

# CURJ

SPRING 2004  
VOL. 4 NO. 1

CALTECH UNDERGRADUATE RESEARCH JOURNAL

## KILLING THE MESSENGER WITH RNA INTERFERENCE

PLUS

SOLAR  
PROMINENCES

FOUR COLOR  
THEOREM

LINE-OF-SIGHT  
ALGORITHMS

THERMAL  
BIMORPHS

# CURJ



## COVER FEATURE

- 24 KILLING THE MESSENGER** by Vincent Auyeung  
RNA interference is driving a revolution in genetic analysis and may become a major antiviral therapy.

## LETTERS

- 4 COLLABORATIONS BETWEEN THE ARTS AND SCIENCES** by Ramone Muñoz
- 6 REVISITING ARISTOTLE** by Ed McCaffery

## REVIEWS

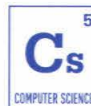
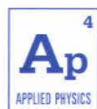
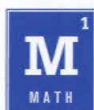
- 8 HOW MANY COLORS ARE ENOUGH?** by Andrew Yang  
It took the advent of computers to solve a problem that humans could not.
- 16 THE ELECTROMAGNETIC THEORY BEHIND SOLAR PROMINENCES** by Eric Lin  
Beautiful phenomena are explained using principles of electromagnetism.

## RESEARCH

- 32 FILMS THAT BEND** by Azriel C. Epilepsia  
Bimorphs: apply heat and they bend. The author explores the issues and difficulties of making and using a new technological building block.
- 36 LEAPING OVER CHASMS** by Joseph Gonzalez  
A new algorithm speeds up the evaluation of line-of-sight on digital elevation maps.

## FINIS

- 44 TECHNOSPHERE** by Seth Drenner





## Make Beautiful Bits at Finisar!

Finisar hires exceptional individuals who excel in an innovative, high-energy environment where the standards of high quality and excellence inspire employees to realize aggressive company goals and personal achievement. Currently full-time and intern candidates are being considered for positions within the following areas: Optical Data Links, Test Instruments and Optical Networking.

General Qualifications: B.S., M.S. or Ph.D in Electrical Engineering, Applied Physics, Physics, Computer Science, Computer Engineering and Mechanical Engineering. Excellent written and oral communication skills. Ability to thrive in an innovative, high-energy atmosphere where tasks can be accomplished simply and directly.

Send resumes to: [hr@finisar.com](mailto:hr@finisar.com)

**Finisar**

[www.finisar.com](http://www.finisar.com)

THE CALTECH  
UNDERGRADUATE  
RESEARCH JOURNAL  
INVITES REVIEW  
AND RESEARCH  
**SUBMISSIONS** FROM  
UNDERGRADUATES  
AT ANY INSTITUTION.

**CURJ**

for more information:

[www.curj.caltech.edu](http://www.curj.caltech.edu)

## we seek out-of-the-box thinkers.

Sandia National Laboratories wants people who are eager to tackle the grand scientific and engineering challenges of the 21st century. People who desire to make America, and the world, a better and safer place—and who have the determination to create the technological means to make it that way. Join the team that is changing the world.

**We have exciting opportunities for college graduates at the Bachelor's, Master's, and Ph.D. levels in:**

- Electrical engineering
- Mechanical engineering
- Information Technologies/  
Information Systems
- Computer science
- Computing engineering
- Chemistry
- Nuclear engineering...and more

**We also offer internship, co-op, and post-doctoral programs.**

[www.sandia.gov](http://www.sandia.gov)

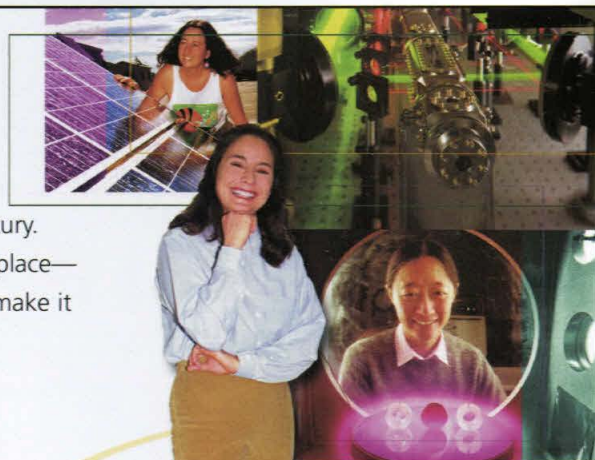


**Sandia  
National  
Laboratories**

Operated for the  
Department of Energy by  
Lockheed Martin Corp.



Sandia is an equal opportunity employer.  
We maintain a drug-free workplace.



**Your  
world  
will  
never  
be the  
same.**



## FROM THE EDITOR

The Caltech Undergraduate Research Journal's mission is not only to present great science in a readable, engaging form but also to foster better writing in the Caltech community and beyond. CURJ is not assembled in a vacuum; each article you read starts as a submission that gets reviewed by a group of our editors. These editors are looking not for perfect prose; instead, they are looking for ideas and science that serve as the foundation for a solid article.

Once an article is selected, one of our editors begins a one-on-one relationship with the author of the submission. We firmly believe that the authors should be involved as much as possible in the editing process, or the voice of the author can get lost in the numerous revisions that each article requires to coalesce into a final draft that fits in with CURJ's established style. It also is helpful for the individuals involved, since it ensures that the editors completely understand the underlying science and that the authors learn how to better compose inviting scientific writing.

Our efforts do not end there, though. We continually work with the Summer Undergraduate Research Fellowship (SURF) office here at Caltech to provide useful and understandable guidelines for the undergraduates from around the world who conduct research at Caltech and hopefully will submit to CURJ after completing their research. Our final product serves as an example that all levels of science can be presented in a clear and interesting manner for a wide audience. Enjoy!



Jordan Boyd-Graber

Executive Editor

## EXECUTIVE EDITORS

**Editor-in-Chief**  
Philip Wong

**Executive Content Editors**  
Jordan Boyd-Graber  
Shaun P. Lee

**Art Director/Designer**  
Steven Neumann

**Executive Online Editor**  
Joseph Johnson

**Operations Manager**  
Grant Chang-Chien  
Tim Tirrell

## ASSOCIATE EDITORS

**Associate Content Editors**  
Ewen Chao  
Mithun Diwakar  
Jack Lee  
Kevin Peng

**Associate Online Editor**  
Zhizhang Xia

**Associate Operations Managers**  
Sanjeeb Bose  
Warner Leedy

**Associate Print Editor**  
Michelle Giron

## STAFF COLLABORATORS

Ramone Muñoz  
Graphics Advisor, Art Center College of Design

Carolyn Merkel  
Director, Student Faculty Programs

Gillian Pierce  
Science Writing Requirement Coordinator

## COVER IMAGE

*Restraint* Steven Neumann

## CURJ DONORS

Caltech Administration  
Typecraft Wood & Jones

CALTECH UNDERGRADUATE  
RESEARCH JOURNAL VOL. IV NO. 1



**ArtCenter**  
College of Design

Original CURJ Design by Aniko Hullner Grau  
©2004 CURJ. ALL RIGHTS RESERVED.  
THIS ISSUE NOT FOR RESALE



## COLLABORATIONS BETWEEN THE ARTS AND SCIENCES

BY RAMONE MUÑOZ

The Caltech Undergraduate Research Journal began in 2001 as a collaboration between students at the California Institute of Technology and the Art Center College of Design. I remember being contacted in 2000 by Lakshminarayan "Ram" Srinivasan, the first editor of CURJ, about getting help from Art Center's Graphic Design department in designing the editorial components of the new research journal. We were immediately excited about continuing the connections which had been ongoing between the two institutions.

Three years and seven issues later, we continue to collaborate on the journal, which has received praise from many respected scientific institutions. I have received requests from other universities, including Stanford, for help in the creation of their own journals, although to date, due to its geographic proximity to Art Center, Caltech has been the most practical and accessible partnership.

Designing a scientific journal has presented some interesting challenges for our graphic designers. Each issue is designed by one graphic student who works with a team of Caltech student editors. From the very beginning of this collaboration it was crucial that the designer understand the research addressed in the journal. Many discussions focus on how to support the articles graphically. This presents challenges to graphic designers whose talents reflect right brain thinking. It also challenges the left-brained student scientist who must communicate the complexities of research to the graphic designers so that they can best present information that is accurate as well as visually exciting.

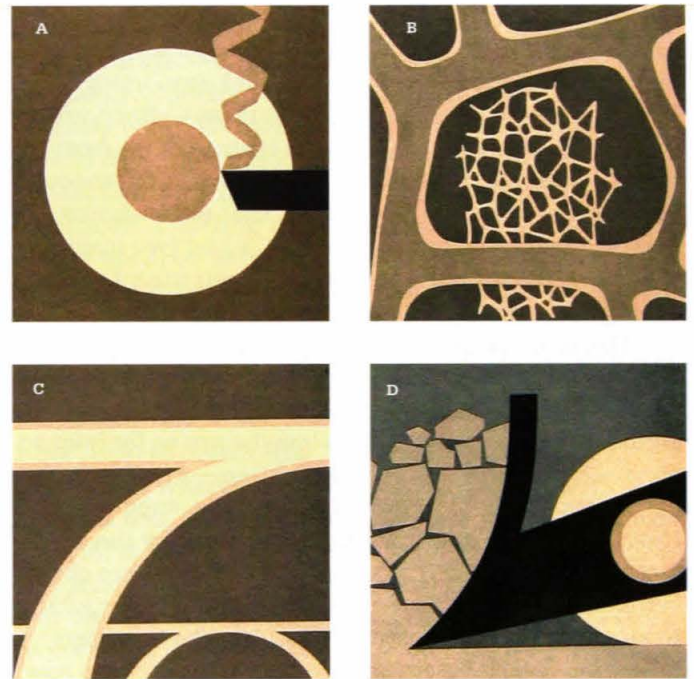
I can remember during my own education, in the early seventies, taking science courses which were conceptually interesting but supplemented with excerpts from Scientific American, and other journals which were deadly boring in appearance. My professors would place transparencies on their overhead projectors, visuals so poorly designed that their diagrams got in the way of their lectures rather than helping clarify ideas. Graphic design is still in short supply in the design of textbooks for K-12 education as well as college textbooks, which I'm convinced diminishes the pleasure of absorbing knowledge.

Graphic design students at Art Center take classes focused on the design of information and have benefited from the work of such notable designers as Edward Tufte and the information architect Richard Saul Wurman. The article in this issue of CURJ titled "How Many Colors Are Enough" is a great example of crossover research that interests Art Center as well as Caltech. One of the first graphic designers to impact the look of complex information was the German painter and graphic designer Anton Stankowski. He explored, in great depth, the subject of color theory and symbolism as it applies to visual communications. In the mid 1960s he published one of the most influential books on information graphics, *The Visual Presentation of Invisible Processes*. The book presented some of the finest examples of graphic design symbolically depicting processes that are not inherently visual. In the book's introduction, Stankowski notes that for graphic designers "it is the task of elucidating technical functions instead of illustrating objects, of refer-

**AT RIGHT** Abstract images by Anton Stankowski depicting (A) machining, (B) the structure of lubrication, (C) building and roads and (D) dredging.

ring to the common denominator of production rather than reproducing outward forms." Abstract graphic elements help readers understand the fields of science, medicine, aerospace, telecommunications and computer generated information. The principles addressed in this classic book are used in contemporary graphic design today and all of the issues of CURJ. Stankowski believed that graphic designers of the future would collaborate with technicians and scientists whenever complex communication projects required the formal design of signs and symbols to clarify ideas. Today, the application of aesthetic sensibilities can enhance all information beyond mere utilitarian functions of communication. Information can be inviting, exciting and even seductive.

CURJ is the journal I wish I had when I attended science classes in college. It represents the kind of collaborations which can generate better understanding and deeper appreciation for the similarities between the Arts and Sciences and how they can enhance each other. I congratulate Philip Wong, the Editor-in-Chief of CURJ, and his exceptional team of Caltech contributors and encourage other colleges and universities with important science departments to consider collaborations with designers. The two sensibilities, often considered a reflection of completely different mindsets, can in fact result in an amazing synergy and greatly enhance learning.



Ramone Muñoz has taught graphic design at the Art Center College of Design for twenty years and serves as the graphic design advisor to the designers of CURJ. He recommends the following books, all of which are considered very useful in the field of information design:

*Information Anxiety 2*, by Richard Saul Wurman  
*Information Architects*, by Richard Saul Wurman  
*Follow the Yellow Brick Road*, by Richard Saul Wurman  
*The Visual Display of Quantitative Information*, by Edward Tufte  
*Information Design*, edited by Robert Jacobson



## REVISITING ARISTOTLE

BY ED McCAFFERY

I find myself in my middle age teaching courses covering the interactions between law and economics and between law and technology to curious Caltech undergraduates and revisiting lessons learned from Aristotle, of all people. No, not the stuff about spontaneous generation: even lawyers are not that scientifically unaware. I realize rather that Aristotle had two insights that lie at the foundation of the modern trinity of law, economics, and technology.

The first insight is that private property works. Aristotle responded to his teacher, Plato, who had famously argued in the *Republic* that private property would corrupt the ruling class (maybe not so far-fetched an idea?), and that a society's children—its most valuable asset—should be raised in common, such that a true meritocracy might emerge, untainted by the effects of familial favoritism. Invoking his signature observational method, Aristotle had his doubts. "[T]hat which is common to the greatest number has the least care bestowed on it," he wrote in the *Politics*. Aristotle saw private property as central to the freedom and prosperity of a "stakeholder" society.

Aristotle was anticipating the now-classic argument of Ronald Coase, whose *The Problem of Social Cost*, published in 1960, ushered in the modern law & economics movement and ultimately gained its author the Nobel Prize in Economic Science. Coase focused on the problem of the possible divergence between private and social costs—what economists call "externalities." I teach my students that externalities are simply unpriced effects, which can be negative or positive. A negative externality (such as pollution or second-hand smoke) occurs when a harm imposed by the production or consumption of a good is not factored into its price. If we do not make people pay for the harms they impose, they will impose too many: a tragedy of the commons will result (simply observe any common eating area on any college campus). Economists before

Coase (most famously A.C. Pigou) had assumed that the way to deal with a negative externality was to impose a tax on the good or activity, set and collected by government. The genius of Coase was to see that no such intervention was needed. Strong private property rights and freedom of contract lead private parties to solve their own externalities. Aristotle redux. An era of deregulation lay before us.

Coase's focus was on negative externalities. But positive externalities form a mirror image. Here the problem is that rational persons collectively will not produce enough of a good thing if they are not paid adequately. Once again, there is a problem of the commons, but here the problem is one of the underdevelopment of a commons—the cost is an opportunity cost, one of the potential benefits foregone. The most important commons we have is the realm of ideas, which we are all free to share once they are discovered and explained. But who—saints and the occasional professor aside—will forage in the treacherous domain of intellectual discovery, without some promise of adequate recompense? Once again private property is an answer. If ideas are indeed held in common, the "least care" will be spent on producing them. Hence the intellectual property system, with its core components, the law of patents and copyright, can bring greater good for the whole through strong private property rights to some.

But there is more to intellectual property than this happy tale of private property works. Here I have been reflecting on Aristotle's second insight (still in the middle of my middle age): the celebrated golden mean. "Excellence is a mean between two vices," as he put it in the *Nichomachean Ethics*. What we too easily forget is that the intellectual property system, especially patents, solves one market failure, that of positive externalities, at the expense of another market failure, that of monopolistic power. There are two vices, not one, and the "most excellent" public policy answer lies somewhere in the middle of these two vices.

Herein lies a considerable difference between the domains of intellectual property and more traditional, normal property. A strong property regime, freedom of contract, and minimal government can achieve welfare



maximization in the usual cases. The same cannot be said for intellectual property, because of the conjunction of market failures. An intellectual property regime thus requires a strong government presence—not only to enforce the rights on behalf of their holders (as we are doing vis-à-vis digital rights for content providers, as in the Digital Millennium Copyright Act), but also to police these rights-holders themselves, who are constantly tempted to expand the domain of their rights and hence their power, giving us a market failure (a monopoly) without, at the margin, the corresponding benefits of innovation.

