcomere. Collectively, their contractions make the entire muscle fiber contract.

FORCE GENERATED by the sarcomeres' contractions is conducted out of the fiber by proteins that span the cell membrane, connecting to extracellular matrix tissue. Among these, dystrophin also serves as a shock absorber, protecting the cell membrane from damage.

JEN CHRISTIANSEN

contributing their nuclei to the cell. Both progrowth and antigrowth factors are involved in regulating this process. Satellite cells respond to insulinlike growth factor I, or IGF-I, by undergoing a greater number of cell divisions, whereas a different growth-regulating factor, myostatin, inhibits their proliferation.

With these mechanisms in mind, about seven years ago my group at the University of Pennsylvania, in collaboration with Nadia Rosenthal and her colleagues at Harvard University, began to assess the possibility of using IGF-I to alter muscle function. We knew that if we injected the IGF-I protein alone, it would dissipate within hours. But once a gene enters a cell, it should keep functioning

BELGIAN BLUE BULL demonstrates the effect of blocking the antigrowth factor myostatin. A natural genetic mutation in this breed produces a truncated, ineffective form of myostatin, which allows muscle growth to go unchecked. The absence of myostatin also interferes with fat deposition, making these "double-muscled" cattle exceptionally lean.

for the life of that cell, and muscle fibers are very long-lived. A single dose of the IGF-I gene in elderly humans would probably last for the rest of their lives. So we turned our attention to finding a way to deliver the IGF-I gene directly to muscle tissue.

Donning New Genes

THEN AS NOW, a major obstacle to successful gene therapy was the difficulty of getting a chosen gene into the desired tissue. Like many other researchers, we selected a virus as our delivery vehicle, or vector, because viruses are skilled at smuggling genes into cells. They survive and propagate by tricking the cells of a host organism into bringing the virus inside, rather like a biological Trojan horse. Once within the nucleus of a host cell, the virus uses the cellular machinery to replicate its genes and produce proteins. Gene therapists capitalize on this ability by loading a synthetic gene into the virus and removing any genes the virus could use to cause disease or to replicate itself. We selected a tiny virus called adeno-associated virus (AAV) as our vector, in part because it infects human muscle readily but does not cause any known disease.

We modified it with a synthetic gene that would produce IGF-I only in skeletal muscle and began by trying it out in normal mice. After injecting this AAV-IGF-I combination into young mice, we saw that the muscles' overall size and the rate at which they grew were 15 to 30 percent greater than normal, even though the mice were sedentary. Further, when we injected the gene into the muscles of middle-aged mice and then allowed them to reach old age, their muscles did not get any weaker.

To further evaluate this approach and its safety, Rosenthal created mice genetically engineered to overproduce IGF-I throughout their skeletal muscle. Encouragingly, they developed normally except for having skeletal muscles that ranged from 20 to 50 percent larger than those of regular mice. As these transgenic mice aged, their muscles retained a regenerative capacity typical of younger animals. Equally important, their IGF-I levels were elevated only in the muscles, not in the bloodstream, an important distinction because high circulating levels of IGF-I can cause cardiac problems and increase cancer risk. Subsequent experiments showed that IGF-I overproduction hastens muscle repair, even in mice with a severe form of muscular dystrophy.

Raising local IGF-I production allows us to achieve a central goal of gene therapy to combat muscle-wasting diseases: breaking the close connection between muscle use and its size. Simulating the results of muscle exercise in this manner also has obvious appeal to the elite athlete. Indeed, the rate of muscle growth in young sedentary animals suggested that this treatment could also be used to genetically enhance performance of healthy muscle. Recently my laboratory worked with an exercise physiology group headed by Roger P. Farrar of the University of Texas at Austin to test this theory.

We injected AAV-IGF-I into the muscle in just one leg of each of our lab rats and then subjected the animals to an eight-week weight-training protocol. At

THE AUTHOR

H. LEE SWEENEY is professor and chairman of physiology at the University of Pennsylvania School of Medicine. He is a member of the Board of Scientific Councilors for the National Institute of Arthritis and Musculoskeletal Diseases, scientific director for Parent Project Muscular Dystrophy, and a member of the Muscular Dystrophy Association's Translational Research Advisory Council. His research ranges from basic investigation of structures that allow cells to move and generate force, particularly the myosin family of molecular motors, to translating insights about muscle cell design and behavior into gene therapy interventions for diseases, including Duchenne muscular dystrophy. He took part in a 2002 symposium on the prospect of gene doping organized by the World Anti-Doping Agency.

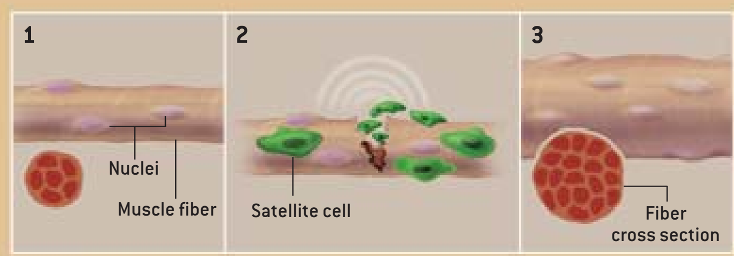
the end of the training, the AAV-IGF-I-injected muscles had gained nearly twice as much strength as the uninjected legs in the same animals. After training stopped, the injected muscles lost strength much more slowly than the unenhanced muscle. Even in sedentary rats, AAV-IGF-I provided a 15 percent strength increase, similar to what we saw in the earlier mouse experiments.

We plan to continue our studies of IGF-I gene therapy in dogs because the golden retriever breed is susceptible to a particularly severe form of muscular dystrophy. We will also do parallel studies in healthy dogs to further test the effects and safety of inducing IGF-I overproduction. It is a potent growth and signaling factor, to which tumors also respond.

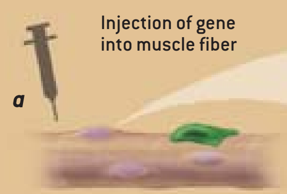
Safety concerns as well as unresolved questions about whether it is better to de-

PUMPING UP WITH GENES

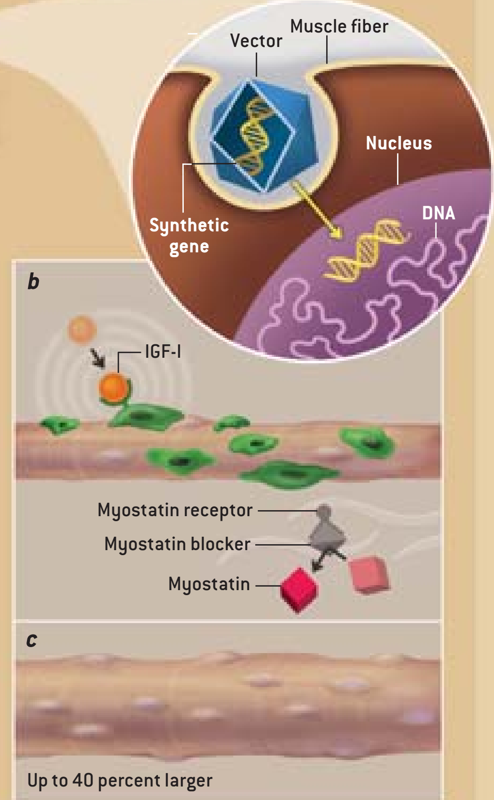
Building athletes' muscle, tweaking its composition, and boosting endurance are enhancements theoretically possible with gene therapy. Using a synthetic gene to simulate an injury signal spurs repair activity by stem cells (*right*), leaving muscle fibers bigger and stronger. Activating a dormant gene or adding a new one could change muscle fiber types (*below*). Unlike systemic drugs, gene therapy also allows key muscle subgroups to be targeted based on the biomechanics of a given sport.



IN NORMAL MUSCLE, a fiber's multiple nuclei (*1*) are responsible for driving the manufacture of new proteins. When repair is needed, chemical signals from the wound draw satellite cells, which proliferate before fusing with the fiber to contribute their nuclei to the effort (*2*). The addition of more nuclei and fresh myofibrils leaves a repaired fiber bulkier than before it was injured (*3*).



GENE THERAPY can stimulate and augment normal repair by manipulating the chemical signals involved. A synthetic gene can be added to muscle inside a delivery vehicle, or vector. This vector will carry the gene into a nucleus, where it can begin directing the fiber to produce a protein (*a*). In normal muscle repair, a protein called IGF-I signals satellite cells to proliferate; another protein called myostatin tells them to stop. Introducing an IGF-I gene or one for a protein that blocks myostatin from signaling to satellite cells will yield the same effect: more satellite cell proliferation (*b*) and a beefed-up muscle fiber (*c*).



JEN CHRISTIANSEN

Natural Advantage

As this article went to press, the *New England Journal of Medicine* was about to release the first documented description of a human being with a genetic mutation that wipes out myostatin production. Such cases have been discussed in scientific circles but never published because the subjects and their families usually do not wish to risk being identified. At least one of those families is rumored to include a European weight-lifting champion, which, if true, would not be surprising, given the tremendous advantage in muscle building and strength that a natural myostatin-suppressing mutation would confer.

But would it constitute an unfair advantage in an athlete, and would it justify other competitors using myostatin-inhibiting drugs or gene therapy simply to level the playing field? These questions are bound to be raised in continuing debate over the possibility of athletes using new muscle therapies to enhance their performance.

Natural "mutants" among athletes have been documented, among them an Olympic gold medalist. Finnish cross-country skier Eero Mäntyranta won two gold medals in the 1964 Winter Olympics. But it was not until decades later that Finnish scientists identified a genetic mutation in Mäntyranta's entire family that causes an excessive response to erythropoietin, leading to extraordinarily high numbers of oxygen-carrying red blood cells. Several of his family members, it turns out, were also champion endurance athletes.

In addition to mutations with dramatic effects, investigators have also begun to discover natural gene variants that more subtly favor certain kinds of athletic activity. For example, last year Australian researchers examined a gene called *ACTN3* in a group of male and female elite sprinters. Nearly 20 percent of people lack a functional version of this gene that gives rise to a protein specific to fast muscle fibers, although a less effective protein normally compensates for its absence. The scientists found an unusually high frequency of the working *ACTN3* gene in the sprinters, however. In particular, more of the female sprinters had two copies of the gene than would be expected in a randomly selected group.

Many research groups are trying to identify other gene variants that give athletes an edge by maximizing oxygen uptake, heart efficiency, power output, endurance or other traits. More than 90 genes or chromosomal locations have been associated with athletic performance so far, and this research is already provoking its own ethical controversies. Critics fear that based on their genetic makeup, children will be recruited into certain sports or, if they lack the right gene mix, denied a chance to advance to the elite level of sports training. Even selective breeding for superathletes has been predicted.

A more certain result of scanning athletes' genomes will be the discovery that some of them, like Mäntyranta's, contain true genetic mutations that amount to genetic enhancement. Such revelations will add still more complexity to ethical arguments over the prospect of gene doping in sports. *—H.L.S.*

EERO MÄNTYRANTA won two Olympic gold medals in 1964. Years later scientists found the source of the Finnish cross-country skier's endurance. A genetic mutation gave his family higher than normal levels of oxygen-carrying red blood cells—higher even than could be achieved with EPO.

liver AAV in humans through the bloodstream or by direct injection into muscle mean that approved gene therapy treatments using AAV-IGF-I could be as much as a decade away. In the shorter term, human trials of gene transfer to replace the dystrophin gene are already in planning stages, and the Muscular Dystrophy Association will soon begin a clinical trial of IGF-I injections to treat myotonic dystrophy, a condition that causes prolonged muscle contraction and, hence, damage.

A still more immediate approach to driving muscle hypertrophy may come from drugs designed to block myostatin. Precisely how myostatin inhibition builds muscle is still unclear, but myostatin seems to limit muscle growth throughout embryonic development and adult life. It acts as a brake on normal muscle growth and possibly as a promoter of atrophy when functional demands on muscle decrease. Experiments on genetically engineered mice indicate that the absence of this antigrowth factor results in considerably larger muscles because of both muscle fiber hypertrophy and hyperplasia, an excessive number of muscle fibers.

Making Muscle and More

PHARMACEUTICAL AND biotechnology companies are working on a variety of myostatin inhibitors. Initially, the possibility of producing meatier food animals piqued commercial interest. Nature has already provided examples of the effects of myostatin blockade in the Belgian Blue and Piedmontese cattle breeds, both of which have an inherited genetic mutation that produces a truncated, ineffective version of myostatin. These cattle are often called double-muscled, and their exaggerated musculature is all the more impressive because an absence of myostatin also interferes with fat deposition, giving the animals a lean, sculpted appearance.

The first myostatin-blocking drugs to have been developed are antibodies against myostatin, one of which may soon undergo clinical testing in muscular dystrophy patients. A different approach mimics the cattle mutation by creating a smaller version of myostatin, which lacks

AP PHOTO

the normal molecule's signaling ability while retaining the structures that dock near satellite cells. This smaller protein, or peptide, essentially caps those docking locations, preventing myostatin from attaching to them. Injecting the peptide into mice produces skeletal muscle hypertrophy, and my colleagues and I will be attempting to create the same effect in our dog models by transferring a synthetic gene for the peptide.

Myostatin-blocking therapies also have obvious appeal to healthy people seeking rapid muscle growth. Although systemic drugs cannot target specific muscles, as gene transfer can, drugs have the benefit of easy delivery, and they can immediately be discontinued if a problem arises. On the other hand, such drugs would be relatively easy for sport regulatory agencies to detect with a blood test.

But what if athletes were to use a gene therapy approach similar to our AAV-IGF-I strategy? The product of the gene would be found just in the muscle, not in the blood or urine, and would be identical to its natural counterpart. Only a muscle biopsy could test for the presence of a particular synthetic gene or of a vector. But in the case of AAV, many people may be naturally infected with this harmless virus, so the test would not be conclusive for doping. Moreover, because most athletes would be unwilling to undergo an invasive biopsy before a competition, this type of genetic enhancement would remain virtually invisible.

And what of the safety of rapidly increasing muscle mass by 20 to 40 percent? Could an athlete sporting genetically inflated musculature exert enough force to snap his or her own bones or tendons? Probably not. We worry more about building muscle in elderly patients with bones weakened by osteoporosis. In a healthy young person, muscle growth occurring over weeks or months would give supporting skeletal elements time to grow to meet their new demands.

This safety question, however, is just one of the many that need further study in animals before these treatments can even be considered for mere enhancement of healthy humans. Nevertheless, with gene therapy poised to finally become a viable medical treatment, gene doping cannot be far behind, and overall muscle enlargement is but one way that it could be used [*see illustration on page 67*]. In sports such as sprinting, tweaking genes to convert muscle fibers to the fast type might also be desirable. For a marathoner, boosting endurance might be paramount.

Muscle is most likely to be the first tissue subject to genetic enhancement, but others could eventually follow. For example, endurance is also affected by the amount of oxygen reaching muscles. Erythropoietin is a naturally occurring protein that spurs development of oxygen-carrying red blood cells. Its synthetic form, a drug called Epoietin, or simply EPO, was developed to treat anemia but has been widely abused by athletes—most publicly by cyclists in the 1998 Tour de France. An entire team was excluded from that race when their EPO use was uncovered, yet EPO abuse in sports continues.

Gene transfer to raise erythropoietin production has already been tried in animals, with results that illustrate the potential dangers of prematurely attempting such enhancements in humans. In 1997 and 1998 scientists tried transferring synthetic erythropoietin genes into monkeys and baboons. In both experiments, the animals' red blood cell counts nearly doubled within 10 weeks, producing blood so thick that it had to be regularly diluted to keep their hearts from failing.

The technology necessary to abuse gene transfer is certainly not yet within reach of the average athlete. Still, officials in the athletic community fear that just as technically skilled individuals have turned to the manufacture and sale of so-called designer steroids, someday soon a market in genetic enhancement may emerge. Policing such abuse will be much harder than monitoring drug use, because detection will be difficult.

It is also likely, however, that in the decades to come, some of these gene therapies will be proved safe and will become available to the general population. If the time does arrive when genetic enhancement is widely used to improve quality of life, society's ethical stance on manipulating our genes will probably be much different than it is today. Sports authorities already acknowledge that muscle-regenerating therapies may be useful in helping athletes to recover from injuries.

So will we one day be engineering superathletes or simply bettering the health of the entire population with gene transfer? Even in its infancy, this technology clearly has tremendous potential to change both sports and our society. The ethical issues surrounding genetic enhancement are many and complex. But for once, we have time to discuss and debate them before the ability to use this power is upon us. SA

With gene therapy poised to become a viable medical treatment, gene doping cannot be far behind.

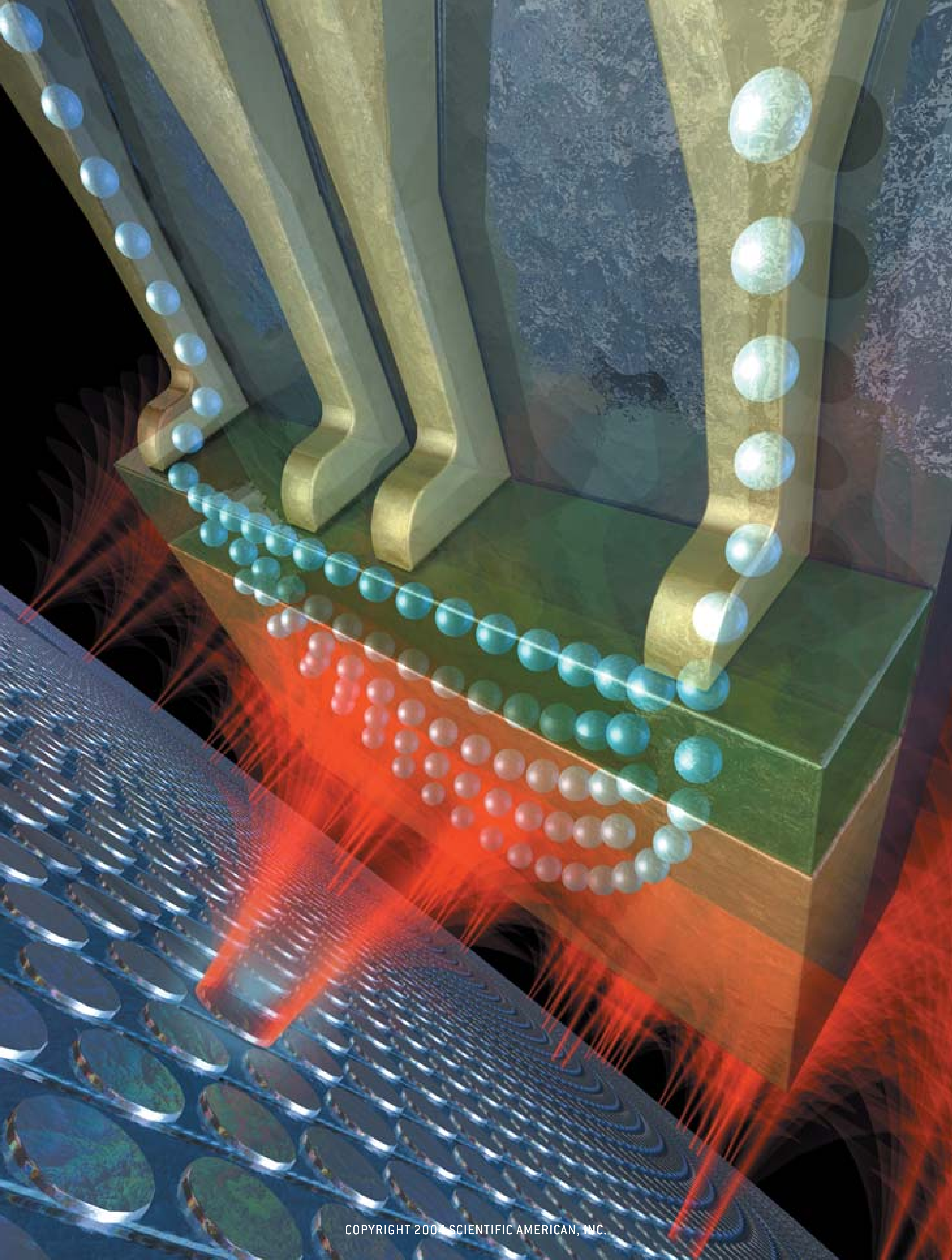
MORE TO EXPLORE

Viral Mediated Expression of Insulin-like Growth Factor I Blocks the Aging-Related Loss of Skeletal Muscle Function. Elisabeth R. Barton-Davis et al. in *Proceedings of the National Academy of Sciences USA,* Vol. 95, No. 26, pages 15,603–15,607; December 22, 1998.

Muscle, Genes and Athletic Performance. Jasper L. Anderson, Peter Schjerling and Bengt Saltin in *Scientific American,* Vol. 283, No. 3, pages 48–55; September 2000.

Toward Molecular Talent Scouting. Gary Taubes in *Scientific American Presents: Building the Elite Athlete,* Vol. 11, No. 3, pages 26–31; Fall 2000.

Viral Expression of Insulin-like Growth Factor-I Enhances Muscle Hypertrophy in Resistance-Trained Rats. Sukho Lee et al. in *Journal of Applied Physiology,* Vol. 96, No. 3, pages 1097–1104; March 2004.



Magnetic Field Nanosensors

Tiny devices that take advantage of a recently discovered physical effect called extraordinary magnetoresistance could be used in blazingly fast computer disk drives with huge capacities and in dozens of other applications involving the sensing of magnetic fields

By Stuart A. Solin

COMPUTER DISK DRIVES of the future could use nanoscopic read heads based on so-called extraordinary magnetoresistance (EMR) to detect the magnetic fields (*red*) of the ones and zeros recorded on the disk. The EMR effect works by changing the paths of electrons (*white balls*) traveling through the device.

BRYAN CHRISTIE DESIGN

It often happens in science that research focused on one phenomenon results in the unexpected discovery of a new effect that is much more exciting and important. In 1995 a case of this serendipity occurred with my research group, then at the NEC Research Institute in Princeton, N.J.

We were studying the properties of a microelectronics structure called a semiconductor superlattice, which consisted of layers of gallium arsenide and gallium aluminum arsenide stacked like a club sandwich. We knew that this superlattice had very interesting electrical properties. In particular, we were investigating how the thickness of the layers determined whether the superlattice behaved as a metal, with low electrical resistance, or as an insulator, with high resistance. We immersed the system in a magnetic field, a procedure that enabled us to study dynamical processes involving the electrons in the superlattice.

To our great surprise, we saw that the resistance of the superlattice increased dramatically when we strengthened the magnetic field. Such behavior would be expected in a magnetic material but not in something like our superlattice, which was made entirely of nonmagnetic constituents. The percentage increase in resistance, called the magnetoresistance (MR), was so big and so unusual that my team immediately redirected its efforts to explain the fundamental physics of this new effect.

By 1997 we had developed a sound basic understanding of this new kind of large MR. Moreover, we predicted that considerably greater MR might be obtained from a much simpler structure fabricated from a nonmagnetic metal such as gold (Au) and a single layer of the semiconductor indium antimonide (InSb). In 1998, working with Jean Heremans of Ohio University, we built an InSb-Au structure that fulfilled our prediction [*see box on page 75*]. In a magnetic field of five teslas, the MR was about 1 million percent at room temperature, thousands of times as great as any MR previously observed at that temperature. We dubbed this phenomenon "extraordinary magnetoresistance," or EMR, and we knew that it could be useful for a wide variety of technologies that require sophisticated magnetic field sensors, such as ultrahigh-density data recording, automotive control systems, industrial applications, medical devices and consumer electronics.

What's So Special about EMR

EMR WAS FAR FROM being the first large MR effect discovered. In recent decades, a series of such phenomena have been observed, in part as a response to the demand for small yet highly sensitive magnetic field sensors [*see box on opposite page*]. EMR is unique among large MRs, however, in that it does not require a magnetic material as part of the structure—a property that is scientifically intriguing and also advantageous for certain applications, as I will explain later.

The other big MR effects occur when the magnetic field of a material's atoms interacts with the intrinsic magnetism of the electrons flowing through the material. You can picture the innate magnetism of the electrons by imagining that each one contains a tiny bar magnet with a north and a south pole. Normally these magnets point in random directions and have no effect on the flow of current. But in a magnetic material, the electrons tend to become polarized, with their magnets aligned with the material's magnetic field. Once an electric current is polarized, it flows more easily through a material whose magnetic field is parallel with its polarization than through one that is antiparallel.

Thus, MR structures typically have one layer of magnetic material that polarizes the current and a second layer that has a controllable magnetization; the second layer impedes or lets through the current depending on how its field is oriented with respect to the first layer. Devices that utilize the electron's magnetism in this way are called either magnetoelectronics or spintronics, the latter because the magnetism is closely related to a quantity called spin [see "Spintronics," by David D. Awschalom, Michael E. Flatté and Nitin Samarth; SCIENTIFIC AMERICAN, June 2002].

How, then, does EMR work in the absence of a magnetic material? The answer lies in a second way that magnetic fields interact with moving electrons. When a charged particle, such as an electron, travels through a magnetic field, the field exerts a transverse force on the particle, curving its trajectory [*see box on page 74*]. This effect is what causes ordinary MR. The field curves the electrons' trajectories, even turning them into helices if it is strong enough. Because the electrons travel along longer, winding paths, their net motion from one end of the material to the other is slowed down. In this way, the current is reduced; the resistance is enhanced. To be more precise, the electrons actually travel along random zigzag paths because of collisions with impurities or other defects in the material. Nevertheless, the magnetic field turns each straight zig or zag into a curve, increasing the total path length traveled as the electrons make their erratic way through the material.

The very much larger effect of EMR also depends on the magnetic field curving the electrons' paths. EMR's great magnitude, however, is caused by the interplay of the curved paths and the detailed geometry of the EMR device at the nanometer scale (billionths of a meter). The shape, location and electrical properties of the elements of the device—such as electrical contacts and regions of different materials—can all contribute to this geometrically based MR.

For magnetic MR devices and ordinary MR in nonmagnetic semiconductors, the geometric contribution to the MR is insignificant compared with the physical contribution (the part that depends on the intrinsic physical properties of the material, such as the number of electrons per unit volume available to carry a current). By designing novel hybrid structures of nonmagnetic semiconductors and metals, my group was able to construct devices in which the geometric contribution to the MR far exceeded the physical contribution.

Understanding EMR

TO UNDERSTAND EMR, consider the device shown at the bottom of the illustration on page 74. The key part is a disk of gold that is embedded in a thin slab of narrow-gap semiconductor, such as InSb. (The "gap" of a semiconductor is a band of quantum states forbidden to its electrons, the size of which influences many of the material's electronic properties.) The conductivity of the metal is about 2,000 times greater than that of the semiconductor.

When we apply a voltage across electrical contacts at each end of the slab, current flows through the device. The current flows along electric field lines that the voltage establishes in the material. A property of electric field lines is that they tend to align themselves at right angles to the outside surface of a good conductor. This effect causes them to curve inward and concentrate on the metal disk. The current is thus funneled through the highly conductive metal, which causes the device as a whole to have a low resistance. The exact value of the resistance will depend on the geometry—that is, on the relative dimensions and shape of the metal and semiconductor.

