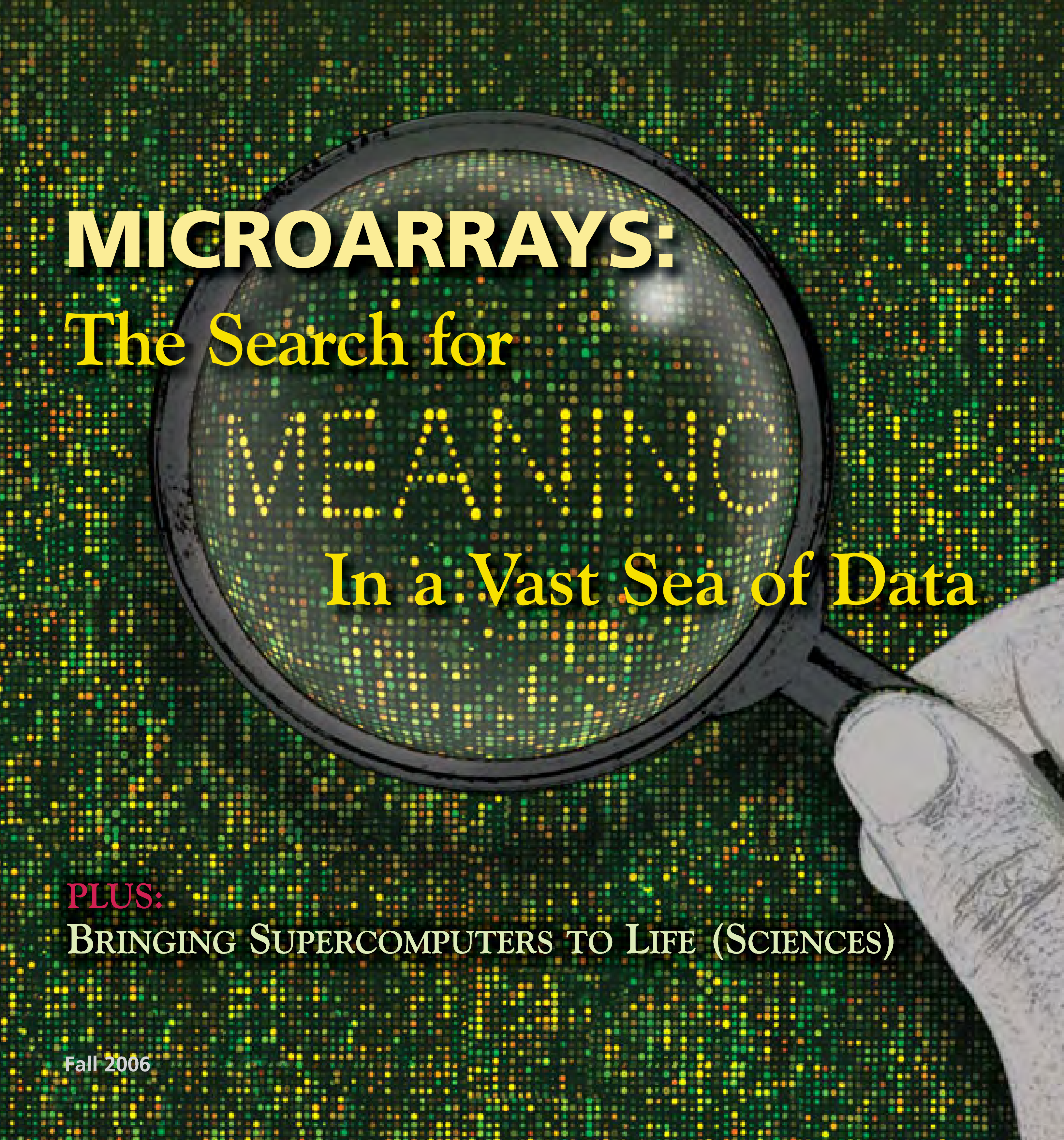
DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

MICROARRAYS: The Search for Meaning In a Vast Sea of Data



PLUS:

Bringing Supercomputers to Life (Sciences)

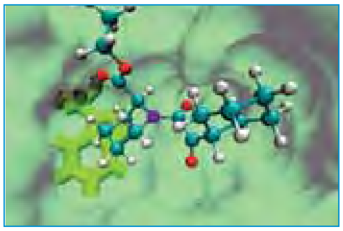
Fall 2006

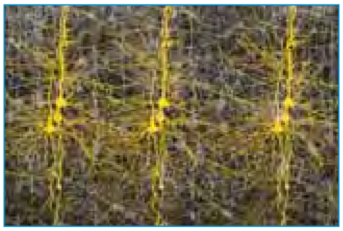
contents

ContentsFall 2006

Correction:
In the Summer 2006 issue of *BCR*, the Human vs. Machine feature story mistakenly stated that CASP only evaluates protein structure prediction methods that include human input. Prior to 2004, CASP collaborated with the former CAFASP (Critical Assessment of Fully Automated Structure Prediction) in evaluating fully automatic methods. Since then, CASP provides the only blind test of fully computational structure modeling methods.

On the Cover:
Cover Art by Rachel C. Jones of Affiliated Design





Fall 2006
Volume 2, Issue 4
ISSN 1557-3192

Executive Editor
David Paik, PhD

Managing Editor
Katharine Miller

Science Writers
Kristin Cobb, PhD
Louisa Dalton
Hannah Hickey
Katharine Miller
Regina Nuzzo, PhD

Community Contributors
David Paik, PhD
Zachary Pincus

Layout and Design
Affiliated Design

Printing
Advanced Printing

Editorial Advisory Board
Russ Altman, MD, PhD
Brian Athey, PhD
Andrea Califano, PhD
Valerie Daggett, PhD
Scott Delp, PhD
Eric Jakobsson, PhD
Ron Kikinis, MD
Isaac Kohane, MD, PhD
Paul Mitiguy, PhD
Mark Musen, MD, PhD
Tamar Schlick, PhD
Jeanette Schmidt, PhD
Michael Sherman
Arthur Toga, PhD
Shoshana Wodak, PhD
John C. Wooley, PhD

For general inquiries, subscriptions, or letters to the editor,
visit our website at
www.biomedicalcomputationreview.org

Office
Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

Biomedical Computation Review is published quarterly by Simbios National Center for Biomedical Computing and supported by the National Institutes of Health through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Jennie Larkin, PhD (NHLBI)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Charles Friedman, PhD (NLM)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Wen Masters, PhD (NSF)
Richard Morris, PhD (NIAID)
Grace Peng, PhD (NIBIB)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)

From The Editor

DAVID PAIK, PhD, EXECUTIVE EDITOR

A Review of the *Review*

This, the sixth issue of this magazine (and the final issue of Volume 2), provides a good opportunity for reflection on where we've been and where we're headed. Three years ago, the NIH requested applications to establish National Centers for Biomedical Computing. Each center would be required to "disseminate software, data and new discoveries to the national community." But, the NIH specified, "[journal] publications and a good website...may not be sufficient."

At the time, it was striking that there existed no major communication media—conferences, journals or magazines—devoted to the entire gamut of biomedical computation. To fill that gap, we proposed launching a new magazine, *Biomedical Computation Review*, which would specifically aim to communicate ideas among a very diverse group of researchers who share a common interest in biomedical computing, whether at the molecular, cellular, organism or population scale. Opening up new channels of communication could only lead to improving the rate of cross-fertilization of ideas between fields.

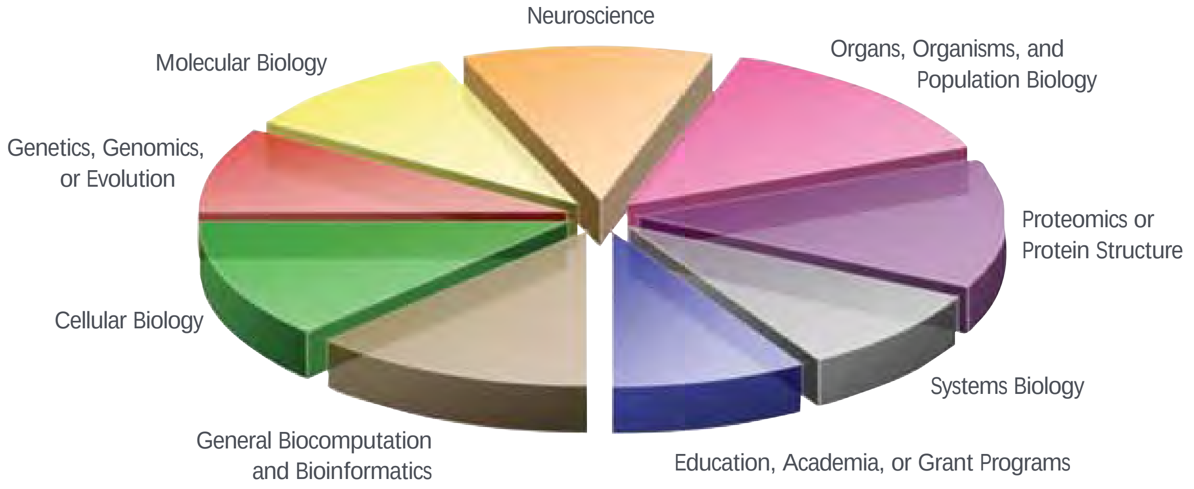
After only six issues, we've covered a very broad swath of biomedical computing. The six tables of contents read like a veritable smorgasbord of scientific fields. We've selected a mix of cross-cutting issues and in-depth articles that includes 12 feature length articles, 37 news articles, seven editorials/guest editorials, three editor's picks, six Under the Hood mini-tutorials, one book review, five Seeing Science articles exploring the boundaries of biomedical computation and the arts, and three Featured Labs. In those articles, *BCR* has highlighted the work of 274 different researchers from 109 institutions (about one in five outside the United States).

About 85% of the magazine content is written by professional science writers and the other 15% by community researchers. The editorial board consists of 15 members and the 12 program and science officers for Simbios provide additional valuable input and direction. Affiliated Design of Livingston, Montana, directs the magazine's layout and design, and our debut issue (June 2005) recently won an award for Excellence in Communication and Graphic Design from *Graphic Design USA*.

About 3,000 people receive the magazine in print and the website has had even more unique visitors from all over the world, with the numbers increasing every month.

As we look forward to future issues of *BCR*, we have lots of great new topics to cover. There's so much interesting material to cover in a limited number of pages so please bear with us until we get a chance to highlight the research areas that are nearest and dearest to you. Biomedical computing is advancing so rapidly on so many fronts and so we look to you, our readers, for additional input and guidance. Let us know what we're doing right or where we might do better. And proposals for written contributions are welcome. □

Topics Covered by *BCR*



(Feature stories that significantly touched on more than one field are counted more than one time.)

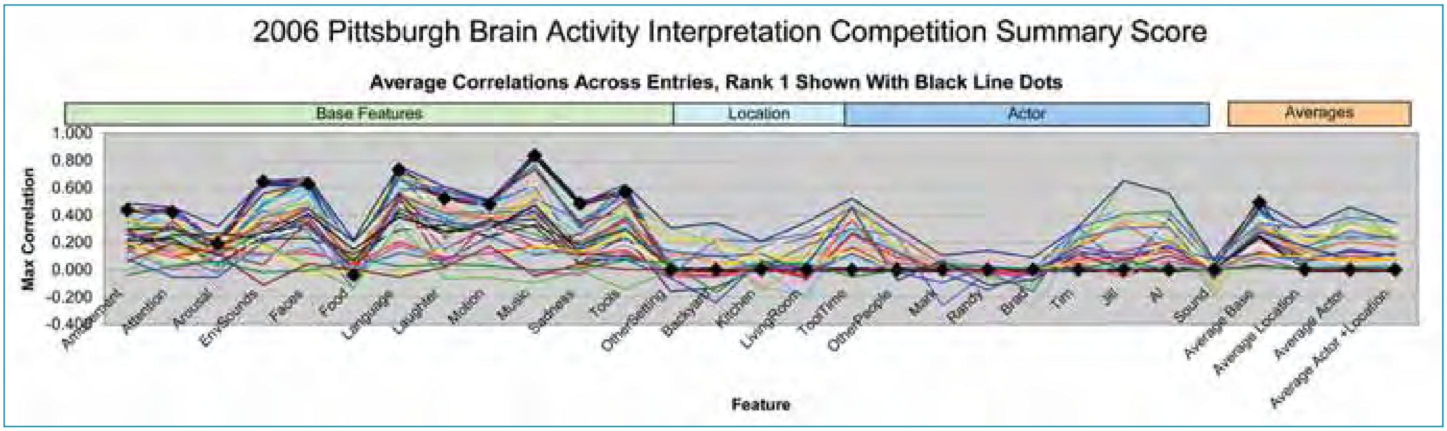
NewsBytes

Neurocomputation of Music, Faces and Belly Laughs

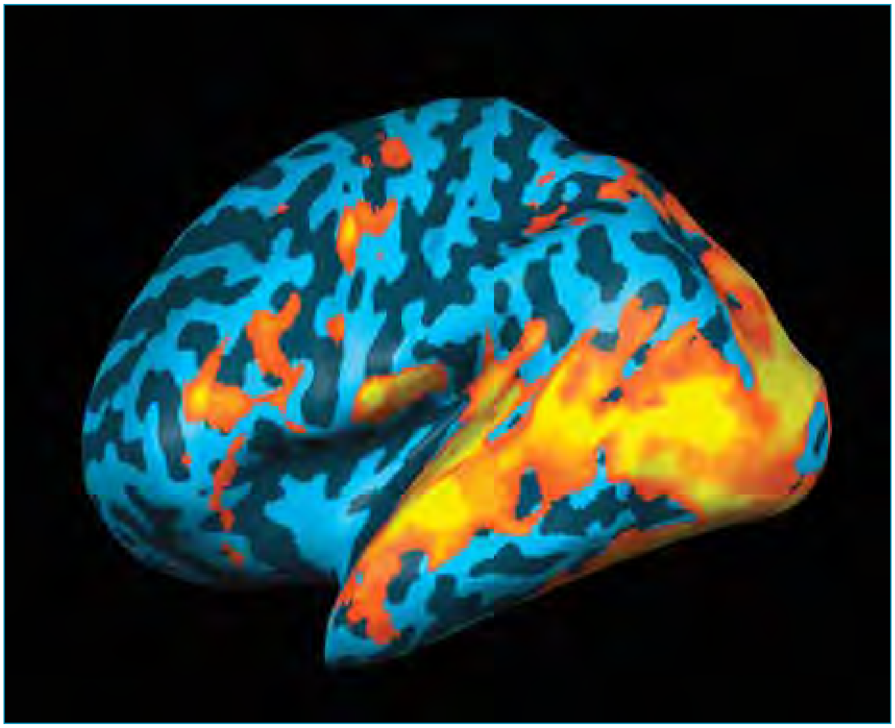
Peek inside the skull of a couch potato watching reruns on TV and you'll see non-stop patterns of blood flow throughout the brain. If you learn to pick out which activity patterns match up with, say, a good belly laugh, then you might be on your way to reading the viewer's internal experiences. Recently, experts from a variety of fields competed to glean subjective perceptions like humor from functional MRIs of TV viewers. They were surprisingly successful.