Examples of overreaching abound. Broadly, they occur in the two dimensions of time and (intellectual) space. Consider them in turn.

First, time. The definition of property in Western societies is simple and absolute: property owners own their goods infinitely. This is efficient, as Harold Demsetz persuasively argued decades ago, because it solves a problem of inter-temporal waste—the tendency of a present owner to cheat the future, as by not replenishing the soil (an externality, again). By conferring infinite terms, an owner will maximize value over and through time, even if this means selling his goods or land to someone better able to do so. But intellectual property, because of its monopoly power, is of limited time (as the U.S. Constitution, in Art. I, Section 8, actually requires of patents and copyrights). Yet owners of existing patents and copyrights frequently seek to extend their term. They can do this by lobbying Congress, as Disney effectively did in helping to bring about the Copyright Term Extension Act, giving Mickey Mouse a longer term as its indentured servant. The constitutionality of the Act was upheld by the U.S. Supreme Court, notwithstanding a persuasive argument that extending existing copyrights cannot plausibly incentive new ones. In parallel fashion, many pharmaceutical companies have engaged in questionable business practices to attempt to extend the term of their patents, as by filing “new” patents on slightly altered old patents, seeking to get another twenty-year bite at the monopoly apple (Bristol-Meyers Squibb was held liable for large sums for doing this two years ago with an anxiety treatment drug).

Second, space. Traditional property, both real and personal, is well-defined by its metes and bounds, the occasional border dispute aside. But where does one idea end and another begin? Many companies attempt to use items in their intellectual property portfolio as swords to deter any possible innovation in related but distinct areas—innovation that might bring with it competitive pricing. A variant on this theme occurred recently when Warner-Lambert settled criminal charges and civil liabilities in the amount of \$430 million in connection with the “off-label” promotion of the drug Neurontin. The drug had been approved by the FDA for use by epilepsy patients, but Parke-Davis (a division of Warner-Lambert) engaged in a range of business practices to extend the domain of their drug into other, unrelated and unproven uses. Media reports suggest the practice is widespread. The basic syndrome is the attempt to extend the monopoly in one “space” (the “on-label” use) to others (the “off-label” ones).

Here is where I get back to the golden mean. America has, and needs, a strong intellectual property system, to protect and give incentives to its many artists, innovators and scientists. But we also need—and less often have—a strong government and regulatory presence (in almost all government agencies, such as the PTO, FTC, FCC, and FDA), backed by a vigilant court system (to watch the regulators!), and to monitor and limit the monopolistic players from going too far in either space or in time.

Unfortunately, little of the law of intellectual property cashes itself out into clear, bright-line rules (for example, what is and is not patentable, or what related but non-identical ideas do or do not infringe). We must live with less certain standards, and we must depend on human judgment to monitor the extremes. Living in the middle is not always the analytically clearest thing to do. But, as Aristotle counseled, it may just be the “most excellent” course of action.

*Ed McCaffery is the Robert C. Packard Trustee Chair in Law and Political Science at USC Law School and Visiting Professor of Law & Economics at Caltech. He also directs the Caltech Law & Technology program. The author of several books, he is working on another, A New Understanding of Property (U. Chicago Press).*



# HOW MANY COLORS ARE ENOUGH?

BY ANDREW YANG

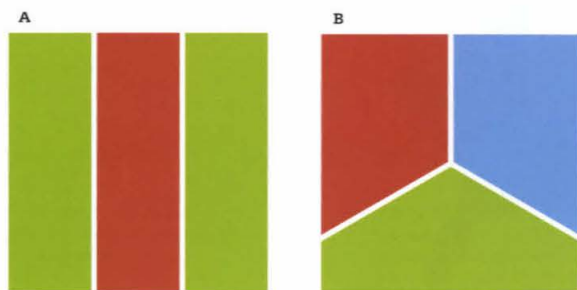
A YOUNG CHILD SITS DOWN TO A COLORING book and finds an empty map of Europe. She picks out her three favorite crayons and colors Germany, Belgium, and Luxembourg. She goes to color France, but realizes that using any of the colors that she has in her hand would make it impossible to tell one country from another [FIGURE 1]. A natural question to ask is how many different colors she needs to color the whole map. By coloring carefully, it turns out that four colors are enough. Determining whether this fact holds for maps in general, however, proves to be a difficult problem. Indeed, it took mathematicians more than a century to finally conclude that it was true, but the eventual solution itself created new questions about the nature of mathematical inquiry.

The Four Color Theorem states that any two-dimensional map can be colored with four colors so that any two adjacent regions are different colors. Unlike other mathematical problems, the Four Color Theorem has a statement so simple that non-mathematicians easily understand it. Much like Fermat's famous Last Theorem, many amateurs have attempted, but failed, to prove the theorem. Its accessibility to a wide audience has only added to its reputation as an interesting and worthy mathematical problem.



**FIGURE 1** After Germany, Belgium, and Luxembourg are colored, France cannot be colored red, green, or blue without being the same color as a neighboring nation.

**FIGURE 2** Maps for which two or three colors suffice, (A) and (B) respectively



**FIGURE 3** Map for which four colors are needed

Centuries ago, cartographers must have recognized that four colors were enough to color most maps, but attempts to prove it did not take root in the mathematical community until the 1850s. Early attempts to solve the problem failed, although partial progress was made toward a resolution of the problem. It was not until 1976 that Wolfgang Haken and Kenneth Appel proved the theorem using considerably more advanced techniques and equipment than their predecessors. Their proof involved computer-assisted calculations, a highly unusual and controversial technique at the time. Despite its simple formulation, the problem is complex enough to be an important part of 20th century mathematics.

#### HOW SMALL A PALETTE?

Like any mathematical problem, the Four Color Theorem should be examined closely to see exactly what it claims and what it does not claim. The theorem does not state that four colors are always necessary; in special cases three, or even two, colors may suffice. For example, the maps in **FIGURE 2A** and **FIGURE 2B** show instances where two and three colors, respectively, suffice.

In these special cases, four colors are not necessary, but to show that four colors are sometimes necessary, consider the map in **FIGURE 3**. This map is like the map in Figure 2b, except that a circular region has been placed in the center. The key feature of this

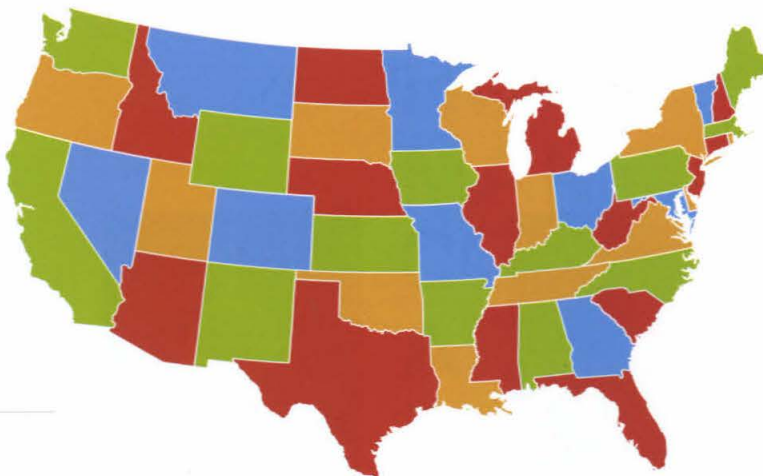
map is the fact that every region is adjacent to every other region on the map. If two regions are the same color, there will be adjacent regions that are the same color. Thus, to satisfy the condition that adjacent regions are different colors, each region must be colored differently—that is, four colors are necessary to color the map, as in Figure 1.

A map is called  $n$ -colorable if it can be colored using at most  $n$  colors in a way such that any two adjacent regions are different colors. Thus, the map in Figure 2A is 2-colorable, 3-colorable, 4-colorable, and is  $n$ -colorable for any number  $n$  greater than or equal to 2. The map in Figure 2B is 3-colorable, but not 2-colorable. The map in Figure 3 is 4-colorable, but not 3-colorable.

Another way of stating the Four Color Theorem is that every map is 4-colorable. There are an infinite number of distinct maps, so it is impossible to check this statement by examining every map and verifying that they are each 4-colorable. Nevertheless, there seems to exist some common property of maps that ensures that, regardless of how they are drawn, they are 4-colorable. One can verify this, for instance, in the case of maps that are as complicated as that of the United States [**FIGURE 4**]. While there are several mathematical methods of handling the proof of a statement about an infinite number of objects, the final proof, in some sense, did reduce the number of objects considered to a finite number, as we shall see later.



**FIGURE 4** A map of the United States with only four colors



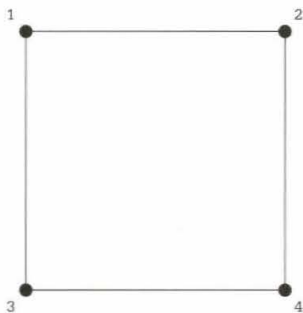
#### WHAT EXACTLY IS A MAP?

Words are often tossed about in everyday conversation with imprecise or ill-defined meanings. For example, when we say map, what do we really mean? And what does adjacent mean? The Four Color Theorem can be translated into more precise terms using the language of a branch of mathematics known as graph theory.

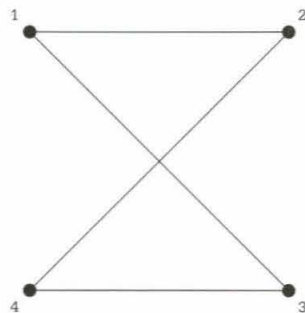
Graph theory deals with objects called graphs, which consist of a set of vertices and edges. We consider only finite graphs, that is, graphs with a limited number of both vertices and edges. The vertices of a graph can be any objects, but they are usually depicted as points in a plane. Edges are line segments connecting two vertices. For example, the graph consisting of the vertices 1, 2, 3, 4, and edges between 1 and 2, 1 and 3, 2 and 4, and 3 and 4 can be represented pictorially, as in **FIGURE 5**.

We focus on graphs where a pair of vertices can be joined by at most one edge, and more specifically, on a class of graphs known as planar graphs. These graphs can be drawn in the plane in a way where no two edges intersect except at a vertex. The graph in Figure 5 is planar, but this does not mean that every representation of this graph will have no intersecting edges, as **FIGURE 6** shows. One can check that the graph in Figure 6 represents the same graph defined above, but the edge connecting 1 and 3 intersects the edge connecting 2 and 4. Nevertheless, Figure 5 demonstrates that there is indeed an arrangement such that the edges do not intersect anymore.

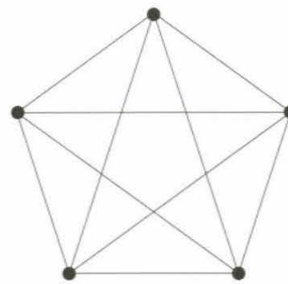
When we say that a graph is planar, we only require that there exists one representation of the graph where no edges intersect. If someone says that a graph is planar, and presents a pictorial representation of the graph supporting his claim, the claim can be verified by just examining the picture and checking that no edges intersect; for example, Figure 5 shows that the graph is planar. However, it is more difficult to verify that a graph is not planar, because this involves looking at a possibly infinite number of different representations and then checking if edges intersect or not. To demonstrate that not all graphs are planar, consider the graph in **FIGURE 7**. No matter how you rearrange the lines or the points, you can't keep one edge from crossing another—try it!



**FIGURE 5** An example of a planar graph with no intersecting edges



**FIGURE 6** We can redraw a planar graph so that its edges cross, but that does not mean that the graph is non-planar.



**FIGURE 7** A non-planar graph

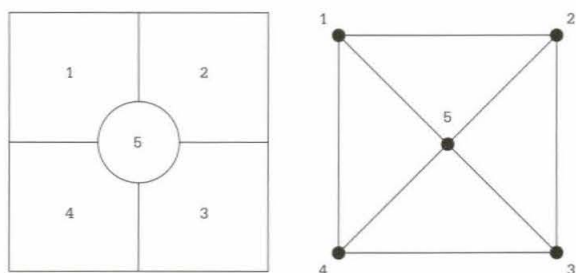
A map of regions can be converted into a graph by associating a vertex to each region, and connecting an edge between two vertices if and only if the corresponding regions are adjacent in the original map. This procedure is illustrated in **FIGURE 8**, where this conversion is performed on a fairly simple map.

Did you notice that the graph obtained via this conversion is planar? This is always true when this conversion is performed. Suppose the vertices can be assigned colors so that any edge has endpoints of different colors. If this coloring is used to color the corresponding map, the property of this vertex coloring ensures that any two adjacent regions in the original map have different colors. Similarly, if a map is colored in a way so that no two adjacent regions share the same color, the vertices on the induced graph can be assigned colors in a way so that any edge connects vertices of different colors [**FIGURE 9**].

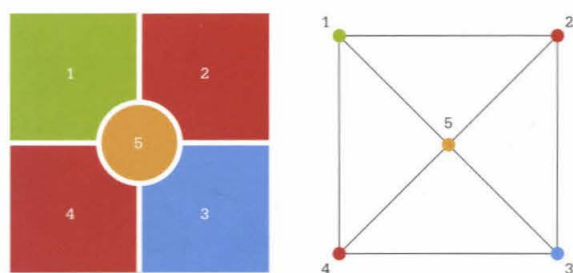
We should mention here that the proof discussed here does not apply to every arbitrary map in the real world. For example, consider the case of Poland. Because of complicated European history, the Soviet Union took control of the initially German city of Königsberg, which they renamed Kaliningrad and turned into a major naval base. When the Soviet Union broke up, however, Russia kept control of the port city, so Kaliningrad still borders Poland. Thus, the graph of the countries around Poland is not planar [**FIGURE 10**]. Consequently, real cartographers sometimes need more than four colors to indicate non-contiguous regions.

Because proving the Four Color Theorem is equivalent to proving that any planar graph is 4-colorable and because mathematicians find it easier to work with graphs than maps, they approached the issue from a graph-theoretic perspective. Unfortunately, this problem, when stated in graph-theoretic terms, is just as difficult as the original problem. The advantage of translating the problem into the language of graph theory lies in the fact that generally known facts about graphs (which appear in a wide range of problems, not just map coloring) can be applied to help prove the Four Color Theorem.





**FIGURE 8** We convert a map to a planar graph by making each region a vertex and drawing an edge between two vertices only if the corresponding regions are adjacent. Turning a map into a graph preserves the coloring.



**FIGURE 9** We can still assign colors to a map after conversion to a graph.

#### AFTER EARLY FAILURES, COMPUTERS GIVE IT A TRY

The Four Color Theorem made its appearance in the mathematical community in 1852 when an undergraduate student in London, Francis Guthrie, asked his older brother Frederick in a letter whether it was possible to color any map using only four colors so that adjacent regions would be different colors. His brother did not know the answer, so Francis asked a mathematics professor, Augustus De Morgan, whether he knew the answer to this problem. De Morgan did not, and he began to circulate the problem amongst various colleagues in the mathematical community. Gradually, mathematicians realized this was a problem worthy of study because no one seemed to be able to find an answer.

In 1879, the mathematician Alfred Kempe announced that he had proved the theorem. Although this proof was shown to be wrong in 1890, the basic ideas in Kempe's proof would be used by future generations to solve the problem. Kempe approached the proof via the method of proof by contradiction. In this method of proof, the theorem is assumed to be false, and then a contradiction with some already known fact of mathematics is found. The only way out of this dilemma is to say that the original assumption that the theorem is false is incorrect, so the theorem must be true.