Now consider what happens if we apply a magnetic field perpendicular to the slab. As in ordinary MR, the field produces an additional force on the charges, deflecting them like a plane in a cross-

Overview/*A New Magnetoresistance*

- Magnetoresistance is the phenomenon in which the electrical resistance of a metal or a semiconductor increases or decreases in response to a magnetic field. A variant of this effect thousands of times as great as any previously observed was discovered in 1998 and dubbed extraordinary magnetoresistance (EMR). EMR depends on the detailed geometry of a device made out of conductive metal and semiconductor.
- EMR could be used to make computer disk-drive read heads that are faster and capable of handling higher densities of data than present-day read heads (which depend on giant magnetoresistance, or GMR). Because EMR read heads contain no magnetic materials, they would have lower noise than GMR read heads, thereby improving performance.
- Many other applications could benefit from EMR, including position-sensing robots, other speed and position sensors in industry, antilock brakes, "smart" shock absorbers, ignition timing and control systems, flip-phone switches, and nonvolatile memory in low-cost appliances.

wind, so that they travel at an angle to the electric field lines. Given a strong enough magnetic field, they can be deflected a full 90 degrees at the boundary of the metal disk. In other words, the current flows around the perimeter of the metal disk instead of entering it. (There is a little more to the story than that, as is explained in the illustration, but the current's avoiding of the metal disk is the key end result.)

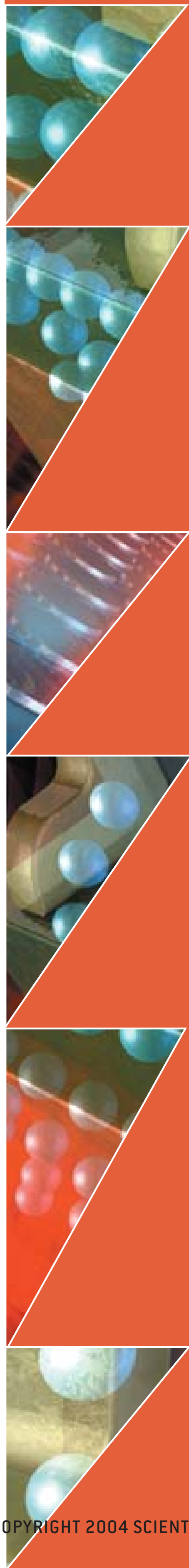
The current behaves exactly as if the metal disk was replaced by a big cavity in the semiconductor that had to be circumnavigated. Such a structure—a semiconductor with a disk cut out of it, which squeezes the current through the two narrow channels of semiconductor—has a much higher resistance than an uninterrupted slab of semiconductor (which itself has higher resistance than a semiconductor with a disk of metal embedded in it). Thus, at zero magnetic field the metal disk acts as a short circuit (very low resistance), and in a strong enough field it is equivalent to an empty space—an open circuit (very high resistance). This change in state produces the EMR effect.

Achieving this understanding of EMR was greatly stimulated by the pioneering work of Charles Wolf and Lester Stillman of the University of Illinois, who in the 1970s studied structures made of semiconductor and metal (called hybrid structures). In particular, they looked at the hybrid structures' carrier mobility, which is a measure of the ease with which the carriers of electric current move in an electric field. Carriers can be electrons or holes. A hole is the absence of an electron from a sea of electrons that behaves in many ways like a positively charged particle.

Building on Wolf and Stillman's work, my colleagues and I realized that the EMR in weak magnetic fields ought to be much bigger for semiconductors with higher carrier mobility. Narrow-gap materials such as InSb have the requisite high mobility. This prediction prompted efforts to develop semiconductor materials with increased mobility. Lesley Cohen and the late Tony Stradling of Imperial College London made significant progress preparing high-mobility, ultrathin films of InSb. Michael Santos of the University of Oklahoma has successfully fo-

BRYAN CHRISTIE DESIGN

The Magnetoresistance Zoo



If a material's electrical resistance increases or decreases when a magnetic field is applied, the change in resistance is known as **magnetoresistance (MR).** This effect, discovered in 1857 by British physicist William Thomson (Lord Kelvin), is negligible in metals and very small to moderate in semiconductors. In recent decades, a number of much larger MR effects have been discovered—including some with multibillion-dollar technological applications.

GMR (giant MR): In one design, two ferromagnetic layers (such as cobalt or iron) sandwich a layer of nonmagnetic metal. The magnetism of one ferromagnetic layer is held in a fixed orientation, whereas the other can be reoriented by an outside magnetic field. Maximum current flows when the two layers have parallel magnetism, and the minimum flows when they are antiparallel. The read heads of modern magnetic disk drives use GMR to detect the magnetic fields of bits of information on the disk. Discovered in 1988 by Peter A. Grünberg of the Jülich Research Center in Germany and Albert Fert of the University of Paris-Sud.

TMR (tunneling MR): Similar to GMR, but instead of using a metal as the meat of the sandwich, TMR uses a very thin insulator through which current flows by quantum tunneling of electrons. The next generation of disk read heads will probably use TMR. Both GMR and TMR are sensitive to very small changes in the applied magnetic field. First measured by Michel Jullière of INSA at Rennes in 1975, this effect drew renewed interest in the early 1990s.

CMR (colossal MR): Occurs in insulating manganese oxide crystals called manganites. The applied magnetic field causes the material to change from nonmagnetic and insulating to ferromagnetic and metallic—that is, the magnetic field *reduces* the resistance. The transition generally occurs only at temperatures below 150 kelvins and for magnetic fields of several teslas. First discovered in the 1950s by G. H. Jonker and J. H. van Santen of Philips in the Netherlands and rediscovered in 1994 by Sung-Ho Jin of Bell Laboratories and collaborators, CMR is no longer a serious contender for read-head applications.

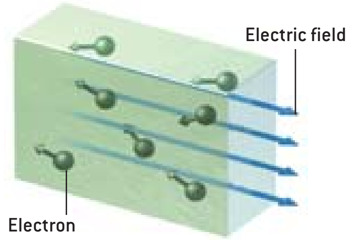
BMR (ballistic MR): Observed when a needle and a wire, both of which are metallic and magnetic, are joined by a nanoscopic contact. The applied magnetic field changes the magnetism of the two components from parallel (low resistance) to antiparallel (high resistance). The effect depends on the ballistic travel of electrons across the nanocontact (as opposed to diffusive transport that occurs over longer distances in a metal). Discovered in 1999 by N. Garcia, M. Muñoz and Y.-W. Zhao of the Council for Scientific Research (CSIC) in Madrid. In late 2003 questions arose about the validity of some BMR results, which some researchers believe were caused by a different effect. Use of BMR in read heads is a remote possibility.

EMR (extraordinary MR): Like ordinary MR and unlike all the other large MR effects, EMR is produced in a structure that does not include any magnetic material. See the main text for further discussion of EMR.

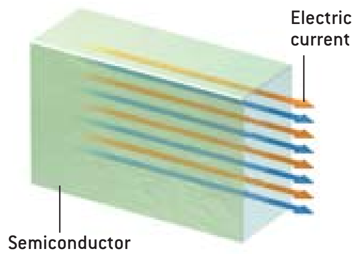
PRODUCING THE EXTRAORDINARY

Magnetoresistance arises when a magnetic field changes the configuration of electric fields in a piece of semiconductor, reducing the current flow. "Extraordinary" magnetoresistance (EMR) occurs when the geometry of a metal-semiconductor device interacts with the electric fields to greatly accentuate the effect.

CURRENT FLOW IN A SEMICONDUCTOR

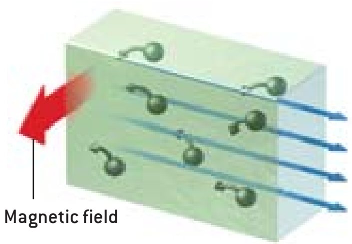


When a voltage is applied across the ends of a slab of semiconductor, it sets up an electric field that causes randomly moving electrons to drift along the slab on average.

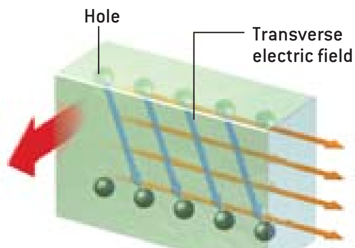


That drift of electrons adds up to an electric current flowing in parallel to the electric field lines. By convention, the direction of current flow is opposite to that of the electron drift.

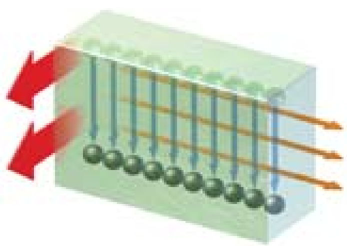
A MAGNETIC CROSSWIND



When a magnetic field is applied transversely, it curves the electrons' trajectories. That effect increases the path length that the electrons travel, which reduces their average drift velocity along the slab. Thus, the current is reduced, producing magnetoresistance.

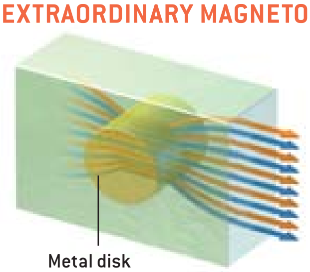


The curved paths also cause electrons to accumulate on the bottom surface of the slab and holes (positively charged absences of electrons) to accumulate on top. Together those charges generate a transverse electric field. The current flow is at an angle to the total electric field, deflected by the magnetic field.

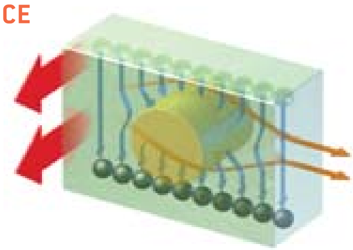


When the magnetic field is strong enough, the transverse electric field overwhelms the original electric field. Current now flows at a right angle to the electric field lines.

EXTRAORDINARY MAGNETORESISTANCE



A metal disk implanted in the semiconductor distorts the electric field lines, which are approximately perpendicular to a conductor's outer surface. The field lines and the current flow are concentrated through the metal disk, so more current flows through the device than when the disk is absent. That is, the device resistance is very low.



When a strong magnetic field is applied, again charges build up at top and bottom, the vertical electric field overwhelms the original field, and the electric field is distorted to be perpendicular to the metal disk. Again the current flow is at right angles to the electric field lines, which now causes the current to skirt around the disk. Being squeezed through narrow strips of semiconductor drastically reduces the current; the device's resistance is greatly increased—extraordinary magnetoresistance.

BRYAN CHRISTIE DESIGN

cused on heterostructures (structures combining two types of semiconductor) made of InSb and indium aluminum antimonide (InAlSb) layers. Dirk Grundler and his colleagues at Hamburg University in Germany have carried out extensive studies of EMR in indium arsenide–metal hybrid structures.

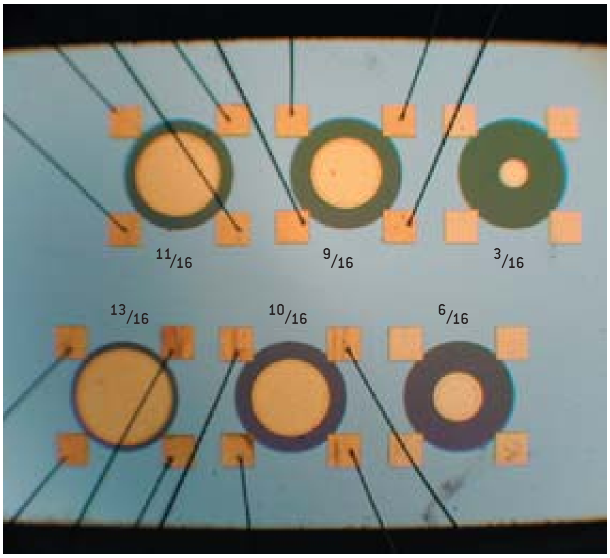
Myriad Applications

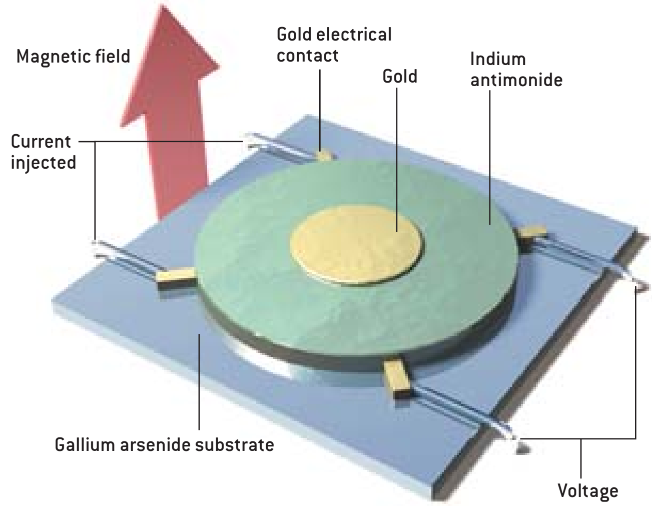
TWO FACTORS HAVE DRIVEN the discovery and study of MR phenomena during the past 16 years: pure intellectual curiosity and the promise of technological applications for magnetic sensors. That promise has certainly been realized in the case of giant MR, or GMR, which was discovered in 1988. Most, if not all, of the magnetic disk drives used in today's computers employ GMR read-head sensors to detect the magnetic bits of stored information.

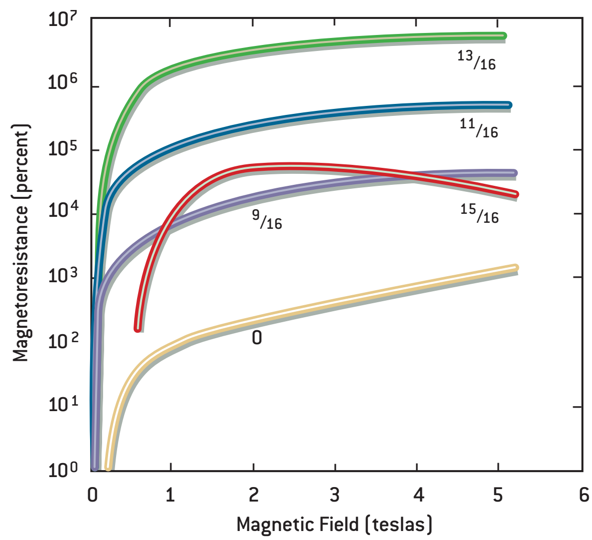
MR sensors in general and EMR sensors in particular have myriad potential applications. Industrial ones include process monitors, position-sensing robots for factory production lines, magnetic-field testing for machinery and engines, speed sensing for gears, and position sensors for ferromagnetic parts. Some automotive applications are antilock brakes, "smart" shock absorbers, vehicle counting systems, and ignition timing and control systems. Consumers benefit from the sensors in flip-phone switches, solid-state compasses, nonvolatile memory in low-cost appliances, elevator control switches, and noiseless motor controls in disk drives. They might see the gadgets used in banks to do currency sorting and counting based on magnetic inks. Sensors of both low and high fields could also find uses in medical devices.

Of these applications, computer disk-drive read heads are probably the most technologically challenging. Disk drives have three key components: the magnetic disk medium, which stores the information, the write-head element, which writes information onto the disk, and the read-head element, which reads the information. All three components will have to be improved significantly to satiate the ongoing demand for low-cost, high-speed storage at ever greater densities.

PROOF OF PRINCIPLE







Extraordinary magnetoresistance (EMR) was first demonstrated in disks (*photograph, above left; schematic, above right*) known as van der Pauw disks, after L. J. van der Pauw of Philips Laboratories. In the late 1950s he first studied current flow through various semiconductor shapes in the presence of a magnetic field. Gold disks were embedded in one-millimeter-radius disks of tellurium-doped indium antimonide. The disks were about 1.5 microns thick, deposited on a substrate of gallium arsenide. Numbers indicate the radius of the gold disk as a fraction of a millimeter. The voltage across two gold contacts on each disk was determined for a fixed amount of current flowing through two others. From those data, the EMR was computed and was found to improve with increasing gold-disk radius up to 13/16 of a millimeter (*graph, right*). The experiment was done at room temperature.

STUART A. SOLIN (*photograph*); BRYAN CHRISTIE DESIGN (*schematic*); LUCY READING (*graph*)

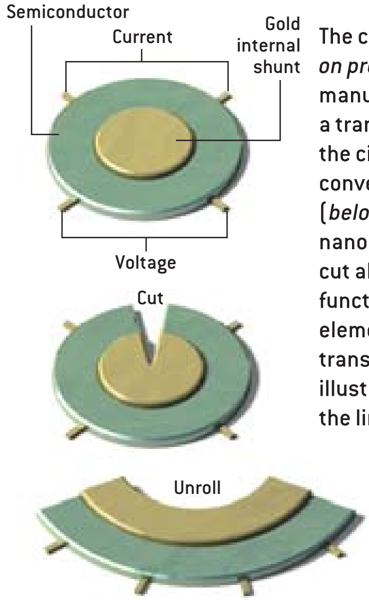
A bit of information on a magnetic disk consists of a small magnetized region that produces a minute magnetic field just above that spot on the disk. For simplicity, you can think of a field pointing up out of the disk as a "one" and a field pointing down into the disk as a "zero." Every square inch of a modern disk drive has about 20 billion of these bits, for a density of 20 gigabits per square inch (Gb/in^2). As the size of the bit is reduced to increase storage density, the read head must be shrunk and its sensitivity must increase to detect the weaker magnetic field of the smaller bit. The head must also respond to the field faster, because a smaller bit on the rotating disk spends less time under it.

In evaluating a read head, what really matters is not the raw magnitude of the MR effect but the head's signal-to-noise ratio, which depends on a number of other factors in addition to the amount of MR. A major source of noise for all MR sensors that use magnetized materials—that is, all but EMR—is magnetic noise. This effect occurs because the magnetism in the material is generated by innumerable magnetic atoms, like a host of tiny bar magnets, all roughly aligned but randomly fluctuating like compass needles being jiggled about. For large volumes of material, the fluctuations average out to be negligible relative to the total magnetism. As the sensor volume decreases, however, the proportion of noise increases. Magnetic noise might fundamentally limit read heads based on mag-

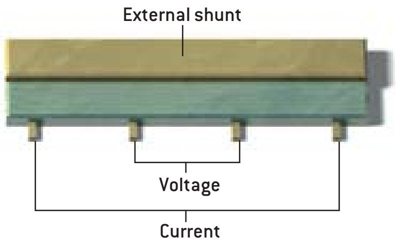
THE AUTHOR

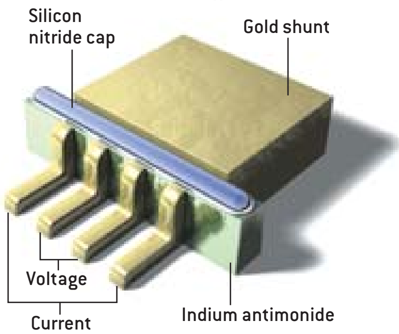
STUART A. SOLIN became Charles M. Hohenberg Professor of Experimental Physics in the College of Arts & Sciences at Washington University in 2002. Before that, he was a Fellow at the NEC Research Institute in Princeton, N.J., which awarded him the Best Patent Award for 1998 and the Technology Impact Award for 2000. He received an honorary doctor of science degree from Purdue University in 2003 (to add to his 1969 Ph.D. from there). Throughout his career, his research has focused on fundamental physical phenomena in ordered and disordered solids. He has published more than 230 scientific articles and been granted 15 patents. His hobbies are reading American history and playing the drums, with a preference for rock and jazz.

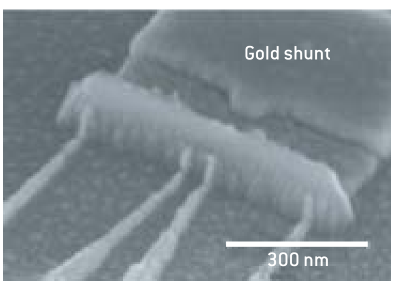
PRACTICAL DEVICES



The circular design of an EMR device [*see illustration on preceding page*] is not practical for commercial manufacture of nanoscopic electronics. By performing a transformation called a conformal mapping, however, the circular design with an internal shunt can be converted into a linear design with an external shunt (*below*) that is readily manufactured with existing nanoelectronic technology. In concept, the disks are cut along a radius and unrolled. The linear device functions exactly like a circular one, just with all its elements and the electric and magnetic fields transformed by the same mapping. (This schematic illustration is simplified for clarity; the proportions of the linear device are not as shown.)







A realistic design for a disk-drive read head based on a linear external shunt (*left*) was constructed in prototype (*below left*). The silicon nitride cap insulates the gold contacts on one side from the gold shunt on the other. The density of data storage possible with this read head is determined by the thickness of the indium antimonide (InSb) mesa and the separation of the two voltage contacts. The prototype read head could handle more than 700 gigabits per square inch, about three times the estimated density of next-generation TMR read heads. The quasiperiodic ripples on the side walls of the InSb layer were unintended but turned out to be crucial for the device to function.

netic materials to data densities of a few hundred Gb/in^2, yet the five-year target for the magnetic recording industry is 1,000 Gb/in^2, or one terabit per square inch (Tb/in^2). Because read heads based on EMR use nonmagnetic materials, they would not suffer from magnetic noise limitations. They thus represent a possible option for the design of a 1-Tb/in^2 read head.

But a density of 1 Tb/in^2 corresponds to each bit occupying a square 25 nanometers on a side. The read head needs to be of similar size. Fabricating a viable, nanoscopic EMR device is a significant scientific and technological challenge. The scientific problem is to design a nanoscopic device with sufficient EMR to be of practical use. When one crosses into the nanoscopic regime, the physics of the electrical conduction process changes in a way that significantly reduces EMR. Technologically, one must design a workable EMR nanostructure that can be fabricated using convenient techniques. Unfortunately, the disk structures used in the first demonstration of EMR are not easily scalable to nanoscopic dimensions. To meet these challenges, the NEC Princeton team was expanded to include physicists J. Shen Tsai and Yu. A. Pashkin of NEC Japan, who are, respectively, experts in the electrical conduction of nanostructures and in the methods of electron-beam lithography used to fabricate them.

The basic EMR device that I described earlier is said to be internally shunted—the metal disk (the shunt) lies inside a ring of semiconductor, like an island surrounded by a circular moat. My expanded NEC team realized that this geometry could be rearranged by a mathematical process called conformal mapping to form an externally shunted device [*see illustration at left*]. The end result is that the semiconductor and metal form two strips side by side, with all the electrical contacts—two for sensing a voltage, two for passing a current in and out—along the free edge of the semiconductor. The shunt (a strip of conductive metal) is now on the outside, across from all the electrical contacts. It is not as easy to describe in words what shape the electric field lines take or how the flowing carriers get deflected from those lines by a magnetic field (and hence avoid the conducting metal). But the magic of a conformal mapping is that the device is guaranteed to operate in identical fashion to the circular version, just with the geometry of all the critical elements (semiconductor, metal, field lines and current flows) transformed by the mapping. The linear, external-shunt version has the added advantage that it can be readily constructed and operated at nanoscopic sizes.

To make our devices, we used state-of-the-art electron-beam lithography—and airplanes: our samples traveled back and forth between Princeton to Tsukuba four times during their fabrication. The final result was worth all the miles that were clocked. We ended up with EMR devices—made from the Santos type of heterostructures—that had a raw data density of about 700 Gb/in^2. The practicalities of incorporating this proof-of-principle device into a working read head will probably roughly halve that density.

BRYAN CHRISTIE DESIGN; STUART A. SOLIN (*micrograph*)

To date, our group's design has achieved MR values in excess of 35 percent at a field of 0.05 tesla—good enough, I believe, for practical technology.

A Puzzle

WHILE STRIVING to grasp the fundamental physics of the EMR effect in the nanoscopic-size regime, we realized that the device that we had fabricated *should not work*. It "should" have an EMR of less than 1 percent. The reason has to do with how the electrons (or holes) travel in random zigzag paths that drift in the direction of the current flow (called diffusive transport). The average length of each straight section (the distance traveled between collisions with defects) is called the mean free path. Some of the elements of our nanoscopic structure are smaller than the mean free path. Consequently, a charge carrier is much more likely to strike the sidewall of the structure before it has a chance to ricochet from a defect. Therefore, a carrier's motion through the device is ballistic—it traverses the device in a straight line, not by zigzags. It works out that a magnetic field deflects ballistic carriers that are confined in nanoscopic structures much less than it does diffusive carriers traversing macroscopic structures. Thus, the smaller deflection at the metal–semiconductor interface in a nanoscopic device significantly reduces the EMR.

Fortunately, and not by design, our fabrication process produced rippled sidewalls with an approximate periodicity that enhanced the scattering of carriers striking them [*see micrograph in box on opposite page*]. This scattering transformed the carrier motion from ballistic to diffusive, and the large EMR associated with diffusive transport resulted. Serendipity, which led to the original discovery of the EMR phenomenon, was repeated in the process of fabricating a viable nanoscopic prototype! Happily, we now know why these ripples form, and we believe that we can control their size.

EMR read heads would have many good features in addition to the high data densities and low magnetic noise [*see table above*]. Their intrinsic response speed could be more than 100 times as fast as that of other read heads. They can be integrated onto semiconductor substrates easily and should have low fabrication costs. A disadvantage is the need to operate at relatively low temperatures—not much warmer than room temperature. One feature is both an advantage and a disadvantage: the response of the EMR read heads increases with the square of the magnetic field strength. That behavior is "nonlinear" (which is bad) but makes for high sensitivity (which is good).

The biggest challenge at present is that EMR sensors are still a new and unproved technology. Significant technological and economic barriers must be overcome if EMR is to be commercially successful for magnetic recording. Such obstacles are not atypical in the development of potentially disruptive technologies. Indeed, other disruptive technologies could render an EMR read head obsolete before it is even developed. Heat-assisted magnetic recording (HAMR), which is being worked on at Seagate Corporation, and Millipede nonmagnetic recording, an IBM project, are examples of competing technologies [see "Avoiding a Data Crunch," by Jon William Toigo; SCIENTIFIC AMERICAN, May 2000; and "The Nanodrive Project," by Peter Vettiger and Gerd Binnig; SCIENTIFIC AMERICAN, January 2003]. Yet even if that happens, EMR's discoverers hope to see EMR employed for a number of the other applications I have cited. What lies ahead is the hard work necessary to bring serendipity's two generous gifts to fruition. SA

HOW DO THEY PERFORM?

Different types of MR offer various combinations of attributes that may be useful for commercial applications. Here is how the MRs' expected maximal performance in delivered products stacks up against the five-year target of the magnetic recording industry. EMR read heads in computer hard drives, for instance, would have higher data densities and lower magnetic noise than today's GMR-based machines and could operate more than 100 times as fast. The data presented are based on public information; some companies may keep their latest results proprietary.