"Our goal is to know how the brain represents information," says Walter Schneider, PhD, professor of psychology at the University of Pittsburgh and principal investigator of the Experience Based Cognition group, which sponsored the competition. "In theory, if we can understand the information in the activity of somebody's brain, then we can understand what they perceived."

In the competition, 40 teams of researchers from nine countries developed pattern-classification methods for interpreting fMRI data. They used training data derived from three volunteers watching scenes from two episodes of "Home Improvement" 14 times each—



"In theory, if we can understand the information in the activity of somebody's brain, then we can understand what they perceived," says Walter Schneider.



Above: Using brain activity patterns like those shown here (derived from fMRI), the winning team (Veeramachaneni's group) had the highest weighted average correlation for the prediction of various features of the volunteers' subjective viewing experience. Top: On the graph, the correlations of the team's predictions are shown in bold black; other teams' scores are shown in color.

once in an MRI scanner and 13 times while reporting their perceptions. The teams tested their methods on fMRI images of the volunteers watching a third set of scenes from the TV show. The goal was to decipher each individual's brain activation patterns and then describe his or her TV-watching experience in a way that would closely match the volunteer's real-time impressions. Winners were announced in June at the Organization for Human Brain Mapping meeting in Florence, Italy.

Overall, predictions were remarkably accurate, Schneider says. The easiest patterns to pick out in the fMRI data were those that occurred when volunteers heard background music. The top group's prediction for music perception was "almost right on top" of the volunteers' own ratings, he says, with an average correlation of 0.84. Patterns for faces, language, and environmental sounds were also generally easy to detect, and some groups excelled at identifying when the volunteers recognized specific actors in the scenes. On the other hand, nearly all groups stumbled at figuring out when food was visible on the screen. Perhaps the mere sight of food doesn't evoke strong signals in the brain, Schneider says, "although one subject did skip lunch, and we got better responses for him."

The top group, led by **Sriharsha Veeramachaneni, PhD**, a researcher at the Center for Scientific and Technological Research at the Istituto Trentino Di Cultura in Italy (ITC-IRST) with a background in computer engineering, built a model with recurrent neural networks. Despite knowing "practically nothing" about analyzing brain images, Veeramachaneni says, the researchers soon realized they could treat these signals as generic data for purposes of this competition.

The second-place team, led by **Denis Chigirev**, a physics doctoral student at Princeton University, concentrated on extensive preprocessing of the data across space and time—an approach that reflects the group's perspective. "Physicists pay careful attention to what is signal and what is noise," Chigirev says. "We wanted to let the signal tell us what to do."

Alexis Battle, a computer science doctoral student at Stanford University, led the third group which explicitly modeled correlations in the dataset. "We thought about the relationships in the data that we could exploit," Battle says. "We chose to encode the relationships in a formal probabilistic framework."

Schneider is already "playing matchmaker" to help facilitate new multidisciplinary collaborations next year. According to **Daphne Koller, PhD**, professor of computer science at Stanford University and principal investigator for Battle's team, "The fMRI field is at the point that genomics was 10 years ago. There's a tremendous opportunity now for us to integrate computational methods with the understanding that's being developed by the brain scientists."

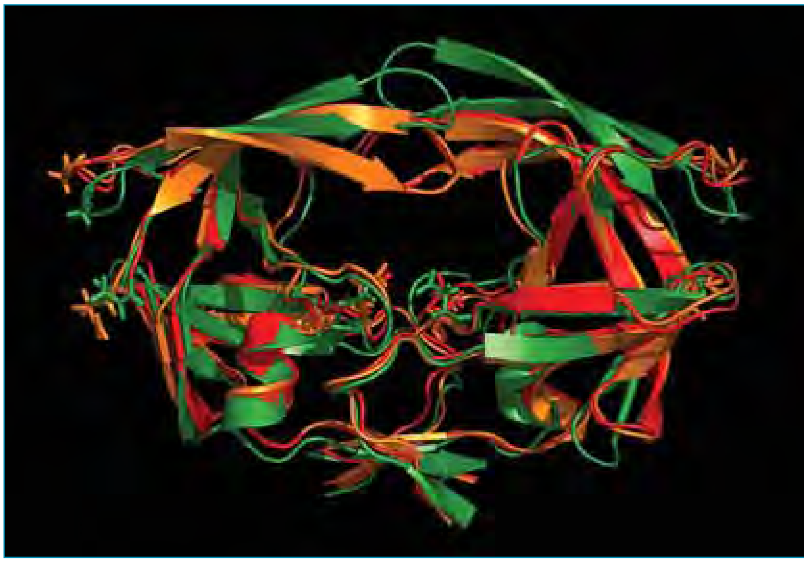
—Regina Nuzzo, PhD

Simulations Find Possible HIV Achilles' Heel

A blindside attack on HIV-1 protease might just combat drug-resistant strains of HIV, according to simulations run by researchers at the University of California, San Diego. When the simulations shut down an exposed movement on the side of the enzyme, the active site shut down as well. The work was published in *Biopolymers* in June 2006.

HIV-1 protease is an indispensable workhorse of the HIV virus: It cuts viral protein chains into building blocks ready for assembly into new virus particles. Many of today's anti-HIV drugs target this enzyme, generally by plugging up its active site and permanently closing two flaps over that area. In HIV strains resistant to these drugs, HIV-1 protease developed flaps

Perryman and his colleagues suggest designing drugs to target flap movement on HIV-1 protease instead of (or in addition to) the protein's active site.



A new target for anti-HIV drugs may be the allosteric grooves on the side of HIV-1 protease (see gaps in the middle of the right and left sides). When those are pinched together (see green protein, right and left sides), the flaps over the active site (top) can open. The flaps remain closed when the groove is propped open (red and orange versions). Courtesy of Alexander Perryman.

that are harder to latch shut. So now some researchers are suggesting targeting flap movement instead of (or in addition to) the active site.

That's why Alexander Perryman, PhD, now a postdoctoral fellow at California Institute of Technology, Andrew McCammon, PhD, professor of theoretical chemistry and pharmacology at UCSD, and their coworkers were very curious when they noticed an interesting movement on a side surface of HIV-1 protease in molecular dynamics simulations performed in 2004. When the protease closed its flaps across the active site, a groove on the peripheral surface expanded. Conversely, as the active site flaps opened, that same groove, called the allosteric groove, shrunk. It looked as if the movements were directly linked.

So the researchers hypothesized that inhibiting the movement of the allosteric groove would inhibit the movement of the active site flaps as well. In simulations that invoked an imaginary force or drug acting on the allosteric groove, they found their hypothesis was correct. When the allosteric groove is propped open by an imaginary drug, the flaps that guard the active site stay closed. And when the groove is pinched together slightly, these flaps will open.

It is still entirely unknown whether an actual drug exists, or could be created, that would apply the same force as the imaginary drug in the UCSD simulations. Celia Schiffer, PhD, associate professor of biochemistry and molecular pharmacology at the University of Massachusetts Medical School, thinks the groove movements are important for protease function, yet she is not convinced that the allosteric groove is a viable drug target. "I think practically that would be a very difficult place for inhibitors to bind in a specific and high-affinity manner."

But Carlos Simmerling, PhD, associate professor of chemistry at State University of New York, Stony Brook, is impressed by the UCSD strategy of finding a new drug target by observing enzyme movement. "The idea of targeting the mechanism is a lot more powerful than targeting the shape of the binding pocket, which is what current drugs do," he says.

—Louisa Dalton

Lung Tumors Recap Developmental Patterns

Researchers have long speculated that many of the genetic programs responsible for rapid growth of tumors are also important for the growth that occurs during normal embryonic development.

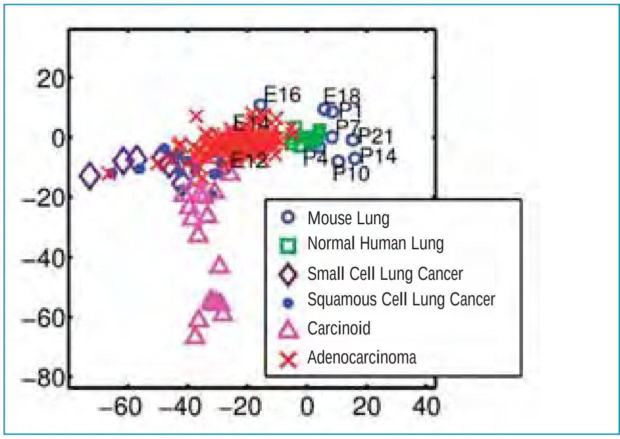
Now, researchers at the Children's Hospital Informatics Program at Harvard have found not only a relationship between tumors and lung development, but also a trend: The tumors with genetic profiles that resemble early lung development are deadlier than those with profiles that resemble later lung development. Separating out the least aggressive tumors from the more dangerous ones might help some lung cancer patients avoid unnecessary toxic chemotherapy. The work was published in *PLoS Medicine* in July 2006.

"Until now, lung cancers were classified through clustering of gene expression data, without seeing the trend from the point of view of development," says Hongye Liu, PhD, research fellow in the Children's Hospital Informatics Program at Harvard and MIT. "But we've found that the development trend can predict which cancer is worse."

Earlier work by Liu's co-authors, Alvin Kho, PhD, and Isaac Kohane, MD, PhD, showed that the gene expression profiles for each of several different types of brain tumors form distinct clusters when projected onto the gene expression profile of mouse genomic cerebellar development. The work by Liu and colleagues confirms these findings in the lung cancer context and takes them one step further by finding a connection between tumors, development and prognosis.

Charles Powell, MD, professor of clinical medicine at Columbia University College of Physicians and Surgeons, says Liu's work is important in emphasizing the link between cancer and development, but prognostic indicators in this paper need to be tested prospectively. More interesting, he says, is the potential

Tumors with genetic profiles that resemble early lung development are deadlier than those with profiles that resemble later lung development.



Principal components of gene expression data for mouse lung and normal human lung compared to that of various types of human lung cancer. The mouse lung development profile (blue dots) marches to the right over time. The most malignant forms of lung cancer (small cell lung cancer) more closely resemble early lung development in the mouse, while the least malignant forms (adenocarcinomas) more closely resemble later lung development in the mouse and normal human lung tissue. Carcinoids (purple triangles) are known to be quite different from the other types of cancers and have a pattern of gene expression that clusters perpendicular to and below the others. Carcinoids can look like small cell lung cancer under a microscope, but the two types of cancer require different treatments. This gene expression tool might help to distinguish them.

for insight into the origins of lung cancer. “The steps that transform a damaged cell into lung cancer of one type or another are likely to be similar to normal development in the lung,” he says. “If we can follow-up this paper to understand those steps then we should be able to discover novel insights into lung carcinogenesis.”
—Kathy Miller

Chain “A” of the protein Ubiquitin Hydrolase, which contains the most complicated knot that Grosberg and Lua found in a protein. It has a knot with at least five crossings in it when viewed as a flat object. Courtesy of Rhonald Lua.

Proteins in Knots? NOT!

When you accidentally twist a shoelace, garden hose, or necklace, it can get annoyingly tangled into intractable knots. On the microscopic level, biopolymers—string-like molecules such as DNA—also form knots, with one mysterious exception: knotted proteins are rare. Physicists have now used computational methods to quantify just how rare in the May 2006 issue of *PLoS Computational Biology*.

“We found that the proportion of proteins with knots is several orders of magnitude smaller than chance would predict,” says author Alexander Grosberg, PhD, professor of physics at the University of Minnesota. “The degree of it is spectacular.”

To envision a knot in a protein, imagine grasping the ends of an amino acid chain (the N-terminus and C-terminus), one end in each hand, and then stretching it out. If you can’t stretch it into a straight line, then it contains a knot.

“We found that the proportion of proteins with knots is several orders of magnitude smaller than chance would predict,” says Alexander Grosberg.

Of course, finding knots in real proteins requires a computer rather than a pair of hands. Grosberg and his co-author, postdoc Rhonald Lua, PhD, developed a knot-detecting algorithm that they used to scan 4716 proteins with known shapes from the Protein Data Bank. They found only 19 proteins (0.4 percent) with knots. Bolstering their findings, two other groups (from MIT and Italy) independently arrived at almost the same list of knotted proteins (they missed two of Grosberg’s).

Grosberg and Lua next set out to quantify how often proteins would be expected to form knots if only chance was at work. They simulated the shapes of random polymers with chains of equal length, density, and flexibility as proteins using a statistical technique—random walk on a lattice. Starting at a single point, this algorithm draws a path in three dimensions by randomly moving one unit at a time in one of six possible directions: up, down, forward, backward, right, or left. The end result is a randomly crinkled chain that may or may not contain knots. The proportion of these random polymers with knots trounced that found in real proteins: Simulated polymers at lengths of a typical protein (200-500 amino acids) formed knots 15-60 percent of the time.