Kempe reduced the problem to considering what are known as normal maps—maps where no region is completely contained by another region and no more than three regions meet at any point. Kempe correctly proved that if there were a graph that disproved the Four Color Theorem, that is, a map that required five colors for its coloring, then there would also be a

## “Real cartographers sometimes need more than four colors”

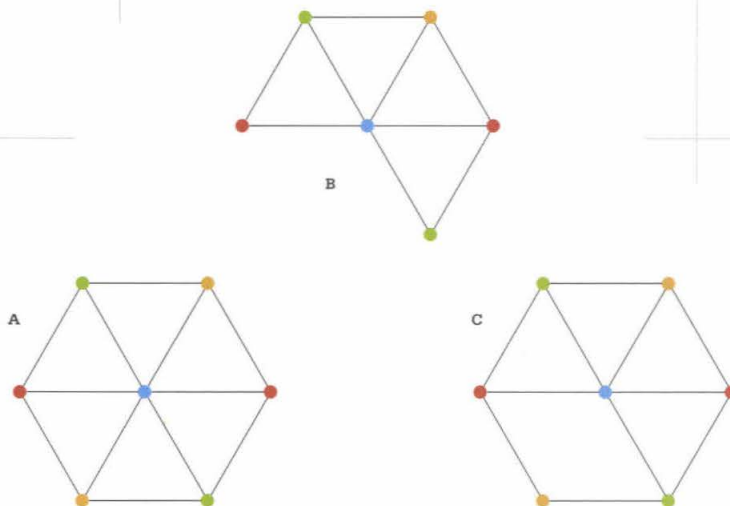
normal counterexample, a normal map that is also a counterexample to the Four Color Theorem. Normal counterexamples are better behaved than general counterexamples, so from this point on we will always mean normal counterexample when we write counterexample. So, if the theorem were false, there would be at least one counterexample to the Four Color Theorem. There must be some counterexample with the fewest number of regions.

Kempe proved that any normal map contained some region with fewer than six adjacent regions. He then went on to show (this time incorrectly) that from a minimal counterexample containing some region that had fewer than six neighbors (which every normal map had), one could construct a smaller counterexample, which contradicts the fact that the minimal counterexample was the smallest possible one. If this last argument were correct, the problem would have been resolved, but his proof that the existence of a region with five or fewer neighbors in a counterexample would lead to a smaller counterexample was wrong.

Despite the flaws in this “proof,” we can already see the key ideas that underlie the correct proof. We want to find a set of subgraphs (a subset of the vertices and all the edges joining them [FIGURE 11])



**FIGURE 10** An example of a non-planar graph in the real world; because there is a ring of areas surrounding Poland, there is no way to show that Russia borders Poland without crossing an edge.



**FIGURE 11** (B) is a subgraph of (A), but (C) is not a subgraph of (A) because it is missing an edge.

where at least one configuration from that set is present in every normal map. Such a set is called unavoidable, since some configuration from that set must be found in every normal map. Once we have one counterexample in hand, we want to show that it leads to a smaller counterexample. If we can, then the configuration is called reducible. To prove the theorem, it suffices to find an unavoidable set of reducible configurations. For any proposed minimal counterexample, we can look at the unavoidable configuration, reduce it, and then have a new "minimal" configuration. That means, however, that the proposed minimal solution wasn't actually the true minimal configuration, thus leading to a contradiction.

Kempe proved that the set of configurations consisting of a region with two, three, four, or five adjacent regions is unavoidable. Where he failed was in showing that each of these was reducible. Thus, the difficulty of the proof seemed to lie in finding an unavoidable set where each configuration was reducible. Although mathematicians were able to prove the theorem for relatively small maps (those with fewer than about 35 regions) using generalizations of these techniques by 1950, no one really knew how to approach the general problem.



“It involved both an incredible amount of  
raw computation and sophisticated  
mathematical analyses of the output  
to arrive at the final unavoidable set.”

The crucial development came in the late 1960s, when the German mathematician Heinrich Heesch introduced the method of discharging. Discharging is the assignment of a number called a “charge” to each vertex in a configuration, either positive, zero, or negative, and then moving those charges around. You can take some charge from one vertex and then spread it around to neighboring vertices in such a way that the total charge is preserved.

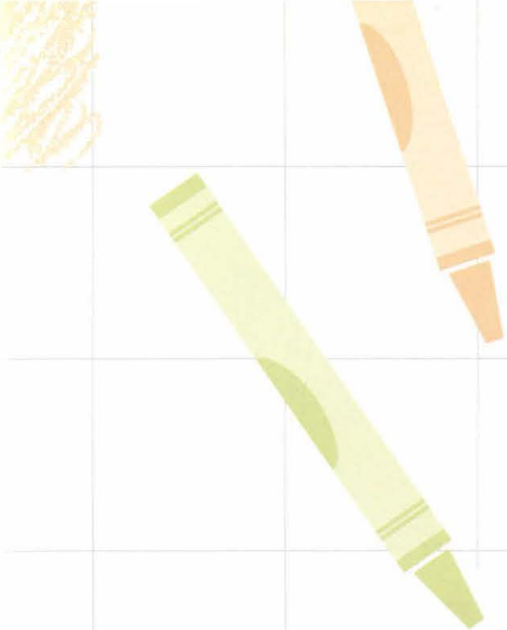
Mathematicians proved that, given some discharging procedure, some set of unavoidable configurations could be constructed. Discharging was an important invention because it gave mathematicians a way to construct many different unavoidable sets. However, these configurations were not necessarily all reducible, and an unavoidable set of reducible configurations was needed to prove the Four Color Theorem. Mathematicians hoped that it was possible to create sets of configurations and then use modified discharging methods to create more and more configurations that would be reducible. But Heesch was not content with only pencil and paper methods; he was also developing computer methods to rapidly determine if a configuration was reducible.

In the early 1970s, the mathematicians Wolfgang Haken and Kenneth Appel began to study Heesch’s work in more depth and realized that there seemed to be certain easily identifiable characteristics of configurations that made them reducible. By using computers, they looked for unavoidable sets consisting of these types of configurations. They constantly had to update the program with each output to change the discharg-

ing procedure to yield more and more reducible configurations. They also needed to use Heesch’s computer methods to test whether the configurations that were produced by their discharging methods were reducible. By 1976, after proceeding in this fashion for several years, Appel and Haken succeeded in finding an unavoidable set of approximately fifteen hundred reducible configurations. This discovery gave a proof of the theorem, since the existence of such a set meant that there were no minimal counterexamples, and thus, no counterexample whatsoever. The final proof was a tour de force of mathematics and computer science, as it involved both an incredible amount of raw computation and sophisticated mathematical analyses of the output to arrive at the final unavoidable set. After more than 125 years, the theorem had finally been proven.

#### A COMPUTER WITH CRAYONS

Despite the successful resolution of a mathematical problem, the use of a computer in the proof of the Four Color Theorem disturbed a significant number of mathematicians. Before this proof, every resolved mathematical problem of note had been solved through a series of logical arguments, each of which could be checked by hand. Although mathematical proofs have become incredibly complicated, it is still usually possible for a trained person (normally a mathematician) to read through a proof, check the validity of each statement, and thus verify the truth of the proof. Indeed, this is what mathematicians spend most of their time doing, and the process of




checking proofs is a key component of the review process used by mathematical journals.

A person, however, cannot check the calculations of a computer because it is impossible to know exactly what a computer is doing at every step of a calculation. It is possible to verify that Appel and Haken's program, when used by an infallible computer, produces a valid proof, but infallible computers do not exist. Due to errors in the hardware of a computer and idiosyncrasies of the software on a computer, it is impossible to know with absolute certainty that a computer is performing calculations correctly. Errors can be introduced in the calculations that lead to an output of "true," despite the fact that a statement is "false."

Despite this objection, the most commonly accepted view of the proof now is that, because it can be verified many times on a wide array of computers that use different hardware and software, the proof is correct. It is believed to be highly unlikely that all these computers would share some common flaw that would give an incorrect result in the computer calculations of the proof. Also, simpler proofs have been published since 1976, that have a smaller unavoidable set to check (near 650, as opposed to 1500), thereby reducing the possibilities for computer error. Currently, it is possible for any person to download the proof and the software used for the proof and verify the theorem for himself independently on his home computer, and the fact that no one has successfully been able to dispute the validity of the proof in the past twenty-five years has led to near

universal acceptance of the proof of the Four Color Theorem among the mathematical community.

Although the Four Color Theorem is a problem encountered by three-year-olds and the problem can be understood by practically anyone, it took more than a century for mathematicians to discover the solution. Even more surprising was that this solution required tools that were available only in the past few decades. Although there are no practical consequences of this proof, its creation is one of the great mathematical achievements of the 20th century because it ushered in a wide array of mathematical and programming techniques. 

*Andrew Yang is currently a fourth year undergraduate at the California Institute of Technology majoring in Mathematics. Next year he plans on entering the Ph.D. program at Princeton University to study number theory. This paper was edited with the assistance of Gillian Pierce.*

#### FURTHER READING

- 1 K. Appel, W. Haken. The Solution of the Four-Color-Map Problem. *Scientific American* 237, (1977).
- 2 J. Casti. *Mathematical Mountaintops: The Five Most Famous Problems of All Time* (Oxford University Press, New York, 2001).





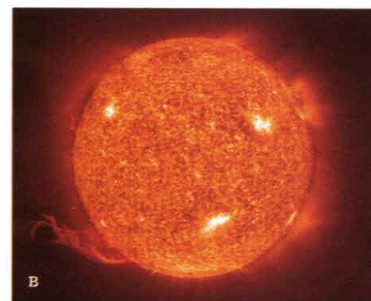
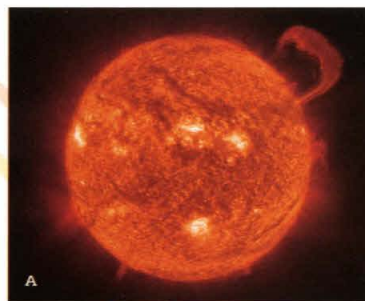
# THE ELECTROMAGNETIC THEORY BEHIND SOLAR PROMINENCES



BY ERIC LIN

ON THE MORNING OF JANUARY 11, 1997, television stations across the United States stopped transmission as a cloud of particles hit and disabled a communications satellite in orbit. These particles came from a solar prominence, a beautiful display that periodically bursts from the Sun [FIGURE 1]. These huge arch-shaped structures, also known as solar protuberances or solar filaments, are hundreds of thousands of kilometers in length and can be seen from the Earth during solar eclipses. The prominences may last for weeks or even months before finally becoming unstable and erupting, ejecting large masses of charged particles that sometimes disrupt orbiting spacecraft or terrestrial power grids. Hence, astronomers study these solar formations to help prepare for these outbursts and to learn how to counteract the damage they cause.

Descriptions of these phenomena appeared throughout the last millennium, though it was not until the 1860s that the proper tools allowed astronomers to deduce that solar filaments consisted of masses of glowing gas. Since then, astronomers have developed better instruments for studying them, but in spite of technological advancements, astronomers continue to puzzle over the precise nature and causes of these common solar occurrences.



**FIGURE 1** (A) A photograph of a solar prominence, the arch-shaped structure on the Sun's upper right. The prominence is about 30 times the size of the Earth. (B) Another photograph of a solar filament.  
Source: SOHO (ESA & NASA).

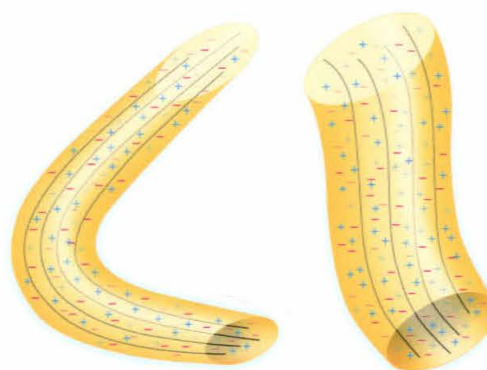
#### A HISTORY OF PROMINENCE OBSERVATIONS

Ever since observations were first recorded in the early thirteenth century, the origin and exact mechanisms of solar protuberances have been shrouded in mystery. The Swedish astronomer Bilger Wassenius observed three or four of these solar occurrences during an eclipse in 1733 and described them as "red flames," erroneously believing they were clouds in the lunar atmosphere.

In the mid-1800s, the invention of the spectrograph, a device that tells us the wavelengths of light emitted by a source, allowed astronomers to conclude that protuberances were masses of glowing gas. Because different gases emit light at unique wavelengths, studying the wavelengths at which these peaks occur gave astronomers information about the composition of solar prominences. Since then, astronomers have developed additional instruments, such as magnetographs, which analyze the polarization of light and in doing so estimate the strength of the prominences' magnetic fields.

Scientists now know that solar protuberances are composed of plasma, a mixture of ionized gases that forms when electrons are stripped from neutral atoms at high temperatures, leaving behind positive ions. By combining observed data with knowledge about the properties of plasma, physicists have been able to develop models that explain the behavior of solar filaments.





**FIGURE 3** Diagram of plasma containing gas particles with magnetic field lines passing through it. The particles include electrons (shown as - signs in the figure) and positive ions (shown as + signs in the figure). The magnetic field lines (the lines that curve through the plasma) are frozen in place in the plasma. Source: Bellan, 2000.

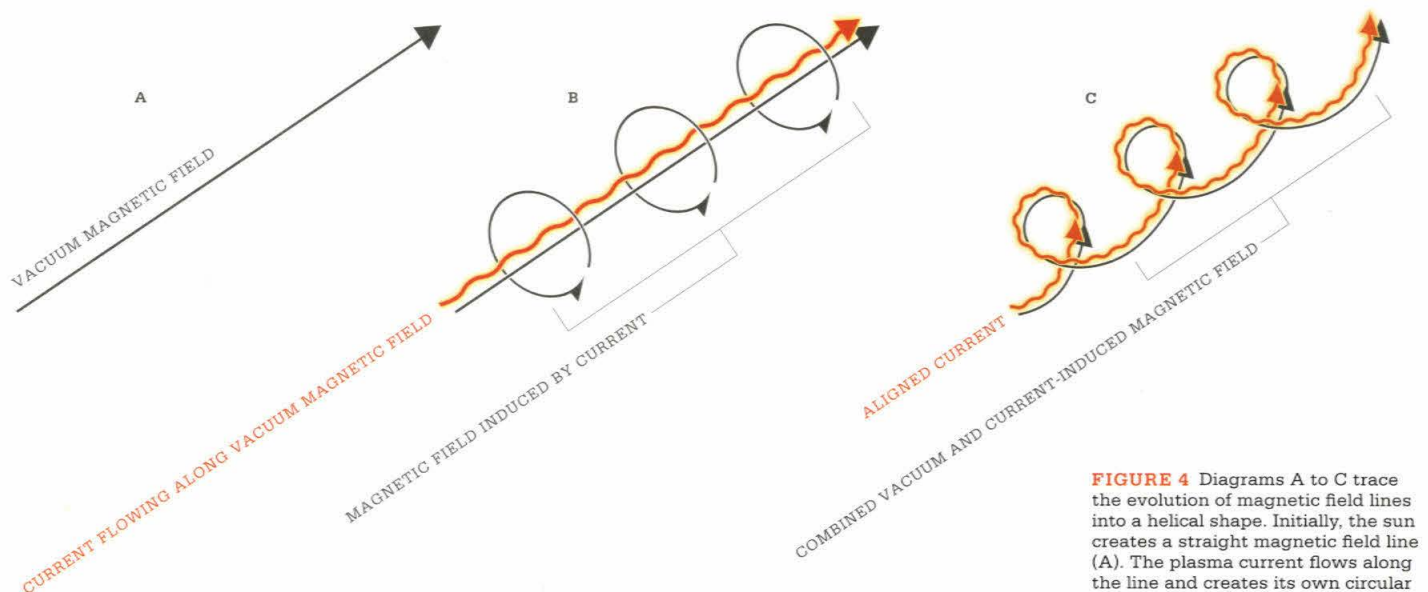
#### ELECTROMAGNETISM DESCRIBES THE STRUCTURE OF PROMINENCES

Though today's increasingly detailed studies of prominences use complex mathematical equations to describe aspects of plasma physics, the main characteristics of the solar formations can be explained by the fundamentals of electromagnetism.

The enlarged photo of a solar prominence in **FIGURE 2** shows its large, arched shape, a feature caused by the sun's arched magnetic field. The plasma of the prominence is constrained to match the shape of the magnetic field. Plasma is an excellent conductor of electricity because the electrons and ions in it are able to move easily through the gas. In a prominence, however, the path of the plasma is affected by two laws. Faraday's law of induction states that any time the plasma moves across a magnetic field line in an electrical conductor, an electric field will be created within the plasma. Additionally, Gauss's law states that charged particles in a conductor, such as plasma, will always rearrange themselves to cancel out internal electric fields. The only way to satisfy both requirements is for the magnetic field lines to be frozen inside the plasma (**FIGURE 3**), much as strands of string are caught in a thick gel. This means that the arched shape of the sun's magnetic field lines fixes the plasma to be arch-shaped as well.