TYPE OF MR EFFECT USED	MR AT 300 KELVINS (percent)	DATA DENSITY (Gb/in^2)	SIGNAL-TO-NOISE RATIO (decibels) (larger is better)	TIME CONSTANT (nanoseconds) (smaller is faster)	MAGNETIC FIELD NEEDED (teslas) (smaller is better)
Target	4–10	100–1,000	30–40	0.01–0.1	0.005–0.05
EMR	> 35	> 300	43	< 0.001	0.05
GMR	10	125	29	0.1	0.005
TMR	15	200 estimated	34	0.1	0.001
CMR	0.4	100 estimated	–17	1.0	0.05
BMR	3,000	> 1,000	10	0.1	0.03

MORE TO EXPLORE

Special Issue on Magnetoelectronics. *Physics Today,* Vol. 48, No. 4; April 1995.

Colossal Magnetoresistance. Josep Fontcuberta in *Physics World,* Vol. 12, No. 2, pages 33–38; February 1999.

Enhanced Room-Temperature Geometric Magnetoresistance in Inhomogeneous Narrow-Gap Semiconductors. S. A. Solin, Tineke Thio, D. R. Hines and J. J. Heremans in *Science,* Vol. 289, pages 1530–1532; September 1, 2000.

Layered Magnetic Structures: History, Highlights, Applications. Peter Grünberg in *Physics Today,* Vol. 54, No. 5, pages 31–37; May 2001.

Semiconductor Fridges Get into Shape. Stuart A. Solin in *Physics World,* Vol. 14, No. 6, pages 29–30; June 2001.

Nonmagnetic Semiconductors as Read-Head Sensors for Ultra-High-Density Magnetic Recording. S. A. Solin et al. in *Applied Physics Letters,* Vol. 80, Issue 21, pages 4012–4014; May 27, 2002.

Today methane-producing microbes are confined to oxygen-free settings,

When Methane


METHANE-INDUCED HAZE on Earth two billion years ago.

such as the guts of cows, but in Earth's distant past, they ruled the world

Made Climate

By James F. Kasting

About 2.3 billion years ago

unusual microbes breathed new life into young Planet Earth by filling its skies with oxygen. Without those prolific organisms, called cyanobacteria, most of the life that we see around us would never have evolved.

Now many scientists think another group of single-celled microbes were making the planet habitable long before that time. In this view, oxygen-detesting methanogens reigned supreme during the first two billion years of Earth's history, and the greenhouse effect of the methane they produced had profound consequences for climate.

Scientists first began to suspect methane's dramatic role more than 20 years ago, but only during the past four years have the various pieces of the ancient methane story come together. Computer simulations now reveal that the gas—which survives about 10 years in today's atmosphere—could have endured for as long as 10,000 years in an oxygen-free world. No fossil remains exist from that time, but many microbiologists believe that methanogens were some of the first life-forms to evolve. In their prime, these microbes could have generated methane in quantities large enough to stave off a global deep freeze. The sun was considerably dimmer then, so the added greenhouse influence of methane could have been exactly what the planet needed to keep warm. But the methanogens did not dominate forever. The plummeting temperatures associated with their fading glory could explain Earth's first global ice age and perhaps others as well.

The prevalence of methane also means that a pinkish-orange haze may have shrouded the planet, as it does Saturn's largest moon, Titan. Although Titan's methane almost certainly comes from a nonbiological source, that moon's similarities to the early Earth could help reveal how greenhouse gases regulated climate in our planet's distant past.

Faint Sun Foiled

WHEN EARTH FORMED some 4.6 billion years ago, the sun burned only 70 percent as brightly as it does today [see "How Climate Evolved on the Terrestrial Planets," by James F. Kasting, Owen B. Toon and James B. Pollack; SCIENTIFIC AMERICAN, February 1988]. Yet the geologic record contains no convincing evidence for widespread glaciation until about 2.3 billion years ago, which means that the planet was probably even warmer than during the modern cycle of ice ages of the past 100,000 years. Thus, not only did greenhouse gases have to make up for a fainter sun, they also had to maintain average temperatures considerably higher than today's.

Methane was far from scientists' first choice as an explanation of how the young Earth avoided a deep freeze. Because ammonia is a much stronger greenhouse gas than methane, Carl Sagan and George H. Mullen of Cornell University suggested in the early 1970s that it was the culprit. But later research showed that the sun's ultraviolet rays rapidly destroy ammonia in an oxygen-free atmosphere. So this explanation did not work.

Another obvious candidate was carbon dioxide (CO_2), one of the primary gases spewing from the volcanoes abundant at that time. Although they debated the details, most scientists assumed for more than 20 years that this gas played the dominant role. In 1995, however, Harvard University researchers uncovered evidence that convinced many people that CO_2 levels were too low to have kept the early Earth warm.

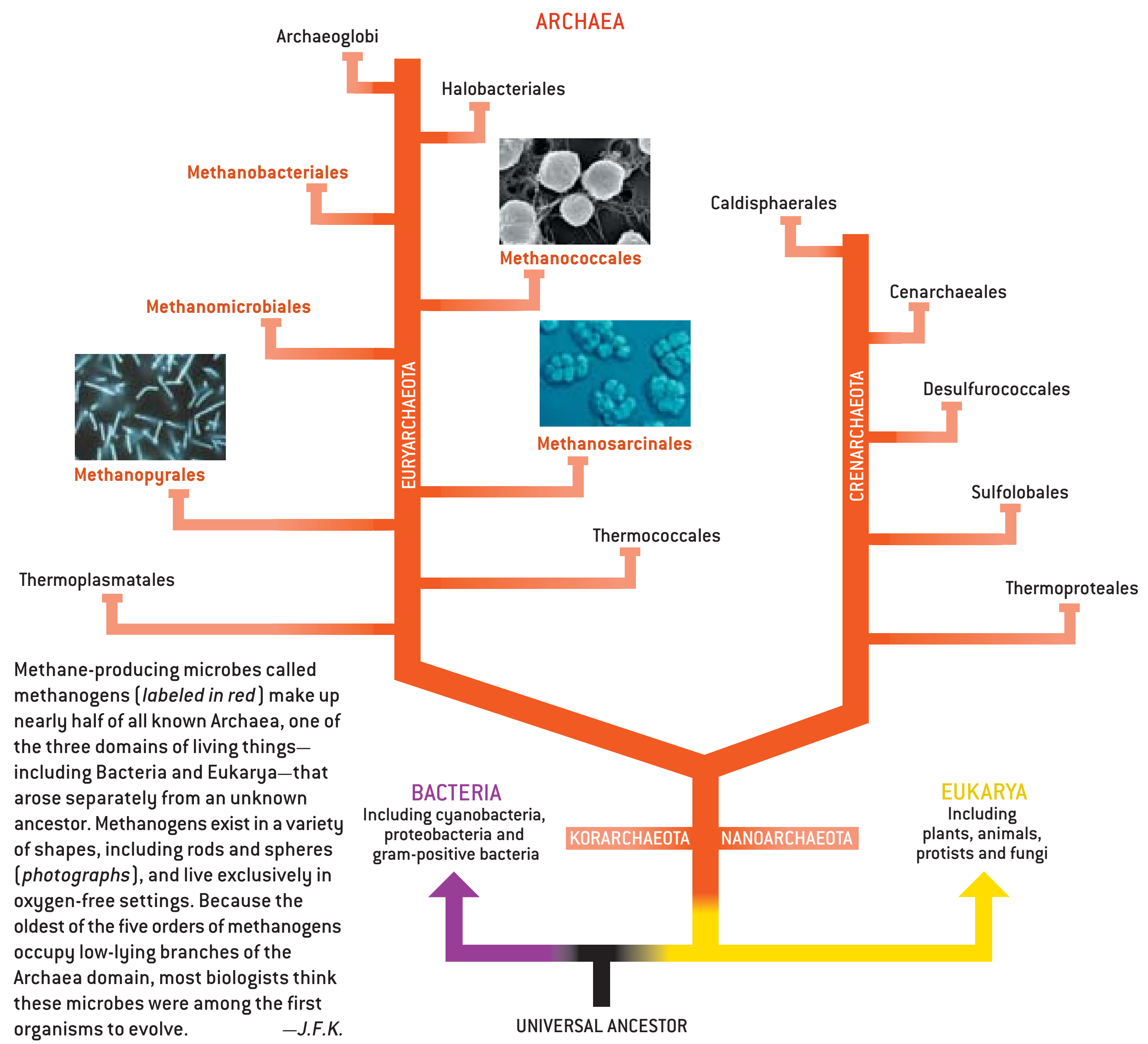
The Harvard team, led by Rob Rye, knew from previous studies that if the atmospheric concentrations of CO_2 had ex-

Overview/*The Microbe That Roared*

- Before about 2.3 billion years ago, Earth's atmosphere and oceans were virtually devoid of oxygen, making the world a nirvana for oxygen-intolerant microbes such as methanogens.
- Scientists now think that methanogens—so named because they give off methane gas as a waste product—could have filled the ancient skies with nearly 600 times as much methane as they do today.
- That extra methane would have produced a greenhouse effect powerful enough to warm the planet even though the sun burned less brightly at that time. Such was the case until the atmosphere filled with oxygen and drove the methanogens into seclusion.

DON DIXON (*preceding pages*)

METHANE MAKERS ON THE TREE OF LIFE



Methane-producing microbes called methanogens (*labeled in red*) make up nearly half of all known Archaea, one of the three domains of living things—including Bacteria and Eukarya—that arose separately from an unknown ancestor. Methanogens exist in a variety of shapes, including rods and spheres (*photographs*), and live exclusively in oxygen-free settings. Because the oldest of the five orders of methanogens occupy low-lying branches of the Archaea domain, most biologists think these microbes were among the first organisms to evolve. —*J.F.K.*

JOHNNY JOHNSON (*illustration*); BOONYARATANAKORNKIT AND D. S. CLARK *Chemical Engineering* AND G. VRDOLJAK *Electron Microscope Lab, University of California, Berkeley* (*top*); KARL O. STETTER *University of Regensburg* (*left*); GENOME NEWS NETWORK (*right*)

ceeded about eight times the present-day value of around 380 parts per million (ppm), the mineral siderite ($FeCO_3$) would have formed in the top layers of the soil as iron reacted with CO_2 in the oxygen-free air. But when the investigators studied samples of ancient soils from between 2.8 billion and 2.2 billion years ago, they found no trace of siderite. Its absence implied that the CO_2 concentration must have been far less than would have been needed to keep the planet's surface from freezing.

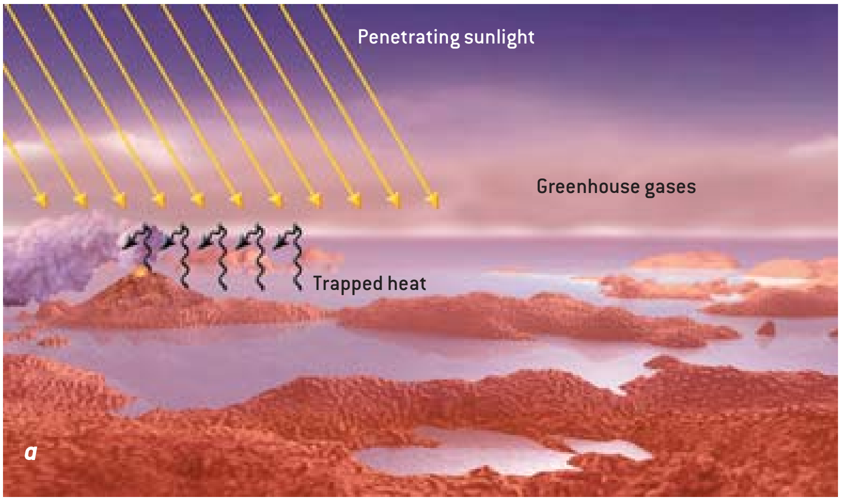
Even before CO_2 lost top billing as the key greenhouse gas, researchers had begun to explore an alternative explanation. By the late 1980s, scientists had learned that methane traps more heat than an equivalent concentration of CO_2 because it absorbs a wider range of wavelengths of Earth's outgoing radiation. But those early studies underestimated methane's influence. My group at Pennsylvania State University turned to methane because we knew that it would have had a much longer lifetime in the ancient atmosphere.

In today's oxygen-rich atmosphere, the carbon in methane is much happier teaming up with the oxygen in hydroxyl radicals to produce CO_2 and carbon monoxide (CO), releasing water vapor in the process. Consequently, methane remains in the atmosphere a mere 10 years and plays just a bit part in warming the planet. Indeed, the gas exists in minuscule concentrations of only about 1.7 ppm; CO_2 is roughly 220 times as concentrated at the planet's surface and water vapor 6,000 times.

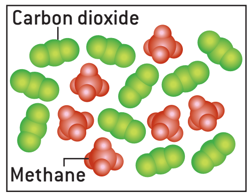
To determine how much higher those methane concentrations must have been to warm the early Earth, my students and

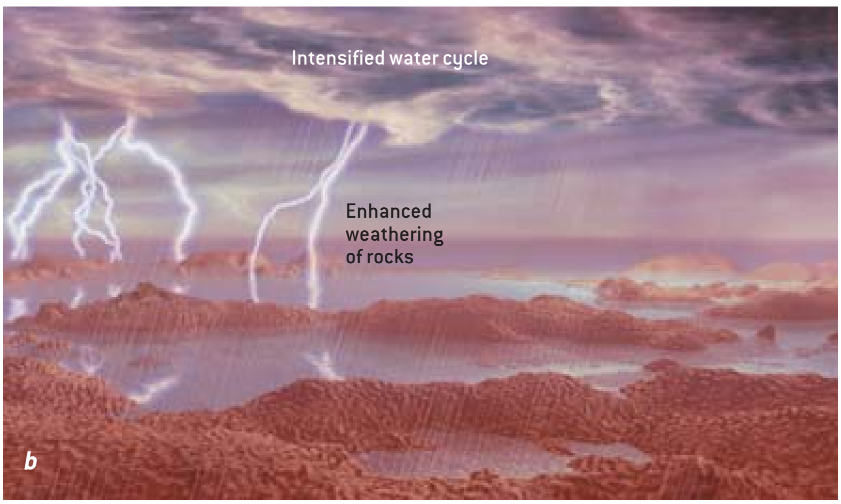
HOW HAZE FORMS

A methane-induced haze of hydrocarbon particles may have held the ancient Earth in a delicate balance between a hothouse and a deep freeze. The concentration of methane would have increased (*a*)—thereby intensifying the greenhouse effect (*b*)—for no more than a few tens of thousands of years before the climate-cooling haze would have developed (*c*).

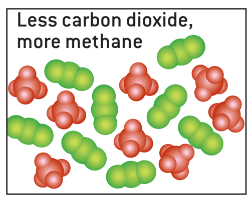


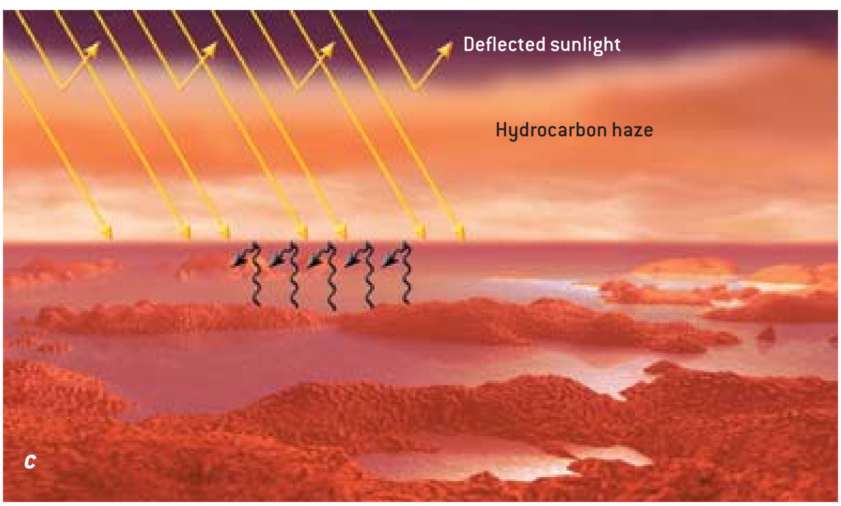
METHANE'S STARRING ROLE in Earth's atmosphere may have begun almost as soon as life originated more than 3.5 billion years ago. Single-celled ocean dwellers called methanogens would have thrived in a world virtually devoid of oxygen—as Earth was at that time—and the methane they produced would have survived in the atmosphere much longer than it does today. This methane—along with another, more abundant greenhouse gas, carbon dioxide (CO_2) from volcanoes (*inset*)—would have warmed the planet's surface by trapping Earth's outgoing heat (*black arrows*) while allowing sunlight (*yellow arrows*) to pass through.



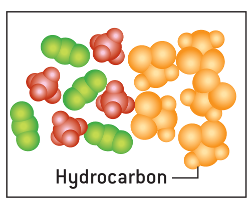


A HUMID GREENHOUSE is the preferred climate for many methanogens; the warmer the world became, the more methane they would have produced. This positive feedback loop would have strengthened the greenhouse effect, pushing surface temperatures even higher. Warmer conditions would have intensified the water cycle and enhanced the weathering of rocks on the continents—a process that pulls CO_2 out of the atmosphere. Concentrations of CO_2 would have dropped as those of methane continued to rise, until the two gases existed in nearly equal amounts (*inset*). Under such conditions, the behavior of methane would have altered dramatically.





CHANGING CHEMISTRY would have kept the rising methane levels from turning Earth into a hothouse. Some of the methane would have linked together to form complex hydrocarbons (*inset*) that then condensed into dustlike particles. A high-altitude haze of these particles would have offset the intense greenhouse effect by absorbing the visible wavelengths of incoming sunlight and reradiating them back to space, thereby reducing the total amount of warmth that reached the planet's surface. Fewer heat-loving methanogens could have survived in the cooler climate; the haze would thus have put a cap on overall methane production.



DON DIXON

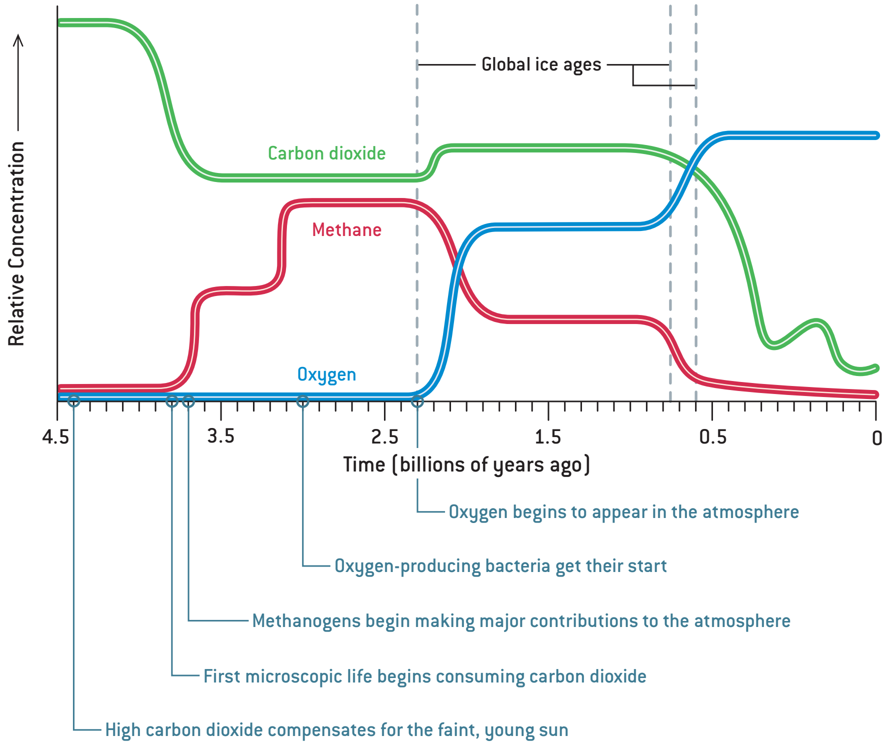
I collaborated with researchers from the NASA Ames Research Center to simulate the ancient climate. When we assumed that the sun was 80 percent as bright as today, which is the value expected 2.8 billion years ago, an atmosphere with no methane at all would have had to contain a whopping 20,000 ppm of CO_2 to keep the surface temperature above freezing. That concentration is 50 times as high as modern values and seven times as high as the upper limit on CO_2 that the studies of ancient soils revealed. When the simulations calculated CO_2 at its maximum possible value, the atmosphere required the help of 1,000 ppm of methane to keep the mean surface temperature above freezing—in other words, 0.1 percent of the atmosphere needed to be methane.

Up to the Task?

THE EARLY ATMOSPHERE could have maintained such high concentrations only if methane was being produced at rates comparable to today. Were primordial methanogens up to the task? My colleagues and I teamed up with microbiologist Janet L. Siefert of Rice University to try to find out.

Biologists have several reasons to suspect that such high methane levels could have been maintained. Siefert and others think that methane-producing microbes were some of the first microorganisms to evolve. They also suggest that methanogens would have filled niches that oxygen producers and sulfate reducers now occupy, giving them a much more prominent biological and climatic role than they have in the modern world.

Methanogens would have thrived in an environment fueled by volcanic eruptions. Many methanogens feed directly on hydrogen gas (H_2) and CO_2 and belch methane as a waste product; others consume acetate and other compounds that form as organic matter decays in the absence of oxygen. That is why today's methanogens can live only in oxygen-free environments such as the stomachs of cows and the mud under flooded rice paddies. On the early Earth, however, the entire atmosphere was devoid of oxygen, and volcanoes released significant amounts of H_2. With no oxygen available to form water,



RELATIVE CONCENTRATIONS of major atmospheric gases may explain why global ice ages (*dashed lines*) occurred in Earth's distant past. Methane-producing microorganisms flourished initially, but as oxygen skyrocketed about 2.3 billion years ago, these microbes suddenly found few environments where they could survive. The accompanying decrease in methane—a potent greenhouse gas—could have chilled the entire planet. The role of carbon dioxide, the most notable greenhouse gas in today's atmosphere, was probably much less dramatic.

H_2 probably accumulated in the atmosphere and oceans in concentrations high enough for methanogens to use.

Based on these and other considerations, some scientists have proposed that methanogens living on geologically derived hydrogen might form the base of underground microbial ecosystems on Mars and on Jupiter's ice-covered moon, Europa. Indeed, a recent report from the European Space Agency's Mars Express spacecraft suggests that the present Martian atmosphere may contain approximately 10 parts per billion of methane. If verified, this finding would be consistent with having methanogens living below the surface of Mars.

Geochemists estimate that on the early Earth H_2 reached concentrations of hundreds to thousands of parts per million—that is, until methanogens evolved and converted most of it to methane. Thermodynamic calculations reveal that if other essential nutrients, such as phosphorus and nitrogen, were available, methanogens would have used most of the available H_2 to make methane. (Most scientists agree that sufficient phosphorus would have come from the chemical breakdown of rocks and that various

THE AUTHOR

JAMES F. KASTING studies the origin and evolution of planetary atmospheres, especially those of Earth and its nearest neighbors, Venus and Mars. Since earning his Ph.D. in atmospheric science at the University of Michigan at Ann Arbor in 1979, he has used theoretical computer models to investigate atmospheric chemistry and to calculate the greenhouse effect of different gases and particles in both the present day and the distant past. Recently Kasting has begun exploring the question of whether Earth-like planets might exist around other stars in our galaxy. He is working with several other scientists to develop the theoretical foundation for NASA's Terrestrial Planet Finder, a space-based telescope designed to locate planets around other stars and scan their atmospheres for signs of life.

JOHNNY JOHNSON

Methane would never again

exert a dominating effect on climate, but it could still have been an important influence at later times.

ocean-dwelling microorganisms were producing plenty of nitrogen.) In this scenario, the methanogens would have yielded the roughly 1,000 ppm of methane called for by the computer models to keep the planet warm.

Even more evidence for the primordial dominance of methanogens surfaced when microbiologists considered how today's methanogens would have reacted to a steamy climate. Most methanogens grow best at temperatures above 40 degrees Celsius; some even prefer at least 85 degrees C. Those that thrive at higher temperatures grow faster, so as the intensifying greenhouse effect raised temperatures at the planet's surface, more of these faster-growing, heat-loving specialists would have survived. As they made up a larger proportion of the methanogen population, more methane molecules would have accumulated in the atmosphere, making the surface temperature still warmer—in fact, hotter than today's climate, despite the dimmer sun.

Smog Saves the Day

AS A RESULT of that positive feedback loop, the world could have eventually become such a hothouse that life itself would have been difficult for all but the most extreme heat–loving microbes. This upward spiral could not have continued indefinitely, however. Once atmospheric methane becomes more abundant than CO_2, methane's reaction to sunlight changes. Instead of being oxidized to CO or CO_2, it polymerizes, or links together, to form complex hydrocarbons that then condense into particles, forming an organic haze. Planetary scientists observe a similar haze in the atmosphere of Saturn's largest moon: Titan's atmosphere consists primarily of molecular nitrogen, N_2, along with a small percentage of methane [*see box on opposite page*]. The scientists hope to learn more when NASA's Cassini spacecraft arrives at Saturn in July [see "Saturn at Last!" by Jonathan I. Lunine; SCIENTIFIC AMERICAN, June].

The possible formation of organic haze in Earth's young atmosphere adds a new wrinkle to the climate story. Because they form at high altitudes, these particles have the opposite effect on climate that greenhouse gases do. A greenhouse gas allows most visible solar radiation to pass through, but it absorbs and reradiates outgoing infrared radiation, thereby warming the surface. In contrast, high-altitude organic haze absorbs incoming sunlight and reradiates it back into space, thereby reducing the total amount of radiation that reaches the surface. On Titan, this so-called antigreenhouse effect cools the surface by seven degrees C or so. A similar haze layer on the ancient Earth would have also cooled the climate, thus shifting the methanogen population back toward those slower-growing species that prefer cooler weather and thereby limiting further increases in methane production. This powerful negative feedback loop would have tended to stabilize Earth's temperature and atmospheric composition at exactly the point at which the layer of organic haze began to form.

Nothing Lasts Forever

METHANE-INDUCED SMOG kept the young Earth comfortably warm—but not forever. Global ice ages occurred at least three times in the period known as the Proterozoic eon, first at 2.3 billion years ago and again at 750 million and 600 million years ago. The circumstances surrounding these glaciations were long unexplained, but the methane hypothesis provides compelling answers here as well.

The first of these glacial periods is often called the Huronian glaciation because it is well exposed in rocks just north of Lake Huron in southern Canada. Like the better-studied late Proterozoic glaciations, the Huronian event appears to have been global, based on interpretations that some of the continents were near the equator at the time ice covered them.

This cold snap formed layers of jumbled rocks and other materials that a glacier carried and then dropped to the ground when the ice melted sometime between 2.45 billion and 2.2 billion years ago. In the older rocks below these glacial deposits are detrital uraninite and pyrite, two minerals considered evidence for very low levels of atmospheric oxygen. Above the glacial layers sits a red sandstone containing hematite—a mineral that forms only under oxygen-rich skies. (Hematite has also been found at the landing site of the Mars rover Opportunity. This hematite is gray, however, because the grain size is larger.) The layering of these distinct rock types indicates that the Huronian glaciations occurred precisely when atmospheric oxygen levels first rose.