Marc Mansfield, PhD, a professor of chemistry and chemical biology at the Stevens Institute of Technology, did pioneering work on knotted proteins in the early 1990s. He says the researchers’ method of generating random polymers produces some bias, but the bias did not significantly affect the result and had no impact on the study’s overall conclusions.

As to the mystery of why proteins avoid knots, Grosberg says “it has to be a product of evolution.” Mansfield agrees: “My money is still on the explanation that a knotted protein just would not fold well, so nature doesn’t use them.”
—Kristin Cobb, PhD

Simulating Wheelchair Posture

Implanting electrodes into paralyzed torso muscles can help individuals with spinal cord injury balance in their seats. So say researchers at Case Western Reserve University, who have built a three-dimensional biomechanical model that predicts how effectively functional electrical stimulation (FES) stabilizes seated postures.

In 2003, the late actor Christopher Reeve received implanted electrodes for FES to help him breathe, and various other

For those with spinal cord injury, hooking one arm over an armrest for stability is a common strategy to maintain balance when reaching. Courtesy of Cleveland FES Center.

types of FES are under investigation for help in bowel and bladder control, coughing, walking, and standing. However, relatively little attention has been paid the subtle muscle movements of torso stabilization required for balanced, steady sitting, says **Ari Wilkenfeld, MD, PhD**, first author of the study that appeared in the March/April issue of the *Journal of Rehabilitation Research & Development*.

A stable seated position means being able to reach with one or both hands and not fall over, Wilkenfeld says. A healthy posture also prevents skeletal deformities, pressure wounds, and too much pressure on internal organs.

The Cleveland group's model of the human torso simulates how three muscle groups work in synergy to rotate the spine and bend it forward and sideways. Knowing from previous research that a paralyzed muscle stimulated by FES produces, at most, about 50 percent of the force of a non-paralyzed muscle, Wilkenfeld, along with investigators **Ronald Triolo, PhD**, and **Musa Audu, PhD**, at the Cleveland FES Center, used the model to calculate the largest range of stable movement that a paralyzed torso could attain under ideal FES.

They found that with the help of FES, paralyzed individuals can hold the weight of one or two bricks at arm's length, bend forward enough to extend their reach by almost a foot, and bend to the side a bit more.

In addition to creating the model, the Cleveland researchers compared its predictions to the actual sitting of a test volunteer with one pair of implanted spine electrodes. They found that one pair is not ideal because it does not fully activate even one of the sets of muscles. Yet they found that the model describes seated postures well.

"It is a promising start," says **Jason Gillette, PhD**, an assistant professor who specializes in biomechanics and motor control at Iowa State University. He suggests testing more individuals and expanding the tests to include active reaching, not just still postures.

Additionally, says Wilkenfeld, they'll need a more sophisticated system of FES implanted electrodes to get the kind of results predicted by the model. Yet, now that they have a model that shows two-handed reach and the stable sitting postures theoretically possible, they can work on the practical details for attaining them.

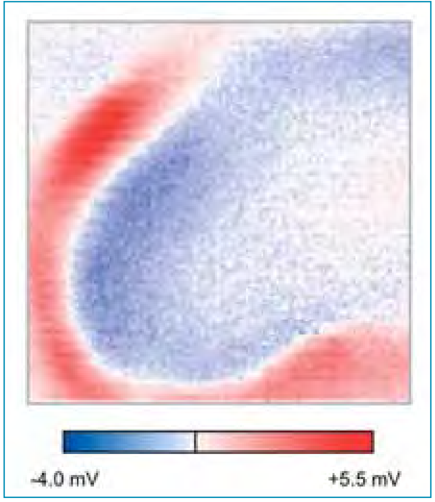
—*Louisa Dalton*

Brain Chips

Neurons are tough cells to study. There are a staggering number of them in most animals, and they are constantly talking with one another. One way to look at groups of neurons in real-time is to take a slice of brain, stimulate it electrically, and measure responses across the slice. Now a new tool may give researchers more neuronal data in the span of a few milliseconds than ever before.

A team headed by **Peter Fromherz, PhD**, a director at Max Planck Institute for Biochemistry in Munich, has developed a computer chip that can measure the activity of thousands of neurons at a time. "We can get a movie of a complete electroactivity map in space and time, with a resolution of eight micrometers," Fromherz says. The work was published in the September 2006 issue of the *Journal of Neurophysiology*.

Fromherz's group worked with Infineon Technologies in Munich to create a special 1-square-millimeter silicon chip containing more than 16,000 transistors. To prepare the device for data collection, the researchers first culture a thin slice of rat hippocampus onto the chip for a few days. Then they stimulate the slice with microelectrodes and take an electrical snapshot every half-millisecond. "Transistors in the chip measure the voltages that arise in the slice, so we can see how electrical activity propagates in the tissue," Fromherz says.



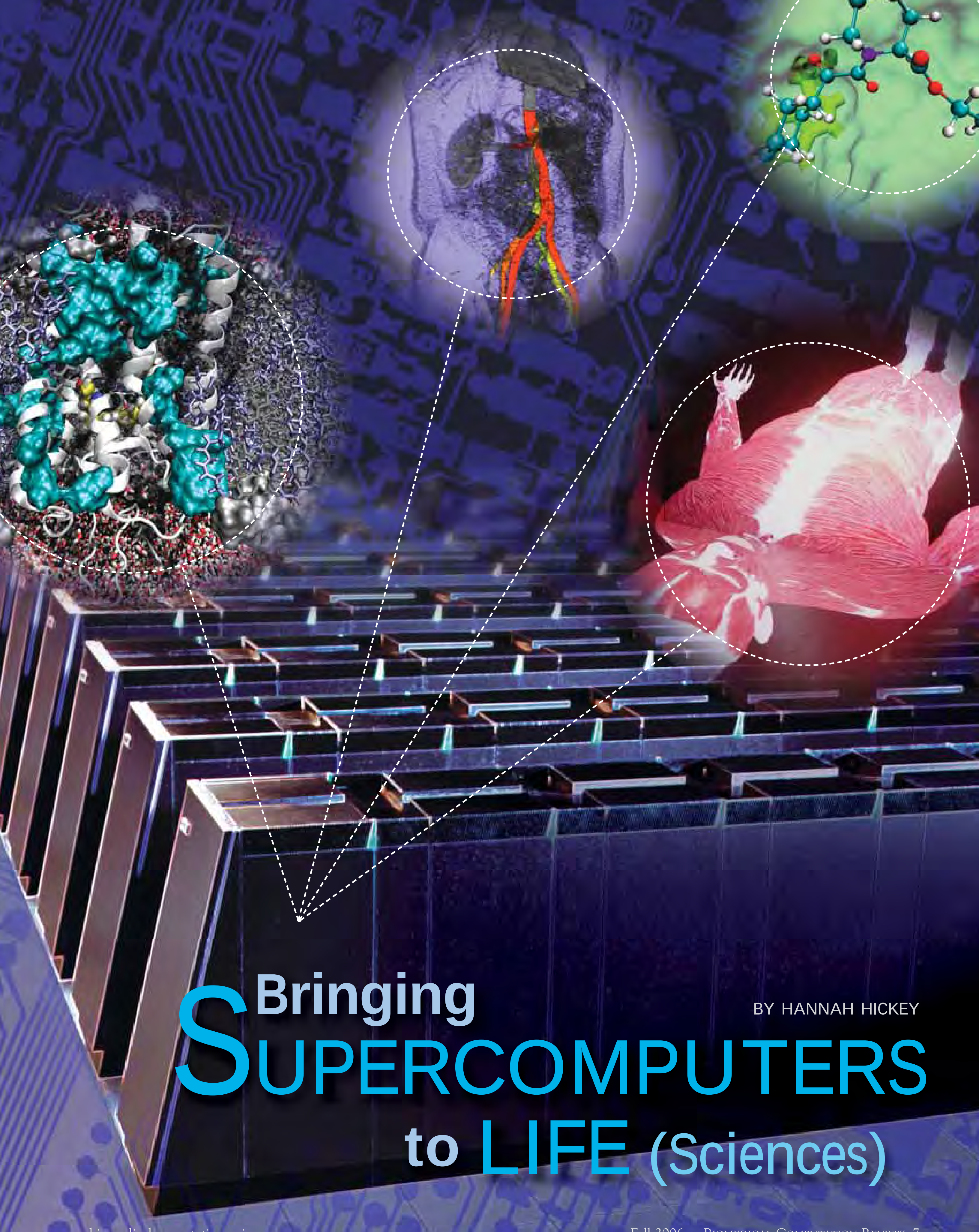
Although the chips themselves are relatively simple, Fromherz says, the computer technology behind it is rather complicated. His team is retooling the apparatus so that it can run off a PC rather than the specialized computers used now. After that, they'll work to make the entire system commercially available for other scientists.

Fromherz's long-term goal is neurocomputing, a coupling of both brain and silicon. He hopes that semiconductor technology can eventually benefit from the brain's powerful ability to store memories. "Right now, that is a little bit science fictiony, I know," Fromherz says. But Fromherz has less lofty goals for the near future. He'd like to see the brain chip help pharmaceutical researchers expand their study of drug effects on the brain by providing data on thousands of neurons at a time. And he hopes that the technology will prove useful to neuroscientists who are open to new technology. "Now the neuroscientists have a new tool, and they will need to think about completely new questions," Fromzherz says.

Indeed, it remains to be seen how useful this chip will turn out to be for brain researchers, says **Arthur Toga, PhD**, professor of neurology at the University of California, Los Angeles. "But I'm a firm believer that almost every leap forward in neuroscience has been preceded by a technological innovation, one that allows us to pose questions that couldn't be posed before," he says. "That's been true all the way from the microscope to the MRI."

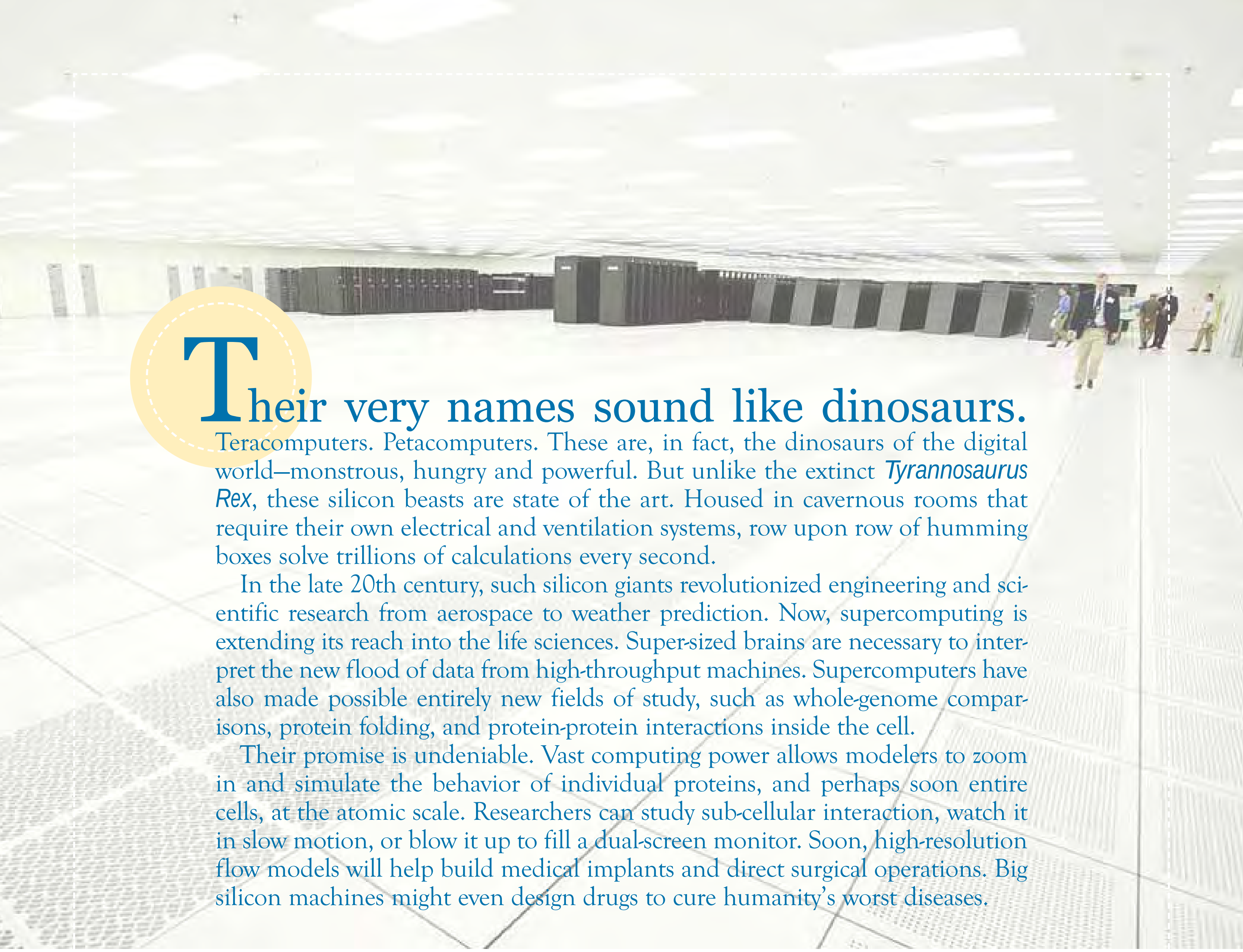
—*Regina Nuzzo, PhD* □

Fromherz and his colleagues used more than 16,000 transistors on a 1-square-millimeter silicon chip to measure field potentials from a slice of rat brain every half-millisecond after stimulation with electrodes. This image shows those potentials after 5 milliseconds have elapsed. Red regions indicate positive voltage; negative signals are in blue. The gray curve traces the structure of the **cornu ammonis** ***in the hippocampus.***



Bringing SUPERCOMPUTERS to LIFE (Sciences)

BY HANNAH HICKEY



Their very names sound like dinosaurs.

Teracomputers. Petacomputers. These are, in fact, the dinosaurs of the digital world—monstrous, hungry and powerful. But unlike the extinct *Tyrannosaurus Rex*, these silicon beasts are state of the art. Housed in cavernous rooms that require their own electrical and ventilation systems, row upon row of humming boxes solve trillions of calculations every second.