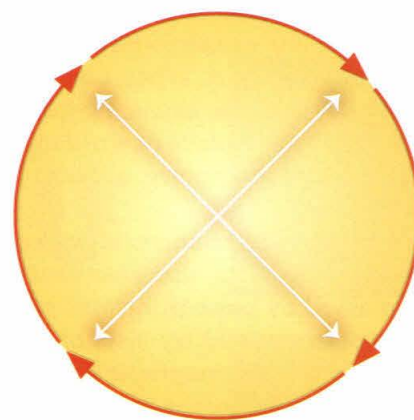
**FIGURE 2** A close-up photo of a prominence on the surface of the sun. The prominence has an overall arched shape and contains helical loops. Source: Big Bear Solar Observatory, New Jersey Institute of Technology.



**FIGURE 4** Diagrams A to C trace the evolution of magnetic field lines into a helical shape. Initially, the sun creates a straight magnetic field line (A). The plasma current flows along the line and creates its own circular magnetic field around the current (B). The original field from the sun can combine with the local magnetic field created by the plasma current to create a helical shape for the overall magnetic field lines (C). Source: Bellan, 2000.

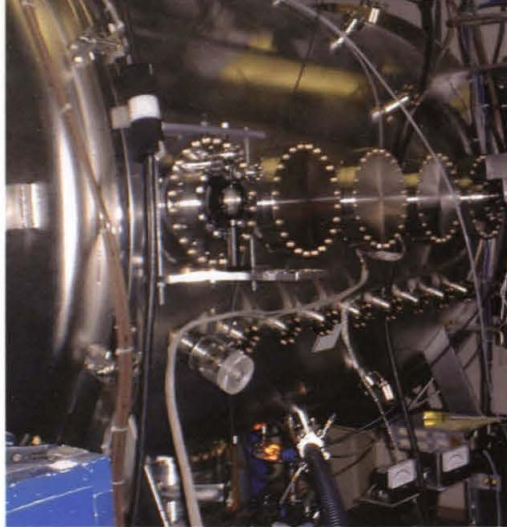
Another aspect of the shape of these protuberances is the formation of small loops along the main arches of the structure. The key to seeing why these loops arise is to examine the effect of the plasma current on the magnetic fields. As a result of the Biot-Savart law, circular magnetic fields are generated around a line of flowing current [FIGURE 4]. These newly-formed circles of magnetic field add to the original stable magnetic field lines of the sun. The two sources of magnetic field lines combine to make a new, stable helical field with loops. The plasma current follows this new shape, giving solar prominences their looping helical structure.

The loops in the prominence sometimes bulge outwards until the whole structure violently erupts. This bulging occurs because antiparallel currents repel each other with a strength proportional to the amount of current. In this case, when the plasma currents flow in a hoop shape, the currents on opposite sides of the hoop flow in opposite directions. This means that opposite sides of the hoop repel each other, causing the hoop of plasma to bulge out [FIGURE 5]. The stronger the plasma current becomes, the more the hoop bulges. If the plasma currents are strong enough, this bulging hoop overcomes the elastic tendency of the prominence's magnetic fields to return to their stable states. As a result, the filament violently erupts and plasma particles spew into space in all directions.



**FIGURE 5** Diagram of a hoop of current, illustrating the forces that cause the hoop to bulge out. On opposite segments of the hoop, the current flows in opposite directions. Antiparallel currents repel, so there will be a repulsive force between any opposite segments. Because of this, the hoop will bulge outwards.





“Particles in  
a solar storm  
caused a  
power outage  
in Quebec.”

**FIGURE 6** A photograph of a vacuum chamber as viewed from the side. This chamber is used to simulate prominences. Source: Bellan, 2000.

#### THE IMPACT OF PROMINENCE ERUPTIONS ON EARTH

When protuberances erupt and the magnetized plasma is discharged, the plasma cloud sometimes drifts towards Earth. Fortunately, the average density of a plasma cloud is only about a hundred particles per cubic centimeter (by comparison, the air on the surface of Earth is more than a trillion times as dense), so these particle clouds cause no physical harm to our planet when they strike it. When they hit the magnetosphere, they sometimes cause a geomagnetic storm, resulting in brilliant displays of auroras; however, they can also cause temporary, noticeable damage to electronic devices.

While the exact mechanism is not completely understood by scientists, this electrical damage can also be explained by basic principles of electromagnetism. The plasma cloud consists of charged particles with an associated magnetic field. The change in magnetic field caused by the presence of the plasma cloud can induce currents in a conductor. In this way, plasma particles, carrying their associated magnetic fields, can induce extra currents to flow within terrestrial power grids in their path and thus temporarily disrupt power on Earth. A memorable example

occurred on March 13, 1989, when particles in a solar storm caused a power outage in Quebec. The plasma particles can also strike spacecraft, damaging their electronic components and creating spurious signals and data.

The fact that these filaments on the surface of the Sun occur at such great distances from Earth can limit the ability of astronomers to observe and understand their structures and behavior. Historically, astronomers have used equipment on Earth to record data from signals and waves coming from protuberances. Because these measurements are often sparse and because the theories in this area are imperfect, these solar phenomena are still poorly understood.

#### PROMINENCES IN THE LABORATORY

Rather than simply building better instruments to study real solar protuberances, Paul Bellan at Caltech is approaching the problem by simulating protuberances in the laboratory so that scientists can take a close-up look at the processes that control the behavior of protuberances. By simulating solar prominences, scientists can obtain more thorough data about their behavior by controlling different variables through the



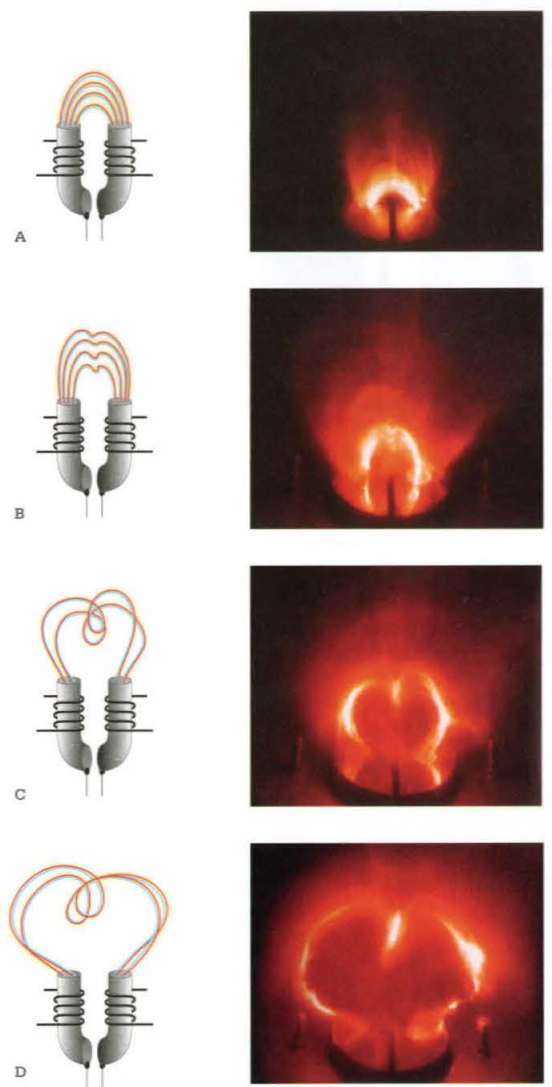
**FIGURE 7** Schematic and photograph of the evolution of a laboratory simulation of a prominence. The current of the moving plasma particles creates the glowing arch-shaped image that can be seen below. In diagram (A), the plasma initially follows the U-shaped field created by the bar magnet. As the current increases from (A) to (D), additional loops can be seen to form. The loops also bulge out as the current increases. These laboratory simulations display the overall arched shape, the loops, and the bulging curved shapes typical of solar prominences. Source: Bellan, 2000.

course of the experiment. These studies lead to a better understanding of the physics behind the behavior of protuberances.

The basic structure used in a simulation apparatus consists of an arrangement of plasma and magnetic fields in a vacuum chamber [FIGURE 6]. The magnetic fields are used to trap plasma in a technique that uses the same laws of electromagnetism as those that govern the trapping of plasma by magnetic fields in solar protuberances.

The arrangement of the electromagnet in this apparatus creates arch-shaped magnetic fields that are similar in shape to those found in the Sun. The diagrams in FIGURE 7 show the typical observed behavior inside the chamber as the current is increased. The additional loops and extra bulging that form in this simulation as the current increases are also observed in solar filaments.

These experiments show that miniaturized versions of some of the most beautiful aspects of solar prominences can be recreated in the laboratory. However, scientists' understanding of these protuberances is still incomplete, and a better understanding of them will require the development of a more complete theory.





“Plasma in the spheromak structure generates self-contained magnetic fields and **spontaneously organizes itself** into the desired configuration.”




**FIGURE 8** Picture of the Sustained Spheromak Physics Experiment (SSPX) apparatus. Plasma and magnetic fields are created in a toroidal arrangement within this structure. Source: David Hill, Lawrence Livermore National Laboratories.

## MAGNETIC CONFINEMENT FUSION HOLDS POTENTIAL FOR A NEW ENERGY SOURCE

The same technology and concepts used to study solar prominences have also been widely considered for use in nuclear-fusion-driven energy-production devices. One application of the continued study of solar filaments and their simulations is the construction of fusion-harnessing structures. There are many similarities between the technology used in magnetic field fusion and that used to simulate solar filaments. Physicists are attempting to use magnetic fields to confine plasma for nuclear fusion by the same process in which plasma is confined by magnetic fields in solar filaments. Also, the same mathematical equations appear in descriptions of both systems, so an improved understanding of how solar prominences behave may improve the ability to control magnetic fields for fusion.

Structures have been proposed in which magnetic fields could be created to hold plasma. In one type of structure, the plasma is arranged in a doughnut-like shape called a spheromak. These shapes are excellent candidates for fusion because plasma in the spheromak structure generates self-contained magnetic fields and spontaneously organizes itself into the desired configuration. This is similar to what occurs with solar prominences, in which the magnetic field lines tend to return spontaneously to their stable shapes. Other currently used structures do not have this advantage and require that the confining magnetic fields be externally generated, which can be costly.

In January of 1999, a structure based on spheromak theory was built at the Lawrence Livermore National Laboratories. The research effort, called the Sustained Spheromak Physics Experiment, consists of a series of experiments designed to determine the spheromak's potential in efficiently containing hot plasmas of fusion fuel [FIGURE 8]. Hopefully, prominence research, along with studies of the spheromak structure, will make nuclear fusion a viable energy source in the next few decades. 

*Eric Lin is a fourth year undergraduate in Physics at the California Institute of Technology. He would like to thank his mentor Paul Bellan, Professor of Applied Physics at the California Institute of Technology, for his helpful direction.*

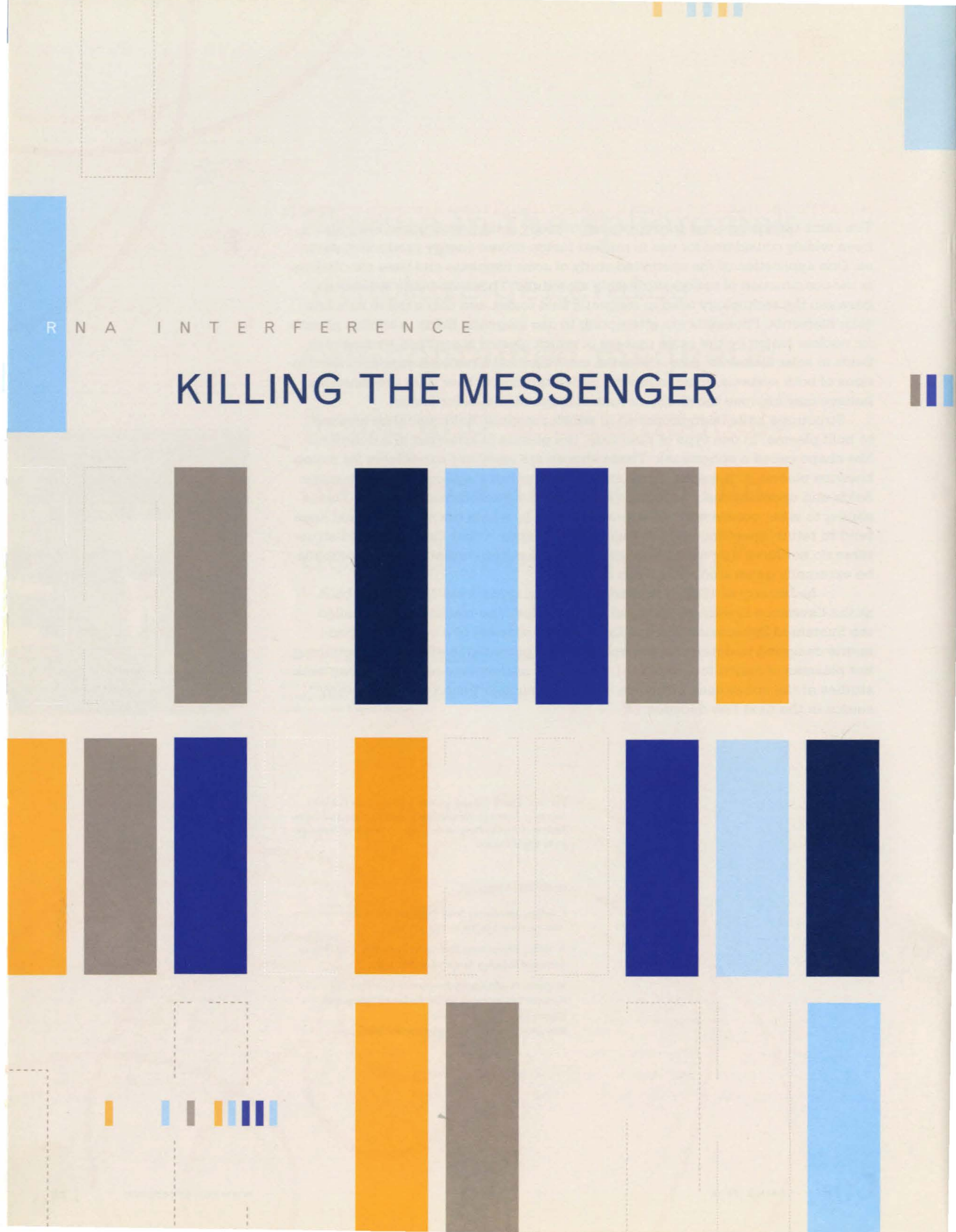
### FURTHER READING

- 1 P. Bellan. Simulating Solar Prominences in the Laboratory. *American Scientist* **88**, 136-143 (MA-APR 2000).
- 2 A. Heller. Experiment Mimics Nature's Way with Plasma. *Science and Technology Review*, **18-20** (DEC 1999).
- 3 Magnetic Fusion Energy Program at Lawrence Livermore National Laboratories. SSPX-Sustained Spheromak Physics Experiment. (JAN 2002). <http://www.mfescience.org/sspx/index.html>



R N A I N T E R F E R E N C E

# KILLING THE MESSENGER



BY VINCENT C. AUYEUNG

**EFFORTS LIKE THE HUMAN GENOME PROJECT** make it possible to predict the existence and structures of new proteins, but this new information sheds limited light on the proteins' function. By carefully tinkering with levels of protein expression—by adjusting protein amounts in the cell or by knocking out a protein's function—biologists can gain further understanding into protein function, offering a more complete picture of how different genes interact in complex organism like human beings.

Currently, proteins are added to cells by introducing the corresponding genes and inducing cells to “over-express” the genes. The opposite effect, “knocking out” a protein, is also used to study the function of proteins but is more difficult. However, the recent discovery of a new technique known as RNA interference (RNAi) has made it possible to rapidly knock out virtually any protein. In this process, short segments of double-stranded RNA trigger the degradation of corresponding messenger RNA segments. RNAi can also target genes that play key roles in viral infections. It is therefore possible to use it to fight infection on a cellular level.