This apparent coincidence remained unexplained until recently: if we hypothesize that methane kept the ancient climate warm, then we can predict a global ice age at 2.3 billion years ago because it would have been a natural consequence of the rise of oxygen. Many of the methanogens and other anaerobic organisms that dominated the planet before the rise of oxygen would have either perished in this revolution or found themselves confined to increasingly restricted habitats.

Although this finale sounds as if it is the end of the methane story, that is not necessarily the case. Methane never again exerted a dominating effect on climate, but it could still have been an important influence at later times—during the late Proterozoic, for example, when some scientists suggest that the oceans froze over entirely during a series of so-called snowball Earth episodes [see "Snowball Earth," by Paul F. Hoffman and Daniel P. Schrag; SCIENTIFIC AMERICAN, January 2000].

Indeed, methane concentrations could have remained significantly higher than today's during much of the Proterozoic

Hazy Comparison

Saturn's moon Titan gets its characteristic orange glow from a dense layer of hydrocarbon particles that forms as sunlight destroys methane high in its atmosphere. Scientists have recently begun to think that a comparable haze cloaked Earth before 2.3 billion years ago. But fortunately for the planet's earliest inhabitants, the similarity stops there.

A chilling –179 degrees Celsius, Titan's atmosphere is dramatically colder than Earth's has ever been. On Earth an organic haze as thick as Titan's would have deflected enough sunlight to counteract the warming effect of methane—a potent greenhouse gas. The planet's surface would have frozen solid, thereby killing the single-celled microbes that created the methane in the first place.

NASA/JPL-CALTECH

Researchers speculate that Titan's haze grows thick because methane gas evaporates readily from an abundant ocean of liquid methane, nitrogen and ethane. Earth's ancient microbes released a mere puff of the gas by comparison, which helped to keep its haze layer relatively thin.

What the primordial Earth lacked in methane it made up for in carbon dioxide and liquid water—two ingredients that made possible the evolution of life. Because scientists have seen no sign of either compound on Titan, they conclude that life as we know it could not have evolved there. But that does not mean Saturn's largest moon cannot tell us something about biological evolution.

Much of the same chemistry that takes place in Titan's atmosphere probably also occurred on the young Earth. The European Space Agency plans to test that theory with its Huygens probe, which is scheduled to arrive at Saturn onboard NASA's Cassini spacecraft in July. If the probe enters Titan's atmosphere successfully next year, investigators will obtain the first direct evidence of what methane-induced smog is really like. Such observations could provide clues about how Earth maintained its delicate balance between the climate-cooling haze and the methane-dominated greenhouse that kept the planet habitable for well over a billion years. *—J.F.K.*

eon, which ended about 600 million years ago, if atmospheric oxygen had continued to be somewhat lower and the deep oceans were still anoxic and low in sulfate, a dissolved salt common in modern seawater. The rate at which methane escaped from the seas to the atmosphere could still have been up to 10 times as high as it is now, and the concentration of methane in the atmosphere could have been as high as 100 ppm. This scenario might explain why the Proterozoic remained ice-free for almost a billion and a half years despite the fact that the sun was still relatively dim. My colleagues and I have speculated that a second rise in atmospheric oxygen, or in dissolved sulfate, could conceivably have triggered the snowball Earth episodes as well—once again by decreasing the warming presence of methane.

Extraterrestrial Methane

AS COMPELLING AS this story of methanogens once ruling the world may sound, scientists are forced to be content with no direct evidence to back it up. Finding a rock that contains bubbles of ancient atmosphere would provide absolute proof, but such a revelation is unlikely. The best we can say is that the hypothesis is consistent with several indirect pieces of evidence—most notably, the low atmospheric CO_2 levels inferred from ancient soils and the timing of the first planet-encompassing ice age.

Although we may never be able to verify this hypothesis on Earth, we may be able to test it indirectly by observing Earth-like planets orbiting other stars. Both NASA and the European Space Agency are designing large space-based telescopes to search for Earth-size planets orbiting some 120 nearby stars. If such planets exist, these missions—NASA's Terrestrial Planet Finder and ESA's Darwin—should be able to scan their atmospheres for the presence of gases that would indicate the existence of life.

Oxygen at any appreciable abundance would almost certainly indicate biology comparable to that of modern Earth, provided that the planet was also endowed with the liquid water necessary for life. High levels of methane, too, would suggest some form of life. As far as we know, on planets with Earth-like surface temperatures only living organisms can produce methane at high levels. The latter discovery might be one of the best ways for scientists to gain a better understanding of what our own planet was like during the nascent stages of its history. SA

MORE TO EXPLORE

Greenhouse Warming by CH_4 in the Atmosphere of Early Earth. Alexander A. Pavlov, James F. Kasting, Lisa L. Brown, Kathy A. Rages and Richard Freedman in *Journal of Geophysical Research—Planets,* Vol. 105, No. E5, pages 11,981–11,990; May 2000.

Life and the Evolution of Earth's Atmosphere. James F. Kasting and Janet L. Siefert in *Science,* Vol. 296, pages 1066–1068; May 10, 2002.

Methane-Rich Proterozoic Atmosphere? Alexander A. Pavlov, Matthew T. Hurtgen, James F. Kasting and Michael A. Arthur in *Geology,* Vol. 31, No. 1, pages 87–90; January 2003.

Detecting MAD COW Disease

New tests can rapidly identify the presence of dangerous prions—the agents responsible for the malady—and several compounds offer hope for treatment

By Stanley B. Prusiner

Last December mad cow disease made its U.S. debut when federal officials announced that a holstein from Mabton, Wash., had been stricken with what is formally known as bovine spongiform encephalopathy (BSE). The news kept scientists, government officials, the cattle industry and the media scrambling for information well past New Year's. Yet the discovery of the sick animal came as no surprise to many of us who study mad cow disease and related fatal disorders that devastate the brain. The strange nature of the prion—the pathogen at the root of these conditions—made us realize long ago that controlling these illnesses and ensuring the safety of the food supply would be difficult.

As researchers learn more about the challenges posed by prions—which can incubate without symptoms for years, even decades—they uncover strategies that could better forestall epidemics. Key among these tools are highly sensitive tests, some available and some under development, that can detect prions even in asymptomatic individuals; currently BSE is diagnosed only after an animal has died naturally or been slaughtered. Researchers have also made some headway in treating a human prion disorder called Creutzfeldt-Jakob disease (CJD), which today is uniformly fatal.

Identifying the Cause

ALTHOUGH MASS CONCERN over mad cow disease is new in the U.S., scientific efforts to understand and combat it and related disorders began heating up some time ago. In the early 1980s I proposed that the infectious pathogen causing scrapie (the sheep analogue of BSE) and CJD consists only of a protein, which I termed the prion. The prion theory was greeted with great skepticism in most quarters and with outright disdain in others, as it ran counter to the conventional wisdom that pathogens capable of reproducing must contain DNA or RNA [see "Prions," by Stanley B. Prusiner; SCIENTIFIC AMERICAN, October 1984]. The doubt I encountered was healthy and important, because most dramatically novel ideas are eventually shown to be incorrect. Nevertheless, the prion concept prevailed.

In the years since my proposal, investigators have made substantial progress in deciphering this fascinating protein. We know that in addition to causing scrapie and CJD, prions cause other spongiform encephalopathies, including BSE and chronic wasting disease in deer and elk [see "The Prion Diseases," by Stanley B. Prusiner; SCIENTIFIC AMERICAN, January 1995]. But perhaps the most startling finding has been that the prion protein, or PrP, is not always bad. In fact, all animals studied so far have a gene that codes for PrP. The normal form of the protein, now called PrP^{C} (C for cellular), appears predominantly in nerve cells and may help maintain neuronal functioning. But the protein can twist into an abnormal, disease-causing shape, denoted PrP^{Sc} (Sc for scrapie, the prion disease that until recently was the most studied).

Unlike the normal version, PrP^{Sc} tends to form difficult-to-dissolve clumps that resist heat, radiation and chemicals that would kill other pathogens. A few minutes of boiling wipes out bacteria, viruses and molds, but not PrP^{Sc}. This molecule makes more of itself by converting normal prion proteins into abnormal forms: PrP^{Sc} can induce PrP^{C} to refold and become PrP^{Sc}.

YUN SUK-BONG *EPA*

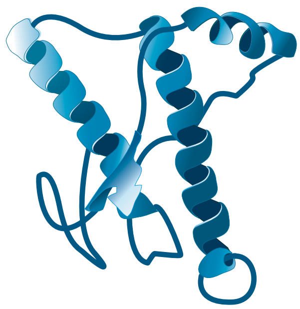
MEAT GETS TESTED in South Korea to determine its origin after the December 2003 discovery that a cow from Washington State had bovine spongiform encephalopathy (BSE). The American beef market has lost millions of dollars since the finding because of import bans enacted by many countries.

Cells have the ability to break down and eliminate misfolded proteins, but if they have difficulty clearing PrP^{Sc} faster than it forms, PrP^{Sc} builds up, ruptures cells and creates the characteristic pathology of these diseases—namely, masses of protein and microscopic holes in the brain, which begins to resemble a sponge. Disease symptoms appear as a result.

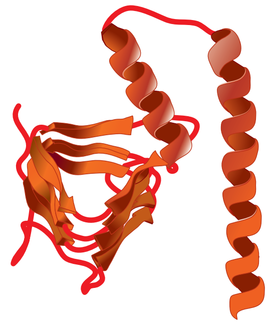
Prion diseases can afflict people and animals in various ways. Most often the diseases are "sporadic"—that is, they happen spontaneously for no apparent reason. Sporadic CJD is the most common prion disease among humans, striking approximately one in a million, mostly older, people. Prion diseases may also result from a mutation in the gene that codes for the prion protein; many families are known to pass on CJD and two other disorders, Gerstmann-Sträussler-Scheinker disease and fatal insomnia. To date, researchers have uncovered more than 30 different mutations in the PrP gene that lead to the hereditary forms of the sickness—all of which are rare, occurring about once in every 10 million people. Finally, prion disease may result from an infection, through, for instance, the consumption of bovine prions.

A BAD INFLUENCE

Prion proteins exist in at least two forms (*below*)—the normal, or cellular, version (PrP^{C}) and the disease-causing one (PrP^{Sc}). In a process that is not well understood, PrP^{Sc} changes PrP^{C} into more PrP^{Sc}; the newly altered prions, in turn, can corrupt other normal ones (*bottom*). Usually the body eliminates PrP^{Sc} before too much of it accumulates. But if it does build up and is not successfully removed by the cell's machinery, illness can result.

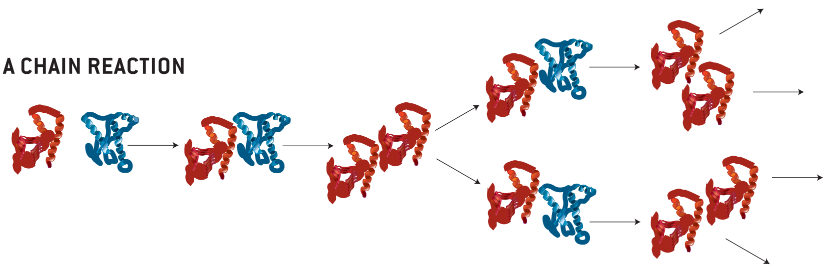


NORMAL PRION PROTEIN (PrP^{C})



DISEASE-CAUSING PRION (PrP^{Sc})

A CHAIN REACTION



LUCY READING

Tracing the Mad Cow Epidemic

THE WORLD AWOKE to the dangers of prion disease in cows after the BSE outbreak that began ravaging the British beef industry in the mid-1980s. The truly novel concepts emerging from prion science forced researchers and society to think in unusual ways and made coping with the epidemic difficult. Investigators eventually learned that prions were being transmitted to cattle through meat-and-bone meal, a dietary supplement prepared from the parts of sheep, cattle, pigs and chickens that are processed, or rendered, for industrial use. High heat eliminated conventional pathogens, but PrP^{Sc} survived and went on to infect cattle.

As infected cattle became food for other cattle, BSE began appearing throughout the U.K. cattle population, reaching a high of 37,280 identified cases in 1992. The British authorities instituted some feed bans beginning in 1989, but it was not until 1996 that a strict ban on cannibalistic feeding finally brought BSE under control in the U.K.; the country saw 612 cases last year. Overall the U.K. has identified about 180,000 mad cows, and epidemiological models suggest that another 1.9 million were infected but undetected.

For many people, the regulations came too late. Despite the British government's early assurances to the contrary, mad cow disease proved transmissible to humans. In March 1996 Robert Will, James Ironside and Jeanne Bell, who were working in the National CJD Surveillance Unit in Edinburgh, reported that 11 British teenagers and young adults had died of a variant of Creutzfeldt-Jakob disease (vCJD). In these young patients the patterns of PrP^{Sc} deposition in the brain differed markedly from that found in typical CJD patients.

Many scientists, including myself, were initially dubious of the presumed link between BSE and vCJD. I eventually changed my mind, under the weight of many studies. One of these was conducted by my colleagues at the University of California at San Francisco, Michael

Overview/*Rooting Out Prions*

- The recent discoveries of cows afflicted with mad cow disease, or bovine spongiform encephalopathy (BSE), in the U.S., Canada and elsewhere emphasize the need for better tests and policies so that infection does not spread and the public is reassured.
- Several new tests make such identification much more rapid than it has been, and some are able to identify low levels of dangerous prions so officials will not have to wait until cattle are sick before they know there is a problem.
- Invention of a blood test would allow diagnoses to be made for living animals and for people and could be important if some of the therapies now being investigated to treat prion diseases in humans prove to be effective.

Scott and Stephen DeArmond, who collected data in mice genetically engineered to resemble cattle, at least from a prion protein point of view (the PrP gene from cattle was inserted into the mouse genome). These mice became ill approximately nine months after receiving injections of prions from cattle with BSE or people with vCJD, and the resulting disease looked the same whether the prions originated from cows or vCJD patients.

As of February 2004, 146 people have been diagnosed with vCJD in the U.K. and another 10 elsewhere. No one knows exactly how many other people are incubating prions that cause vCJD. Epidemiological models suggest that only a few dozen more individuals will develop vCJD, but these models are based on assumptions that may prove wrong. One assumption, for example, is that vCJD affects only those with a particular genetic makeup. Because prions incubate for so long, it will take some time before we know the ultimate number of vCJD cases and whether they share similar genetics.

In vCJD, PrP^{Sc} builds up, not just in the brain but also in the lymphoid system, such as the tonsils and appendix, suggesting that PrP^{Sc} enters the bloodstream at some point. Animal studies have shown that prions can be transmitted to healthy animals through blood transfusions from infected animals. In response to this information, many nations have enacted stricter blood donation rules. In the U.K., people born after 1996, when the tough feed ban came into force, can receive blood only from overseas (those born before are considered already exposed). In the U.S., those who spent three months or more in the U.K. between 1980 and 1996 cannot give blood.

Although such restrictions have contributed to periodic blood shortages, the measures appear justified. Last December the U.K. announced the vCJD death of one of 15 individuals who received transfusions from donors who later developed vCJD. The victim received the transfusion seven and a half years before his death. It is possible that he became infected through prion-tainted food, but his age argues against that: at 69, he was much older than the 29 years of typical vCJD patients. Thus, it seems fairly likely that vCJD is not limited to those who have eaten prion-infected beef.

Since the detection of mad cow disease in the U.K., two dozen other nations have uncovered cases. Canada and the U.S. are the latest entrants. On May 20, 2003, Canadian officials reported BSE in an eight-and-a-half-year-old cow that had spent its life in Alberta and Saskatchewan. (The country's only previous mad cow had arrived as a U.K. import 10 years earlier.) Although the animal had been slaughtered in January 2003, slow processing meant that officials did not test the cow remains until April. By then, the carcass had been turned into pet food exported to the U.S.

Seven months later, on December 23, the U.S. Department of Agriculture announced the country's first case of BSE, in Washington State. The six-and-a-half-year-old dairy cow had entered the U.S. at the age of four. The discovery means that U.S. officials can no longer labor under the misconception that the nation is free of BSE. Like Canada, U.S. agricultural interests want the BSE problem to disappear. Financial woes stem primarily from reduced beef exports: 58 other countries are keeping their borders shut, and a $3-billion export market has largely evaporated.

A WORLDVIEW

Many countries have reported cattle afflicted with BSE, but cases of human infection thought to stem from eating meat from sick cows—variant Creutzfeldt-Jakob disease (vCJD)—remain low, for now at least.

COUNTRY	BSE CASES	vCJD DEATHS (current cases)
Austria	1	0
Belgium	125	0
Canada	2	1
Czech Republic	9	0
Denmark	13	0
Falkland Islands	1	0
Finland	1	0
France	891	6
Germany	312	0
Greece	1	0
Hong Kong	0	1*
Ireland	1,353	1
Israel	1	0
Italy	117	1
Japan	11	0
Lichtenstein	2	0
Luxembourg	2	0
Netherlands	75	0
Oman	2	0
Poland	14	0
Portugal	875	0
Slovakia	15	0
Slovenia	4	0
Spain	412	0
Switzerland	453	0
U.S.	1	0 (1)†
U.K.	183,803	141 (5)
Worldwide		151(6)

*Awaiting confirmation †British subject

DATA AS OF APRIL 30, 2004; SOURCES: INTERNATIONAL ORGANIZATION FOR EPIZOOTIC DISEASES; WORLD HEALTH ORGANIZATION; UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Designing Diagnostics

THE MOST straightforward way to provide this assurance—both for foreign nations and at home—appears to be simple: just test the animals being slaughtered for food and then stop the infected ones from entering the food supply, where they could transmit pathogenic prions to humans. But testing is not easy. The USDA uses immunohistochemistry, an old technique that is cumbersome and extremely time-consuming (taking a few days to complete) and so is impractical for universal application.

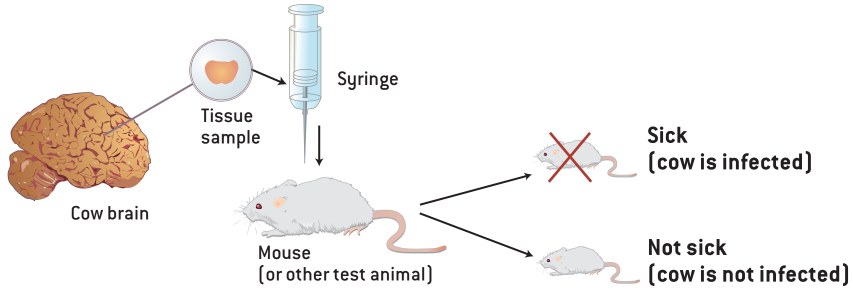
Accordingly, others and I have been working on alternatives. In the mid-1980s researchers at my lab and elsewhere produced new kinds of antibodies that can help identify dangerous prions in the brain more efficiently. These antibodies, similar to those used in the standard test, recognize any prion—normal or otherwise. To detect PrP^{Sc}, we need to first remove any trace of PrP^{C}, which is done by applying a protease (protein-degrading enzyme) to a brain sample. Because PrP^{Sc} is generally resistant to the actions of proteases, much of it remains intact. Antibodies then added to the sample will reveal the amount of PrP^{Sc} present. Using a similar approach, a handful of companies, including Prionics in Switzerland and Bio-Rad in France, have developed their own antibodies and commercial kits. The results can be obtained in a few hours, which is why such

TESTING FOR MAD COW

Four kinds of tests are now used to detect dangerous prions (PrP^{Sc}) in brain tissue from dead cattle. By identifying infected animals, public health officials and farmers can remove them from the food supply. Some of these tests, however, are time-consuming and expensive, so researchers are working to develop the ideal diagnostic: one that could quickly detect even tiny amounts of PrP^{Sc} in blood and urine and thus could work for live animals and people. The hope is to forestall outbreaks by catching infection as early as possible and, eventually, to treat infection before it progresses to disease.

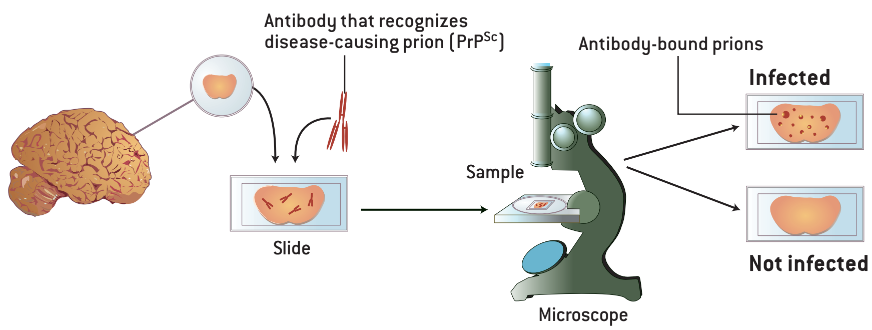
BIOASSAY

The bioassay can take up to 36 months to provide results and can be very expensive. Its advantage is that it can reveal particular strains of prions as well as how infectious the sample is based on the time it takes for the test animal to become sick.



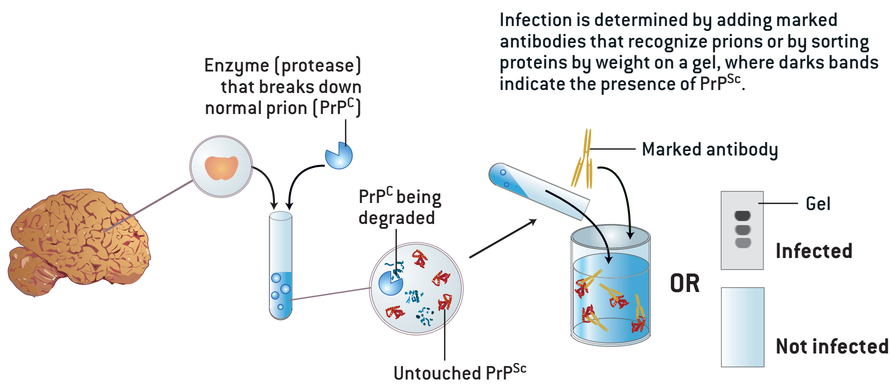
IMMUNOHISTOCHEMISTRY

The first test used to specifically detect prions, immunohistochemistry is considered the gold standard that other tests must meet. But because technicians must examine each slide, the process is very time-consuming, often taking as many as seven days, and is not useful for mass screenings.



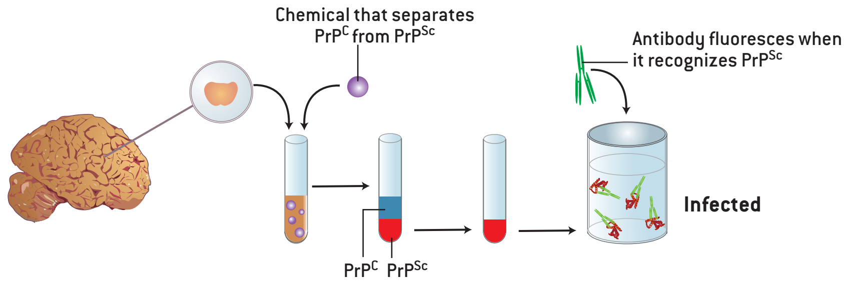
IMMUNOASSAY

Many companies produce immunoassays, and these rapid tests are now in wide-spread use in Europe; they have just been introduced to the U.S. Results can be had within eight hours, and hundreds of samples can be run simultaneously. These tests work well only with high levels of PrP^{Sc}.



CONFORMATION-DEPENDENT IMMUNOASSAY (CDI)

This automated test can detect very low levels of PrP^{Sc} and reveal how much of the dangerous prion is present in the sample without first having to degrade PrP^{C}. It informs experts about the animal's level of infection within five hours or so. It has been approved for testing in Europe. CDI is being tested on tissue from live animals and might one day serve as a blood test.



LUCY READING

kits are proving useful in the mass screenings under way in Europe and Japan. (Japan discovered its first case of BSE in 2001 and by this past April had reported 11 infected cows in total.)

These rapid tests, however, have limitations. They depend on PrP^{Sc} accumulating to detectable amounts—quite often, relatively high levels—in an animal's brain. Yet because BSE often takes three to five years to develop, most slaughter-age cattle, which tend to be younger than two years, usually do not test positive, even if they are infected. Therefore, these tests are generally most reliable for older bovines, regardless of whether they look healthy or are "downers." At the moment, downer cattle, which cannot stand on their own, are the group most likely to be tested.

Until new regulations came out in January, the U.S. annually sent roughly 200,000 downers to slaughter for human consumption. Of these, only a fraction were tested. Over the next year, however, the USDA will examine at least 200,000 cows for BSE. (Whether milk can be affected remains open; my laboratory is testing milk from BSE-infected cattle.)

Because of the limitations of the existing tests, developing one that is able to detect prions in the bulk of the beef supply—that is, in asymptomatic young animals destined for slaughter—continues to be one of the most important weapons in confronting prion disease outbreaks.

Scientists have pursued several strategies. One tries to boost the amount of PrP^{Sc} in a sample so the prions are less likely to escape detection. If such an amplification system can be created, it might prove useful in developing a blood test for use in place of ones that require an animal's sacrifice (not enough PrP^{Sc} circulates in the blood to be detectable by current methods). Claudio Soto of Serono Pharmaceuticals and his colleagues have attempted to carry this out. They mixed brain preparations from normal and scrapie-infected hamsters and then subjected the mixture to sound pulses to break apart clumps of PrP^{Sc} so it could convert the normal form of the prion protein into the rogue version. The experiment resulted in a 10-fold increase in protease-resistant prion protein. Surachai Supattapone of Dartmouth Medical School has obtained similar results.

Another strategy focuses on the intricacies of the protein shape instead of trying to bolster amounts. A test I developed with my U.C.S.F. colleague Jiri Safar, for example, is based on the ability of some antibodies to react with either PrP^{C} or PrP^{Sc}, but not both. Specifically, the antibody targets a portion of the prion protein that is accessible in one conformation but that is tucked away in the alternative conformation, much like a fitted corner of a sheet is hidden when folded. This specificity means that test samples do not have to be subjected to proteases. Removing the protease step is important because we now know that a form of PrP^{Sc} is sensitive to the action of proteases, which means that tests that eliminate PrP^{C} probably also eliminate most or all of the protease-sensitive PrP^{Sc}—and so these tests could well underestimate the amount of PrP^{Sc} present by as much as 90 percent.

Spontaneous Prion Disease

Because prion diseases have aspects that resemble those caused by viruses, many people use viral analogies when talking about prions. But these analogies can sow confusion. An example is the presumed origin of the mad cows in Canada and the U.S. Although it is true that bovine spongiform encephalopathy (BSE) first appeared in the U.K. and then spread elsewhere through exported prion-contaminated feed, the idea of a traditional bacterial or viral epidemic is only partly helpful. In those situations, quarantines or bans can curb the spread of disease. But prions can arise spontaneously, which is an extremely important characteristic that distinguishes prions from viruses. In fact, any mammal is capable of producing prions spontaneously.