In the late 20th century, such silicon giants revolutionized engineering and scientific research from aerospace to weather prediction. Now, supercomputing is extending its reach into the life sciences. Super-sized brains are necessary to interpret the new flood of data from high-throughput machines. Supercomputers have also made possible entirely new fields of study, such as whole-genome comparisons, protein folding, and protein-protein interactions inside the cell.

Their promise is undeniable. Vast computing power allows modelers to zoom in and simulate the behavior of individual proteins, and perhaps soon entire cells, at the atomic scale. Researchers can study sub-cellular interaction, watch it in slow motion, or blow it up to fill a dual-screen monitor. Soon, high-resolution flow models will help build medical implants and direct surgical operations. Big silicon machines might even design drugs to cure humanity's worst diseases.

Supercomputing in Science: A Timeline

1950s to 1960s
The roots of supercomputing

1955 Physicists devise computer code for a global circulation model, and by the mid-1960s are using the largest available computers to run global-scale climate simulations.

1960s The term "supercomputer" enters the lexicon as IBM rolls out the 7030 (aka "Stretch") and Control Data Corporation releases its CDC 6600.

1970s to 1980s
Supercomputers integrated into climatology, astrophysics, and aeronautics

1976 The legendary Cray-1 supercomputer is installed at Los Alamos National Laboratory where it is used to simulate nuclear explosions.

1977 National Center for Atmospheric Research purchases a Cray-1 supercomputer which operates for the next 12 years running climate simulations.

Early 1980s Astrophysicists use supercomputers to simulate galaxy formation.

1980s Large-scale computing provides an alternative to wind tunnels in aeronautics research. By the 1990s, computers have virtually replaced wind tunnels.

While ordinary computers have already changed the study of life, supercomputers open up new horizons, offering the possibility of discovering new ways to understand life's complexity.

From Big Iron to Armies of Ants

To solve mammoth calculations, scientists have traditionally booked time on "Big Iron" custom machines housed at national supercomputing centers or universities. Today the landscape is shifting. These mammoth machines, though not extinct, are facing tough competition.

"What we [scientists] wanted was one computer that was much faster. What we got was a lot of computers," comments Vijay Pande.

"It used to be that the power of those machines [at supercomputing centers] was many orders of magnitude more than what anybody had access to," says **Philip Bourne, PhD**, professor of pharmacology at the University of California in San Diego and editor-in-chief of *PLoS Computational Biology*. "Now that's not true anymore—computing is really cheap." Alternatives exist in a thriving range of home-built, borrowed or networked systems. Many researchers choose to buy a cluster of off-the-shelf processors rather than wait for time on a "Big Iron" machine.

In 2003, students at Virginia Polytechnic Institute in Blacksburg, Virginia, helped build one of the world's fastest machines by assembling 1,100 PowerMac G5 processors. At the time it was the third-fastest computer in the world, and the $7 million price was a bargain compared to a retail price of more than $200 million for an equivalent big iron computer. Similar clusters continue to sprout up every year. The most recent Top500 list, a biannual tally of the world's 500 fastest computers, shows that networked, off-the-shelf processors now claim 72 percent of the positions.

Driving this trend is the frustrating evolution of supercomputers. Since the 1990s, spurred by economics, supercomputers themselves became vast assemblages of small processors. "What we [scientists] wanted was one computer that was much faster. What we got was a lot of computers," comments **Vijay Pande, PhD**, associate professor of chemistry and of structural biology at Stanford University. The world's fastest machine, IBM's Blue Gene, now incorporates a whopping 131,072 individual processors. Each one is relatively slow, even compared to what's offered in new laptops, but it's energy-efficient, which allows them to be packed into a small space without overheating.

Massively parallel machines have many downsides. For one thing, the total speed of a single processor is sometimes less important than how quickly individual processors can communicate. This shuffling back and forth of information becomes a bottleneck for the speed of the system. It also means that the entire system runs only as quickly as the slowest processor on the machine—a weakest-link rule known as Amdahl's Law.

Supercomputers today are like "armies of ants," says **Allan Snavely, PhD**, director of the Performance, Modeling and Characterization Laboratory at the San Diego Supercomputing Center. To enlist these ants, computer code will first have to be parallelized—split up into instructions that multiple processors can handle simultaneously. The difficulty of dividing up the problem means a supercomputer with 100 processors won't be able to solve a problem 100 times as fast. And today's "massively parallel" supercomputers don't just incorporate 100 processors, but thousands of processors. Running on these machines often means tweaking the code yet again, says **Mark Miller, PhD**, a

1996 to 2006 and beyond

Supercomputers extend their reach to biology

1985
The U.S. National Science Foundation establishes five national supercomputing centers: The Cornell Theory Center at Cornell University; The National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign; The Pittsburgh Supercomputing Center at Carnegie Mellon University and the University of Pittsburgh; The San Diego Supercomputer Center at the University of California, San Diego; The John von Neumann Center at Princeton University.

1996
Research Collaboratory for Structural Bioinformatics sets up shop in the UCSD Supercomputer Center.

2000
Folding@Home project launched to investigate protein folding mechanisms. It's now the world's most powerful grid computing endeavor, solving up to 200 trillion calculations per second.

2004
Blue Gene, IBM's flagship machine, outranks Japan's Earth Simulator computer as the world's fastest computer. It can sustain 36 trillion calculations per second.

2005
IBM and the Ecole Polytechnique Fédérale de Lausanne launch "Blue Brain," which aims to model the human neocortex on one of the BlueGene/L machines.

2010
National Science Foundation's targeted date to roll out a petaflop computer for science and engineering. The machine will solve a quadrillion (10^{15}) calculations per second.

biology researcher at the San Diego Supercomputing Center.

Large-scale supercomputing centers' importance will shift from renting time on computers to offering technical expertise, Bourne predicts, helping scientists run code on a parallel machine. Also, as journals increasingly require placing data in a public database, supercomputing centers can fill that void. "The ability to store large amounts of data, that value has increased dramatically," Bourne says.

Spreading the Load to Volunteer Computers

Today, many of the most crushingly difficult scientific computing problems aren't being solved in supercomputing centers or on university clusters. They're as likely to be solved in your living room. Take, for example, the quest to unlock the mysteries of protein folding: Predicting how a string of amino acids will curl up into the same structure every time is one of biology's holy grails. If we could do this, we might design drugs to fit particular targets, understand diseases of protein misfolding, and be able to visualize unknown proteins from their amino acid sequence.

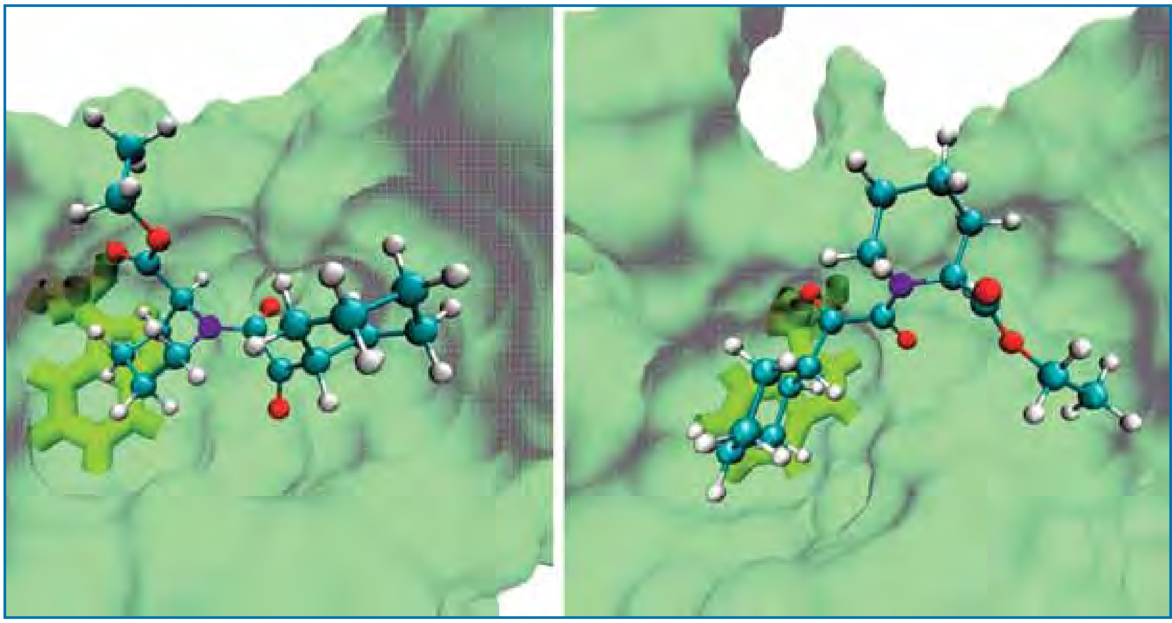
To run models of protein folding at an atomic scale requires making calculations every femtosecond—one billionth of a microsecond—in order to capture atomic vibrations. But the folding process, like many things in biology, happens much more slowly—on the order of microseconds or milliseconds. This means an atomic model of protein folding from start to finish requires a billion to a trillion steps. Also, the typical protein comprises hundreds of amino acids, each of which exerts a force on every other amino acid. Finding the lowest energy configuration for all of these amino acids is what's called an NP-hard problem. Such problems become exponentially more difficult with every extra piece of data and so approximate solutions are typically sought.

Some enterprising protein-folding projects recruit volunteers' unused PC processing time—an idea pioneered by the SETI@home project and now referred to as "grid computing."

The Human Proteome Project recently finished rough predictions for all the proteins in the human genome in a single year—a job that would have taken a century on the available laboratory cluster.

"It's probably best thought of as a supercomputer but with radically different architecture," says Vijay Pande, who leads the Folding@Home project, now the largest grid computing venture in the world. With more than 180,000 member CPUs, Folding@Home commands more raw FLOPS (floating point operations per second, a measure of computer power) than all the supercomputing centers combined—up to 200 trillion calculations per second—and transfers 50 gigabytes of data every day. Pande wants to understand the nature of protein folding to better understand why proteins sometimes misfold, causing diseases like Alzheimer's and cystic fibrosis.

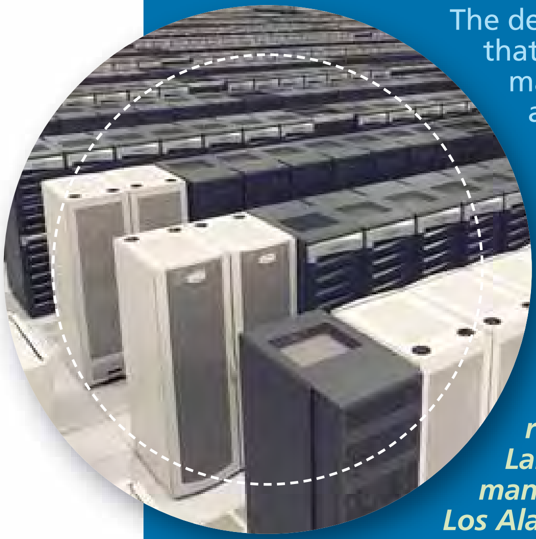
Other protein-prediction codes running in a home office near you include Rosetta@home, based at the University of Washington in Seattle, which predicts structures for proteins of unknown function; Predictor@home, based at the Scripps Research Institute in San Diego, which compares different structure-prediction algorithms; and the Human Proteome Project, out of



With collaborators at Fujitsu, Folding@Home published results showing the initial modeled structure of a protein that is the target of immunosuppressive drugs (FKBP) in complex with a small molecule ligand (left); and the final structure after a 20 nanosecond simulation (right). In this and other work, Folding@Home has demonstrated that atomistic models of biologically relevant systems can be calculated with a useful level of precision and accuracy by bringing several orders of magnitude more computational power to the problem. This work is allowing important advances in rigorous physical drug-binding prediction. Courtesy of Hideaki Fujitani, Fujitsu.

What's a supercomputer?

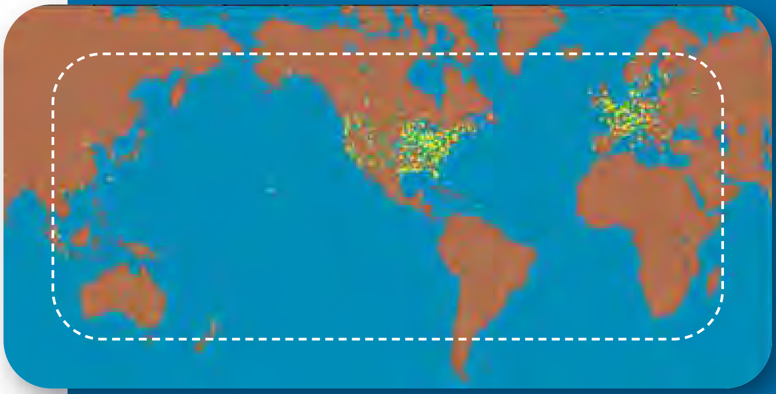
The definition of "supercomputer" is fluid—it just means a machine that's among the world's fastest. Not only is the world's fastest machine always changing, but so is the architecture for creating a supersized computer.



◀ **"BIG IRON"** supercomputers are the traditional supercomputers: custom-built machines housed in refrigerator-like boxes. They first emerged in the 1980s, produced by Cray, Inc. These custom supercomputers still lead the Top500 list of the world's fastest machines. Because they share information and data quickly between processors, they can tackle the most complex problems. ***IMAGE: The "Q" supercomputer, used by researchers at Los Alamos National Laboratory to simulate a ribosome manufacturing a protein. Courtesy of Los Alamos National Laboratory.***

▶ **CLUSTERS** connect tens, hundreds, and in some cases thousands of off-the-shelf PCs. Software codes, typically written in LINUX, provide communication. These are sometimes called "PC farms," or "Beowulf clusters," after the first systems of this type. Clusters are a much cheaper way to boost computing power. ***IMAGE: Photos of a team assembling the 1,100-processor cluster at Virginia Polytechnic Institute in 2003. Courtesy of Ken Wieringo, VPI.***

AN IN-HOUSE NETWORK (not pictured) is created when an organization connects its computers together, letting users borrow each others' computing power. Such a system is a type of in-house "grid" in analogy with the electrical grid, which shares a resource between many intermittent users. Many businesses, including pharmaceutical companies, digital animation studios and financial-investment firms, have networked employees' desktop machines to create an in-house supercomputer, essentially for free.