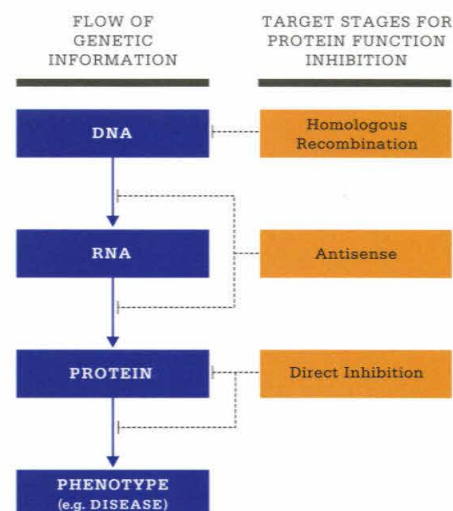
#### THE PATH OF GENETIC INFORMATION

The central dogma of biology holds that genetic information flows from DNA to messenger RNA (mRNA) to proteins. An analogous situation is a factory in which there is a manual containing all of the factory's product specifications, kept in a central office. From this office,

all the specifications are copied and distributed to the factory floor where products are produced. In cells, DNA—the master manual—is kept in the nucleus and information is distributed in the form of mRNA copies bearing instructions for synthesizing proteins.

In order to eliminate the function of a protein, biologists can theoretically target any step along this flow of information [FIGURE 1]. For example, it is possible in rare situations to find an inhibitor that directly interferes with target protein. However, there is no way to systematically generate inhibitors for specific proteins.

The gold standard for eliminating the function of a protein has traditionally been a genetic knockout, in which the gene coding for the protein is physically removed from DNA. In this way, the target protein is not synthesized at all. Biologists have used this process for decades, by randomly removing chunks of DNA using chemical mutagens or radiation. These techniques, however, are not specific enough for biologists to control what parts of the DNA are removed and to manipulate specific genes. Homologous recombination, a technique developed in the past 20 years, has allowed eliminating specific genes via targeted shuffling between similar stretches of DNA. This process, however, is extremely inefficient. Recombination occurs in only one out of every hundred treated cells and 99.99% of these recombinant cells die due to misdirected recombinations. Despite being a standard technique in yeast and mice, homologous recombination is not a suitable process for application in humans.



**FIGURE 1** A protein's effect can be inhibited at many points in the flow of genetic information.





## “RNAi is characterized by its high efficiency and exquisite specificity”

### KILLING THE MESSENGER

What about the second step in the process of encoding proteins? If the production of mRNA for a particular protein could be inhibited, ribosomes would be unable to synthesize the protein. The disadvantage of targeting mRNA is that multiple copies of mRNA are transcribed to synthesize a protein; unless every copy is inhibited, the cell will still be able to synthesize some the protein. Nevertheless, there are two methods to specifically target mRNA strands.

The first relies on the single-stranded nature of mRNA. The double-helical (“twisted ladder”) structure of DNA deduced by Watson and Crick consists of two strands held together by hydrogen bonds. The strands are composed of complementary base pairs: guanine on one strand pairs with cytosine on the other and adenine pairs with thymine. Since these nucleotides always pair up, the strands are actually redundant because the sequence of one strand codes the sequence of the other.

In contrast, mRNA is single-stranded, a critical property that allows ribosomes to read the information from the mRNA. Thus, one way to inhibit a particular mRNA is to introduce a complementary “antisense” strand of RNA. This will match up with the target mRNA, making it unreadable, and inhibiting protein synthesis. Antisense RNAs can also interfere with the transcription of mRNA in the nucleus. This technology, however, achieved only limited success primarily because it failed to inactivate many target genes.

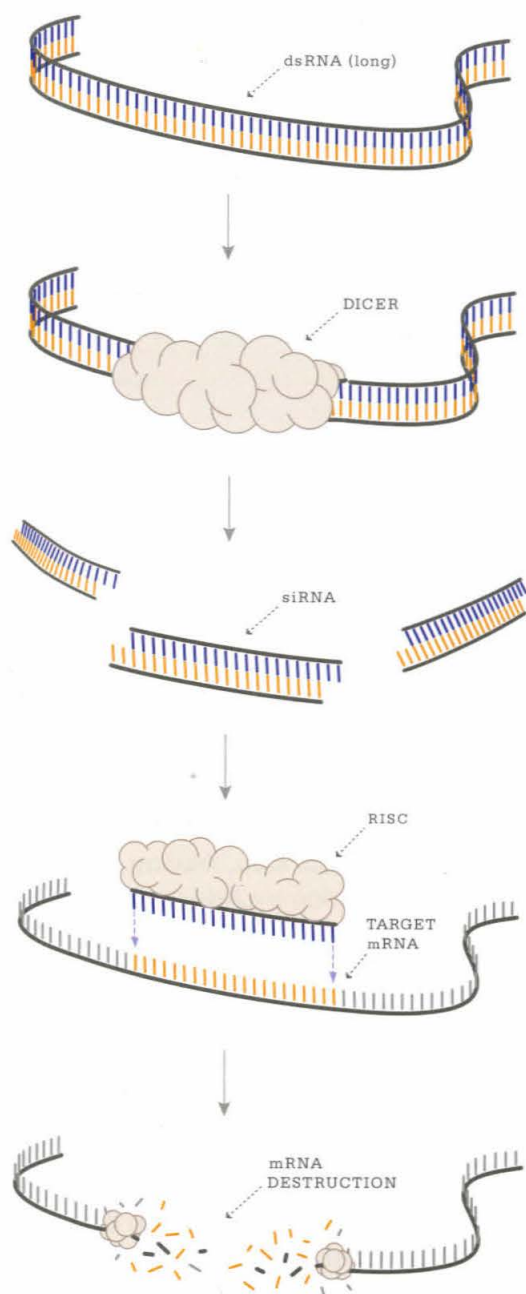
A more exotic method for targeting mRNA was discovered in the 1980s. Researchers found that certain RNA strands have catalytic properties such as the ability to cleave other RNA molecules. These RNA enzymes, better known as ribozymes, catalyze the destruction of target mRNA strands. Ribozymes also achieved only limited success, despite attempts to use them as therapeutic agents to destroy RNAs that code for HIV proteins.

### RNA INTERFERENCE

RNA interference, or RNAi, is a new tool in which double-stranded RNA (dsRNA) acts as a mediator to suppress the function of specific mRNA strands. Unlike ribozymes, the dsRNA does not directly catalyze mRNA cleavage; instead, the double-stranded RNA is incorporated into a protein complex that, through the function of enzymes, cleaves the target mRNA. RNAi is characterized by its high efficiency and exquisite specificity; in theory, only mRNAs that match the dsRNA guide perfectly are degraded. Other mRNAs remain untouched.

The RNAi phenomenon was first encountered by accident during studies of plants in the 1990s. A group of researchers attempted to make purple petunias purpler by inserting an extra copy of the gene encoding the enzyme that synthesizes purple pigment. Many of the plants, however, became white; instead of supplementing the original gene, the added gene somehow silenced its effect. Evidence supporting this claim was the discovery of a small RNA fragment in the white plants that matched up with the added gene but wasn't found in normal plants. At the same time, researchers performing RNA experiments on nematodes (*C. elegans*) found that injections of mRNA also inhibited protein synthesis. A breakthrough in understanding came when Andrew Fire, Craig Mello, and their collaborators discovered that the mRNA preparations were contaminated with small quantities of double-stranded RNA formed by pairing the mRNA and small amounts of the antisense RNA.

We now know that dsRNA is the common trigger in both cases of silenced genes and that a set of cellular proteins catalyzes the destruction of mRNA. This intracellular machinery appears to be present in nearly every eukaryotic organism studied from fungi to humans. In lower organisms, such as nematodes

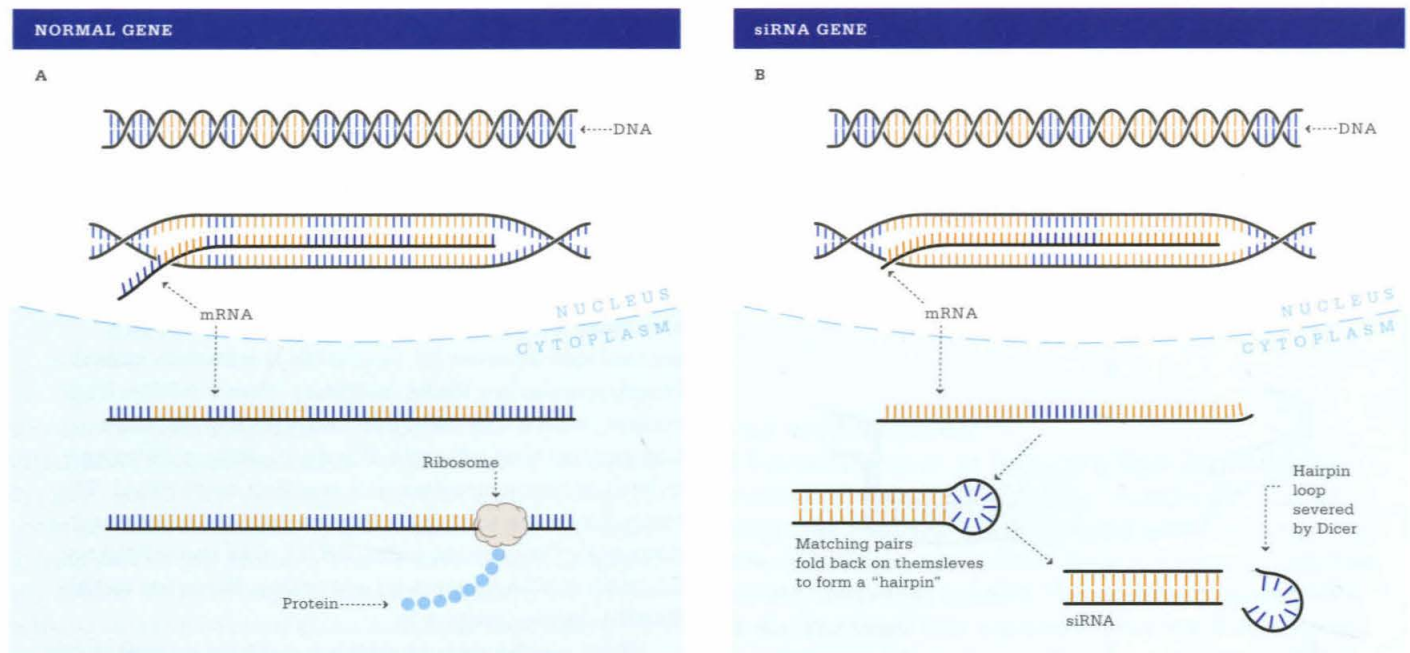


**FIGURE 2** RNAi is activated by dsRNA. An enzyme called Dicer chops up long dsRNA into small fragments siRNA. These fragments guide RISC to the target mRNA, where RISC catalyzes its destruction.

and fruit flies, the introduction of long double-stranded RNA segments triggers the RNAi machinery. Introducing the dsRNA into nematodes can be as simple as synthesizing it *in vitro* and feeding it to the worms or even soaking the worm in a dsRNA solution. Once the long dsRNAs are inside the cells, an enzyme called Dicer cuts the long strand into shorter 21-nucleotide fragments called short-interfering RNAs (siRNAs). These siRNA fragments, which are incorporated into the enzyme complexes that execute the degradation of target mRNA strands, are the true mediators of RNAi. The complexes are appropriately named RNA Induced Silencing Complexes (RISC). RISC use the siRNA to identify mRNA targets by matching them up to the dsRNA guide [FIGURE 2].

This mechanism to destroy mRNAs by matching them to a dsRNA sequence evolved in eukaryotic cells and has been carried over to eukaryotic organisms. A major function of the RNAi machinery may be to resist viruses and other intracellular pathogens, some of which have genomes encoded in RNA instead of DNA. Indeed, some plant viruses carry inhibitors of the plant RNAi machinery; if these inhibitors are missing, infection is severely hindered. There is also evidence that RNAi plays a major role in the control of gene expression during early development. Additionally, in flies and worms, short genes have been identified that encode small double-stranded RNAs called micro RNAs (miRNAs). The fruit fly gene *bantam* appears to encode a miRNA that triggers the degradation of the gene that causes a cell to commit suicide. If the *bantam* gene is inactivated, larvae die shortly after pupation, demonstrating the vital role of dsRNA in the cellular regulation of development.





**FIGURE 3** Artificially designed siRNA genes provide an intracellular source of dsRNA. Natural genes (A) that code for protein make mRNA segments that are long and single-stranded.

#### RNAi IN MAMMALS

Although RNAi is relatively easy to use in lower organisms, there were several major barriers to overcome before RNAi could be used in higher organisms such as mammals. In worms and fruit flies, long dsRNAs can be used to activate RNAi. Mammalian cells, however, have evolved a different response to long strands of dsRNA. Because these strands are often signs of viral infection, a class of molecules called interferons cause apoptosis, or cell suicide, an undesired outcome for RNAi applications. Shorter siRNA fragments, however, are small enough to escape the notice of the interferon system. Since the long dsRNA is normally processed into these smaller fragments anyway, mammalian cells are simply treated with siRNAs instead of full-length dsRNAs.

Another problem was inserting these siRNAs into mammalian cells. As in worms and fruit flies, siRNAs can be artificially produced and directly delivered to cells. But artificial synthesis of RNA is expensive; further, RNA is unstable and the effects of RNAi quickly wear off as cells divide and the dsRNA decays. An ideal solution would be to design genes that synthesize the desired siRNA and insert these genes into cells, which would then continually replenish the supply of dsRNA. But when mRNAs and

In contrast, RNA transcribed from a siRNA gene (B) is designed to fold back on itself to form a double-strand "hairpin" that is processed in a siRNA by the cell.

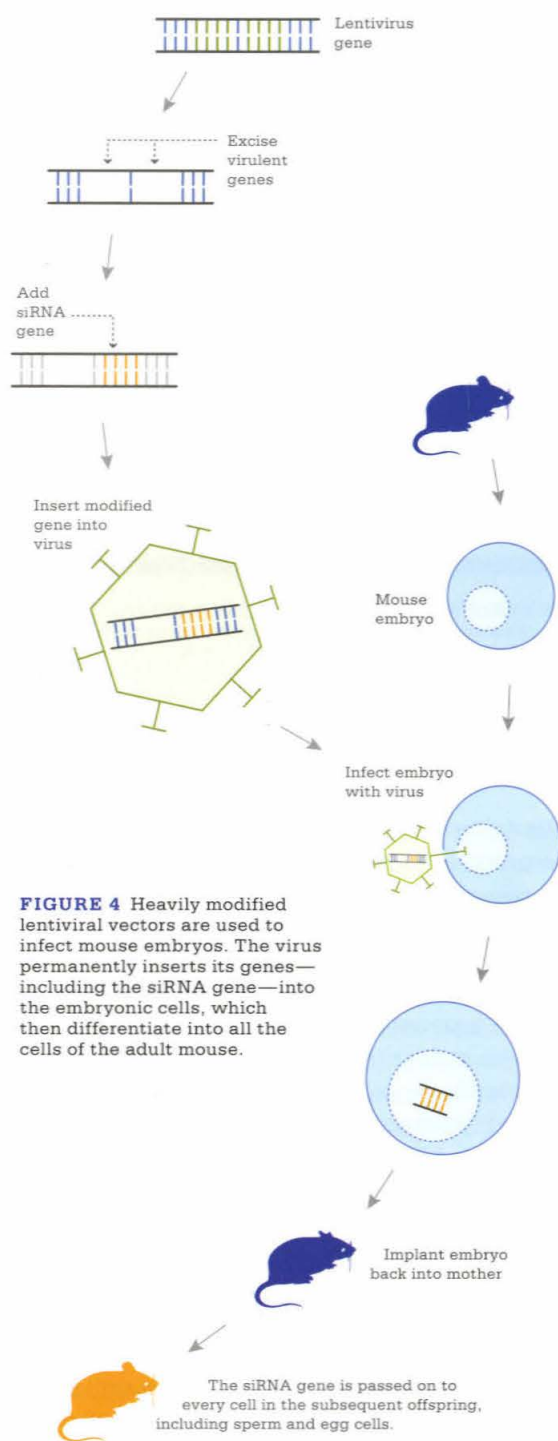
other RNAs are transcribed from RNA, the product is single-stranded instead of double-stranded. The solution is to design siRNA genes so that they fold back on themselves to form hairpin shapes consisting of a stem and a loop. Conveniently, cellular Dicer enzymes recognize and cut off the loop, leaving a short fragment of dsRNA that triggers RNAi [FIGURE 3]. Such siRNA genes can be delivered into cells using several well-established techniques, although most of these techniques do not permanently integrate the siRNA gene into the cellular RNA.