Spontaneous prion disease, for instance, is thought to have triggered the epidemic of kuru, which decimated a group called the Fore in New Guinea in the past century. The theory is that Creutzfeldt-Jakob disease occurred in an individual, whose brain was then consumed by his or her fellow Fore in a funerary rite involving cannibalism. The continued practice created a kuru epidemic.

Similarly, a feed ban that prevents cows from eating the remains of other animals is crucial in containing BSE. But such bans will not eliminate the presence of mad cows when pathogenic prions arise spontaneously. If every year one out of a million humans spontaneously develops a prion disease, why not the same for cows? Indeed, I suspect that the North American BSE cases could well have arisen spontaneously and that afflicted animals have occasionally appeared unrecognized in herds ever since humans started cattle ranching. We have been extraordinarily lucky that a past spontaneous case did not trigger an American BSE epidemic. Or perhaps small epidemics did happen but went undetected.

Still, many prefer the idea that the two mad cows in North America acquired prions from their feed. Such reasoning allows people to equate prions with viruses—that is, to think of prions only as infectious agents (even though they can also be inherited and occur spontaneously)—and to offer a seemingly plausible plan to eradicate BSE by quarantining herds. But ignoring the revolutionary concepts that govern prion biology can only hamper efforts at developing an effective program to protect the American public from exposure to these deadly agents. We must think beyond quarantine and bans and test for prions even in the absence of an epidemic. *—S.B.P.*

THE AUTHOR

STANLEY B. PRUSINER is professor of neurology and biochemistry at the University of California San Francisco School of Medicine. He is a member of the National Academy of Sciences, the Institute of Medicine and the American Philosophical Society . In 1997 he won the Nobel Prize in Physiology or Medicine for his discovery of and research into prions. This is his third article for *Scientific American.* In the spirit of disclosure, Prusiner notes that he founded a company, InPro Biotechnology, which offers several prion tests, some of which are licensed to Beckman Coulter.

Our test—the conformation-dependent immunoassay (CDI)—gained approval for use in Europe in 2003 and might be sensitive enough to detect PrP^{Sc} in blood. The CDI has already shown promise in screening young cows. In the fall of 2003 Japan reported two cases of BSE in cows 21 and 23 months of age. Neither animal showed outward signs of neurological dysfunction. In the case of the 23-month-old cow, two commercially available tests for PrP^{Sc} returned inconclusive, borderline-positive results, but the CDI showed that the brain stem harbored malevolent prions.

Neither of these cases would have been discovered in Europe, where only cattle older than 30 months (24 months in Germany) must be tested if they are destined for human consumption. Initially the Japanese government proposed adopting the European Union's testing protocol. But consumer advocates forced the government to change its policy and test every slaughtered animal. Given that seemingly healthy animals can potentially carry pathogenic prions, I believe that testing all slaughtered animals is the only rational policy. Until now, the tests have been inadequately sensitive. But the advent of rapid, sensitive tests means universal screening can become the norm. (I understand that this statement could seem self-serving because I have a financial interest in the company making the CDI test. But I see no other option for adequately protecting the human food supply.)

Some New Insights

DURING OUR WORK on the CDI, we discovered a surprising fact about the development of prion disease. As alluded to above, we found that the prion is actually a collection of proteins having different degrees of resistance to protease digestion. We also learned that protease-sensitive forms of PrP^{Sc} appear long before the protease-resistant forms appear. Whether protease-sensitive PrP^{Sc} is an intermediate in the formation of protease-resistant PrP^{Sc} remains to be determined. Regardless, a test that could identify protease-sensitive forms should be able to detect infection before symptoms appear, so the food supply can have maximum protection and infected patients can be assisted as early as possible. Fortunately, by using the CDI, my colleagues and I have been able to detect low levels of the protease-sensitive forms of PrP^{Sc} in the blood of rodents and humans.

Hunting for prions in blood led us to another surprise as well. Patrick Bosque, now at the University of Colorado's Health Sciences Center, and I found prions in the hind limb muscles of mice at a level 100,000 times as high as that found in blood; other muscle groups had them, too, but at much lower levels. Michael Beekes and his colleagues at the Robert Koch Institute in Berlin discovered PrP^{Sc} in virtually all muscles after they fed prions to their hamsters, although they report high levels of prions in all muscles, not just in the hind limbs. (We do not know why our results differ or why the hind limbs might be more prone to supporting prions than other muscles are.) These findings were not observed in rodents exclusively but in human patients as well. U.C.S.F. scientists Safar and DeArmond found PrP^{Sc} in the muscles of some CJD patients, and Adriano Aguzzi and his colleagues at the University of Zurich identified PrP^{Sc} in the muscles of 25 percent of the CJD patients they examined.

Of course, the ideal way to test for prions would be a noninvasive method, such as a urine test. Unfortunately, so far the only promising lead—discovery of protease-resistant PrP in urine—could not be confirmed in later studies.

Novel Therapies

ALTHOUGH NEW diagnostics will improve the safety of the food and blood supply, they will undoubtedly distress people who learn that they have a fatal disease. Therefore, many investigators are looking at ways to block prion formation or to boost a cell's ability to clear existing prions. So far researchers have identified more than 20 compounds that can either inhibit prion formation or enhance prion clearance in cultured cells. Several compounds have been shown to extend the lives of mice or hamsters when administered around the time they were inoculated with prions, but none have been shown to alter the course of disease when administered well after the initial

The Significance of Strains

One distinct characteristic that prions do share with viruses is variability; they can come in strains—forms that behave somewhat differently. Laboratory work convincingly shows that prion strains result from different conformations of PrP^{Sc}, but no one has yet figured out exactly how the structure of a given strain influences its particular biological properties. Nevertheless, strains can clearly cause different illnesses. In humans, Creutzfeldt-Jakob disease, kuru, fatal insomnia and Gerstmann-Sträussler-Scheinker disease all result from different strains. Sheep have as many as 20 kinds of scrapie. And BSE may also come in various versions. For instance, a 23-month-old animal in Japan and another cow in Slovakia had a good deal of PrP^{Sc} in the midbrain, whereas in most cases PrP^{Sc} tends to accumulate in the brain stem.

The necessity of recognizing and understanding strains has become clear in studies of sheep exposed to BSE prions. A particular breed of sheep, called the ARR/ARR genotype, resists scrapie, the ovine form of mad cow disease (the letters refer to amino acids on the sheep's prion protein). And so several European countries have created scrapie eradication programs based on breeding flocks with the ARR/ARR genotype. Yet these sheep get sick when inoculated with BSE prions. That these sheep are resistant to scrapie but susceptible to BSE prions has important implications for farming practices. It can be argued that such homogeneous populations will only serve to increase the incidence of BSE prions in sheep. Such a situation could prove to be dangerous to humans who consume lamb and mutton because scrapie strains do not seem to sicken humans, whereas BSE prions do. Experiments with animals should be conducted to see if BSE prions from sheep are as deadly as those from cows before livestock experts continue with selective sheep-breeding programs. *—S.B.P.*

SUNNY DENE RANCH in Mabton, Wash., has been under quarantine since December. This herd and others in the U.S. are now under surveillance for diseased cattle.

KEVIN P. CASEY *EPA*

infection occurred. Furthermore, many of these agents require high doses to exert their effects, suggesting that they would be toxic in animals.

Beyond the problem of potential toxicity that high doses might entail lies the challenge of finding drugs that can cross the blood-brain barrier and travel from the bloodstream into brain tissue. Carsten Korth, now at Heinrich Heine University in Dusseldorf, and I—and, independently, Katsumi Doh-ura of Kyushu University in Japan and Byron Caughey of the National Institute of Allergy and Infectious Diseases—have found that certain drugs known to act in the brain, such as thorazine (used in the treatment of schizophrenia), inhibit prion formation in cultured cells. Another compound, quinacrine, an antimalarial drug with a structure that resembles thorazine, is approximately 10 times as powerful.

Quinacrine has shown some efficacy in animals. My co-workers and I administered quinacrine to mice, starting 60 days after we injected prions into their brain, and found that the incubation time (from the moment of infection to the manifestation of disease) was prolonged nearly 20 percent compared with untreated animals. Such an extension might be quite significant for humans with prion diseases if they could be made to tolerate the high levels of quinacrine needed or if more potent relatives of the drug could be made. My U.C.S.F. colleagues Barnaby May and Fred E. Cohen are pursuing the potency problem. In cell cultures, they have boosted the effectiveness of quinacrine 10-fold by joining two of its molecules together.

Another therapeutic approach involves the use of antibodies that inhibit PrP^{Sc} formation in cultured cells. Several teams have had some success using this strategy. In mice inoculated with prions in the gut and then given antibodies directed against prion proteins, the incubation period was prolonged. So far, however, only a few patients have received antiprion drugs. Quinacrine has been administered orally to patients with vCJD and to individuals who have the sporadic or genetic forms of prion disease. It has not cured them, but it may have slowed the progression of disease; we await further evidence.

Physicians have also administered pentosan polysulfate to vCJD patients. Generally prescribed to treat a bladder condition, the molecule is highly charged and is unlikely to cross the blood-brain barrier, so it has been injected directly into a ventricle of the brain. The drug has apparently slowed the progression of vCJD in one young man, but it seems unlikely that it will diffuse throughout the brain because similarly charged drugs—administered in the same way—have not.

A controlled clinical trial is needed before any assessment of efficacy can be made for quinacrine and other antiprion drugs. Even an initial clinical trial may prove to be insufficient because we have no information about how delivery of the drug should be scheduled. For example, many cancer drugs must be given episodically, where the patient alternates periods on and off the drug, to minimize toxicity.

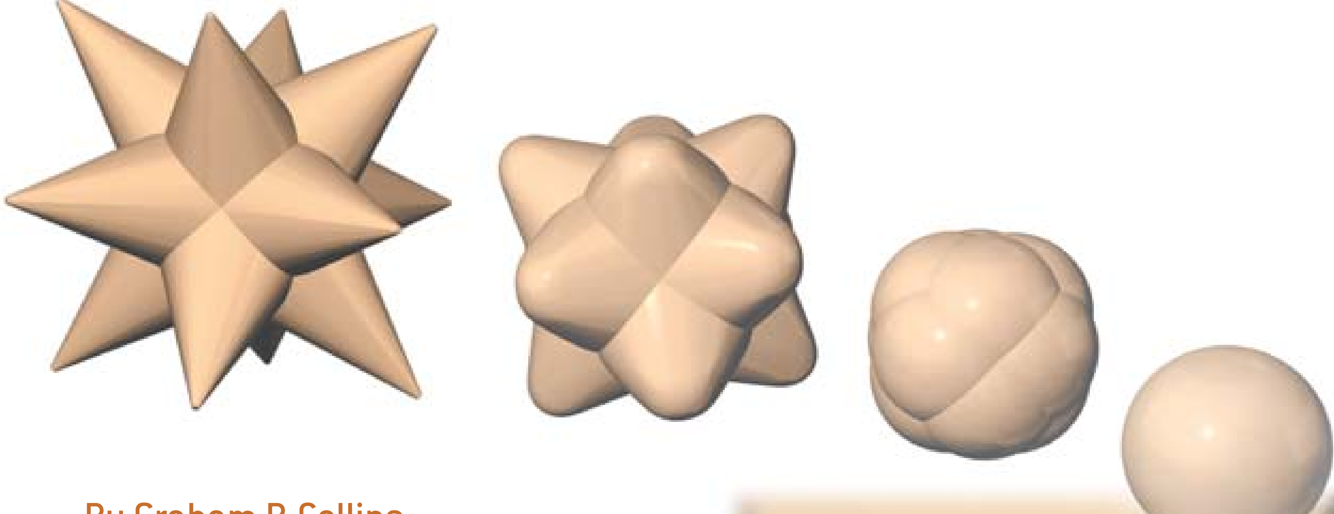
Although the road to a successful treatment seems long, we have promising candidates and strategies that have brought us much further along than we were just five years ago. Investigators are also hopeful that when a successful therapy for prion disease is developed, it will suggest effective therapies for more common neurodegenerative diseases, including Alzheimer's, Parkinson's and amyotrophic lateral sclerosis (ALS). Aberrant, aggregated proteins feature in all these diseases, and so lessons learned from prions may be applicable to them as well. SA

MORE TO EXPLORE

Prions. Stanley B. Prusiner. 1997 Nobel Prize lecture. Available from the Nobel Foundation site at: **www.nobel.se/medicine/laureates/1997/index.html**

Prion Biology and Diseases. Second edition. Edited by Stanley B. Prusiner. Cold Spring Harbor Laboratory Press, 2004.

Advancing Prion Science: Guidance for the National Prion Research Program. Edited by Rick Erdtmann and Laura B. Sivitz. National Academy Press, 2004.



By Graham P. Collins

Stand up and look around. Walk in a circle. Jump in the air. Wave your arms. You are a collection of particles moving about within a small region of a 3-manifold—a three-dimensional space—that extends in all directions for many billions of light-years.

Manifolds are mathematical constructs. The triumph of physics since the time of Galileo and Kepler has been the successful description of reality by mathematics of one flavor or another, such as the mathematics of manifolds. According to physics, everything that happens takes place against the backdrop of three-dimensional space (leaving aside the speculations of string theorists that there are tiny dimensions in addition to the three that are manifest) [see "The Theory Formerly Known as Strings," by Michael J. Duff; Scientific American, February 1998]. Three dimensions means that three numbers are needed to specify the location of a particle. Near Earth, for instance, the three numbers could be latitude, longitude and altitude.

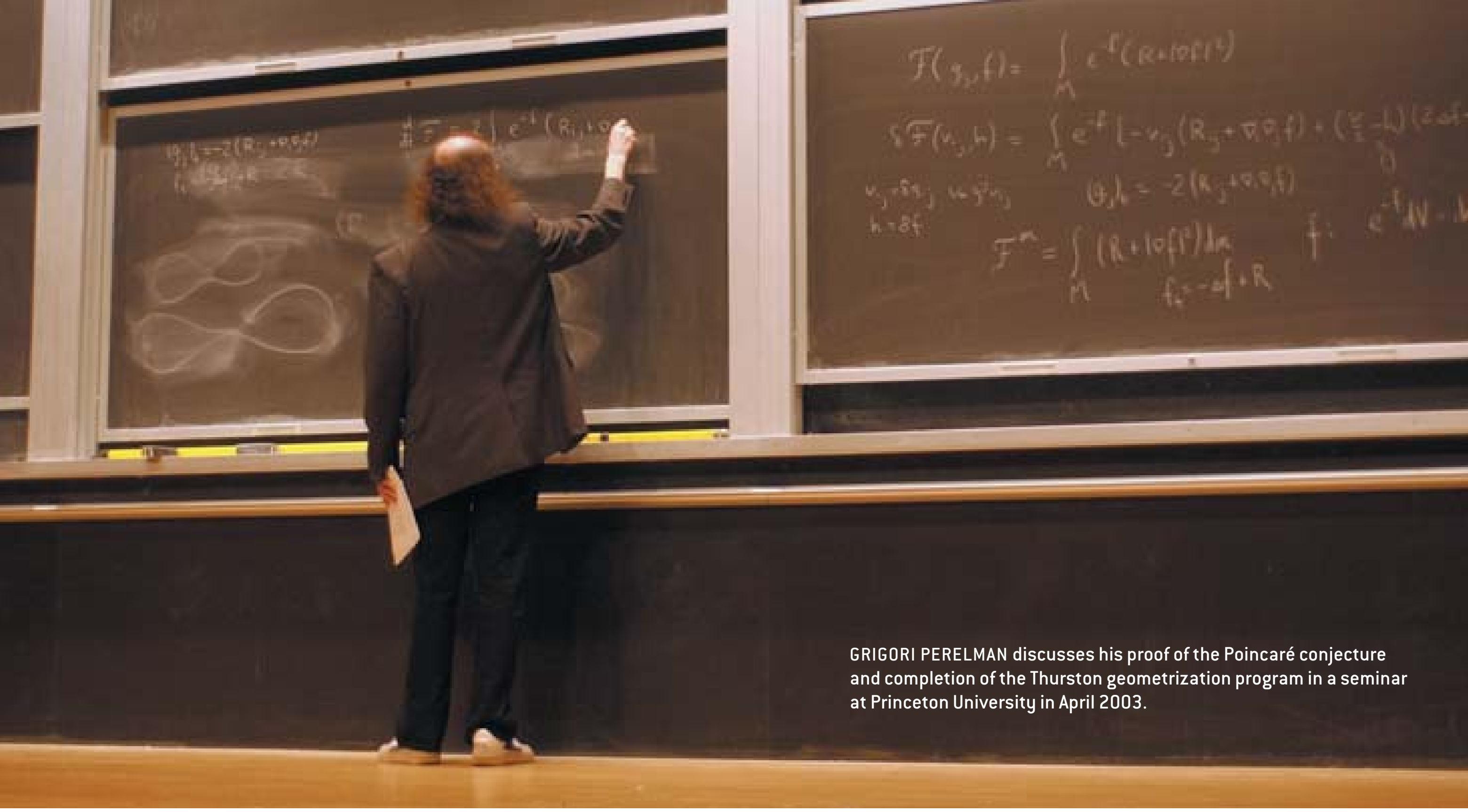
According to Newtonian physics and traditional quantum physics, the three-dimensional space where everything happens is fixed and immutable. Einstein's theory of general relativity, in contrast,

HENRI POINCARÉ conjectured in 1904 that any three-dimensional object that shares certain properties of the three-dimensional sphere can be morphed into a 3-sphere. It took 99 years for mathematicians to prove his conjecture. (Beware: the three-dimensional sphere is probably not what you think it is!)

The Shapes of Space

A Russian mathematician has proved the century-old Poincaré conjecture and completed the catalogue of three-dimensional spaces. He might earn a $1-million prize

AIP EMILIO SEGRÈ VISUAL ARCHIVES (*photograph*); COMPUTER ILLUSTRATION BY JANA BRENNING, WITH 3-D SHAPES BY DON FOLEY (*opposite page*)



GRIGORI PERELMAN discusses his proof of the Poincaré conjecture and completion of the Thurston geometrization program in a seminar at Princeton University in April 2003.

CELENE CHANG *Daily Princetonian*

makes space an active player: the distance from one point to another is influenced by how much matter and energy are nearby and by any gravitational waves that may be passing by [see "Ripples in Spacetime," by W. Wayt Gibbs; Scientific American, April 2002]. But whether we are dealing with Newtonian or Einsteinian physics and whether space is infinite or finite, space is represented by a 3-manifold. Understanding the properties of 3-manifolds is therefore essential for fully comprehending the foundations on which almost all of physics—and all other sciences—are built. (The 4-manifolds are also important: space and time together form a 4-manifold.)

Mathematicians know a lot about 3-manifolds, yet some of the most basic questions have proved to be the hardest. The branch of mathematics that studies manifolds is topology. Among the fundamental questions topologists can ask about 3-manifolds are: What is the simplest type of 3-manifold, the one with the least complicated structure? Does it have many cousins that are equally simple, or is it unique? What kinds of 3-manifolds are there?

The answer to the first of those questions has long been known: a space called the 3-sphere is the simplest compact 3-manifold. (Noncompact manifolds can be thought of as being infinite or having an edge. Hereafter I consider only compact manifolds.) The other two questions have been up for grabs for a century but may have been answered in 2002 by Grigori ("Grisha") Perelman, a Russian mathematician who has most probably proved a theorem known as the Poincaré conjecture.

First postulated by French mathematician Henri Poincaré exactly 100 years ago, the conjecture holds that the 3-sphere is unique among 3-manifolds; no other 3-manifold shares the properties that make it so simple. The 3-manifolds that are more complicated than the 3-sphere have boundaries that you can run up against like a brick wall, or multiple connections from one region to another, like a path through the woods that splits and later rejoins. The Poincaré conjecture states that the 3-sphere is the only compact 3-manifold that lacks all those complications. Any three-dimensional object that shares those properties with the sphere can therefore be morphed into the same shape as a 3-sphere; so far as topologists are concerned, the object *is* just another copy of the 3-sphere. Perelman's proof also answers the third of our questions: it completes work that classifies all the types of 3-manifolds that exist.

It takes some mental gymnastics to imagine what a 3-sphere is like—it is not simply a sphere in the everyday sense of the word [*see box on pages 98 and 99*]. But it has many properties in common with the 2-sphere, which we are all familiar with: If you take a spherical balloon, the rubber of the balloon forms a 2-

Overview/*Proving Poincaré*

- For 100 years, mathematicians have been trying to prove a conjecture that was first proposed by Henri Poincaré relating to an object known as the three-dimensional sphere, or 3-sphere. The conjecture singles out the 3-sphere as being unique among all three-dimensional objects, or manifolds.
- A proof of the Poincaré conjecture has finally come, with the work of a young Russian mathematician, Grigori Perelman. His analysis also completes a major research program that classifies all possible three-dimensional manifolds.
- Our universe might have the shape of a 3-sphere. The mathematics has other intriguing connections to particle physics and Einstein's theory of gravity.

sphere. The 2-sphere is two-dimensional because only two coordinates—latitude and longitude—are needed to specify a point on it. Also, if you take a very small disk of the balloon and examine it with a magnifying glass, the disk looks a lot like one cut from a flat two-dimensional plane of rubber. It just has a slight curvature. To a tiny insect crawling on the balloon, it would seem like a flat plane. Yet if the insect traveled far enough in what would seem to it to be a straight line, eventually it would arrive back at its starting point.

Similarly, a gnat in a 3-sphere—or a person in one as big as our universe!—perceives itself to be in "ordinary" three-dimensional space. But if it flies far enough in a straight line in any direction, eventually it will circumnavigate the 3-sphere and find itself back where it started, just like the insect on the balloon or someone taking a trip around the world.

Spheres exist for dimensions other than three as well. The 1-sphere is also familiar to you: it is just a circle (the rim of a disk, not the disk itself). The n-dimensional sphere is called an n-sphere.

Proving Conjectures

AFTER POINCARÉ proposed his conjecture about the 3-sphere, half a century went by before any real progress was made in proving it. In the 1960s mathematicians proved analogues of the conjecture for spheres of five dimensions or more. In each case, the n-sphere is the unique, simplest manifold of that dimensionality. Paradoxically, this result was easier to prove for higher-dimensional spheres than for those of four or three dimensions. The proof for the particularly difficult case of four dimensions came in 1982. Only the original three-dimensional case involving Poincaré's 3-sphere remained open.

A major step in closing the three-dimensional problem came in November 2002, when Perelman, a mathematician at the Steklov Institute of Mathematics at St. Petersburg, posted a paper on the www.arxiv.org Web server that is widely used by physicists and mathematicians as a clearinghouse of new research. The paper did not mention the Poincaré conjecture by name, but topology experts who looked at it immediately realized the paper's relevance to that theorem. Perelman followed up with a second paper in March 2003, and from April to May that year he visited the U.S. to give a series of seminars on his results at the Massachusetts Institute of Technology and Stony Brook University. Teams of mathematicians at nearly a dozen leading institutes began poring over his papers, verifying their every detail and looking for errors.

At Stony Brook, Perelman gave two weeks of formal and informal lectures, talking from three to six hours a day. "He answered every question that arose, and he was very clear," says mathematician Michael Anderson of Stony Brook. "No one has yet raised any serious doubts." One more comparatively minor step has to be proved to complete the result, Anderson says, "but there are no real doubts about the validity of this final piece." The first paper contains the fundamental ideas and is pretty well accepted as being verified. The second paper contains applications and more technical arguments; its verification has not reached the level of confidence achieved for the first paper.

The Poincaré conjecture has a $1-million reward on offer for its proof: it is one of seven such "Millennium Problems" singled out in 2000 by the Clay Mathematics Institute in Cambridge, Mass. Perelman's proof has to be published and withstand two years of scrutiny before he becomes eligible for the prize. (The institute might well decide that its posting on the Web server qualifies as "published" because the result is undergoing as rigorous a peer review as any paper gets.)

Perelman's work extends and completes a program of research that Richard S. Hamilton of Columbia University explored in the 1990s. The Clay Institute recognized Hamilton's work with a research award in late 2003. Perelman's calculations and analysis blow away several roadblocks that Hamilton ran into and could not overcome.

If, as everyone expects, Perelman's proof is correct, it actually completes a much larger body of work than the Poincaré conjecture. Launched by William P. Thurston—now at Cornell University—the Thurston geometrization conjecture provides a full classification of all possible 3-manifolds. The 3-sphere, unique in its sublime simplicity, anchors the foundation of this magnificent classification. Had the Poincaré conjecture been false—that is, if there were many spaces as "simple" as a sphere—the classification of 3-manifolds would have exploded into something infinitely more complicated than that proposed by Thurston. Instead, with Perelman's and Thurston's results, we now have a complete catalogue of all the possible shapes that a three-dimensional space can take on—all the shapes allowed by mathematics that our universe (considering just space and not time) could have.

TEAMS OF MATHEMATICIANS at nearly a dozen leading institutes began poring over Perelman's papers, verifying every detail and LOOKING FOR ERRORS.

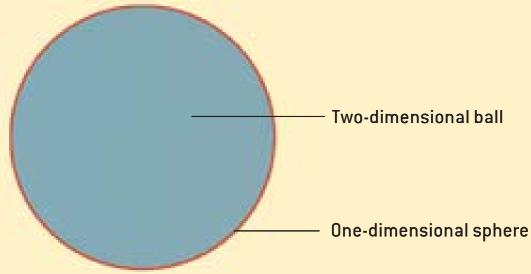
Rubber Doughnuts

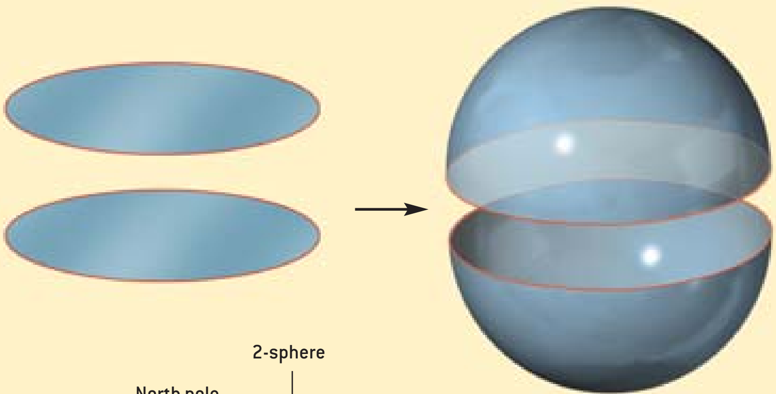
TO UNDERSTAND the Poincaré conjecture and Perelman's proof in greater depth, you have to know something about topology. In that branch of mathematics the exact shape of an object is irrelevant, as if it were made of play dough that you could stretch, compress and bend to any extent. But why should we care about objects or spaces made of imaginary play dough? The reason relates to the fact that the exact shape of an object—the distance from one point on it to another—is a level of structure, which is

Multidimensional Music of Spheres

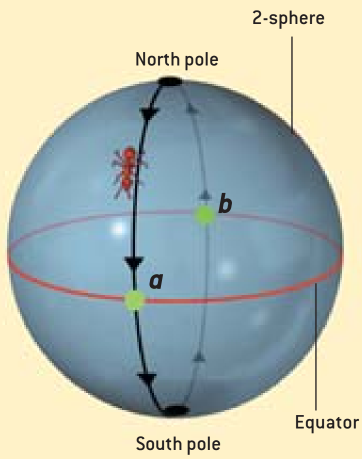
The 3-sphere at the heart of Poincaré's conjecture takes a bit of effort to visualize. Mathematicians who prove theorems about higher-dimensional spaces do not have to visualize them. They make do with abstract properties, guided by intuitive notions based on analogies to lower dimensions (but being careful not to take the analogies literally). Others, too, can form an idea of what higher-dimensional objects are like by working up from familiar lower-dimensional examples. The 3-sphere is a case in point.