◀ **GRID COMPUTING** uses unrelated computers to solve pieces of a giant calculation. Volunteers sign up over the Internet to donate their unused processing cycles. SETI@Home, the pioneer, is still scanning radio waves for signs of intelligent life. Other projects predict the effects of global warming (Climateprediction.net), look for prime numbers (Great Internet Mersenne Prime Search) or detect gravitational waves from spinning neutron stars (Einstein@Home), to name a few. Biology projects include Folding@Home, fightAIDS@Home, and the United Devices Cancer Research Project. CERN plans to use this architecture to store and analyze data from the Large Hadron Collider beginning in 2007. ***IMAGE: Computers all over the world are working on the protein-folding problem. This map shows the distribution of IP addresses as of November, 2004. Courtesy of Vijay Pande, Folding@Home.***

the Institute for Systems Biology in Seattle, which predicts structures for human proteins. In summer 2006 CERN, in Geneva, announced a project to study malaria on the grid, and Israeli scientists hope to map genetic diseases.

The benefits of such a scheme are obvious. The Human Proteome Project recently finished rough predictions for all the proteins in the human genome in a single year—a job that would have taken a century on the available laboratory cluster. Buying equivalent computing time for Folding@Home from a company like Sun Microsystems would cost $1.5 billion a year, Pande says.

But it's an open question how many codes will work on a motley collection of home computers, accommodate unpredictable run times, and tolerate infrequent communication. Problems that work best on the grid are the ones that don't require a lot of back-and-forth communication. SETI@home is a classic example; each user runs the same pattern-recognition algorithm on a different chunk of radio-wave output. In geek speak, this is an "embarrassingly parallel problem"—one that can easily be split into independent tasks on many processors.

Embarrassing or not, many biological computing problems may eventually become parallelized. "In biology you're looking at a very large number of small bits of data," Bourne says. And clever algorithms may succeed in running even complex problems on the grid. "Protein folding was not something that I think people would have thought could be broken up," Pande says. "My gut feeling is that there will be many things that could be suited to this type of technology."

It's a question of being on the "leading edge" of science versus the "bleeding edge," he admits. "A lot of people don't want to get cut by the bleeding edge." Many scientists are wary of investing time in a technology that's in its infancy. To ease the transition, the Berkeley Open Infrastructure for Network Computing (BOINC), which is funded by the NSF, offers free CPU-scavenging code to interested researchers. The Open Grid Form, launched in June 2006, aims to establish standards and promote grid computing in the research community. And the World Community Grid provides free coordination for distributed computing projects that have a humanitarian bent. Since its launch in 2004 the World Community Grid has hosted fightAIDS@home and the Human Proteome Folding Project.

Biological Simulations

Enthusiasm for grid computing must be tempered by realism. Some problems will never run on the grid. In particular, some large-scale simulations and visualizations are just too convoluted to split up. Every component is constantly interacting with every other part. In a recent simulation of the human heart at the San Diego Supercomputing Center, the flag-

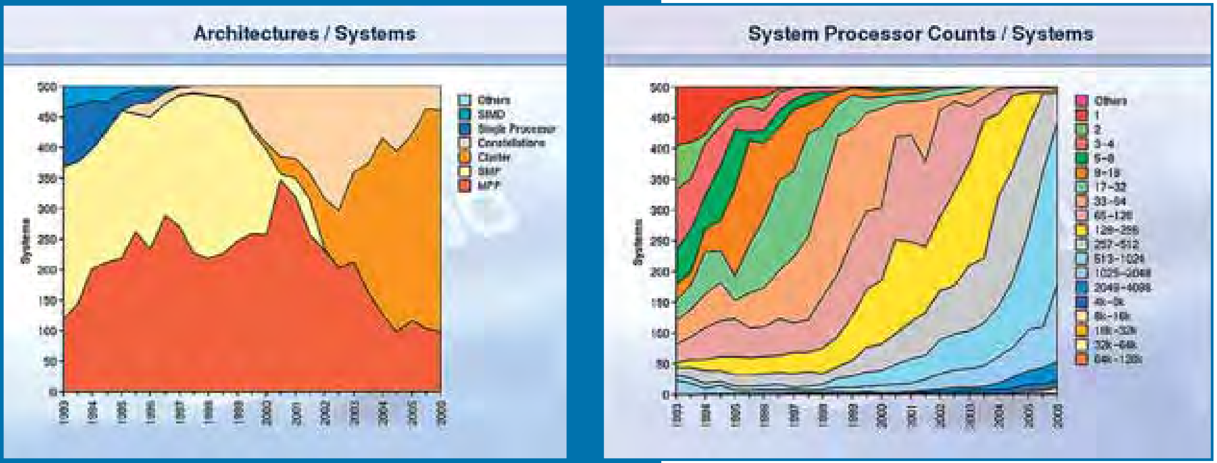
Top of the FLOPS

The widely quoted Moore's Law predicts that processing power will double every 18 months. So far the trend, attributed to Intel cofounder **Gordon Moore**, has held true. Processors continually speed up and supercomputers combine them in ever larger numbers. Today's fastest computers, including the Blue Gene machines, are at the teraflop scale—one trillion calculations every second.

But engineers already have their sights set on the next benchmark: petascale computers, which would be a thousand times faster, performing one quadrillion calculations per second. The National Science Foundation announced it would enable petascale computing for science and engineering by the year 2010. Many scientists say they could occupy a machine of that size with existing calculations.

Some question whether Moore's Law will eventually reach a limit. At some point, computers can't pack more processing power into a small space without overheating the components. On the other hand, machines can't be so widely dispersed that information, which is limited by the speed of light, takes too long to travel from one processor to another.

Quantum computers and DNA computers may someday introduce new technologies, even as today's machines reach their physical limits. "Most likely while we're sitting around debating how much further we can go with silicon computing, some genius is on the verge of a radical new invention," says Allan Snavely.



Graphs of the top 500 computers in the world showing that cluster architectures are becoming more common (left) and that they are made up of an increasing number of individual processors (right). Courtesy of Top500.org.

ship machine spent 99 percent of its time twiddling its thumbs (at a billion cycles per second) waiting to receive its neighbor's results. Running this problem on a grid, where communication takes seconds rather than nanoseconds, would be an exercise in frustration.

In 1995, fewer than one in 20 researchers using the San Diego Supercomputing Center was a biologist. By 2005, that number had quadrupled to almost one in five, and government labs are seeing a similar trend. Last October, researchers at Los Alamos National Laboratory in New Mexico completed the first biological simulation to incorporate more than a million atoms: They used Newton's laws of motion to watch the 2.64 million atoms of the ribosome manufacturing a protein. Such atomic-scale simulations allow researchers to mimic experiments *in silico*, observing processes at slower speeds or at a magnified scale. Biologists at IBM Research now use their Blue Gene machine largely for molecular dynamics applications, says **Robert Germain, PhD**, a staff researcher at IBM TJ Watson Research Center near New York City. A recent detailed simulation of the membrane protein rhodopsin, which used about a third of their machine's mammoth computation power, suggested that water molecules may play an active role in its function.

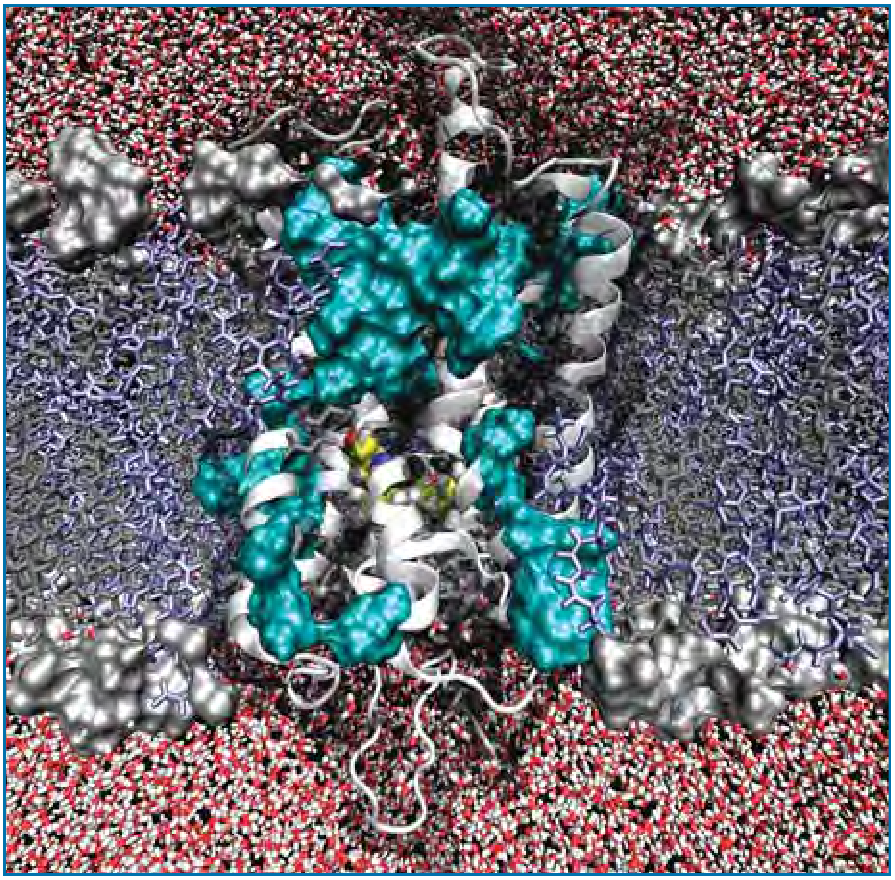
"I think we will model larger and larger biological systems," Germain predicts. He also sees the models themselves improving. While simulating a living thing is not inherently different from recreating a physical event—exploding galaxies, say, or air flowing over an airplane wing—biology has more complex structure. **Kevin Sanbonmatsu, PhD**, the Los Alamos researcher who ran the ribosome simulation, began his career in physics, but appreciates biology's challenges. When writing the code to model a ribosome, Sanbonmatsu says, he had many more types of atoms that had to be placed in specific locations than if he were modeling a semiconductor.

The toughest demands for a combination of size and speed may come from clinical practice. "We have an insatiable appetite for high-performance computing," says **Charles Taylor, PhD** , associate professor of bioengineering and surgery at Stanford University. His group solves fluid-dynamics equations that model blood flow through arteries. Beginning with a 3D image from a patient, Taylor recreates the inner workings of large arteries at millimeter-scale resolution, problems which incorporate 5 million to 10 million variables, each depending on all the others. Someday he hopes a surgeon could compare different options in the computer to decide on the best procedure for a particular patient.

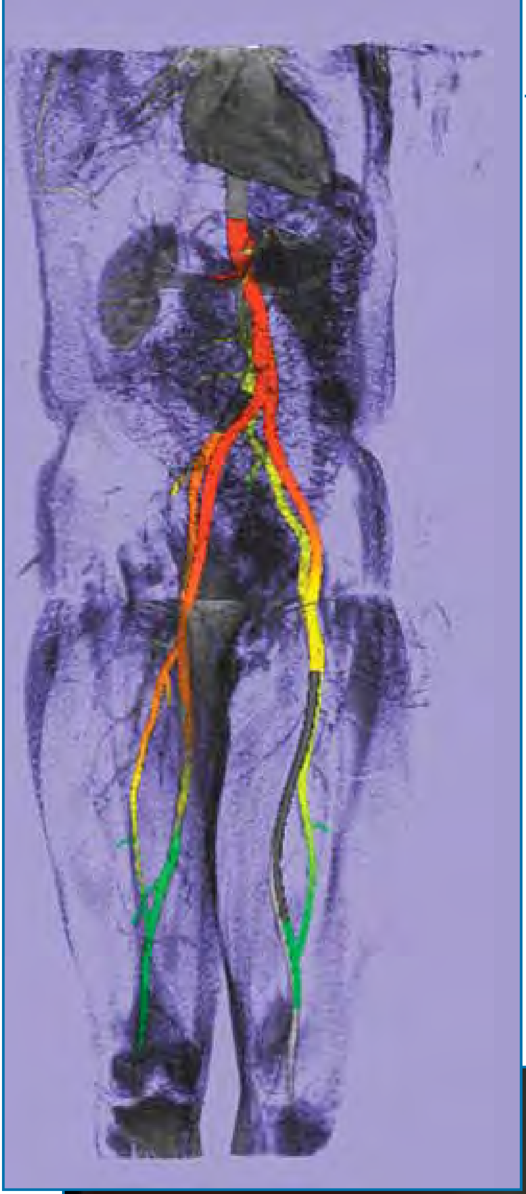
In 1995, fewer than one in 20 researchers using the San Diego Supercomputing Center was a biologist. By 2005, that number had quadrupled to almost one in five.

Unfortunately, today even Taylor's dedicated, 64-processor SGI supercomputer struggles when confronting a scenario with medical complications. An aortic arch with turbulent flow downstream requires calculating every 10 microseconds, meaning it takes 10 thousand or 100 thousand steps to complete a single cardiac cycle.

"You want to be able to turn these around really quickly," he says. Today's computers take days to run the model; doctors would like to compare multiple



IBM researchers ran molecular dynamics simulations on Blue Gene that show the protein rhodopsin (silver ribbon) interacting with specific omega-3 fatty acids in the surrounding membrane. The work suggests that fatty acids play a role in rhodopsin's function as the protein receptor primarily responsible for sensing light. This simulation ran for two million timesteps of one femtosecond (one quadrillionth of a second) each. Membrane-protein research commands one third of the Blue Gene supercomputer's nodes. Courtesy of Michael Pitman, IBM Research.



treatment options in just a few hours. The computing power necessary to do that is likely on the horizon, he says. Taylor serves on a government panel looking to integrate supercomputers in the medical device industry, the way aerospace and car manufacturers did in the past. He says, "I feel pretty confident that ten years from now, we'll look back on this time and we'll find it hard to imagine that these tools were not used in clinical practice."