#### HOW TO DELIVER YOUR OWN siRNA GENE

One method of delivering siRNA genes into mammalian cells is to use lentiviral vectors [FIGURE 4]. Lentiviruses are a family of retroviruses that insert viral genomic DNA permanently into cellular DNA. Laboratories have engineered lentiviruses to carry genes of interest in a safe and efficient manner; sixty percent of the native viral genes have been eliminated, including the genes that code for replication and virulence. Nevertheless, lentiviruses can still efficiently deliver genes into a wide variety of cell types.

Lentiviruses carrying a siRNA gene can infect single-celled mouse embryos, which give rise to all the cell types found in the fully developed animal.





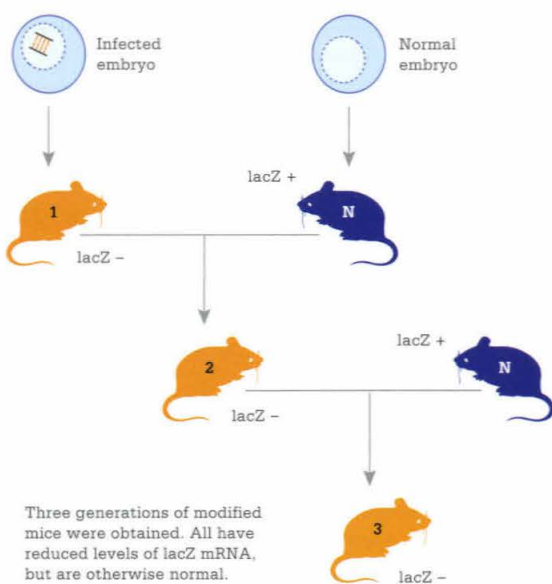
**FIGURE 4** Heavily modified lentiviral vectors are used to infect mouse embryos. The virus permanently inserts its genes—including the siRNA gene—into the embryonic cells, which then differentiate into all the cells of the adult mouse.

Because of this, the integrated siRNA gene is passed on to every cell in the animal, including sperm and egg cells. Thus, a protein functional knockout is present in every cell of the body. Compared to homologous recombination, which makes genetic knockouts of the gene itself, lentiviral delivery of siRNA is faster and more efficient. Infecting a single-cell embryo is as simple as exposing the egg cell surface to a solution containing the engineered virus. Better yet, the success rate of lentiviral gene delivery is often greater than 75%, which is almost a million times more effective than homologous recombination.

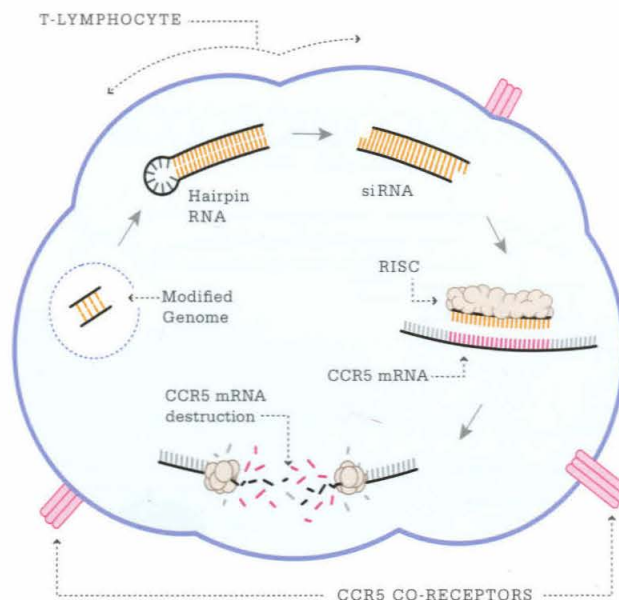
Using lentiviruses, David Baltimore's laboratory has successfully generated transgenic mouse lines that produce siRNAs in every cell; these siRNAs are designed to target a nonessential reporter gene called *lacZ*, which codes for the enzyme beta-galactosidase. For each line we generated, a parent mouse was infected with the engineered virus and mated with a normal mouse; their progeny are considered the first filial generation (F1). Those progeny are then mated to normal mice to produce the second filial generation (F2). Two mouse lines were developed from these initial generations. No gross defects or abnormalities were observed among any of the three generations of either line [FIGURE 5]. This indicates that continuous stimulation of the RNAi machinery is not necessarily fatal and does not grossly interfere with normal growth and development.

We then collected RNA from five organs—the brain, thymus, liver, lungs, and kidneys—to determine whether cells make siRNAs in each of the organs and whether the siRNAs trigger the degradation of *lacZ* mRNAs. Indeed, the siRNAs are expressed in all five organs, although at varying levels. At the same time, *lacZ* mRNAs levels were reduced in all five organs, though the degree of reduction varied. In general, the more siRNA found in the organ, the more *lacZ* levels were reduced, which was consistent with observations that have been published by other researchers. We demonstrated that it is possible to make transgenic mice in which the levels of a particular mRNA (and its corresponding protein) are heavily reduced or effectively absent. This is extremely useful for studies in which a protein's function must be curtailed or eliminated in multiple cell types and locations. For example, proteins that play a role in animal behavior could be studied by using lentivirus-delivered siRNAs to degrade the protein's mRNA in mice and then analyzing the behavior of the mice as they grow to adulthood.





**FIGURE 5** A lacZ knockout line was engineered by infecting mouse embryos with a lentivirus carrying an anti-lacZ siRNA gene.



**FIGURE 6** Using RNAi, David Baltimore's lab engineered T-cells with reduced expression of the CCR5 gene, thereby reducing HIV infection.

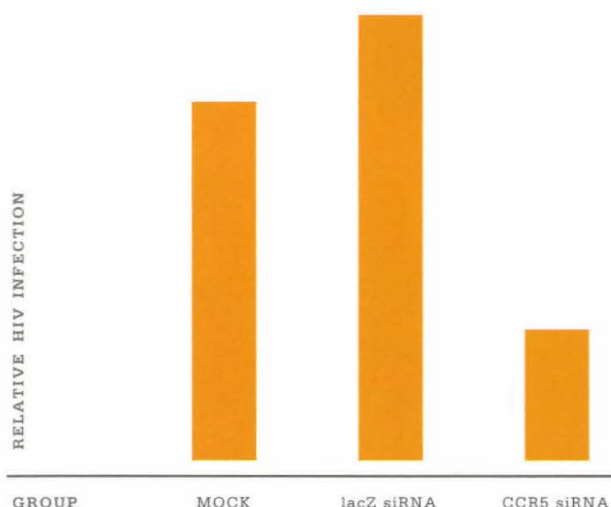
### STOPPING HIV PRODUCTION IN HUMANS

Not all potential applications of lentivirus-mediated RNAi involve the generation of entire transgenic animals. One major application is the use of RNAi as an antiviral therapy. In such an application, siRNAs would be designed to target viral genes or even human genes involved in viral entry or replication. Lentiviruses would then be used to deliver the corresponding siRNA genes into cells vulnerable to a particular viral disease. These cells would degrade any viral RNAs that enter the cell or prevent viral entry altogether. In other words, individual cells would become immune to viral infection.

David Baltimore's lab has demonstrated the feasibility of such a strategy by engineering T-cells that express siRNAs designed to inhibit HIV infection [FIGURE 6]. These cells play a crucial role in the body's immune system. The gene we chose to target is not a viral gene, but rather a human gene that codes for a cell surface protein called CCR5. During the early stages of infection, HIV-1 uses CCR5, along with several other T-cellular proteins, in order to gain access

to T-cells for infection; if CCR5 is missing, HIV-1 infection is significantly hampered. CCR5 does not appear to be necessary for normal immune function. In fact, a small percentage of the human population completely lacks the gene for CCR5; such people are resistant to HIV-1 infection but otherwise appear to have normal, fully functional immune systems. Another major advantage is that CCR5 is a human gene. Therapies that specifically target HIV genes are often vulnerable to viral "escape mutants" because HIV mutates very rapidly and becomes ineffective once the target gene mutates to avoid attack. CCR5 is a stable target because mutations in human genes are extremely unlikely. Overall, CCR5 is an ideal target for RNAi-mediated inhibition of HIV infection.

Using lentiviruses, we inserted a siRNA gene targeting CCR5 into human T-cells. We observed a significant tenfold reduction in CCR5 levels on the surface of the transduced T-cells. This reduction in CCR5 expression was accompanied by a three-to-sevenfold reduction in the level of HIV infection. It is clear that individual T-cells can be made resistant to HIV infec-



**FIGURE 7** Treating T-cells with siRNAs targeted against CCR5 inhibits HIV infection. Relative to control T-cells, treated cells are three to four times less likely to be infected by HIV. Treatment of T-cells with siRNAs targeted against lacZ (unrelated to HIV infection) does not reduce infection, demonstrating the specificity of RNAi.

tion, although this resistance is not perfect [FIGURE 7]. A strategy like this might work as a therapeutic tool in human beings. Bone marrow containing stem cells would be taken from a patient. Lentiviruses could be used to integrate a siRNA gene targeting CCR5 into these cells that ultimately make T-cells in the body. Marrow containing stem cells with the siRNA gene would then be re-implanted into the patient. Thanks to the integrated siRNA, all the T-cells made from the bone marrow would be resistant to HIV infection. Though such a scenario will take years to accomplish, the strategy is not farfetched. The use of modified viruses to deliver genes into stem cells is already major therapy for certain types of severe combined immunodeficiency disease (SCID), a condition where children are born without a functional immune system.

Over the past five years, RNA interference has blossomed from a biological oddity into a rapidly advancing field with a breathtaking variety of applications across many eukaryotic organisms. A tool that quickly and efficiently knocks out specific mRNAs,

“CCR5 is  
an ideal target  
for RNAi-mediated  
inhibition of HIV  
infection.”

RNAi is spearheading a revolution in genetics and protein analysis. The rewards of RNAi may be even greater: coupled with the development of lentiviral vectors, RNAi may one day become the therapy of choice for fighting HIV and other viral diseases.

*Vincent Auyeung is a third year undergraduate at the California Institute of Technology and a student in David Baltimore's lab. His work is supported by the Beckman Scholars program of Arnold and Mabel Beckman.*

#### FURTHER READING

1. A. Fire, S. Xu, M. Montgomery, et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**:806-11 (1998).
2. G. Hannon, ed. *RNAi: A Guide to Gene Silencing* (Cold Spring Harbor Laboratory 2003).
3. Qin, et al. Inhibiting HIV-1 infection in human T cells by lentiviral-mediated delivery of small interfering RNA against CCR5. *Proceedings of the National Academy of Science* **100**: 183-188 (2002).
4. M. Sohail, ed. *Gene Silencing by RNA Interference: Technology and Application* (CRC PRESS 2004).



# FILMS THAT BEND

BY AZIEL C. EPILEPSIA

**MICROFLUIDICS, THE STUDY OF THE MOTION** of fluids on a micron or nanometer scale, is a relatively new, interdisciplinary field. With current applications ranging from miniature biological Petri dishes to inkjet printer nozzles, microfluidics is both widespread and versatile. Research in microfluidics aims to do for chemistry and biology exactly that which the integrated circuit did for electronics: miniaturization. Experimental setups for chemical reactions requiring an entire laboratory bench could be reduced to a chip no larger than the palm of your hand. Technicians who currently conduct and monitor experiments could one day be replaced by a computer automating the experimental process. By scaling down the physical size of such apparatus to a micron level, scientists would be able to decrease raw material expenses and power consumption as well as increase efficiency and resolution.

One promising application of microfluidics lies in the fabrication of transistor-like devices. Transistors, which were invented in the late 1940s, revolutionized electronics by replacing large, unwieldy vacuum tubes as amplifying elements in electrical circuits. They are made of semiconductors and metals and allow engineers to control electrical currents. In a microfluidic circuit, the direction and intensity of liquid current represent signals much like the way direction

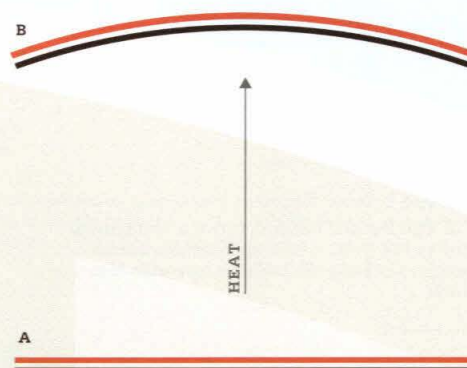


and magnitude of electrical current represent electrical signals in a transistor-based microprocessor chip. In order for such a microfluidic circuit to operate, scientists must be able to control the flow of liquids between parts of the circuit. That's where thermal bimorphs come in—by converting light energy into mechanical energy, thermal bimorphs serve as pumps and valves for micron-scale fluid flow.

#### SILICONE IN A NON-COSMETIC CONTEXT

A thermal bimorph is a thin film consisting of two layers of different materials that expand at different rates when exposed to heat. Thus, when heat is applied to a bimorph, one layer expands more than the other, and the bimorph bends. A bimorph can function as a valve by bending far enough into a channel to seal off the path of a liquid. Also, by bending and displacing a fixed volume of liquid in a microfluidic chamber, a bimorph can also function as a pump.

The two layers of a bimorph are each quantified by a thermal expansion coefficient  $\alpha$ , expressed in units of parts per million per Kelvin (ppm/K), which relates the magnitude of expansion to changes in temperature. Therefore, the difference between the thermal expansion coefficients of the materials composing the two layers of a bimorph determines how much a bimorph bends with applied heat.



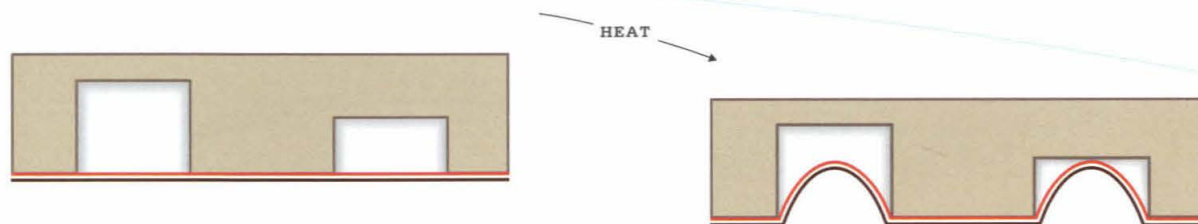
**FIGURE 1** (A) shows a thermal bimorph before heat is applied. The upper layer (orange) has a higher thermal expansion coefficient than the lower layer (brown). (B) When heat is applied, the bimorph bends toward the upper layer.

We studied thermal bimorphs composed of two silicone elastomers, Dow Corning's Sylgard184 ( $\alpha=310$  ppm/K) and General Electric's RTV21 ( $\alpha=200$  ppm/K). The combination of the two individual elastomers results in a thermal expansion coefficient of  $310 - 200 = 110$  ppm/K for the bimorph, the difference between the expansion coefficients of the individual materials. Because the Sylgard184 layer has a higher thermal expansion coefficient, the bimorphs should bend toward the Sylgard184 layer as heat is applied [FIGURE 1].

To fabricate a thermal bimorph, drops of one elastomer are first placed on a silicon wafer. Before the elastomer is allowed to cure from liquid to solid on the surface of the wafer, the wafer is fixed in place by vacuum on the flat upper surface of a cylindrical, vertical axle. The axle is then spun, creating a uniformly thin film of elastomer on the surface of the wafer ranging from 15 to 50 microns in thickness. This first film is allowed to cure, and drops of the second elastomer are subsequently deposited on top of it. The wafer is spun again, and the second layer of elastomer is allowed to cure. This entire process allows us to make a bimorph with a thickness between 20 and 100 microns.