1 Start by considering a disk with a circle forming its boundary. To mathematicians, the disk is a "two-dimensional ball" and the circle is a "one-dimensional sphere." Also, a "ball" of whatever dimensionality is the filled-in object, analogous to a baseball, and a "sphere" is the surface of a ball, analogous to a balloon. The circle is one-dimensional because it takes just a single number to specify a location on it.

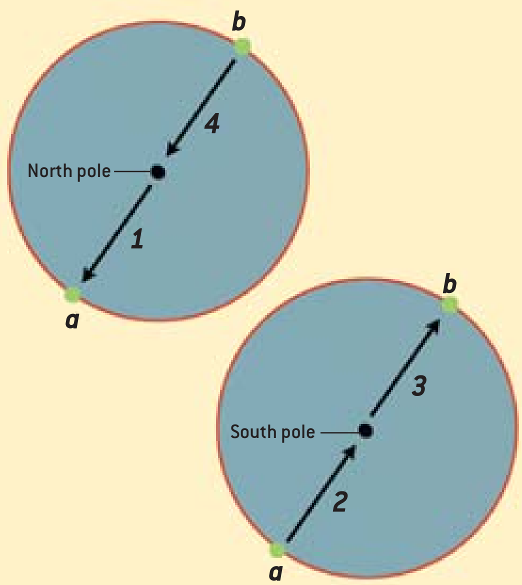




2 Now we can build the two-dimensional sphere out of two copies of the disk. Distort one disk into a hemisphere like the Northern Hemisphere and distort the other to be like the Southern Hemisphere. Then glue the two hemispheres together at their boundary, which becomes the equator. Voilà: the 2-sphere.



3 Imagine an ant starting out from the North Pole, walking along the great circle formed by the International Date Line and the meridian that passes through Greenwich, England (*left*). If we map that path back onto the two disks (*right*), we see that the ant travels in a straight line (*1*) out to the edge of the northern disk (*a*). Then it crosses to the corresponding point on the southern disk and continues across that disk in a straight line (*2* and *3*). When it reaches the edge again (*b*), it crosses back to the northern disk and continues on to its starting point, the North Pole (*4*). We have followed its path as it circumnavigates the 2-sphere by tracking it on the disks. The only tricky part is figuring out that the direction of travel appears to reverse when it crosses from one disk to the other.

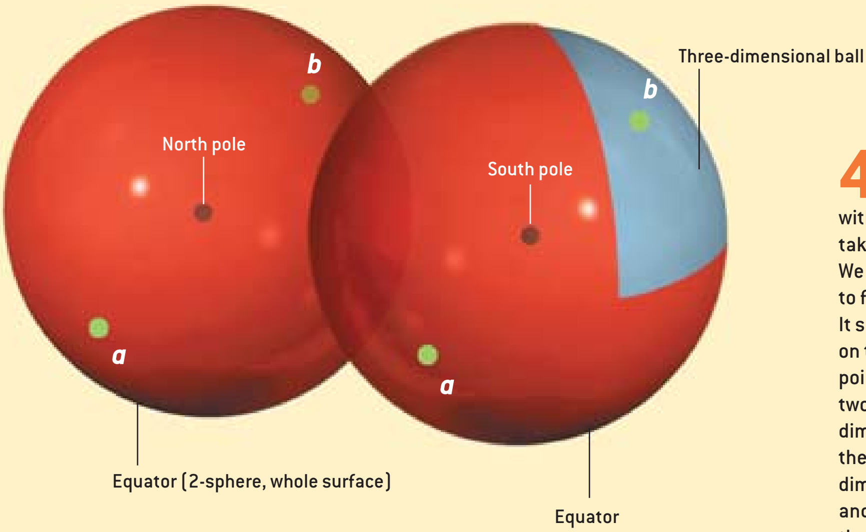


called the geometry of the object. By considering a play-dough object, topologists discover which properties of the object are so fundamental that they exist independently of its geometric structure. Studying topology is like discovering which properties humans have in general by considering the properties of a "play-dough person" who can be morphed into any particular human being.

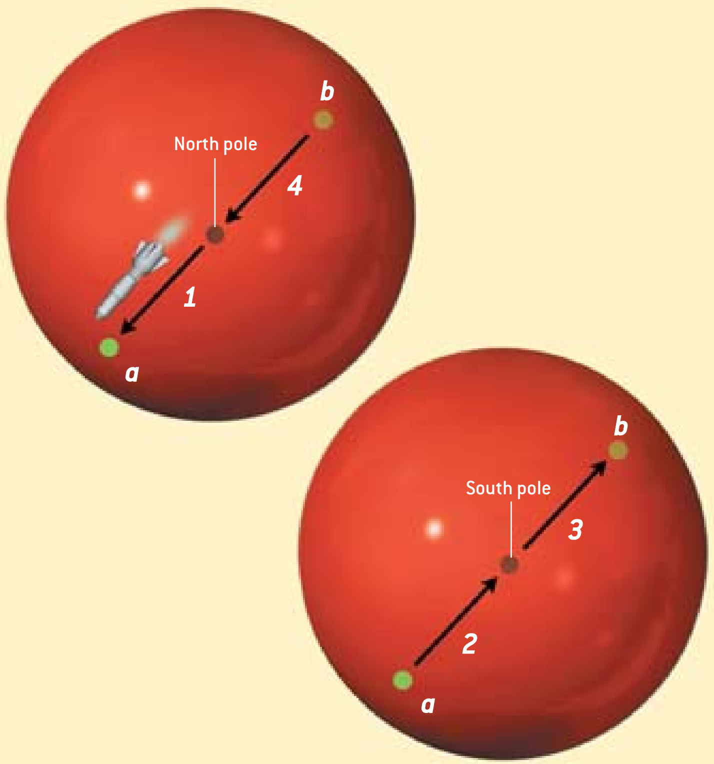
If you have read any popular account of topology, you have probably encountered the hoary old truism that a cup and a doughnut are indistinguishable to a topologist. (The saying refers to a ring-shaped doughnut, not the solid, jam-filled kind.) The point is that you can morph the play-dough cup into a doughnut shape simply by smushing the clay around, without having to cut out any holes or glue any patches together [*see illustration on page 100*]. A ball, on the other hand, can be turned into a doughnut only by either boring a hole through its middle or stretching it into a cylinder and gluing the ends together. Because such cutting or gluing is needed, a ball is not the same as a doughnut to a topologist.

What interests topologists most are the surfaces of the ball and the doughnut, so instead of imagining a solid we should imagine a balloon in both cases. The topologies are still distinct—the spherical balloon cannot be morphed into the ring-shaped balloon, which is called a torus. Topologically, then, a sphere and a torus are distinct entities. Early topologists set out to discover how many other topologically distinct entities exist and how they

DON FOLEY



4 Then consider our 2-sphere and the three-dimensional volume it contains (a "three-dimensional ball") and do with the ball and sphere what we did with the circle and disk: take two copies and glue the boundaries together. We cannot visualize how to distort the balls in four dimensions to form some analogue of hemispheres, but we don't need to. It suffices to know that corresponding points on the surfaces—on the 2-spheres—are joined together, like corresponding points on the circles were. The result of joining together the two balls is the 3-sphere, which is the "surface" of a four-dimensional ball. (In four dimensions—where the 3-sphere and the 4-ball really live—the "surface" of an object is three-dimensional.) We can call one ball the northern hemisphere and the other the southern. The north pole is at the center of the northern ball (like the north pole was marked at the center of the northern disk).



5 Next imagine these balls are large, empty regions of space, and someone sets out in a rocket ship from the north pole. Eventually she reaches the "equator" (*1*), which is the entire sphere surrounding the northern ball. At the equator she crosses over to the southern hemisphere and travels on a straight line through its center (the south pole) to the opposite side of the equator (*2* and *3*). There she crosses back to the northern hemisphere and travels back to the north pole, her starting point (*4*). We have just imagined someone traveling along the surface of a four-dimensional ball, circumnavigating it! This 3-sphere, consisting of two balls joined on their spherical surfaces, is the space to which the Poincaré conjecture applies. Our universe might have the shape of a 3-sphere.

The process can be continued to five dimensions—to make a 4-sphere—but it becomes even harder to visualize what is going on. Similarly, any given n-sphere can be constructed from two n-balls: just glue together the boundaries of the two balls. Each boundary is an (n–1)-sphere, just as the boundary of a disk (a 2-ball) is a circle (a 1-sphere). The result is an n-sphere, which encloses an (n + 1)-ball.

could be characterized. For two-dimensional objects, which are also called surfaces, the answer is neat and tidy: it all boils down to how many "handles" a surface has.

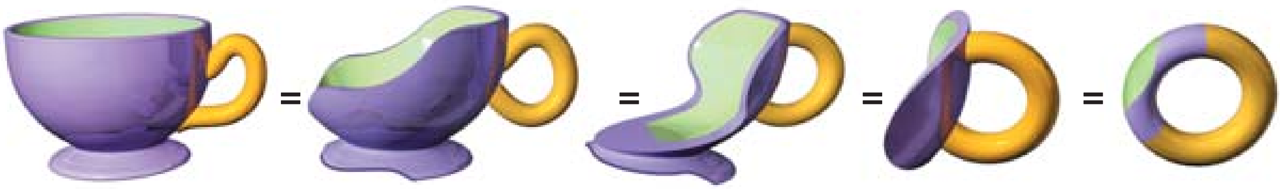
By the end of the 19th century, mathematicians understood how to classify surfaces. Out of all the surfaces, the sphere, they knew, had a unique simplicity. Naturally they started wondering about three-dimensional manifolds. To start with, was the 3-sphere unique in its simplicity, analogous to the 2-sphere? The century-long history that follows from that elementary question is littered with false steps and false proofs.

Henri Poincaré tackled this question head-on. He was one of the two foremost mathematicians who were active at the turn of the 20th century (the other being David Hilbert). Poincaré has been called the last universalist—he was at ease in all branches of mathematics, both pure and applied. In addition to advancing numerous areas of mathematics, he contributed to the theories of celestial mechanics and electromagnetism as well as to the philosophy of science (about which he wrote several widely read popular books).

Poincaré largely created the branch of mathematics called algebraic topology. Around 1900, using techniques from that field, he formulated a measure of an object's topology, called homotopy. To determine a manifold's homotopy, imagine that you embed a closed loop in the manifold [*see box on next page*]. The loop can be wound around the manifold in any possible fashion. We then ask, Can the

TOPOLOGY OF SURFACES

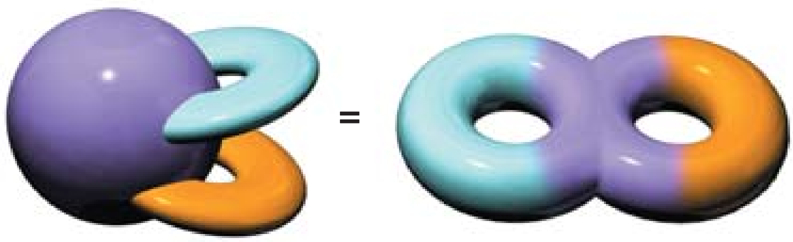
IN TOPOLOGY the exact shape, or geometry, of an object is not important. It is as if everything is made of play dough or rubber and can be morphed by stretching, compressing and twisting. Cutting and joining, however, are forbidden. Thus, in topology any object with a single hole, such as the coffee cup at the far left, is equivalent to the doughnut at the far right.



EVERY POSSIBLE two-dimensional manifold, or surface (restricting to so-called compact, orientable ones), can be constructed by taking a sphere (akin to a balloon, *a*) and adding handles. The addition of one handle yields the genus-1 surface, or torus, which is the surface of the doughnut shape (*above right*). Adding two handles yields the genus-2 surface (*b*) and so on.

a

b



2-SPHERE is unique among surfaces, in that any closed loop embedded on a 2-sphere can be shrunk down to a point (*a*). In contrast, a loop on a torus can get "caught" around the hole in the middle (*b*). Every surface except for the 2-sphere has handles on which the loop can get caught. The Poincaré conjecture proposes that the 3-sphere is similarly unique among all three-dimensional manifolds: any loop on it can be shrunk to a point, but on every other 3-manifold, the loop can get caught.

a

b

DON FOLEY

loop always be shrunk down to a point, just by moving it around, without ever lifting a piece of it out of the manifold? On a torus the answer is no. If the loop runs around the circumference of the torus, it cannot be shrunk to a point—it gets caught on the inner ring of the doughnut. Homotopy is a measure of all the different ways a loop can get caught.

On an *n*-sphere, no matter how convoluted a path the loop takes, it can always be untangled and shrunk to a point. (The loop is allowed to pass through itself during these manipulations.) Poincaré speculated that the only 3-manifold on which every possible loop can be shrunk to a point was the 3-sphere itself, but he could not prove it. In due course this proposal became known as the Poincaré conjecture. Over the decades, many people have announced proofs of the conjecture, only to be proved wrong. (For clarity, here and throughout I ignore two complications: so-called nonorientable manifolds and manifolds with edges. For example, the Möbius band, a ribbon that is twisted and joined in a loop, is nonorientable. A sphere with a disk cut out from it has an edge. The Möbius band also has an edge.)

Geometrization

PERELMAN'S PROOF is the first to withstand close scrutiny. His approach to analyzing three-dimensional manifolds is related to a procedure called geometrization. Geometry relates to the actual shape of an object or manifold: for geometry, an object is made not of play dough but of ceramic. A cup, for example, has a different geometry than a doughnut; its surface curves in different ways. It is said that the cup and the doughnut are two examples of a topological torus (provided the cup has one handle) to which different geometries have been assigned.

To gain a sense of how geometrization served to help Perelman, consider how geometry can be used to classify 2-manifolds, or surfaces. Each topological surface is assigned a special, unique geometry: the one for which the curvature of the surface is spread completely evenly about the manifold. For the sphere, that unique geometry is the perfectly spherical sphere. An eggshell shape is another possible geometry for a topological sphere, but it does not have curvature evenly spread throughout: the small end of the egg is more curved than the big end.

The 2-manifolds form three geometric types [*see box at right*]. The sphere has what is called positive curvature, the shape of a hilltop. The geometrized torus is flat; it has zero curvature, like a plain. All the other manifolds, with two or more handles, have negative curvature. Negative curvature is like the shape of a mountain pass or a saddle: going from front to back, a saddle curves up; from left to right, it curves down. Poincaré (who else?), along with Paul Koebe and Felix Klein (after whom the Klein bottle is named), contributed to this geometric classification, or geometrization, of 2-manifolds.

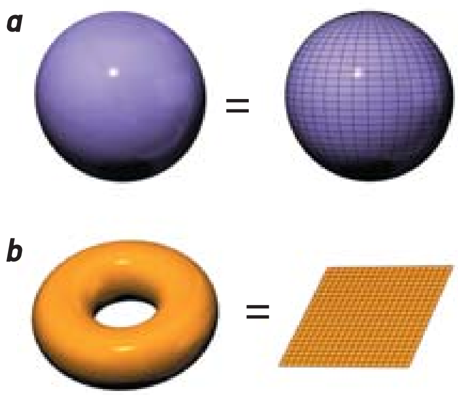
It is natural to try to apply similar methods to 3-manifolds. Is it possible to find unique geometries for each topological 3-manifold, for which curvature is spread evenly throughout the manifold?

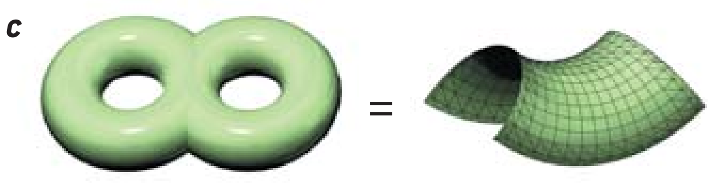
It turns out that 3-manifolds are far more complicated than 2-manifolds. Most 3-manifolds cannot be assigned a uniform geometry. Instead they have to be cut up into pieces, each piece having a different canonical geometry. Furthermore, instead of three basic geometries, as with 2-manifolds, the 3-manifold pieces can take any of eight canonical geometries. The cutting up of each 3-manifold is somewhat analogous to the factorization of a number into a unique product of prime factors.

This classification scheme was first conjectured by Thurston in the late 1970s. He and his colleagues proved large swaths of the conjecture, but crucial points that the entire system depended on remained beyond their grasp, including the part that embodied the Poincaré conjecture. Was the 3-sphere unique? An answer to that question and completion of the

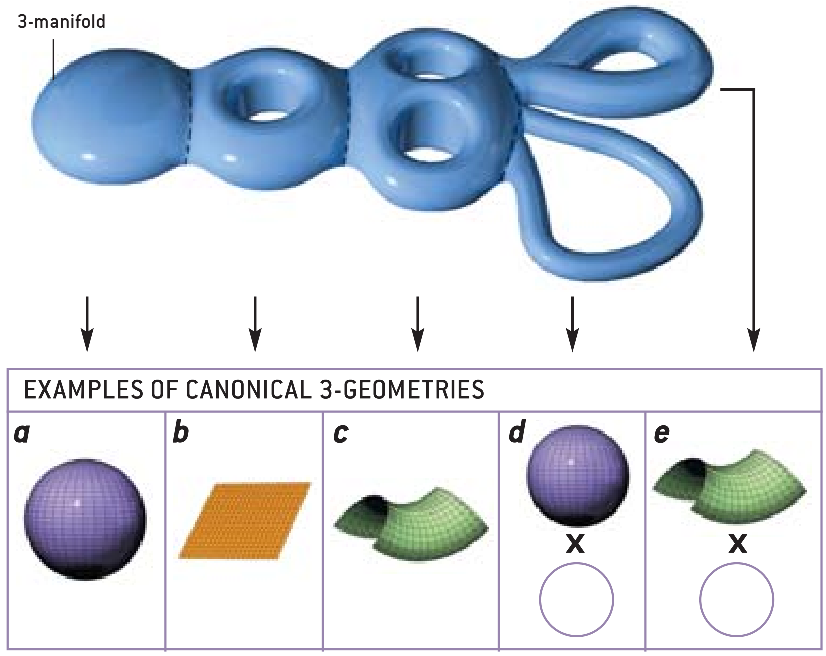
GEOMETRIZATION

2-MANIFOLDS can be classified by "uniformizing" or "geometrizing" them, which means assigning them a specific geometry, or rigid shape. In particular, each can be morphed into a shape that has its curvature evenly distributed. The sphere (*a*) is the unique shape having constant positive curvature, meaning at every point it is curved like a hilltop. The torus (*b*) can be made flat—that is, with zero curvature throughout. To see this, imagine cutting the torus and straightening it out to form a cylinder. Then cut along the cylinder and unroll it to form a flat rectangular plane. The torus has thus been mapped to a flat plane. Surfaces of genus-2 and higher (*c*) can be given constant negative curvature, with other details depending on how many handles are present. Here the constant negative curvature surface is represented by the saddle shape.





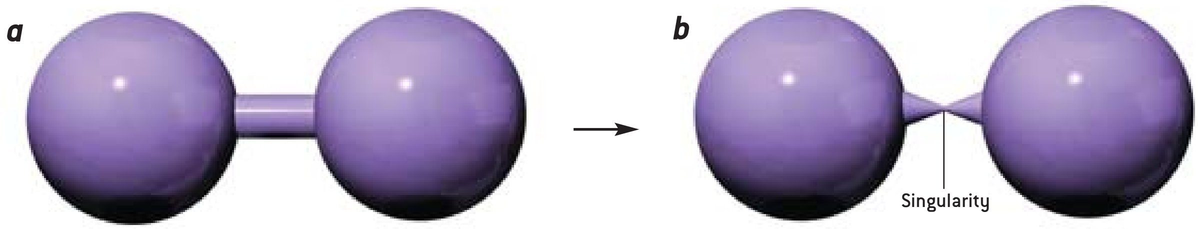
CLASSIFICATION OF 3-MANIFOLDS, which is similar to that of 2-manifolds but far more complicated, is completed by Perelman's work. In general, a 3-manifold has to be divided into pieces, each of which can be morphed into one of eight different canonical three-dimensional geometries. The blue-colored example below (*drawn in cartoon form as a 2-manifold*) consists of equivalents to five of them: constant positive (*a*), zero (*b*) and constant negative (*c*) curvature 3-geometries, as well as the "product" of the 2-sphere and a circle (*d*) and of the negative curvature surface and a circle (*e*).



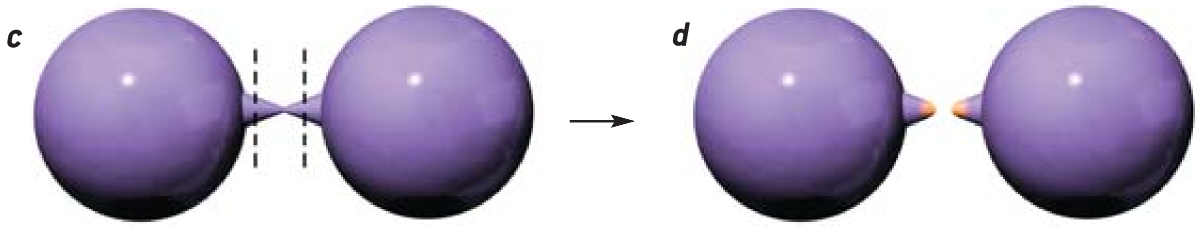
DON FOLEY

DEALING WITH SINGULARITIES

EFFORTS TO APPLY an equation called the Ricci flow to prove Poincaré's conjecture and to geometrize 3-manifolds hit a roadblock before Perelman came along. The Ricci flow, which gradually changes the shape of a 3-manifold, occasionally runs into trouble called singularities. One instance is when part of the manifold has a dumbbell shape—two spheres joined by a tube (*a*). The tube can become pinched down to a point, spoiling the manifold's properties (*b*). Another singularity, called the cigar singularity, was also believed to be possible.



"SURGERY" CAN HANDLE the singularities that arise in the Ricci flow, as shown by Perelman's work. When a region of the manifold starts to pinch down, a small region on each side of the pinch can be cut out (*c*). The cuts are then capped by small spheres, and the Ricci flow is continued (*d*). This process may have to be repeated several times if other regions subsequently pinch down, but Perelman showed that the process terminates eventually. He also showed that cigar singularities never arise.



Thurston program have come only with Perelman's papers.

How might we try to geometrize a manifold—that is, give it a uniform curvature throughout? One way is to start with some arbitrary geometry, perhaps like an eggshell shape with various lumps and indentations, and then smooth out all the irregularities. Hamilton began such a program of analysis for 3-manifolds in the early 1990s, using an equation called the Ricci flow (named after mathematician Gregorio Ricci-Curbastro), which has some similarities to the equation that governs the flow of heat. In a body with hot and cold spots, heat naturally flows from the warmer regions to the cooler ones, until the temperature is uniform everywhere. The Ricci flow equation has a similar effect on curvature, morphing a manifold to even out all the bumps and hollows. If you began with an egg, it would gradually become perfectly spherical.

Hamilton's analysis ran into a stumbling block: in certain situations the Ricci flow would cause a manifold to pinch down to a point. (This is one way that the Ricci flow differs from heat flow. The places that are pinched are like points that manage to acquire infinite temperature.) One example was when the manifold had a dumbbell shape, like two spheres connected by a thin neck. The spheres would grow, in effect drawing material from the neck, which would taper to a point in the middle [*see box above*]. Another possible example arose when a thin rod stuck out from the manifold; the Ricci flow might produce a trouble called a cigar singularity. When a manifold is pinched in this way, it is called singular—it is no longer a true three-dimensional manifold. In a true three-dimensional manifold, a small region around any point looks like a small region of ordinary three-dimensional space, but this property fails at pinched points. A way around this stumbling block had to wait for Perelman.

Perelman came to the U.S. as a postdoctoral student in 1992, spending semesters at New York University and Stony Brook, followed by two years at the University of California at Berkeley. He quickly made a name for himself as a brilliant young star, proving many important and deep results in a particular branch of geometry. He was awarded a prize from the European Mathematical Society (which he declined) and received a prestigious invitation to address the International Congress of Mathematicians (which he accepted). In spring 1995 he was offered positions at a number of outstanding mathematics departments, but he turned them all down to return to his home in St. Petersburg. "Culturally, he is very Russian," commented one American colleague. "He's very unmaterialistic."

Back in St. Petersburg, he essentially disappeared from mathematicians' radar screens. The only signs of activity, after many years, were rare occasions when he e-mailed former colleagues, for example, to point out errors in papers they had posted on the Internet. E-mails inquiring about his pursuits went unanswered.

Finally, in late 2002 several people received an e-mail from him alerting them to the paper he had posted on the mathematics server—just a characteristically brief note saying they might find it of interest. That understatement heralded the first stage of his attack on the Poincaré conjecture. In the preprint, along with his Steklov Institute affiliation, Perelman acknowledged support in the form of money he had saved from his U.S. postdoctoral positions.

DON FOLEY

POINCARÉ (*seated, talking to Marie Curie*) attended the first Solvay Physics Conference in Brussels, Belgium, in October 1911. Behind stand Ernest Rutherford, Heike Kamerlingh Onnes (who discovered superconductivity earlier that year) and Albert Einstein. This conference may have been the only time that Einstein and Poincaré met. Poincaré died nine months later.

HULTON-DEUTSCH COLLECTION/CORBIS

In his paper, Perelman added a new term to the Ricci flow equation. The modified equation did not eliminate the troubles with singularities, but it enabled Perelman to carry the analysis much further. With the dumbbell singularities he showed that "surgery" could be performed: Snip the thin tube on each side of the incipient pinch and seal off the open tube on each dumbbell ball with a spherical cap. Then the Ricci flow could be continued with the surgically altered manifold until the next pinch, for which the same procedure could be applied. He also showed that cigar singularities could not occur. In this way, any 3-manifold could be reduced to a collection of pieces, each having a uniform geometry.

When the Ricci flow and the surgery are applied to all possible 3-manifolds, any manifold that is as "simple" as a 3-sphere (technically, that has the same homotopy as a 3-sphere) necessarily ends up with the same uniform geometry as a 3-sphere. That result means that topologically, the manifold in question *is* a 3-sphere. Rephrasing that, the 3-sphere is unique.

Beyond proving Poincaré's conjecture, Perelman's research is important for the innovative techniques of analysis it has introduced. Already mathematicians are posting papers that build on his work or apply his techniques to other problems. In addition, the mathematics has curious connections to physics. The Ricci flow used by Hamilton and Perelman is related to something called the renormalization group, which specifies how interactions change in strength depending on the energy of a collision. For example, at low energies the electromagnetic interaction has a strength characterized by the number 0.0073 (about 1/137). If two electrons collide head-on at nearly the speed of light, however, the strength is closer to 0.0078.