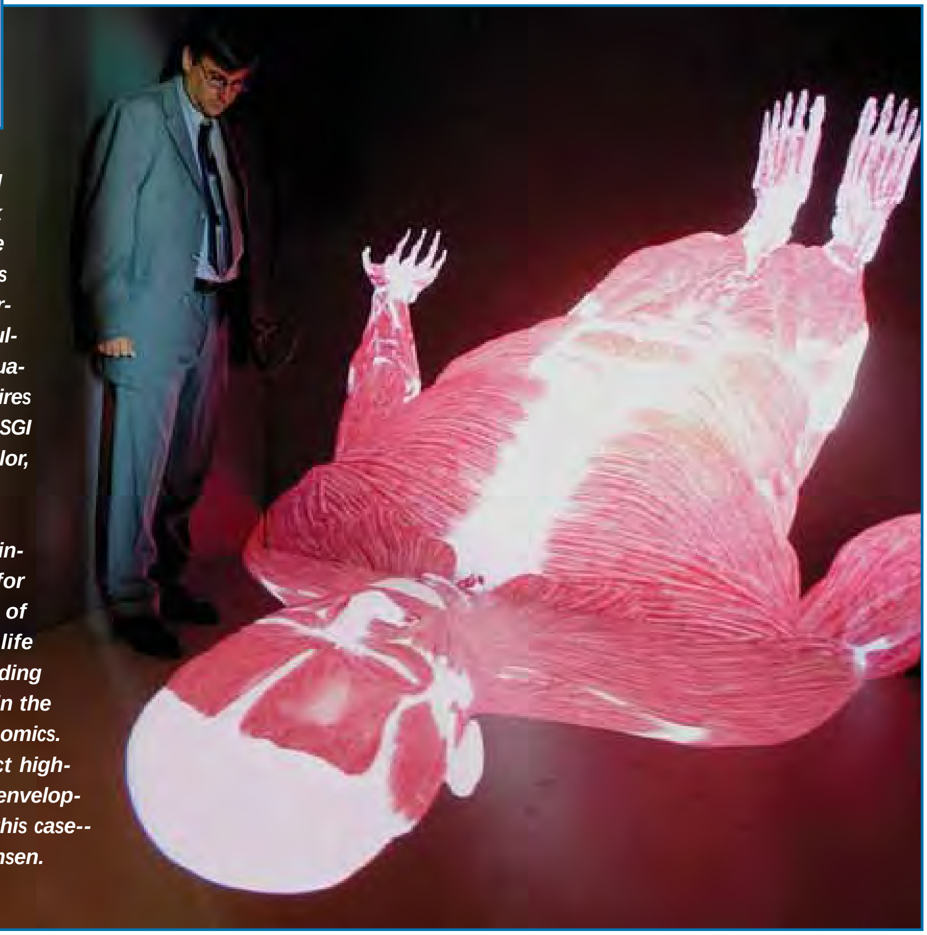
Genetics' Information Overload

Biology is seeing its databases explode. Nowhere is this more dramatic than in genetics. The vast amount of data provided by sequencing the human genome in 2003 was a turning point for biology's use of computers. Bioinformatics researchers can now comb through the sequences looking for patterns and similarities. One of the most promising techniques is whole-genome comparisons where researchers search for portions of the genome that are conserved across species, suggesting they may be important. Again, this turns out to be an NP-hard problem, demanding enormous computing power for genomes that may include billions of base pairs.

And this is only the beginning. Every year it gets cheaper to sequence more genomes.

"The amount of biological data available is increasing much faster than the increase of single processor speeds. It's going much faster than Moore's Law," says **Serafim Batzoglou, PhD**, assistant professor of computer science at Stanford University. Supercomputers will be needed to store, access and analyze this data. The first human genome took years to sequence, and cost millions of dollars. Today every few months a new genome appears. As sequencing technologies get cheaper, it's likely that within a few years we'll have hundreds of human genomes and thousands of different species, Batzoglou predicts.

"The situation has been like quicksand ever since I arrived," laments **Robert**



Above: In this fluid dynamics model of blood flow, the colors display variations in the peak systolic blood pressure from the aorta to the lower extremities. Abrupt pressure changes show regions of relative inefficiency in the circulation. This type of simulation means simultaneously solving millions of nonlinear equations and, for the finest resolution, requires days of computation time on a 64-processor SGI supercomputer. Courtesy of Charles Taylor, Stanford University.

Christoph Sensen, PhD, *professor of bioinformatics and director of the Centre for Advanced Technologies at the University of Calgary, looks down on a larger-than-life image of muscle structures. He is standing inside the CAVE, a 4D virtual environment in the Sun Center of Excellence for Visual Genomics. CAVE computers running JAVA code project high-resolution images at 112 times per second, enveloping visitors in visions of DNA, cells, or--in this case--the human body. Courtesy of Christoph Sensen.*

Petryszak, a technician who for the past three years has managed incoming sequences for the InterPro database at the European Bioinformatics Institute in Cambridge, England. "The horizons have been changing almost monthly." Petryszak adds incoming protein sequences to the database and then annotates the sequences periodically using both an in-home cluster and an external supercomputer. When biologist Craig Venter, PhD, publishes results from his shotgun sequencing project and the sequences go public, Petryszak says, it could triple the Interpro database from its current 600 gigabytes to 1.8 terabytes by the end of 2007. Storage is not a problem, but indexing the sequences and accessing the data quickly to send to users is difficult.

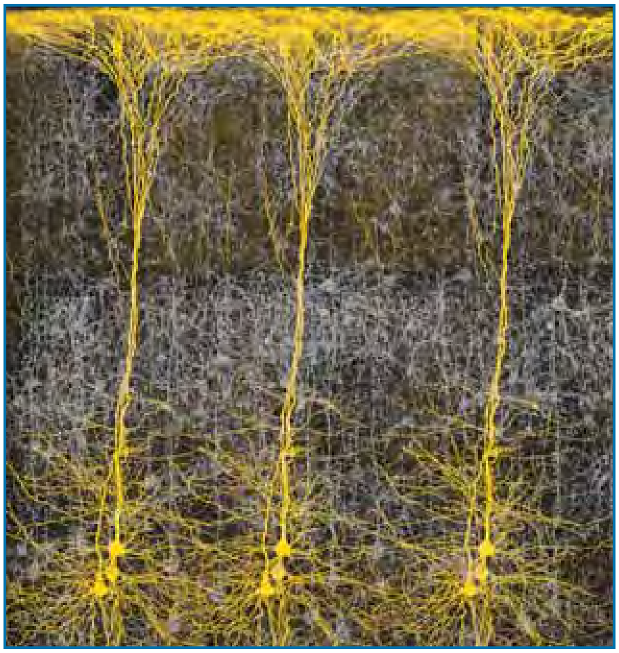
> "Biology is probably going to be the largest user of high-performance computing in the 21st century," Germain predicts.

"The amount of data is just going to be enormous," Petryszak says. "That's going to cause a headache, even for the supposedly heavyweight databases."

Biomedical Computing for the 21st Century

In biology today, supercomputing is the exception. Even computational biologists tend to solve problems using the computers they have on hand. Few dream up questions that would require more resources.

"We have a need for high-performance computing in biology, but there's no demand," says Nathan Goodman, PhD, senior research scientist at the Institute for Systems Biology in Seattle, WA. "If you go to a field like physics, people are always thinking 'What could I do if I had more computing power.' They understand that their ability to analyze data is limited by their computational power." It's a Catch-22, he says. Biologists don't have access to large computers and so they don't propose problems that would require them. Because they don't propose the problems, they don't acquire the resources. Whether it's a question of training or simply the culture of the discipline, biologists are not yet making the most of large-scale computing.



As part of the Blue Brain project, high-performance computers are being used to model the human brain. In preliminary wet-lab research shown here, researchers stained columns of neurons in the neocortex to design a detailed model of its circuitry. Each column contains 10,000 individual neurons; thousands of columns together make up the neocortex. Blue Brain researchers hope to simulate the entire neocortex. In January 2005, the team announced they had simulated 10,000 neurons on the Blue Gene/L machine, a model 10 million times more complex than any previous neural simulation. The project is a collaboration between IBM and the Ecole Polytechnique Federale de Lausanne in Switzerland. Courtesy of IBM Research.

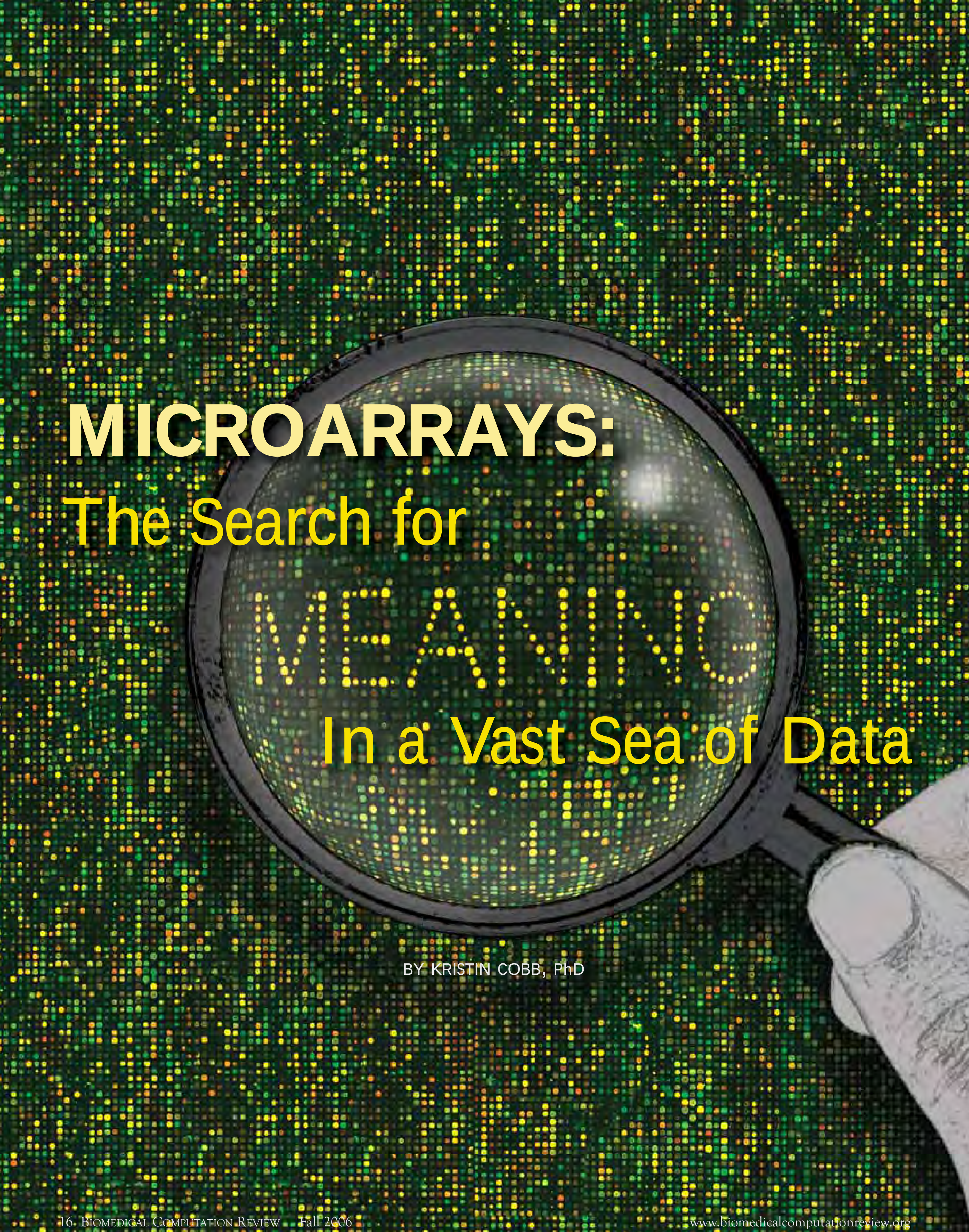
"Why daydream about something you don't have?" Pande says. "But if you give [biologists] the resource, and especially give the students access to it, then they will come up with new algorithms and new uses."

A case in point is geneticist Batzoglou, a convert to large-scale computing. Although his own background is in computer science, he initially shrugged off news that his department had acquired a 600-processor supercomputer for the biosciences. But after the machine arrived, he and his graduate students became some of the biggest users. Last summer, Batzoglou invested $55,000 in grant money to buy his own 100-processor cluster.

"Before we started using it, we didn't realize how useful it is to have such huge computing capabilities," recalls Batzoglou, who writes algorithms to analyze genetic sequences. "If there's anything we've learned it's that the more computing power we have, the more we are going to find ways to use it."

Some fields angle to capitalize on the growth in computing power. The Petascale Collaboratory for the Geosciences, an ad hoc group of scientists established in 2004, draws up questions for the upcoming generation of supercomputers. "I would love to see an analogous effort with biologists," says Snavely, a member of the task force. "To my knowledge there hasn't been this meeting of the minds that says, 'OK, if this is where the technology is going, what important biology problems do we think we could solve?'"