Once the layers of a bimorph had cured, we were able to quantify the bimorph's response to heat. To do so, we first attached a microfluidic channel made of Sylgard184 to the bimorph film on the silicon wafer.





**FIGURE 2** The thermal bimorph (orange and brown) is attached to the bottom of a microfluidic device. Upon heating, the bimorph deflects upwards into the channels.

After attachment, we cut the bimorph into cross sections and placed one section on top of a thermoelectric heater. By increasing the heater's input voltage, we could raise the temperature of the bimorph. To quantify its performance, we observed the entire setup under a microscope and measured the depth of the bimorph's deflection into the channel [FIGURE 2].

Our control bimorph was 55 microns thick and was placed in a microfluidic channel 2 mm wide. We increased the temperature of the bimorph by 101°C and observed a deflection of 343%; that is, the bimorph bent into the channel a distance of 3.43 times the bimorph's thickness. This was a reasonable degree of bending, but in order to optimize the use of silicone elastomers in our bimorph layers, we had to vary some parameters.

#### ALTERING ELASTOMER PROPERTIES

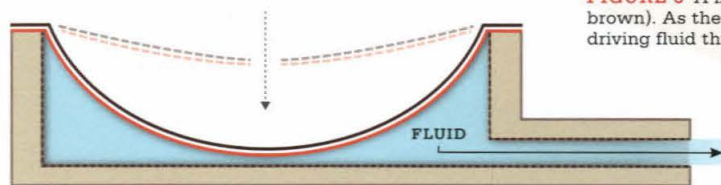
Several intrinsic properties of silicone elastomers prevent them from being easily used in thermal bimorphs. One such property is the low thermal conductivity of silicone elastomers, leading to slow bimorph response times. Previous research has shown that doping an elastomer with carbon has significant effects on its thermal conductivity because it speeds up the rate at

which heat is absorbed. Sylgard184 is normally translucent; the introduction of carbon makes the elastomer opaque. As a result, the elastomer absorbs both light and heat more rapidly. To investigate this effect, we created a bimorph by using Sylgard184 doped with 10% carbon black. The bimorph showed a deflection of 135% when the temperature was increased by 101°C, a small deflection compared with our control bimorph. However, as we expected, the bimorph responded much faster to heating—the deflection was achieved rapidly after increasing the temperature.

Several elastomers, including RTV21, share another undesirable property, namely, high viscosity. A high-viscosity elastomer cannot easily be spun into a thin film. To counteract this, the elastomer in question can be diluted with toluene. Using this method, we were able to produce bimorphs between 8 and 20 microns thick. However, previous research suggested that diluting an elastomer could potentially alter its thermal expansion properties. To investigate this, we created diluted bimorphs and compared them to non-diluted bimorphs. One bimorph, diluted in both layers with 30% toluene, showed much less bending than the control bimorph for a 101°C change in temperature—a total of only 124% deflection was achieved.



“We found that thinner bimorphs generally showed a larger deflection than thicker bimorphs.”




**FIGURE 3** A microfluidic pump using a thermal bimorph (orange and brown). As the bimorph deflects, pressure increases in the large chamber, driving fluid through the channel on the right.

The maximum bimorph bending that we saw was close to 900%; this bimorph was 14 microns thick and made with a 30% toluene dilution. Compared to the control, this bimorph showed a far more successful deflection response. By varying only the thicknesses of different bimorphs and keeping everything else constant, we found that thinner bimorphs generally showed a larger deflection than thicker bimorphs. None of our bimorphs was thinner than 14 microns, but we could expect such bimorphs to have deflections above 900%.

#### BIMORPHS IN COMPUTERS

Our various experiments quantified several different bimorphs' deflections and attempted to optimize the use of silicone elastomers as bimorphs. However, our research did not include the testing of the force generated by the deflection of a bimorph, an important parameter that will be quantified in future research. By knowing the force of a bimorph's deflection, we would be able to predict the volume of liquid displaced by the bimorph's bending as well as the rate at which the liquid would be displaced, precisely the information necessary for scientists to implement thermal bimorphs as pumps controlling fluid flow in a microfluidic circuit [FIGURE 3].

Modern microfluidics research has already seen the creation of integrated microfluidic circuits. Over thirty years ago, random-access memory (RAM) chips represented cutting-edge large-scale integration. In their original form, these chips are obsolete today. Similarly, we can expect future scientists to fabricate astoundingly complex microfluidic circuits for a plethora of applications, including the microfluidic analogs of modern digital circuits. 

*Aziel Epilepsia is a junior in Electrical Engineering at the University of Washington. He wishes to thank his mentor, Dr. Stephen Quake, as well as his co-mentor, Jian Liu, for their guidance over the summer of 2003. He would also like to thank the MURF Program as well as the James Irvine Foundation.*

#### FURTHER READING

1. T. Thorsen, S. J. Maerkl, S. R. Quake. Microfluidic Large-Scale Integration. *Science* 298, 580-584 (2002).
2. M. A. Unger, H. P. Chou, T. Thorsen, A. Scherer, S. R. Quake. Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography. *Science* 288, 113-116 (2000).
3. J. Gaspar, V. Chu, N. Louro, R. Cabeca, J. P. Conde. Thermal actuation of thin film microelectromechanical structures. *Journal of Non-Crystalline Solids* 299-302, 1224-1228 (2002).
4. D. K. Shenoy, D. L. Thomsen III, A. Srinivasan, P. Keller, B. R. Ratna. Carbon coated liquid crystal elastomer film for artificial muscle applications. *Sensors and Actuators A* 96, 184-188 (2002).



LEAP  
LEAP  
AP

ING OVER

CH  
HA  
A

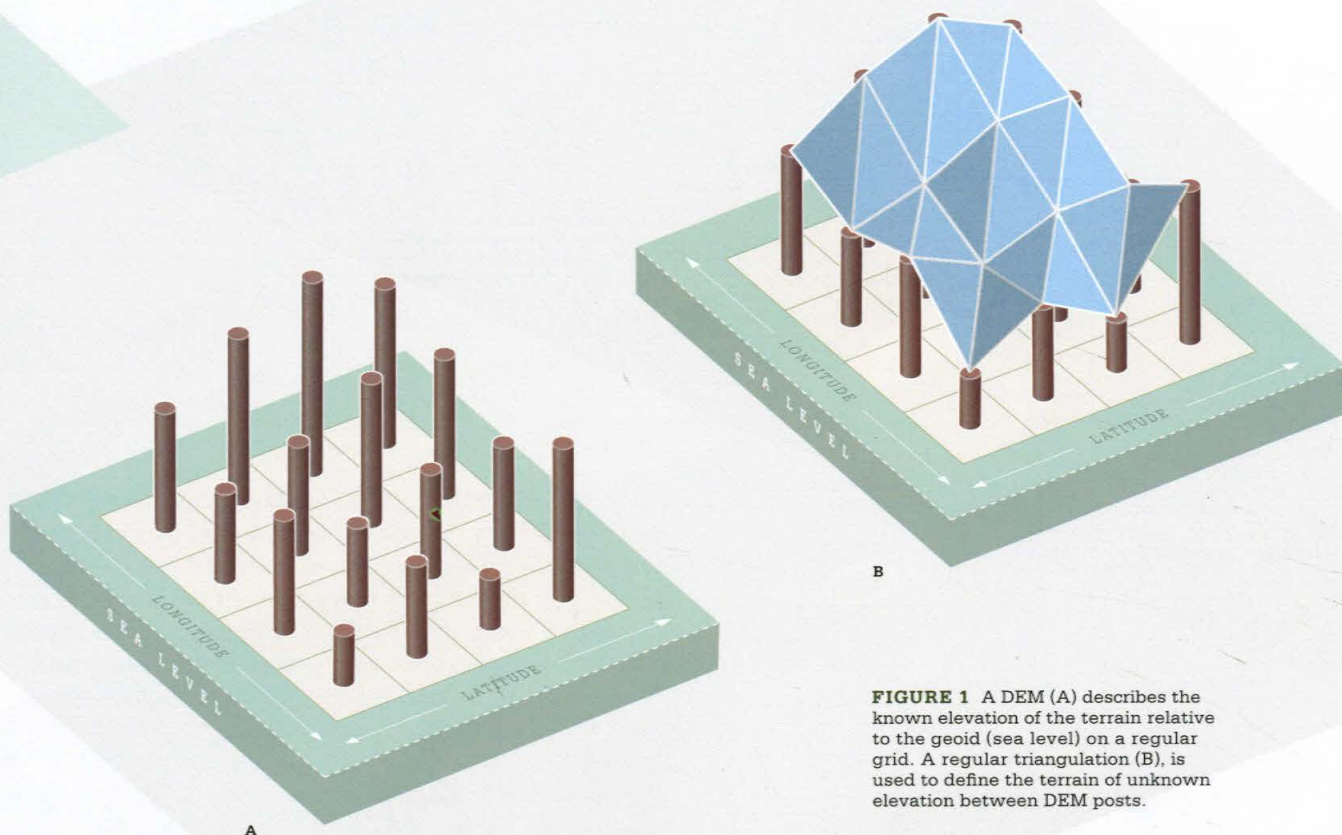
BY JOSEPH E. GONZALEZ

## A NEW ALGORITHM FOR EVALUATING LINE-OF-SIGHT ON DIGITAL ELEVATION MAPS

NEW INTELLIGENCE INFORMATION INDICATES that several key operatives in a radical terrorist cell are currently meeting in an abandoned hut deep in the rugged mountains of northern Afghanistan. Five kilometers away, a Delta Force unit is deployed to capture the operatives. However, if the unit is detected by militants patrolling the mountain trails, the operatives will have sufficient warning to avoid capture. Driving Humvees equipped with mobile workstations, the Delta Force unit can rapidly move to the hut while evading the militants using a powerful new line-of-sight evaluation algorithm in development at NASA's Jet Propulsion Laboratory.

Rapidly and accurately assessing visibility between two points on large terrain maps is critical to the success of military simulations, real time terrain visualization, and the placement of wireless communications systems. Each application imposes different requirements on the speed and accuracy of LOS evaluation and the availability of hardware resources. Military simulations use large terrain maps, often spanning several countries at high resolutions (30 data points per kilometer), and require accurate visibility information for thousands of military platforms every second. These high demands push the current desktop computer architecture to its limits in memory and speed. However, using a technique which divides the problem into more manageable chunks, we can quickly and accurately compute line-of-sight. Consequently, a Delta Force unit could continuously determine what parts of the terrain militants can see and adjust their routes accordingly as the militants moved.





**FIGURE 1** A DEM (A) describes the known elevation of the terrain relative to the geoid (sea level) on a regular grid. A regular triangulation (B), is used to define the terrain of unknown elevation between DEM posts.

#### SQUEEZING THE EARTH INTO A COMPUTER

The topography of the Earth is captured by satellites in the form of a digital elevation map (DEM). A DEM is simply a regular grid of elevation points called DEM posts. While the elevation of the terrain at individual DEM posts is well defined, the terrain between DEM posts is undefined. We chose to use a regular triangulation of the DEM, in which we draw straight lines between adjacent DEM posts. The straight lines form a regular tessellation of triangular planes. These planes compose a regular triangulation and define the terrain between the DEM posts [FIGURE 1].

When it is necessary to reduce the number of DEM posts or edges, an irregular triangulation is of-

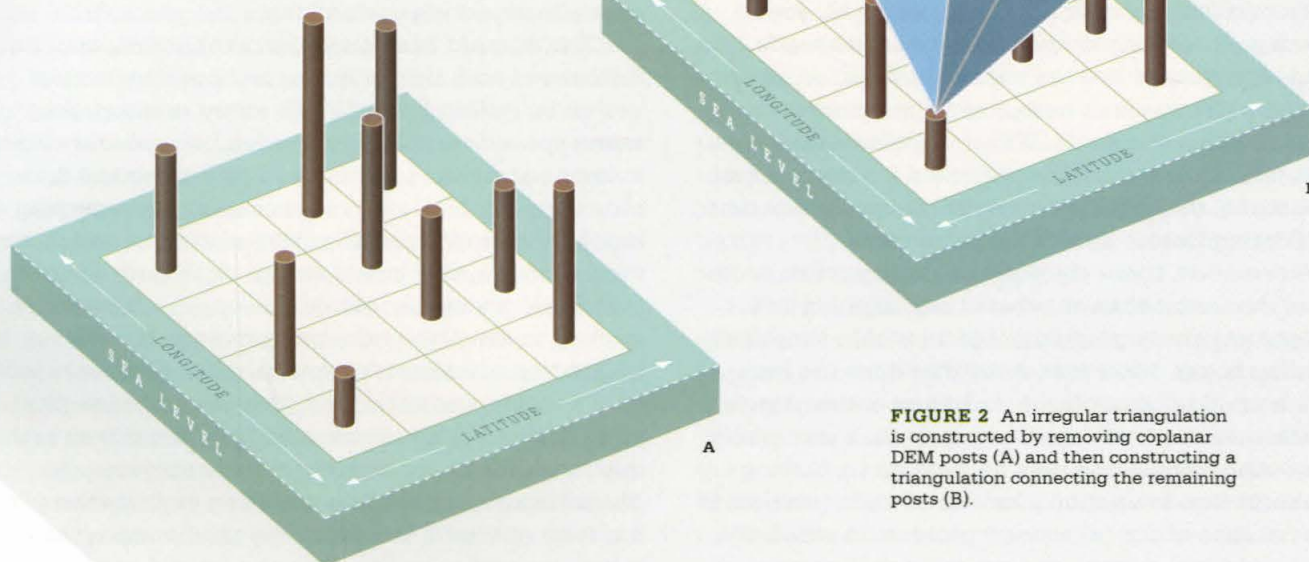
ten used. Irregular triangulations use large triangles to represent regions of nearly coplanar DEM posts [FIGURE 2]. However, while large triangles reduce the number of DEM post they also may remove significant terrain features. Furthermore, it is difficult to find the optimal set of large triangles that best represents the terrain.

Although more complex interpolation techniques exist, they do not necessarily better represent the true surface of the Earth. Consequently, we assume that a regular triangulation of a DEM is the most accurate interpolation. The new algorithm for evaluating LOS uses the regular triangulation of a DEM as its model of the Earth's surface.

#### SEEING THE SLOW WAY

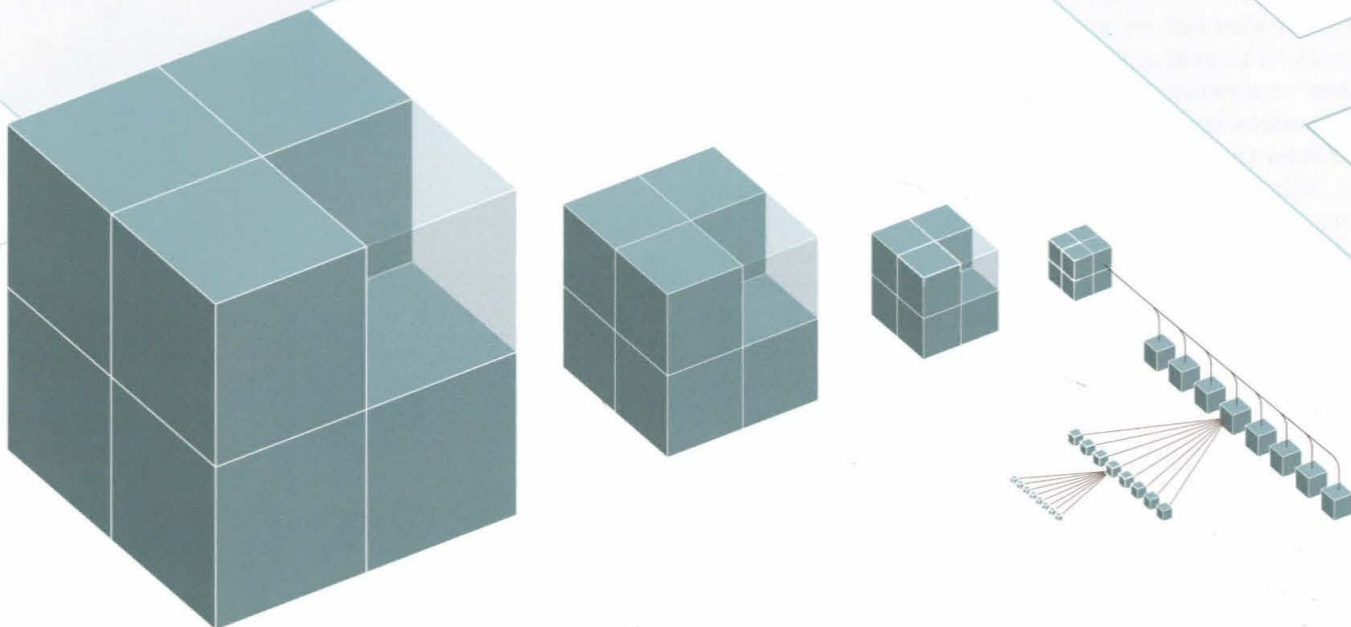
Military simulations currently use one of two major algorithms to evaluate LOS. The "telephone pole algorithm" compares the height of the LOS to the height of the surface model at uniformly spaced steps. If the LOS is below the surface at any step then LOS is obstructed. This algorithm has the advantage of faster computation through the reduction of the number of steps at the expense of decreased accuracy. To reduce the frequency of false results, the number of steps must be increased, thereby slowing the algorithm.