Increasing the collision energy is equivalent to studying the force at a shorter distance scale. The renormalization group is therefore like a microscope with a magnification that can be turned up or down to examine a process at finer or coarser detail. Similarly, the Ricci flow is like a microscope for looking at a manifold at a chosen magnification. Bumps and hollows visible at one magnification disappear at another. Physicists expect that on a scale of about 10^{-35} meter, or the Planck length, the space in which we live will look very different—like a "foam" with many loops and handles and other topological structures [see "Atoms of Space and Time," by Lee Smolin; Scientific American, January]. The mathematics that describes how the physical forces change is very similar to that which describes geometrization of a manifold.

Another connection to physics is that the equations of general relativity, which describe the workings of gravity and the large-scale structure of the universe, are closely related to the Ricci flow equation. Furthermore, the term that Perelman added to the basic flow used by Hamilton arises in string theory, which is a quantum theory of gravity. It remains to be seen if his techniques will reveal interesting new information about general relativity or string theory. If that is the case, Perelman will have taught us not only about the shapes of abstract 3-spaces but also about the shape of the particular space in which we live. SA

Graham P. Collins, a staff writer and editor, has degrees in mathematics and physics. For additional information on the Poincaré conjecture, visit www.sciam.com/ontheweb

MORE TO EXPLORE

The Poincaré Conjecture 99 Years Later: A Progress Report. John W. Milnor. February 2003. Available at **www.math.sunysb.edu/~jack/PREPRINTS/poiproof.pdf**

Jules Henri Poincaré (biography). October 2003. Available at **www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Poincare.html**

Millennium Problems. The Clay Mathematics Institute: **www.claymath.org/millennium/**

Notes and commentary on Perelman's Ricci flow papers. Compiled by Bruce Kleiner and John Lott. Available at **www.math.lsa.umich.edu/research/ricciflow/perelman.html**

Topology. Eric W. Weisstein in *Mathworld—A Wolfram Web Resource.* Available at **mathworld.wolfram.com/Topology.html**

The Mystery of the

VOYNICH MANUSCRIPT

New analysis of a famously cryptic medieval document suggests that it contains nothing but gibberish

By Gordon Rugg

In 1912 Wilfrid Voynich, an American rare-book dealer, made the find of a lifetime in the library of a Jesuit college near Rome: a manuscript some 230 pages long, written in an unusual script and richly illustrated with bizarre images of plants, heavenly spheres and bathing women. Voynich immediately recognized the importance of his new acquisition. Although it superficially resembled the handbook of a medieval alchemist or herbalist, the manuscript appeared to be written entirely in code. Features in the illustrations, such as hairstyles, suggested that the book was produced sometime between 1470 and 1500, and a 17th-century letter accompanying the manuscript stated that it had been purchased by Rudolph II, the Holy Roman Emperor, in 1586. During the 1600s, at least two scholars apparently tried to decipher the manuscript, and then it disappeared for nearly 250 years until Voynich unearthed it.

Voynich asked the leading cryptographers of his day to decode the odd script, which did not match that of any known language. But despite 90 years of effort by some of the world's best code breakers, no one has been able to decipher Voynichese, as the script has become known. The nature and origin of the manuscript remain a mystery. The failure of the code-breaking attempts has raised the suspicion that there may not be any cipher to crack. Voynichese may contain no message at all, and the manuscript may simply be an elaborate hoax.

Critics of this hypothesis have argued that Voynichese is too complex to be nonsense. How could a medieval hoaxer produce 230 pages of script with so many subtle regularities in the structure and distribution of the words? But I have recently discovered that one can replicate many of the remarkable features of Voynichese using a simple coding tool that was available in the 16th century. The text generated by this technique looks much like Voynichese, but it is merely gibberish, with no hidden message. This finding does not prove that the Voynich manuscript is a hoax, but it does bolster the long-held theory that an English adventurer named Edward Kelley may have concocted the document to defraud Rudolph II. (The emperor reportedly paid a sum of 600 ducats—equivalent to about $50,000 today—for the manuscript.)

Perhaps more important, I believe that the methods used

BEINECKE RARE BOOK AND MANUSCRIPT LIBRARY, YALE UNIVERSITY

STRANGE IMAGES of heavenly spheres, fantastic plants and nude women adorn the pages of the Voynich manuscript, which is written in an odd script that does not match that of any known language. The manuscript now resides at the Beinecke Rare Book and Manuscript Library at Yale University.

The Voynich manuscript appeared to be either an unusual code, an unknown language or a sophisticated hoax.

in this analysis of the Voynich mystery can be applied to difficult questions in other areas. Tackling this hoary puzzle requires expertise in several fields, including cryptography, linguistics and medieval history. As a researcher into expert reasoning—the study of the processes used to solve complex problems—I saw my work on the Voynich manuscript as an informal test of an approach that could be used to identify new ways of tackling long-standing scientific questions. The key step is determining the strengths and weaknesses of the expertise in the relevant fields.

Baby God's Eye?

THE FIRST PURPORTED decryption of the Voynich manuscript came in 1921. William R. Newbold, a professor of philosophy at the University of Pennsylvania, claimed that each character in the Voynich script contained tiny pen strokes that could be seen only under magnification and that these strokes formed an ancient Greek shorthand. Based on his reading of the code, Newbold declared that the Voynich manuscript had been written by 13th-century philosopher-scientist Roger Bacon and described discoveries such as the invention of the microscope. Within a decade, however, critics debunked Newbold's solution by showing that the alleged microscopic features of the letters were actually natural cracks in the ink.

Newbold's attempt was just the start of a string of failures. In the 1940s amateur code breakers Joseph M. Feely and Leonell C. Strong used substitution ciphers that assigned Roman letters to the characters in Voynichese, but the purported translations made little sense. At the end of World War II the U.S. military cryptographers who cracked the Japanese Imperial Navy's codes passed some spare time tackling ciphertexts—encrypted texts—from antiquity. The team deciphered every one except the Voynich manuscript.

In 1978 amateur philologist John Stojko claimed that the text was written in Ukrainian with the vowels removed, but his translation—which included sentences such as "Emptiness is that what Baby God's Eye is fighting for"—did not jibe with the manuscript's illustrations nor with Ukrainian history. In 1987 a physician named Leo Levitov asserted that the document had been produced by the Cathars, a heretical sect that flourished in medieval France, and was written in a pidgin composed of words from various languages. Levitov's translation, though, was at odds with the Cathars' well-documented theology.

Furthermore, all these schemes used mechanisms that allowed the same Voynichese word to be translated one way in one part of the manuscript and a different way in another part. For example, one step in Newbold's solution involved the deciphering of anagrams, which is notoriously imprecise: the anagram ADER, for instance, can be interpreted as READ, DARE or DEAR. Most scholars agree that all the attempted decodings of the Voynich manuscript are tainted by an unacceptable degree of ambiguity. Moreover, none of these methods could encode plaintext—that is, a readable message—into a ciphertext with the striking properties of Voynichese.

If the manuscript is not a code, could it be an unidentified language? Even though we cannot decipher the text, we know that it shows an extraordinary amount of regularity. For instance, the most common words often occur two or more times in a row. To represent the words, I will use the European Voynich Alphabet (EVA), a convention for transliterating the characters of Voynichese into Roman letters [*see box on opposite page*]. An example from folio 78R of the manuscript reads: *qokedy qokedy dal qokedy qokedy*. This degree of repetition is not found in any known language. Conversely, Voynichese contains very few phrases where two or three different words regularly occur together. These characteristics make it unlikely that Voynichese is a human language—it is simply too different from all other languages.

The third possibility is that the manuscript was a hoax devised for monetary gain or that it is some mad alchemist's meaningless ramblings. The linguistic complexity of the manuscript seems to argue against this theory. In addition to the repetition of words, there are numerous regularities in the internal structure of the words. The common syllable *qo*, for instance, occurs

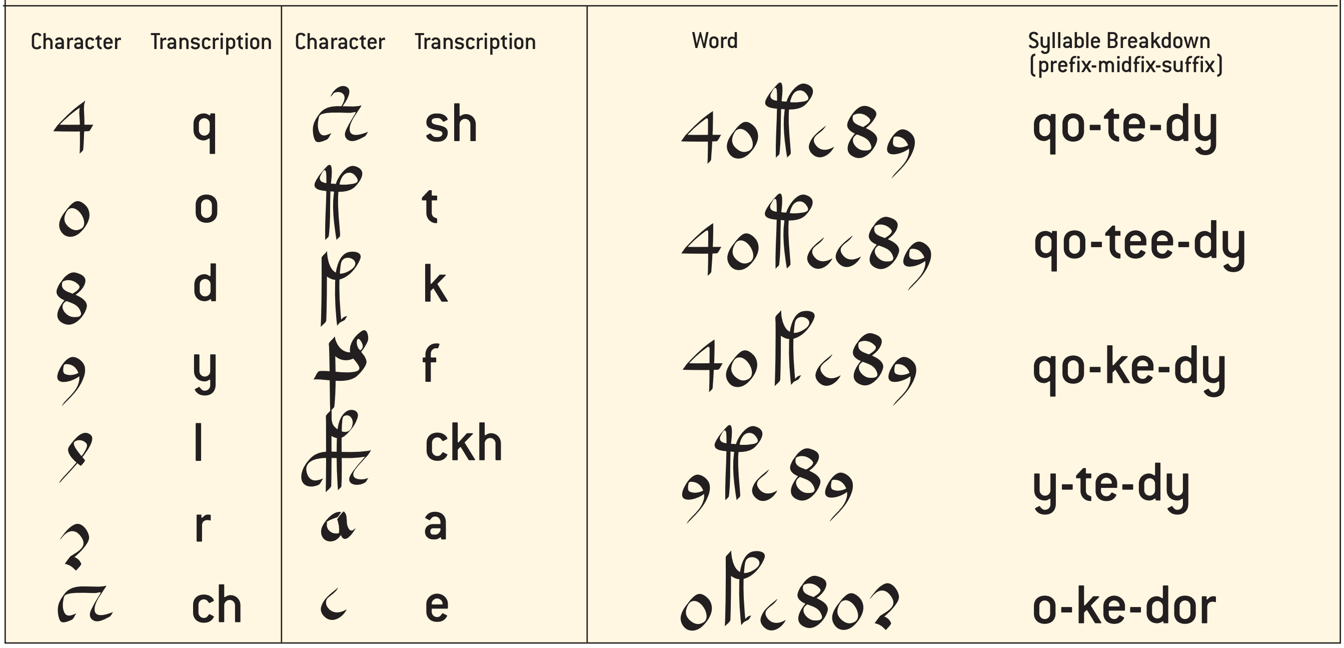
Overview/*A Medieval Mystery*

- A medieval document rediscovered in 1912, the Voynich manuscript was long assumed to be written in code. But the failure of all deciphering attempts has raised the suspicion that the script may not contain a message.
- New analysis has shown that a 16th-century coding tool could produce meaningless text that has most of the complex features of the Voynich script. This finding bolsters the theory that the manuscript is an alchemist's hoax.

BEINECKE RARE BOOK AND MANUSCRIPT LIBRARY, YALE UNIVERSITY

A VOYNICHESE PRIMER

The odd characters of the Voynich script can be transcribed into Roman letters using the European Voynich Alphabet. The transcriptions of 14 common characters are shown here (*below left*). Most Voynichese words can be broken down into prefix, midfix and suffix syllables (*below right*). The repetition of words in Voynichese is much greater than that of any known language.



NINA FINKEL

only at the start of words. The syllable *chek* may appear at the start of a word, but if it occurs in the same word as *qo,* then *qo* always comes before *chek*. The common syllable *dy* usually appears at the end of a word and occasionally at the start but never in the middle.

A simple "pick and mix" hoax that combines the syllables at random could not produce a text with so many regularities. Voynichese is also much more complex than anything found in pathological speech caused by brain damage or psychological disorders. Even if a mad alchemist did construct a grammar for an invented language and then spent years writing a script that employed this grammar, the resulting text would not share the various statistical features of the Voynich manuscript. For example, the word lengths of Voynichese form a binomial distribution—that is, the most common words have five or six characters, and the occurrence of words with greater or fewer characters falls off steeply from that peak in a symmetric bell curve. This kind of distribution is extremely unusual in a human language. In almost all human languages, the distribution of word lengths is broader and asymmetric, with a higher occurrence of relatively long words. It is very unlikely that the binomial distribution of Voynichese could have been a deliberate part of a hoax, because this statistical concept was not invented until centuries after the manuscript was written.

Expert Reasoning

IN SUMMARY, the Voynich manuscript appeared to be either an extremely unusual code, a strange unknown language or a sophisticated hoax, and there was no obvious way to resolve the impasse. It so happened that my colleague Joanne Hyde and I were looking for just such a puzzle a few years ago. We had been developing a method for critically reevaluating the expertise and reasoning used in the investigation of difficult research problems. As a preliminary test, I applied this method to the research on the Voynich manuscript. I started by determining the types of expertise that had previously been applied to the problem.

The assessment that the features of Voynichese are inconsistent with any human language was based on substantial relevant expertise from linguistics. This conclusion appeared sound, so I proceeded to the hoax hypothesis. Most people who have studied the Voynich manuscript agreed that Voynichese was too complex to be a hoax. I found, however, that this assessment was based on opinion rather than firm evidence. There is no body of expertise on how to mimic a long medieval ciphertext, because there are hardly any examples of such texts, let alone hoaxes of this genre.

THE AUTHOR

GORDON RUGG became interested in the Voynich manuscript about four years ago. At first he viewed it as merely an intriguing puzzle, but later he saw it as a test case for reexamining complex problems. He earned his Ph.D. in psychology at the University of Reading in 1987. Now a senior lecturer in the School of Computing and Mathematics at Keele University in England, Rugg is editor in chief of *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks.* His research interests include the nature of expertise and the modeling of information, knowledge and beliefs.

Several researchers, such as Jorge Stolfi of the University of Campinas in Brazil, had wondered whether the Voynich manuscript was produced using random text-generation tables. These tables have cells that contain characters or syllables; the user selects a sequence of cells—perhaps by throwing dice—and combines them to form a word. This technique could generate some of the regularities within Voynichese words. Under Stolfi's method, the table's first column could contain prefix syllables, such as *qo,* that occur only at the start of words; the second column could contain midfixes (syllables appearing in the middle of words) such as *chek,* and the third column could contain suffix syllables such as *y.* Choosing a syllable from each column in sequence would produce words with the characteristic structure of Voynichese. Some of the cells might be empty, so that one could create words lacking a prefix, midfix or suffix.

Other features of Voynichese, however, are not so easily reproduced. For instance, some characters are individually common but rarely occur next to each other. The characters transcribed as *a, e* and *l* are common, as is the combination *al,* but the combination *el* is very rare. This effect cannot be produced by randomly mixing characters from a table, so Stolfi and others rejected this approach. The key term here, though, is "randomly." To modern researchers, randomness is an invaluable concept. Yet it is a concept developed long after the manuscript was created. A medieval hoaxer probably would have used a different way of combining syllables that might not have been random in the strict statistical sense. I began to wonder whether some of the features of Voynichese might be side effects of a long-obsolete device.

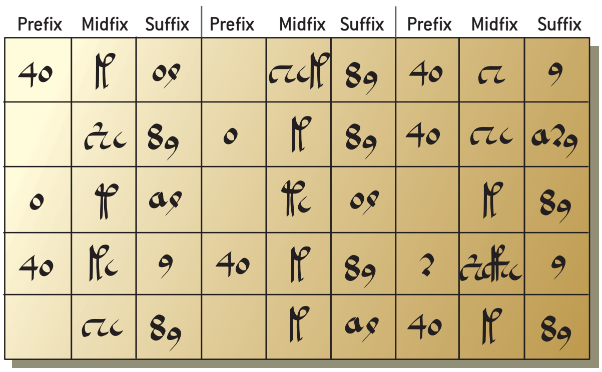
The Cardan Grille

IT LOOKED AS IF the hoax hypothesis deserved further investigation. My next step was to attempt to produce a hoax document to see what side effects emerged. The first question was, Which techniques to use? The answer depended on the date when the manuscript was produced. Having worked in archaeology, a field in which dating artifacts is an important concern, I was wary of the general consensus among Voynich researchers that the manuscript was created before 1500. It was illustrated in the style of the late 1400s, but this attribute did not conclusively pin down the date of its origin; artistic works are often produced in the style of an earlier period, either innocently or to make the document look older. I therefore searched for a coding technique that was available during the widest possible range of origin dates—between 1470 and 1608.

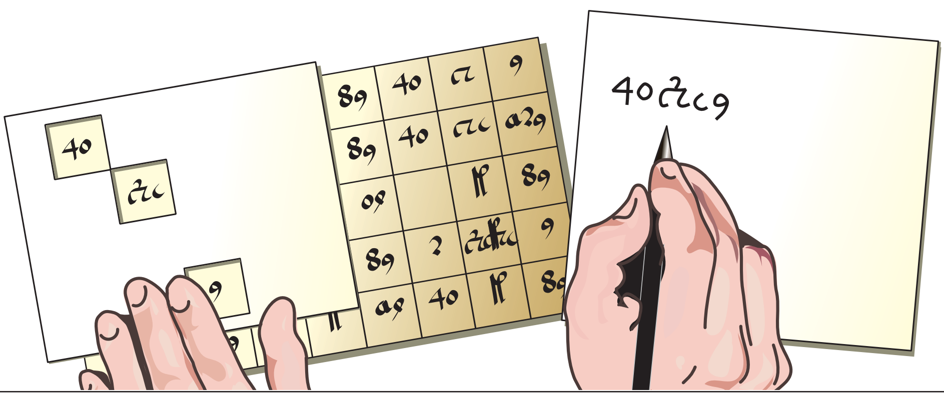
A promising possibility was the Cardan grille, which was introduced by Italian mathematician Girolamo Cardano in 1550. It consists of a card with slots cut in it. When the grille is laid over an apparently innocuous text produced with another copy of the same card, the slots reveal the words of the hidden message. I realized that a Cardan grille with three slots could be used to select permutations of prefixes, midfixes and suffixes from a table to generate Voynichese-style words.

A typical page of the Voynich manuscript contains about

REPRODUCING THE MANUSCRIPT



Prefix	Midfix	Suffix	Prefix	Midfix	Suffix	Prefix	Midfix	Suffix
qo	k	ol		chek	dy	qo	ch	y
	she	dy	o	k	dy	qo	che	ary
o	t	al		te	ol		k	dy
qo	ke	y	qo	k	dy	r	shckhe	y
	che	dy		k	al	qo	k	dy



A medieval hoaxer might have used a coding tool called the Cardan grille to create the Voynich manuscript. First the hoaxer draws a table containing columns of prefix, midfix and suffix syllables [*see part of a sample table and its transcription at top*]. Then the hoaxer places a card with three slots over the table to select the syllables for Voynichese words (*left*).

LUCY READING

English adventurer Edward Kelley may have concocted the document to defraud Rudolph II, the Holy Roman Emperor.

10 to 40 lines, each consisting of about eight to 12 words. Using the three-syllable model of Voynichese, a single table of 36 columns and 40 rows would contain enough syllables to produce an entire manuscript page with a single grille. The first column would list prefixes, the second midfixes and the third suffixes; the following columns would repeat that pattern [*see box on opposite page*]. You can align the grille to the upper left corner of the table to create the first word of Voynichese and then move it three columns to the right to make the next word. Or you can move the grille to a column farther to the right or to a lower row. By successively positioning the grille over different parts of the table, you can create hundreds of Voynichese words. And the same table could then be used with a different grille to make the words of the next page.

I drew up three tables by hand, which took two or three hours per table. Each grille took two or three minutes to cut out. (I made about 10.) After that, I could generate text as fast as I could transcribe it. In all, I produced between 1,000 and 2,000 words this way.

I found that this method could easily reproduce most of the features of Voynichese. For example, you can ensure that some characters never occur together by carefully designing the tables and grilles. If successive grille slots are always on different rows, then the syllables in horizontally adjacent cells in the table will never occur together, even though they may be very common individually. The binomial distribution of word lengths can be generated by mixing short, medium-length and long syllables in the table. Another characteristic of Voynichese—that the first words in a line tend to be longer than later ones—can be reproduced simply by putting most of the longer syllables on the left side of the table.

The Cardan grille method therefore appears to be a mechanism by which the Voynich manuscript could have been created. My reconstructions suggest that one person could have produced the manuscript, including the illustrations, in just three or four months. But a crucial question remains: Does the manuscript contain only meaningless gibberish or a coded message?

I found two ways to employ the grilles and tables to encode and decode plaintext. The first was a substitution cipher that converted plaintext characters to midfix syllables that are then embedded within meaningless prefixes and suffixes using the method described above. The second encoding technique assigned a number to each plaintext character and then used these numbers to specify the placement of the Cardan grille on the table. Both techniques, however, produce scripts with much less repetition of words than Voynichese. This finding indicates that if the Cardan grille was indeed used to make the Voynich manuscript, the author was probably creating cleverly designed nonsense rather than a ciphertext. I found no evidence that the manuscript contains a coded message.

This absence of evidence does not prove that the manuscript was a hoax, but my work shows that the construction of a hoax as complex as the Voynich manuscript was indeed feasible. This explanation dovetails with several intriguing historical facts: Elizabethan scholar John Dee and his disreputable associate Edward Kelley visited the court of Rudolf II during the 1580s. Kelley was a notorious forger, mystic and alchemist who was familiar with Cardan grilles. Some experts on the Voynich manuscript have long suspected that Kelley was the author.

My undergraduate student Laura Aylward is currently investigating whether more complex statistical features of the manuscript can be reproduced using the Cardan grille technique. Answering this question will require producing large amounts of text using different table and grille layouts, so we are writing software to automate the method.

This study yielded valuable insights into the process of reexamining difficult problems to determine whether any possible solutions have been overlooked. A good example of such a problem is the question of what causes Alzheimer's disease. We plan to examine whether our approach could be used to reevaluate previous research into this brain disorder. Our questions will include: Have the investigators neglected any field of relevant expertise? Have the key assumptions been tested sufficiently? And are there subtle misunderstandings between the different disciplines that are involved in this work? If we can use this process to help Alzheimer's researchers find promising new directions, then a medieval manuscript that looks like an alchemist's handbook may actually prove to be a boon to modern medicine. SA

MORE TO EXPLORE

The Voynich Manuscript: An Elegant Enigma. Mary E. D'Imperio. Aegean Park Press, 1978.

The Queen's Conjuror: The Science and Magic of Dr Dee. Benjamin Woolley. Flamingo/HarperCollins, 2002.

The Voynich Manuscript: An Elegant Hoax? Gordon Rugg in *Cryptologia,* Vol. 28, No. 1; January 2004.

More information about the Voynich manuscript can be found on the Web at **www.keele.ac.uk/depts/cs/staff/g.rugg/voynich/; www.voynich.nu/; www.dcc.unicamp.br/~stolfi/voynich/;** and **mysite.freeserve.com/philipneal_vms**

BEINECKE RARE BOOK AND MANUSCRIPT LIBRARY, YALE UNIVERSITY

PIPE ORGANS

Big Air

Pipe organs are the largest, most powerful musical instruments on earth. Yet a mere sliver of wood stands between a thundering roar and dead silence.

Organs have functioned on the same simple physics since they were first constructed around 250 B.C. Air flows from a reservoir up past leather-covered wooden pallets into pipes of various lengths, vibrating to produce different pitches. For 2,000 years, humans—usually boys from a church congregation—squeezed bellows by hand or feet to supply the air, until electric motors and blowers took over.

Through the late 1800s, when an organist pressed a key, a series of wooden linkages called trackers pulled open a pallet to allow air to stream into a specific pipe. As listeners sought a greater variety of sounds from ever larger mechanical wonders to liven ever larger churches and municipal halls, however, greater wind pressure was needed, and players had to press ever harder on a key to open a pallet against that pressure. Designers eventually turned to newfangled electricity to solve the problem: when a key was (effortlessly) depressed, it completed a circuit that activated an electromagnet or a solenoid, which in turn operated a pallet directly.

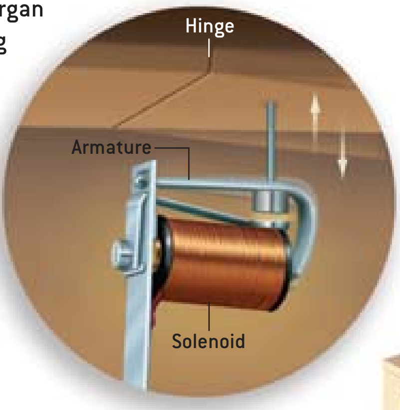
Today organs with each of these types of "actions" are still made. "There are differences in sound," says Stanley Scheer, vice president of Casavant Frères, the venerable organ builder in St. Hyacinthe, Quebec, founded in 1879. There is a kind of refinement in sound quality to certain designs, a kind of excitement to others.

A modest church organ with two keyboards, foot pedals, and 16 or so stops (different tones) may have about 1,200 pipes, cost around $200,000 and take two years to craft by hand. It is completely assembled and tested at the shop, then dismantled and shipped to its final playing space. There, over several weeks, the myriad pieces and trackers or wires are reassembled in an intricate maze. Every pipe is voiced (adding small touches to adjust for the room's acoustics) and tuned. "Each organ is an engineering marvel," Scheer says. And each time an organ is first played for the public, he adds, "we think it's amazing how it results in music."

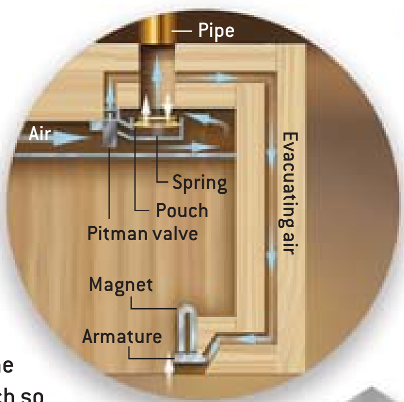
—*Mark Fischetti*

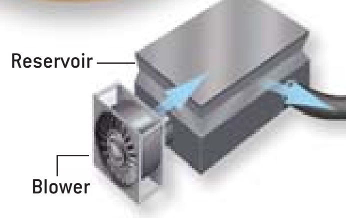
MECHANICAL organ runs on air sent by a blower to a reservoir that regulates the pressure inside a wind-chest. When the organist pulls out a stop, linkages move a wooden slider so its holes line up under pipes in a rank, allowing them to be played. When a key is pressed, trackers pull open a wooden pallet, allowing air to enter a key channel (say, A♭) and stream up into all the pipes above that channel whose stops have been opened.

ELECTROMECHANICAL organ has no trackers. Pressing a key sends current along a wire to a solenoid that pulls open an armature and pallet, letting air into a key channel. A spring (*not shown*) closes the pallet.