"Biology is probably going to be the largest user of high-performance computing in the 21st century," Germain predicts. Sure, this might sound like old news to long-time observers of the biological sciences. But hype in the early 1990s was premature—biological models were still too rough and the computing power was insufficient, says Michael Pitman, PhD, who leads the membrane protein group at the IBM TJ Watson Research Center in Yorktown Heights, New York. Finally, he says, we're nearing the point where supercomputers can live up to the hype. "I've been very encouraged by the kinds of questions we can ask and the quality of answers we're getting," he says. "I do feel that we're in a new era for supercomputers in biology." □



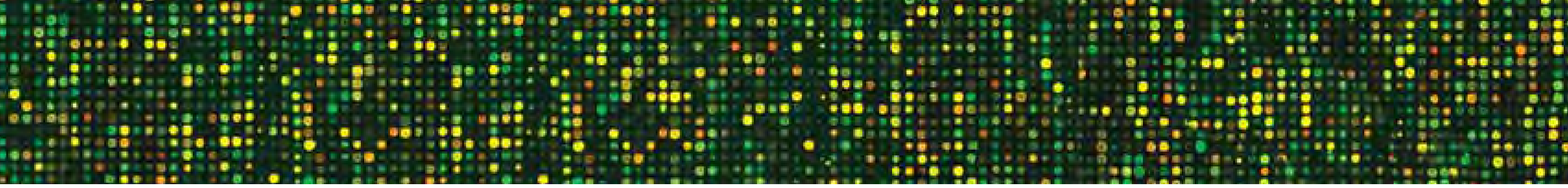
MICROARRAYS: The Search for MEANING In a Vast Sea of Data

BY KRISTIN COBB, PhD

When DNA microarray technology emerged more than a decade ago, it was met with unbridled enthusiasm. By allowing scientists to look at the expression of enormous numbers of genes in the genome at once, microarrays promised to revolutionize our understanding of complex diseases and usher in an era of personalized medicine. Advocates vowed that, someday, with just a finger prick, doctors would instantly know whether patients were having a heart attack, rejecting a transplant, or in the early stages of cancer based on their mRNA patterns, and would tailor treatment accordingly.

In the decade since their introduction, microarrays have permeated virtually all corners of biomedical research; have yielded some useful insights into basic biology and cancer; and are being used, in a preliminary way, to diagnose disease, guide treatment, and streamline drug discovery. But early enthusiasm has been tempered with a dose of reality. Progress has been slower than predicted. And some splashy results in high-profile journals have proven difficult to reproduce, casting a shadow over the real successes.

The shift in perception is palpable in the literature: a 1999 *Nature Genetics* article was entitled "Array of hope," but a 2005 *Nature Reviews* article was entitled "An array of problems."[1,2] One recent paper called microarray studies a "methodological wasteland."[3]



All new technologies have growing pains, and early glitches with the technology itself are partly to blame. But the bigger problem is more fundamental. The huge promise of microarrays is that they give information about every gene, but this is also their huge curse—a crushing onslaught of data. A decade ago, these data were a mismatch with existing statistical tools. Today, there is still no consensus on how to analyze and interpret them. A 2005 survey of microarray users concluded that, “Data interpretation and bioinformatics remain the major hurdles in microarray technology.”[4]

Exponential Advance

Microarrays capture a snapshot of which genes are turned on—or expressed—in a given cell at a given time. Before 1995, scientists could only explore the activity of a few genes at a time. Then two groups of researchers—at Stanford University (led by **Patrick Brown, MD, PhD**) and at Affymetrix—scaled this up thousands-fold with the invention of the microarray. The Stanford microarray is a glass slide coated with a grid of thousands of microscopic spots—each corresponding to a gene—that light up to show which genes are on, which are off, and to what degree they are being expressed.

“At the time it just blew away the next best thing that you could do. It was really a quantum leap ahead,” says **Todd Golub, MD**, director of the Cancer Program at the Broad Institute of Harvard and MIT.

Besides expression microarrays, genotyping microarrays are becoming increasingly popular—these reveal variation in the DNA code rather than in gene activity. Scientists are also working on microarrays that use antibodies to detect proteins, but these have even more technical challenges.

Array of Hope

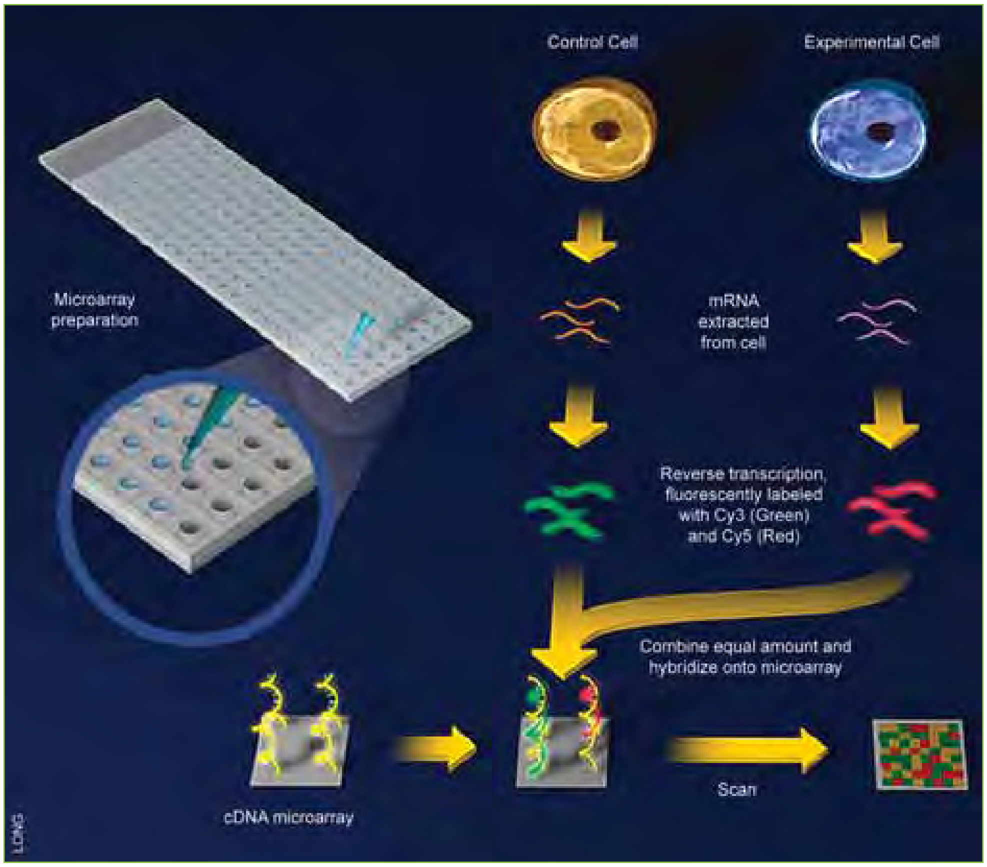
Microarrays are ideally suited to study cancer, a disease of multiple genetic mishaps. They may also yield improved tests for diagnosis and prognosis.

In a 1999 paper in *Science*, Golub automatically and accurately classified leukemia patients into the two main subtypes of the disease using only gene expression patterns.[5] Though these two forms of leukemia were already well recognized and characterized, in principle this strategy could uncover previously unknown subtypes of cancer.

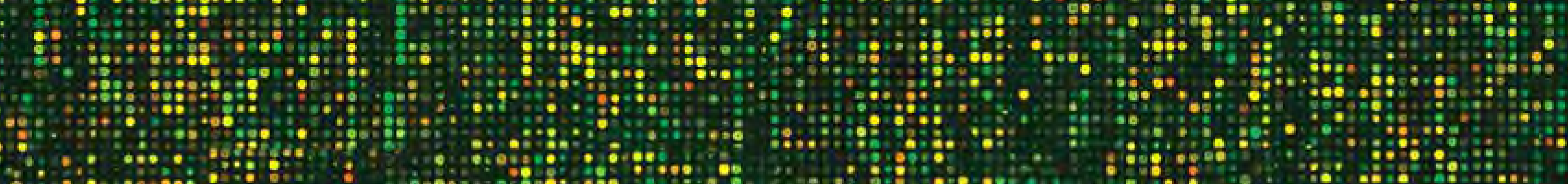
Indeed, in a 2001 *Proceedings of the National Academy of Sciences* (PNAS) paper, researchers identified five unique gene expression patterns in breast cancer and showed that these subtypes were five distinct diseases with different risks of progression.[6]

“This surprised the lab scientists. They said, ‘Wow, look at this—breast cancer isn’t really breast cancer, it’s many types of breast cancer,’” says **Gilbert Chu, MD, PhD**, professor of medicine and biochemistry at Stanford. “But if you talk to anyone who’s been a clinician for many years, they already knew this. They’ve seen breast cancers that looked the same but in some cases vanished with chemotherapy and in others did not. So it’s not a surprise that the gene expression profiles are proving that these cancers are different.”

A natural extension of this work is to isolate the particular genes and expression patterns that are linked to prognosis. For example, Golub derived a 13-gene expression signature that correlated with survival in lymphoma patients; several other groups have isolated gene sig-


Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LONG

***Above:** This schematic portrays an experiment using a spotted or cDNA microarray, which consists of a grid of thousands of microscopic spots on a glass slide; each spot contains cDNA probes for a gene. Here, researchers extract mRNA (made when genes are active) from two types of cells—e.g., tumor cells and control cells—and then label the samples with different fluorescent dyes. When washed over the microarray, these colored transcripts bind to their complementary probes, leaving a trail of informative spots: red for genes turned on in cancer cells, green for genes turned on in normal cells, yellow for genes turned on in both types of cells—with more intense color indicating higher gene activity. Courtesy of* The Science Creative Quarterly, *artist: Jiang Long.*



natures for breast cancer prognosis. These signatures can be used in prognostic tests that gauge if a tumor should be treated aggressively.

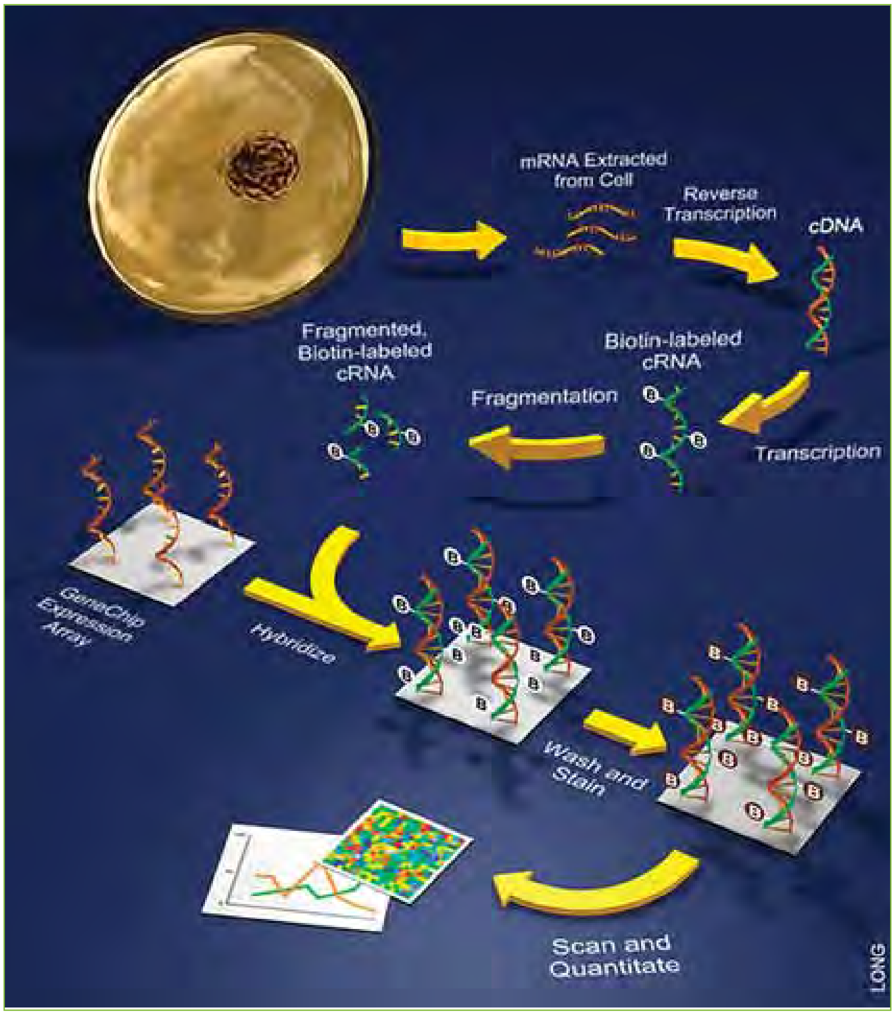
Prognostic genes may also point to novel drug targets. For example, in a recent paper in the *Journal of Clinical Oncology*, Elaina Collie-Duguid, PhD, research fellow at the University of Aberdeen in Scotland, identified one gene that had a 50-fold higher expression in lung cancers that were not responsive to chemotherapy compared with those that were responsive.[7] The gene codes for a protein that prevents tumor cell death; blocking this protein might boost chemotherapy response.

Microarrays may also help tailor a treatment to the person, not just the disease, Chu says. He has identified preliminary gene signatures in the healthy cells of cancer patients that predict which patients will suffer serious side effects from radiation therapy.

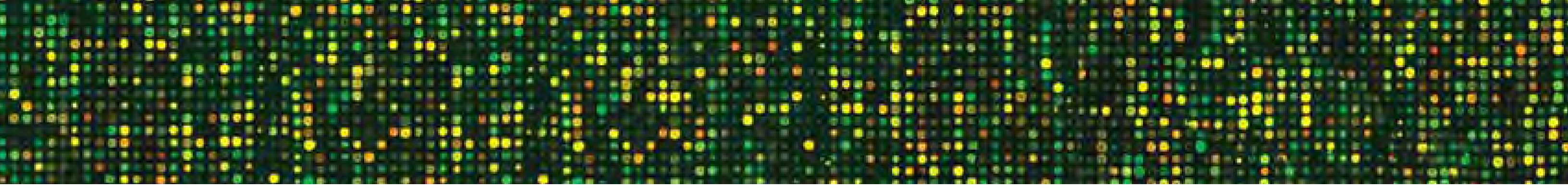
In addition to oncology, microarrays are also being widely applied in heart disease and transplant research. Daniel R. Salomon, MD, associate professor of molecular and experimental medicine at the Scripps Institute in San Diego is working on developing a microarray-based test to quickly tell him if a kidney transplant patient is in acute rejection, chronic rejection, or good condition. And he envisions an even more sophisticated personalized medicine scenario: "What we ultimately want is where the doctor says to these patients, 'I saw you on Thursday and you're doing well but your gene expression analysis tells me you need more immunosuppression, so I'm increasing your dose. Come back and see me in four weeks and we'll draw blood and check your immunosuppression again.'"