The second major line-of-sight algorithm traverses the edges of a piecewise planar model of the terrain. The speed and accuracy of this approach depends on the model of the terrain rather than the algorithm. In this algorithm, LOS is assessed by comparing the line connecting the two points to each edge it crosses. The algorithm indicates that the line-of-sight is blocked if the line falls below any edge. Likewise, it indicates that the line-of-sight is clear if it has reached an end-point without passing under an edge of the terrain. While this method can be applied directly to a regular triangulation of a DEM, the large number of edge crossings makes it prohibitively slow in most applications. Consequently, an irregular triangulation is often used to increase evaluation speed at the expense of accuracy.



**FIGURE 2** An irregular triangulation is constructed by removing coplanar DEM posts (A) and then constructing a triangulation connecting the remaining posts (B).





A OCTREE DATA STRUCTURE

#### ILLUMINATING BOXES IN BOXES

Our new approach to evaluating LOS applies techniques originally developed to accelerate 3D scene rendering. To obtain exceptionally realistic images, rendering algorithms trace the paths of simulated rays of light through complex three-dimensional scenes containing millions of objects. Without optimization, ray-object intersection must be evaluated for every object in the scene, making a detailed, full feature animation like "Finding Nemo" unfeasible to produce.

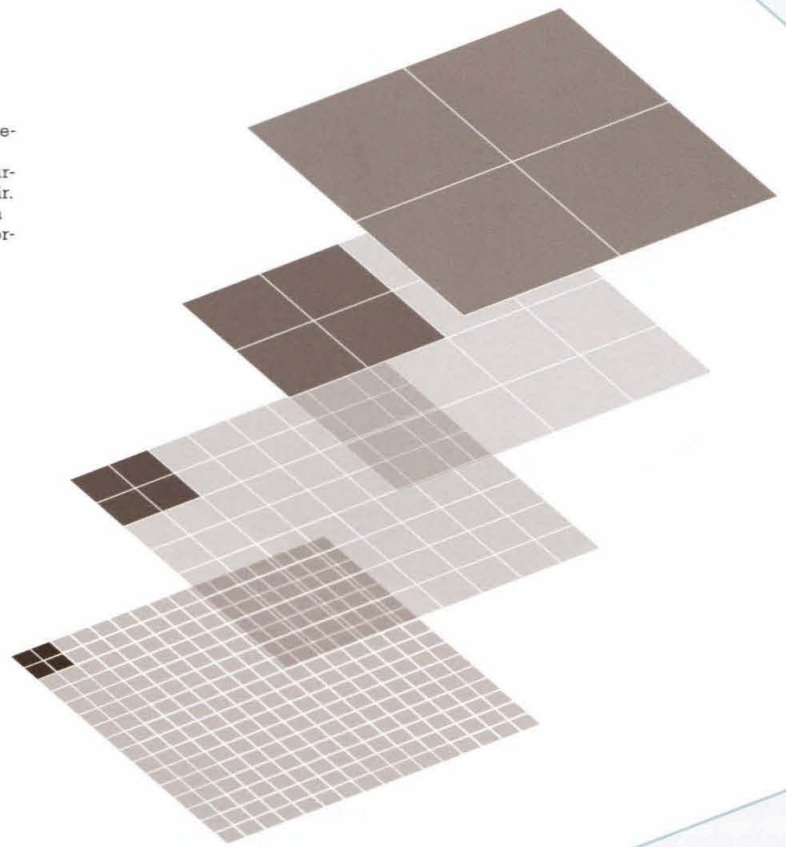
Fortunately, there are powerful optimization techniques that reduce the number of ray-object intersection evaluations by grouping objects within simplified bounding boxes. Moreover, a ray that does not intersect a bounding box will not strike any of the objects contained within that box. Consequently, many costly intersection evaluations may be avoided by making one simple box evaluation. Our line-of-sight problem is a special case of the ray tracing problem in which the line-of-sight is the "ray" and we want to determine if that ray intersects with the terrain. Therefore, by draw

bounding boxes on our map we can also avoid unnecessary intersection evaluations.

The number of intersection evaluations may be further reduced by grouping the bounding boxes within bounding boxes. In 3D scene rendering, the scene space is repeatedly divided into smaller cubical volumes and then reassembled into an octree data structure in which every higher level, representing a box, contains eight smaller levels, each of which contains eight smaller levels and so forth until the data is indivisible [FIGURE 3A]. While this approach could be applied to the LOS evaluation problem, the regular triangulation is a simpler two-dimensional surface with only a single elevation for a given latitude-longitude pair. The two-dimensional analog to the octree is the quadtree, which recursively divides a space into quadrants [FIGURE 3B]. The quadtree exploits the planar form of the DEM by dividing the bounding boxes into four smaller boxes along the latitudinal and longitudinal axes.

**FIGURE 3** An octree data structure (A) recursively divides a three-dimensional space into octants. Unlike most three-dimensional scenes encountered in computer graphics, the DEM is a simple surface with only a single elevation for a given latitude-longitude pair. Therefore, a quadtree data structure (B) that recursively divides a two-dimensional space labeled by latitudinal and longitudinal coordinates into quadrants is better suited.

**B** QUADTREE DATA STRUCTURE



#### GROWING A TREE FROM THE LEAVES

Quadtree spatial partitioning is applied by dividing the DEM along the posts into progressively smaller quadrants until each quadrant contains only 4 posts. Quadrants that contain four DEM posts (and that are therefore indivisible) are represented by four leaves containing the elevations of the four DEM posts. Any quadrant that may be further divided (that is, contains more than 4 posts) is represented as a node containing the highest elevation in that quadrant.

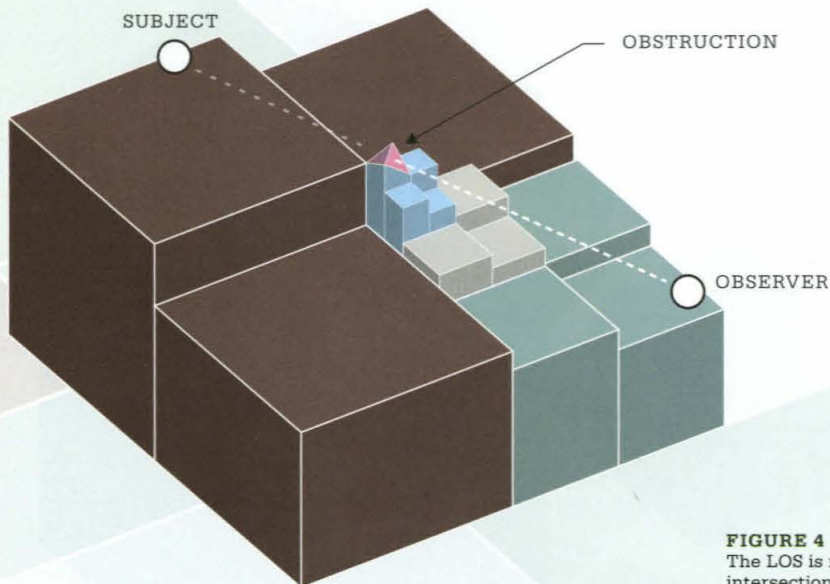
To ensure that the DEM is always evenly divisible, each dimension (rows and columns) of the DEM is enlarged to the nearest integer that can be expressed in the form  $2^n + 1$ , where  $n$  is an integer. However, no data is stored for leaves that are outside of the DEM or for nodes that do not intersect the DEM.

By starting at the leaves, we only need three comparisons per quadrant to determine the maximum elevation within that quadrant rather than comparing all the DEM posts within the quadrant. (Starting with a square 2 posts by 2 posts wide, we need three comparisons to find the highest post in that square. We

can group all of the points on the grid into non-overlapping  $2 \times 2$  squares and store the maximum elevation in each of them. Then, we can group four neighboring  $2 \times 2$  squares into a  $4 \times 4$  block. The highest elevation in the  $4 \times 4$  block is the highest of the four "maximum elevations" of each of the four  $2 \times 2$  blocks, and since there are only four numbers to compare, this only takes three comparisons.) As the quadrants get larger, this dramatically reduces the number of comparisons since the maximum elevation of a large quadrant can be determined from the maximum elevations of the four quadrants that compose it.

The need for memory pointers (common to tree data structures) was also eliminated by using a spatial mapping function. This function maps any quadrant of space on the earth to the unique location in memory at which the maximum elevation for that quadrant is stored. Storing the pointers would have resulted in a quadtree data structure 14 times larger than the original DEM. By using the spatial mapping function the final quadtree data structure was only 33% larger than the original DEM.





**FIGURE 4**

The LOS is reported to be obstructed at the the first intersection of its path with a leaf-level triangulation.

#### IGNORING IRRELEVANT POINTS

The new algorithm quickly evaluates line-of-sight by only checking the points that could affect the answer. If you are trying to evaluate LOS across a valley and you know that every point in the valley is lower than your line, there is no reason to check every point within the valley. By searching the largest bounding boxes first, it is possible to reject LOS-terrain intersection over large regions of a map in relatively few steps. We only search through the more detailed elevation information within boxes that the LOS intersects. Moreover, the terrain is examined at the lowest possible detail necessary to reject intersection.

The algorithm starts by determining if the line-of-sight path intersects any of the top-level bounding boxes. If it does not, then it reports that the path is unobstructed. If the path intersects any of the top-level boxes, then the algorithm repeats for each of the intersecting boxes by checking for intersections between the LOS and each of the lower boxes. If the algorithm ever finds that the LOS intersects a leaf (one of the pairs of triangles in the regular triangulation), then it reports that the path is obstructed; otherwise, it will report that the path is unobstructed. [FIGURE 1]

Because the DEM is always divided into quadrants, the number of computations in this approach is logarithmic in the number of DEM edges along the LOS. In comparison, the complexity of the edge traversal technique is linear in the number of steps and the telephone pole technique has constant time com-

plexity. Therefore, this new technique is faster than the edge traversal technique and slower than the telephone pole technique but doesn't suffer from the telephone pole algorithm's inaccuracy.

To compensate for the curvature of the Earth, we added a few additional values to the dimensions of the bounding boxes. To adjust for the convex top of the bounding boxes, we added a precomputed offset. The distance between a LOS and the surface of the earth increases along the LOS away from the point of tangency. An additional 2nd order approximation of this curvature is subtracted from the top of the box. Prior to subtraction, the second order offset is multiplied by a scaling factor to compensate for the refractive properties of the atmosphere.

Even with this faster LOS algorithm, the size of the maps involved in these computations poses a significant problem. For large data sets the DEM alone can exceed several gigabytes of memory. To make this algorithm practical for large data sets on desktop computers, we need to carefully manage the way we interact with the data. Because each level of the tree is a linear array of binary elevation data describing a relatively smooth surface, compression techniques may be feasible. However, rather than researching techniques to compress the data structure we are focusing on transferring only the necessary information to low-latency memory (RAM) while leaving large portions of the unread data structure on high-latency memory (hard disk). It will often be possible to determine when




# “The terrain is examined at the lowest possible detail necessary to reject intersection.”

LOS is unobstructed by reading from the first few levels of the quadtree data structure. These levels take in substantially small amounts of memory relative to the entire DEM and can easily be stored in RAM. Data from lower levels is needed only when the LOS intersects a bounding box. Because military simulations of events are usually confined to small regions rather than across the whole map, information describing only these regions may be easily copied to RAM.

## FASTER AND MORE ACCURATE

We wrote a prototype implementation of the algorithm and the data structure generator. A quadtree data structure was constructed from a 3-second resolution DEM of Korea. By replaying a previous military simulation running the new quadtree algorithm and again with an edge traversal algorithm using a TIN generated from the same DEM, we assessed the discrepancy between the two techniques. Approximately 19.7% of the 65,000 LOS queries resulted in a discrepancy between the two algorithms, which was probably due to data lost in the TIN description of the terrain. There was a 4% loss in performance of the overall simulation when running the new algorithm, but this was probably due to the configuration of our systems and delays in our computer network.

Borrowing from concepts discovered in the field of ray tracing, we have developed a new algorithm for evaluating LOS. By using a quadtree data structure, the quadtree algorithm can rapidly search for the in-

tersection of the LOS and the terrain. We have reduced the complexity of this search from linear time to logarithmic time while only increasing the size of the data structure by 33%, giving an enormous efficiency gain for very large maps. With the future development of compression and memory management techniques that exploit the new quadtree data structure, we believe that it will be possible to use extremely large high-resolution DEMs, which would make it efficient for small portable computers to easily evaluate line of sight on the fly. A computer with a DEM and this algorithm might some day direct a soldier out of harms way, help a military commander plan a strategy, or assist civilian communications engineers in determining the optimal locations for cellular antennas. 

*Joseph E. Gonzalez is a second year undergraduate in Computer Science at the California Institute of Technology. He would like to thank his mentor Robert Chamberlain, his colleagues at JPL, Professor Alan Barr, and the SURF program.*

## FURTHER READING

1. T. Kay, J. T. Kajiya. Ray Tracing Complex Scenes. *ACM SIGGRAPH* 20(4), 269-268 (1986).
2. R. F. Richbourg, R. J. Graebener, T. Stone, K. Green. Verification and Validation (V & V) of Federation Synthetic Natural Environments. Proceedings of the 2001 Interservice/Industry Training, Simulation, and Education Conference (2001).
3. S. M. Rubin, T. Whitted. A Three Dimensional Representation for Fast Rendering of Complex Scenes. *Computer Graphics* 14(3), 110-116 (1980).



# TECHNOSPHERE NO. 3, NO. 5

BY SETH DRENNER



*Technosphere no. 3, no. 5*  
mixed media, 24" x 36" panels  
Artist's Private Collection



Science helps us understand issues concerning our world through historical patterns, current conditions and physical processes. Our imagination can build on this science an outlook for our future, and it can be the work of an artist to illustrate a particular scenario. These images are meant to illustrate high-technology solutions for our utopian vision, one in which we would transcend today's industrial culture without damaging our planet or ourselves.

Seth Drenner  
Art Center College of Design





**THE WORLD IS GETTING SMALLER.**

**WHICH LEAVES MORE ROOM FOR YOUR IMAGINATION.**

Neil Segil

139-74

To us, imagination has always represented the most exciting frontier. At Northrop Grumman, we use the power of imagination to push our defense and aerospace capabilities years into the future. With projects ranging from the James Webb Space Telescope to Space Based Infra-Red and Millimeter Sensor Technology to the Joint Strike Fighter, we think there are plenty of areas left to explore. Join us and discover a place where the adventures are just beginning. To view our current opportunities, please visit our website at: [www.northropgrumman.com](http://www.northropgrumman.com). U.S. Citizenship is required for most positions. An Equal Opportunity Employer M/F/D/V

***NORTHROP GRUMMAN*** DEFINING THE FUTURE™

[www.northropgrumman.com](http://www.northropgrumman.com)

© 2003 Northrop Grumman Corporation



**Integrated  
Systems**



**Space  
Technology**



**Electronic  
Systems**