ELECTROPNEUMATIC organ operates on differences in air pressure. Pulling out a stop retracts a pitman valve, opening a channel to a small felt disk (the pallet) held in place above a leather pouch and spring. Depressing a key sends current to an electromagnet, which pulls up an armature, allowing high-pressure air to evacuate from the channel, drawing down the pouch so air streams up into a pipe.





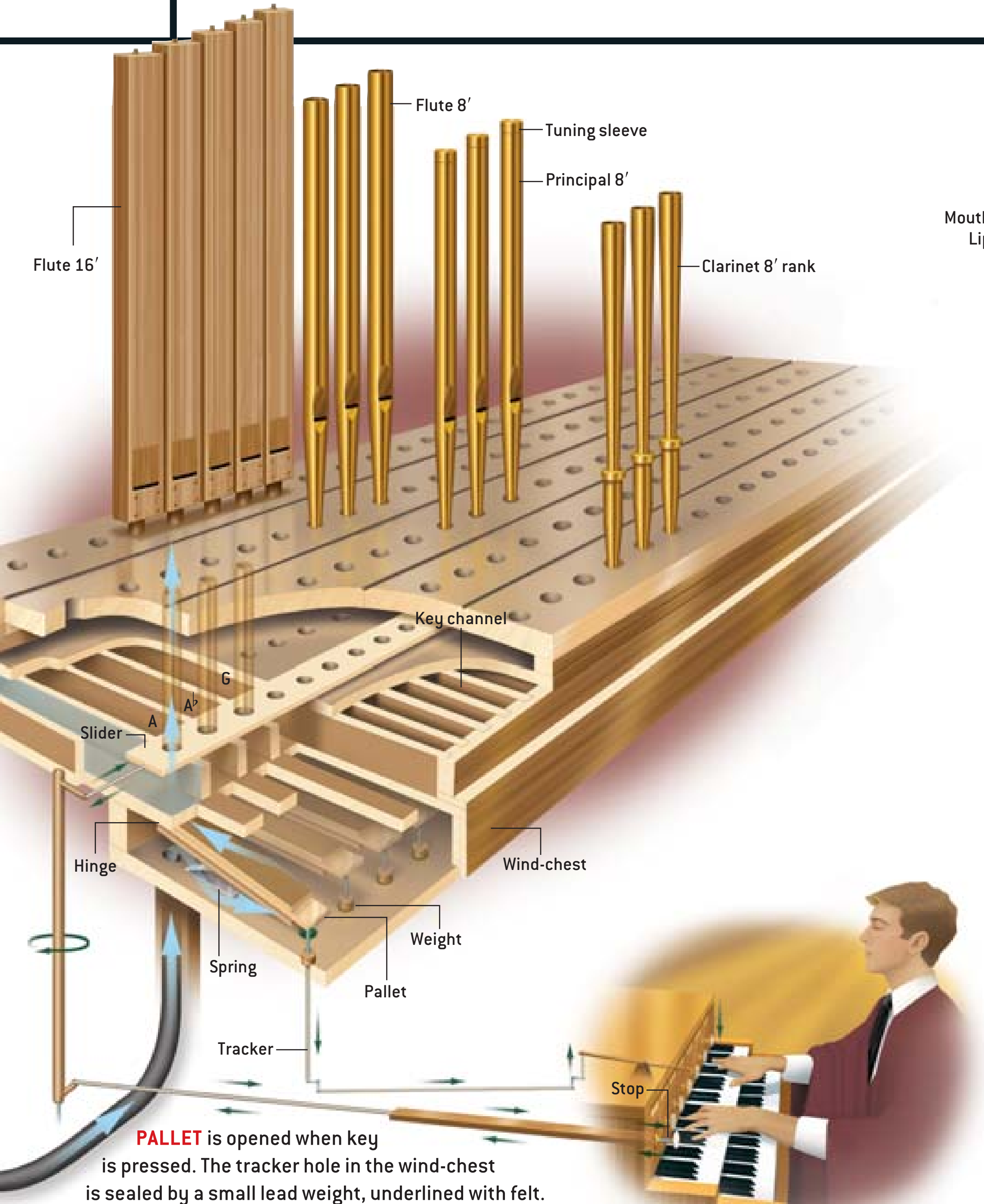
KENT SNODGRASS *Precision Graphics*

DID YOU KNOW . . .

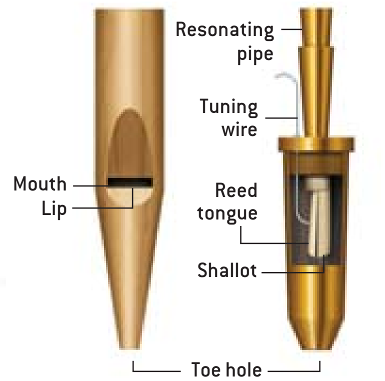
➤ SOUNDS TINNY: Wood is used to make the bigger pipes that emit bass notes, because it can increase resonance while lessening harmonic overtones. In metal pipes, more lead promotes fundamental tone; more tin provides greater harmonics. Zinc and copper may also be used. Different pipe thicknesses influence the timbre, too.

➤ SOUNDS BIG: The vast dimensions of St. Paul's Cathedral in London create a nine-second reverberation for its grand organ, built in the 1690s; organists must learn how to harness the rolling sound yet ignore the distraction as they play. In the Newberry Memorial Organ at Yale University, the "Bombarde" stop opens 32-foot-high pipes that shake the house at 16 cycles a second. The John Wanamaker Grand Organ that fills the seven-story atrium of Hecht's department store in Philadelphia has 28,500 pipes, the most in the world; the biggest is wide enough for a pony to stand in, the smallest is a quarter-inch tall. The Walt Disney Concert Hall organ in Los Angeles, debuting this summer, weighs 40 tons; steel framing supports more than 100 giant pipes curved to follow the swooping walls, and the 6,134 pipes are anchored in heavy metal footings to resist earthquakes.

➤ SOUNDS SWELL: An organ sporting a "swell" feature has sets of pipes enclosed behind large vertical, wooden louvers; the pipes inside sound muffled, but as the organist opens the louvers with a foot pedal, the volume rises, or swells, to a dramatic crescendo.



PALLET is opened when key is pressed. The tracker hole in the wind-chest is sealed by a small lead weight, underlined with felt. When a key is released, a spring closes the pallet, which is quieted by a felt lining and sealed by a leather gasket.



FLUE PIPES (*left*) are hollow; as wind streams up from an open pallet, it hits the upper lip, causing the column of air to vibrate and thus produce sound. A movable sleeve or stopper at the top enables tuning. In a reed pipe (*right*), wind causes a tongue to vibrate against a shallot, which stimulates the air column in the resonating pipe. A tuning wire pinned against the reed determines pitch, like a finger on a violin string. Pipes may be square, round or tapered. Flues tend to be used for organ (diapason, principal), string and flute tones, reeds for trumpets and clarinets. Closed pipes are half the length of open pipes for the same pitch. An open pipe two feet long (from mouth to top) produces middle C (261.6 hertz).

Topic suggested by reader Emily Brown.
Send ideas to workingknowledge@sciam.com

Stars atop a Silent Volcano

THE LARGEST ASTRONOMICAL OBSERVATORY IN THE WORLD SITS ON MAUNA KEA IN HAWAII. THE VIEWS, UP OR DOWN, ARE SPECTACULAR **BY MARGUERITE HOLLOWAY**

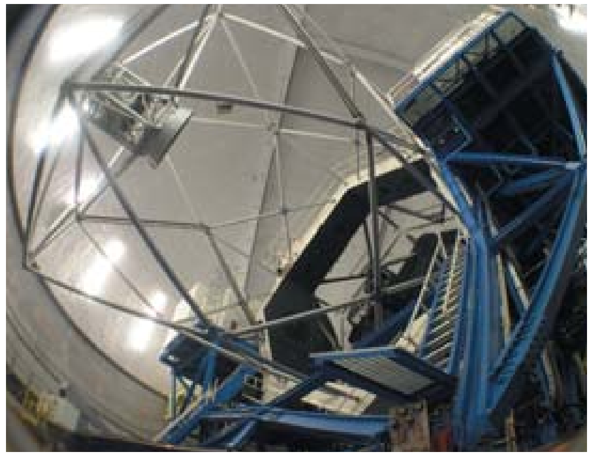
CANADA-FRANCE-HAWAII TELESCOPE was built in 1979. Although the catwalk is not open to visitors, a similar view can be had from the publicly accessible catwalk at the University of Hawaii's 2.2-meter telescope. Both telescopes have been instrumental in recent discoveries of Jupiter's many satellites.

RICHARD J. WAINSCOAT

It is with some trepidation that I turn onto Saddle Road after a half-hour drive north from Kona on Hawaii's Big Island. Only one rental car company—Harper's Car and Truck—will let renters use their vehicles on this road without limiting coverage, so I am expecting a pockmarked dirt nightmare. But Saddle Road, which crosses the top part of the island, is smooth and paved. Its 50 miles are merely narrow, empty of gas stations and stores, and unlit at night. A short distance from two major cities and several smaller ones, Saddle Road seems isolated, cut off from the world.

Turning off at mile marker 28 only intensifies that feeling, as it becomes possible to peer beyond this world. It is here that more than 106,000 people come every year to visit Mauna Kea, a long-quiet volcano, sacred to Hawaiians, atop which astronomers are using the world's largest collection of telescopes to better understand the origin and evolution of the universe and the composition of everything in it.

Mauna Kea, which means "white mountain," is an ideal spot for studying the cosmos. It is unusually clear at night because clouds are drawn away by a downdraft. At 4,200 meters, it is above 40 percent of the earth's atmosphere and also above a layer of clouds that separates it from the lower atmosphere's wet ocean air, so the summit air is free of dust, smog and moisture. Because of these attributes and the fact that the volcano has not been active for 4,500 years or so (the mantle hot spot is now enlivening Mauna Loa 30 miles away as well as a submarine site



KECK I AND II are among the best-known telescopes on Mauna Kea (*left*). Some of the observatory's findings have been made with HIRES (*white box on platform in image at right*). Using this powerful instrument, astronomers have discovered over half of the more than 120 known extrasolar planets.

RICHARD J. WAINSCOAT (*left*); SCOTT KARDEL *Palomar Observatory* (*right*)

even farther to the south), astronomers had their eyes on the mountain for many years. Development started in the late 1960s, and today 13 telescopes of various kinds aid research, including the famous Keck I and Keck II.

Before guests can drive to the summit and see all this powerful equipment, they must spend some time at the visitor center—named after *Challenger* crew member and Hawaiian Ellison S. Onizuka—to acclimate to the altitude. (Official summit tours are offered only on weekends at 1 P.M., but stargazing programs are available every night at 6: see www.ifa.hawaii.edu/mko) This particular Sunday, after watching a film about the history of the site and the ongoing research, a tour group of about 25 people climbs into four-wheel drives for the long, winding, unpaved ride up the volcano. The altitude makes engines inefficient, so filling the tank just before entering Saddle Road is essential; the high elevation also prevents brakes from working well, so the road-gripping ability of a four-wheel drive is mandatory.

The first stop on the barren, brown, wind-scoured mountain is Keck I, one of the twin optical and infrared telescopes that make up the W. M. Keck Observatory. They are the largest such telescopes: each weighs 300 tons and stands eight stories tall. Astronomers have used the instruments to study early stars and how they formed in ancient clouds of gas just after the big bang, some 13.7 billion years ago. They have discovered planets outside our solar system orbiting their own stars. And recently researchers using the Keck II telescope were able to discern methane in the atmosphere of Mars, which could have been introduced by a comet, released by volcanic activity on the planet itself or, some speculate, generated by the metabolism of bacteria.

The summit is chilly, but inside the telescope it seems even colder. Our guide explains that it costs $35,000 a month to keep the Keck domes near freezing—important for eliminating any chance of heat distorting the 10-meter mirror or the metal equipment. (For those who want to visit on weekdays, Keck I's viewing area and gallery are open from 10 A.M. until 4 P.M. Monday through Friday; see www2. keck.hawaii.edu)

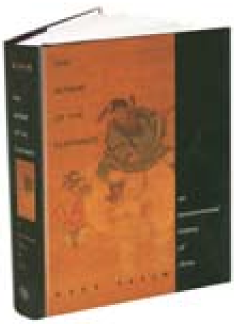
We next visit the University of Hawaii's 2.2-meter telescope, and from the catwalk around it we look down on the white bubble domes of the Kecks and a sea of clouds below and behind them. This is the only other telescope that is open to visitors on the summit. The 2.2-meter was instrumental, along with several others at Mauna Kea, in the discovery of dozens of additional satellites around Jupiter. The planet's total now stands at 63. The telescope was also used in the discovery this February of a nearby—that is, only 33 light-years distant—planet-forming red dwarf star. By examining this star, dubbed AU Mic, astronomers hope to better understand how planets are formed.

Although visitors can't go inside many of the telescopes and can't see the work of astronomers up close, it is still a remarkable feeling to be tiny against these great white structures, to feel a mote against the backdrop of such a monumental search. Many of the tour takers are amateur astronomers who spend every vacation stargazing and visiting observatories. "Mauna Kea Observatory is clearly the best," sighs one as we head back to the cars.

It is late afternoon, the tour is over, and the air is becoming even colder. The four-wheel drives fall into an informal convoy and creep down the mountain. I am heading back along Saddle Road as the sun sets. The light catches small flowers, which blaze like small lanterns along the road. Tufts of white grass glow against the black lava fields; the land looks like a bleached blonde showing her roots. Everywhere light leaps out against the dark. SA

Where the Wild Things Were

FOUR THOUSAND YEARS OF HUMANS VS. NATURE IN CHINA **BY NICHOLAS D. KRISTOF**



THE RETREAT OF THE ELEPHANTS: AN ENVIRONMENTAL HISTORY OF CHINA
By Mark Elvin
Yale University Press, New Haven, Conn. ($39.95)

Who knew that elephant trunk tastes like piglet? Or that more than a millennium ago, a writer declared that Chinese "competed to eat their trunks, the taste of which is said to be fatty and crisp, and to be particularly well suited to being roasted."

Elephants, it turns out, once roamed across nearly all of China, as did rhinoceroses. Indeed, for 1,000 years the standard armor worn by Chinese soldiers was made from rhino hide. Yet these days rhinos are completely extinct in China, and elephants linger only in protected enclaves in the far southwest of the country.

China being China, everything has been carefully documented, so we know that these large mammals retreated gradually over the past 4,000 years, half a step ahead of smaller, two-legged ones. Mark Elvin, an Australian scholar, brilliantly uses that prolonged elephantine trail of tears as the guiding metaphor for his new book, *The Retreat of the Elephants: An Environmental History of China.*

Frankly, I didn't know that I was interested in the history of Chinese elephants, or that I was yearning for an environmental history of China, until I read this book. But Elvin combines an illuminating account of the 4,000-year-long collision of humans and nature with delightful tidbits about everything under the Chinese sun.

One could not have written such an environmental history about, for instance, Britain or Russia. From China's point of view, such countries are modern ingenues with barely any history to speak of. But in China, we hear, for example, that the Duke of Zhou, more than 3,000 years ago, drove "elephants far away" from the Yellow River valley. A record from 548 B.C. describes the ivory trade, and later we begin to get detailed accounts of battles over crops between peasants and elephants in, say, A.D. 962.

Of course, just because something has been recorded does not mean it is true. One account from 1608 reports of trained elephants in the Ming Dynasty court: "If an elephant commits an offense, or injures a human, the imperial command will be issued for him to be beaten.... Only when the beating has been concluded will he rise to his feet to give thanks for the favor received.... In the sixth lunar month they are bathed and mated. The coupling takes place in the water with a female who floats with her face upward, in all respects like a human being." Hmmm. Floating face upward? So that's how Ming Dynasty historians made love.

Elvin is particularly fascinating on the history of China's long wrestling match with water. Chinese civilization may have evolved out of efforts to irrigate the land, and there is an intriguing record of the quest to tame water and land, which would typically succeed for a while until the water rebelled.

The problems were especially acute with the Yellow River, which was not called that in ancient times. Then, a little more than two millennia ago, the Qin and Han dynasties promoted farming along the upper reaches of the river, and the resulting erosion filled the water with sediment that made it muddy and gave it its present name. The sediment raised the riverbed until it was held in place only by man-made dikes that required constant attention—because the water, in essence, flowed aboveground, not below it.

Periodically dikes broke, sometimes catastrophically. A flood in 1117 is said to have killed more than one million people, making it perhaps the worst such disaster since Noah. The Yellow River dramatically changed course in 1194, moving to the south of the Shandong Peninsula, until in 1853 it moved north again. Elvin meticulously recounts China's hydrology, so we learn, for example, that

GERRY ELLIS *Minden Pictures*

ELEPHANTS RANGED **throughout most of the huge area that became China. Today the only Chinese elephants remaining in the wild live in a few protected enclaves near the border with Burma.**

between 1195 and 1578 the Yellow River delta advanced only 39 meters a year (as sediment built up), whereas from 1579 to 1591 it advanced 1,538 meters a year.

Sometimes the sheer weight of detail is numbing, particularly in later chapters offering case studies within China. Readers without an intrinsic fascination with China may find this a book to browse, not to read cover to cover. But as a window into the history of the Middle Kingdom, and an extended account of human interactions with the environment, this is a magisterial work.

What gives this book special resonance is the impact China will have on the global environment in the coming decades. The industrial revolution in the West has been so destructive of nature that we should be wary of what the industrialization of China and India will mean. I congratulate my Chinese friends when they buy their first cars, one after the other, but collectively the result of Chinese industrialization will be to swallow up nonrenewable resources, to increase carbon emissions and presumably global warming, and to send acid rain drizzling down on much of the globe.

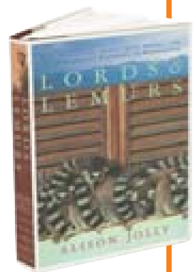
Yet this book does not really illuminate the road ahead. Elvin tells us that it was originally intended to carry us to the present day, but he ends up pretty much grinding to a halt a couple of hundred years ago. The even more gruesome period since—and, brace yourself, the predations still ahead of us—will have to be the subject of a companion volume. Alas, the Chinese elephants have already been driven to the country's fringe and have nowhere else to go. And unless they figure out how to mate even when the female is not floating faceup in a pool of water, they're really in trouble. SA

Nicholas D. Kristof is a columnist for the New York Times. *He won a Pulitzer Prize for his coverage of China and is co-author, with his wife, Sheryl WuDunn, of* China Wakes: The Struggle for the Soul of a Rising Power.

THE EDITORS **RECOMMEND**

LORDS AND LEMURS: MAD SCIENTISTS, KINGS WITH SPEARS, AND THE SURVIVAL OF DIVERSITY IN MADAGASCAR

by Alison Jolly. 24 pages of photographs. Houghton Mifflin, 2004 ($25)



Jolly, a pioneer in the study of primate behavior, first went to Madagascar to observe lemurs 40 years ago. Her research site was at Berenty, a private wildlife refuge that was part of the plantation of an aristocratic French family. The de Heaulmes had come to Berenty in 1928. As they developed their plantation over the years, they also set aside a large area of it for lemurs and other animals and helped the native Tandroy tribe preserve their traditions.

At the beginning of the 21st century, Berenty and its lemurs still flourish because the de Heaulme family are still there—and vice versa. "Forest and family saved each other," Jolly says. The plantation no longer produces sisal commercially; together with the preserved forest and its lemurs, it has become a destination for ecotourists. Woven around the life of the de Heaulme family is the entire history of Madagascar—its geology, its animals and its colonization by humans, beginning with Indonesians and Africans in around A.D. 500. It is an unexpectedly enthralling story, told with great flair.

"THE LEMURS FOUND ME. Staring lemon eyes in a black, heart-shaped face set in a square, white-furred head: a white sifaka clung vertically to a vertical trunk, its tail rolled up like a watch spring. Then, without warning, it leaped. It seemed to double in size. Its hind legs, longer than head and body together, propelled it backward into space in a curve as taut as ballet. It did not jump away, but toward me! It turned in midair to land with both hind feet first, then folded up and clung vertically to another trunk, still watching me. Another followed across the same gap...."

FRANS LANTING *Minden Pictures*

The books reviewed are available for purchase through www.sciam.com

Einstein's Parrot

A GREAT BRAIN AND A BIRD BRAIN SPEND TIME TOGETHER **BY STEVE MIRSKY**

In late April the Associated Press reported the discovery of a diary written by a woman, Johanna Fantova, who was a close friend of Albert Einstein. "The 62-page diary, written in German, was discovered in February in Fantova's files at Princeton University's Firestone Library, where she had worked as a curator," the AP story noted. One fascinating revelation of the diary is that Einstein received a parrot as a 75th-birthday gift. According to the AP, "After deciding the bird was depressed, Einstein tried to alter its mood by telling bad jokes."

Parrots can live for a century. In early May I may (or may not) have encountered a parrot that may (or may not) have been the bird entertained by Einstein. Speaking in German-accented English, the parrot recited a monologue. What follows is a transcript of that monologue:

"How do I order beer in a bar? I say '*Ein stein* for Einstein.' Hey, Parrot, what's the difference between a wild boar and Niels Bohr? When I say that God doesn't play dice, a wild boar doesn't tell me to stop telling God what to do. I hate that. So what do you say to the man who developed the exclusion principle? You say, 'Pauli want a cracker?' Wolfgang Pauli, get it? Hello, is this thing on? Testing, one, two. Hey, Parrot, I had a dream where I made love to Rita Hayworth for an hour. Well, for her it was an hour. For me, 35 seconds. That's relativity. Okay, Newton is standing on the shoulders of a giant, and he says, 'Giant, how do I get down off you?' and the giant says, 'You don't get down off me, you get down off a duck.' I love that one. Parrot, tell me, what is a Lorentz contraction? That's when Mrs. Lorentz knows the baby is coming. It's a timed dilation, not a time dilation, get it? Let's see, two guys walk into an h-bar. An H-BAR. If you knew any physics you'd be on the floor, I swear. Uh, if Ruby Keeler married, uh, John Wheeler, became a doctor and got a job in Vegas, she'd be Ruby Keeler Wheeler the healer dealer. So what would people say if Paul Dirac fell on Jane Russell? They'd say, 'Look at Dirac on Jane Russell.' Oh, they'd say it, trust me. Okay, there are these twins, see. They're 20 years old. And one of them goes zipping around the universe really fast while the other one stays on Earth. The twin who was zipping around comes back, and he's maybe a year older, and he goes to find his brother. And the brother is now 95 years old. And the young twin comes up to him. The old twin looks at the young twin, and tears come to his eyes. And the young twin says, 'Why are you crying?' And the old twin says, 'I'm so happy.' And the young twin says, 'To see me?' And the old twin, he says, 'Yes. The $100 you owed me when you left. It's now $100,000.' From the compounding interest. Oy, these are the jokes, Parrot. What, you don't like living in a cage? Yeah, try being the most famous man in the world. I can't even go out for a haircut. You know, you're a good listener for a parrot. Oh boy, it looks like you just did a Brownian movement. Good thing I lined the cage with my cosmological constant proposal. That proposal was my second biggest mistake. My biggest mistake was my proposal to my first wife. Ba-dum-bum. Parrot, if you had a plastic deer on your lawn covered in Christmas lights, turning them on would give you the faux doe electric effect. Whaddya call it when Leo Szilard and Enrico Fermi pull up an anchor? A chain retraction! Not so good? You should hear me play the violin. So Schrödinger and Heisenberg are driving down the road, and Heisenberg says, 'Hey, I think you just ran over a cat.' And Schrödinger, he says, 'Is it dead?' And Heisenberg says, heh heh, get this: 'I can't be certain.' Okay, so the smartest man in the world is talking to a parrot. Hey, Parrot, that's not a joke, that's my life." SA

ILLUSTRATION BY MATT COLLINS; PHOTOGRAPH BY FRANK VERONSKY

Q Why do people snore?

—G. Alvarado, Iligan City, the Philippines

Lynn A. D'Andrea, a sleep specialist at the University of Michigan Medical School, explains:

Vibrating structures of the upper airway cause snoring. These tissues include the tongue, soft palate, uvula, tonsillar pillars and pharyngeal walls. During sleep, muscle tone throughout the body decreases. Relaxed airway muscles can reduce the size of the airway space, limiting airflow and creating turbulence, especially when inhaling, with snoring as the result.

The reported prevalence of snoring varies and depends on the population studied and the wording of the questionnaire. For instance, data from the 1993 Wisconsin Sleep Cohort Study of 602 people showed that 44 percent of the men surveyed and 28 percent of the women snored habitually. Overall 4 percent of the men and 2 percent of the women had snoring associated with obstructive sleep apnea, a disorder involving repeated pauses or gasps in a person's breathing during sleep when the airway is obstructed. Typically the obstruction ends with an arousal—the snorer wakes briefly—leading to fragmented, less restful sleep. Obstructive sleep apnea can cause excessive daytime sleepiness, decreased attention, poor concentration and decreased energy levels. Vascular complications such as hypertension can also contribute to the disorder. The prevalence of snoring and obstructive sleep apnea seems to increase with age, especially after 65. Additional risk factors associated with snoring include weight gain, alcohol consumption, allergies, nasal obstruction, use of muscle relaxants or sedatives, and smoking (which may increase inflammation, thereby narrowing the upper airways).

Avoiding these risk factors, as well as sleeping on one's side, can help ease snoring. Other remedies run the gamut from simple devices to surgical procedures. An external nasal dilator, for example, attaches to the outside of the nose like a bandage, gently lifting and opening the nasal passages. Additional solutions include oral appliances and devices that create a continuous positive pressure in nasal passages, preventing the tongue and soft palate from collapsing into the upper airway. Finally, surgeons can use a laser to remove some of the vibratory tissues.

Q What kind of patterns does the Search for Extraterrestrial Intelligence (SETI) look for?

Peter R. Backus, observing programs manager at the SETI Institute in Mountain View, Calif., provides this answer:

Before looking for patterns, SETI scientists try to discern a nonrandom signal from the background of cosmic and terrestrial noise. We look for simple signals, ones that are easy to distinguish from astrophysical sources.

Natural radio sources such as quasars and pulsars are "broadband," meaning they emit over a broad range of frequencies. Communications signals may also be broadband but may contain components that are very narrow and easy to distinguish. For example, an analog TV signal spans about six megahertz but carries two signals (video and audio) that are less than one hertz wide. Therefore, SETI programs searching in the radio spectrum look for very narrow bandwidth signals. Likewise, optical SETI programs measure a given star's light in nanosecond intervals, looking for the brief appearance of a slight excess of photons that could originate from a powerful pulsed laser.

Once a signal is detected, the hunt for possible ET patterns will begin. The detected signal may be a simple beacon that points the way to an information channel. Scientists have speculated that a binary code—a pattern of ones and zeros—will be used, perhaps containing a symbolic language or pictures. In 1974, at the rededication of the Arecibo Observatory in Puerto Rico, humans sent a binary-coded picture in the direction of the star cluster M13. The transmission used two frequencies to represent the dark and light pixels of the message. Other encoding schemes have been suggested, but in truth scientists do not know what to expect. That is part of what makes SETI so interesting. SA

For a complete text of these and other answers from scientists in diverse fields, visit www.sciam.com/askexpert

SARA CHEN