Microarrays are also critical in basic biological research that may ultimately have a clinical payoff. For example, by mapping the precise genetic program of embryonic development, Wing Hung Wong, PhD, professor of statistics and

The huge promise of microarrays is that they give information about every gene, but this is also their huge curse—a crushing onslaught of data.



***Left:** The oligonucleotide array (such as the GeneChip from Affymetrix) uses probes that are short strands of DNA synthesized directly on the chip, and a single type of cell is examined at a time. Courtesy of* The Science Creative Quarterly, *artist: Jiang Long. **Right:** InkJet Array. Agilent Technologies, Inc. uses inkjet printing technology to synthesize oligonucleotide probes on a chip. Courtesy of Agilent.*



of health research and policy at Stanford, may be drawing a blueprint for where and when to deliver genes for gene therapy.

Array of Problems

The initial successes in microarrays and their exhilarating promise set off a dizzying flood of microarray studies: fewer than 100 publications in 1999 grew to more than 6000 in 2004. Suddenly investigators were identifying a molecular signature for every disease.

"The statistics and how you analyze the data are still a quagmire," says Greg Engel.

But many publications have since been discredited or have simply fizzled out. Scientists say it's hard to find studies that have led to anything concrete.

"The thing that's surprising to me is that it's taking so long to figure out whether and when the technologies work, and it's taking so long in the face of such enormous enthusiasm," says **David Ransohoff, MD**, professor of medicine and epidemiology at the University of North Carolina, Chapel Hill.

Of the many factors at work—including initial snags with the technology—scientists consistently point to data analysis and interpretation as *the* critical stumbling block.

"For the part of running the experiment, microarrays seem to be working pretty well," says Stanford cardiologist and medical fellow **Greg Engel, MD**. "The informatics part is a whole other area. The statistics and how you analyze the data are still a quagmire."

"I think the greatest challenge at this time remains data interpretation," Golub agrees.

When the data amounted to whether a single gene was on or off, biologists had little need for statisticians. But finding patterns in the activity of 36,000 genes is fundamentally a statistical problem. When microarrays were introduced, even the statisticians were stumped. Existing statistical tools were built to analyze data on a few variables measured on many samples. In microarrays, the situation is reversed: tens of thousands of genes are measured on just a few samples—a phenomenon statisticians are dubbing "p bigger than N" (p features on N samples).

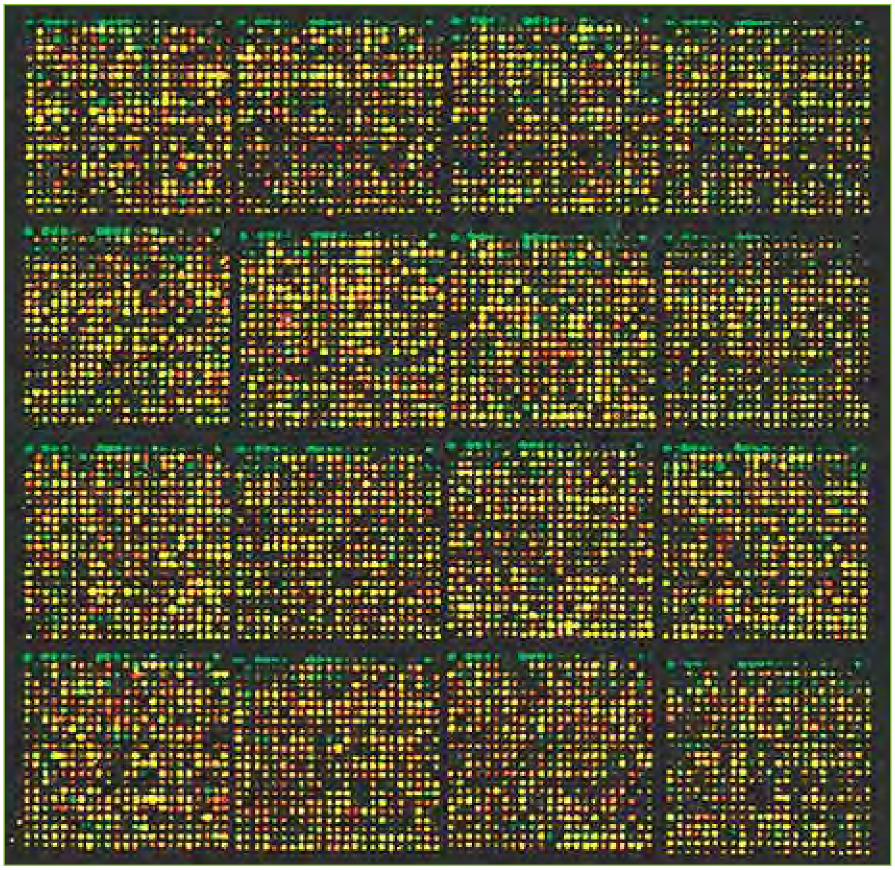
To illustrate the difficulty, imagine that you randomly divide 50 people into two groups and start endlessly measuring their characteristics: age, hair color, favorite food, height, weight, and so forth. Eventually, you will find characteristics that are slightly imbalanced between the two groups just by chance. And the more variables you consider, the more differences you will find. But the pattern of characteristics that separates the two groups is an idiosyncrasy of the sample and has no larger meaning. The same thing happens when you compare 36,000 genes between two sets of 25 cellular samples—some differences in expression may reflect real biological changes, but many more will be false positives.

"The problem here is a deeply profound statistical one," Chu says. "The very nature of microarrays is that they give you tons of data, and very unusual patterns can emerge that are not anything more than noise and statistical fluctuation."

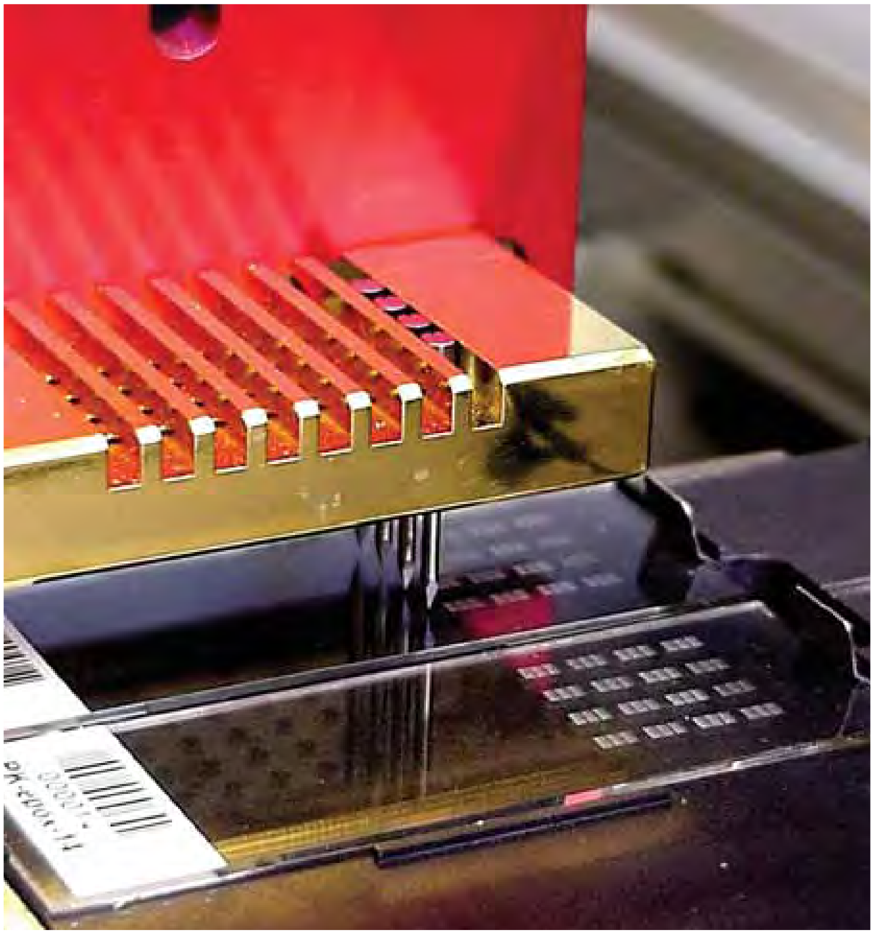
As Engel describes it, "What we're all doing is we're taking a statistical approach and we're all trying it every which way. And you'll even get a pattern. But is that pattern real? That's the major issue for gene chips—is it real?"

Subtle statistical mistakes lead you to find patterns and get published in high-profile journals, Chu says. And it may take years and several expensive follow-up studies for anyone to realize that the finding is not reproducible, unless someone spots the error sooner.

Unfortunately, such sleuthing isn't a job for the casual scientific reader, Chu says. As a perspective in *Nature Genetics* quips, this task requires "forensic statisticians."[8]



Grid of Lights. *Picture of an expression microarray. Courtesy of: Colin Smith, Functional Genomics Laboratory, University of Surrey.*



Making of a microarray: A robot spots a glass slide. Courtesy of NIAID Microarray Research Facility.

Forensic Statisticians

At the request of a colleague, **Robert Tibshirani, PhD**, professor of statistics and health research and policy at Stanford, set out to evaluate a 2004 paper in the *New England Journal of Medicine* that reported a novel gene signature for predicting survival in follicular lymphoma.[9]

Using the data the authors provided online, Tibshirani spent two grueling weeks reconstructing the steps of their analysis and writing a computer program that reproduced their results. Then he put their approach to the test.

To help determine whether a pattern is real or just random noise, statisticians use a trick called split-sample validation: they fit a model only on a portion of the dataset (called the training set) and then test its discriminatory ability on the untouched data (called the validation or test set). If the model only fits noise in the training set, it will usually fall apart when applied to the test set. But even this isn't perfect, because bias can be introduced in choosing the training set and specifying the model.

The authors of the lymphoma study had fit a model (a gene signature) using half the data and found that the model performed well when tested on the remaining half. But when Tibshirani simply swapped these training and test sets and applied the model-fitting program to the new training set, unexpectedly the model did not pop out. In fact, no models popped out, suggesting that their whole finding was spurious.

He also re-ran the computer program on the original training data with tiny changes in the choice of parameters. "Again, the whole thing fell apart like a house of cards," he says. "I also had other colleagues look over my analysis, and they all agree with me: these data look like noise."

"On the broader issue, I think probably a good portion of microarray analyses are wrong," says Robert Tibshirani.

"On the broader issue, I think probably a good portion of microarray analyses are wrong," he says.

A 2005 *Lancet* paper confirms his suspicions.[10] **Stefan Michiels, PhD**, and colleagues at the Institute Gustave-Roussy in France re-analyzed data from the seven largest published studies to report gene expression signatures for cancer prognosis. The papers were published in top peer-reviewed journals, including: *Nature*, *PNAS*, the *Lancet*, and the *New England Journal of Medicine*.

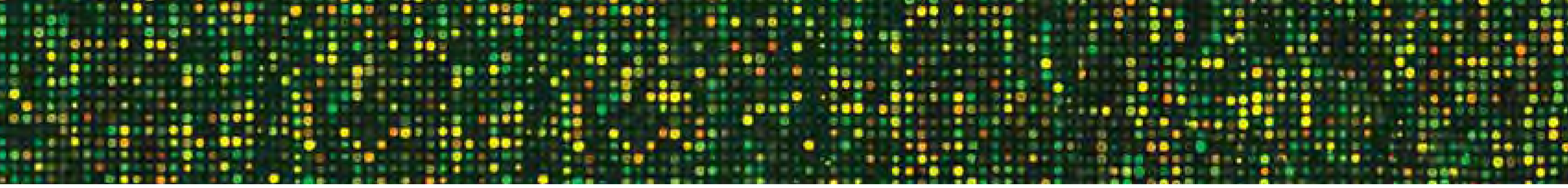
For each of the seven datasets, Michiels' team randomly selected 500 training sets of different sizes; then they built and tested 500 models. What they found: The results were highly dependent on the choice of training set. Every different training set led to a different molecular signature. Moreover, in the majority of trials the signatures selected in the training set had poor or no discriminatory ability in the validation set. Their conclusions: five of the seven studies did not classify patients better than chance, and the remaining two did only slightly better than chance.

"The original investigators may have consciously or unconsciously reported the best performing pair of training-validation data," explains **John P. A. Ioannidis, MD, PhD**, professor and chair of the department of hygiene and epidemiology at the University of Ioannina in Greece, who wrote a commentary for the *Lancet* paper. "I suspect that they probably had some source of selection bias somewhere in the process," he says.

To guard against bias, he recommends using a repeated sampling scheme like that in the *Lancet* paper or having independent groups do the training and validating steps.

Re-Inventing Statistics

The field of statistics moves more slowly than biology, Tibshirani says. So, as the microarray technology raced ahead of the analysis tools, nonstatisticians made up their own statis-



tics to fill in the gaps—which explains a lot of the statistical flaws.

“You don’t see people without training going into labs and doing test-tube experiments. Yet, anybody who has a PC with Excel thinks they can invent statistical methods,” he quips.

Fortunately, statistics is beginning to catch up to the technology. A whole new branch of statistics, “p bigger than N,” has opened up to address the challenges of analyzing microarray data. The resulting innovations will likely be applicable across the burgeoning “-omics” fields.

At the same time, journal editors are tightening standards and requiring authors to follow the MIAME (Minimum Information About a Microarray Experiment) guidelines and to make data available online. They should also encourage authors to provide a script of their analysis, like a statistician’s lab book, Tibshirani says.

“There’s implicit pressure to find positive results. And that’s not a good way to operate,” he says. “A script keeps you honest. It forces you to remember exactly what you did, maybe six months ago. Maybe you’ve forgotten that you’ve actually tried 25 models since last July.” A script also makes it easier for others to evaluate the approach.

Another solution is canned software—such as the packages that he’s developed and made freely available online, SAM and PAM (Significance Analysis of Microarrays and Prediction Analysis of Microarrays). These programs constrain people from simply making the choices that make their data look best. “You need a little bit of a straight-jacket almost,” he says.

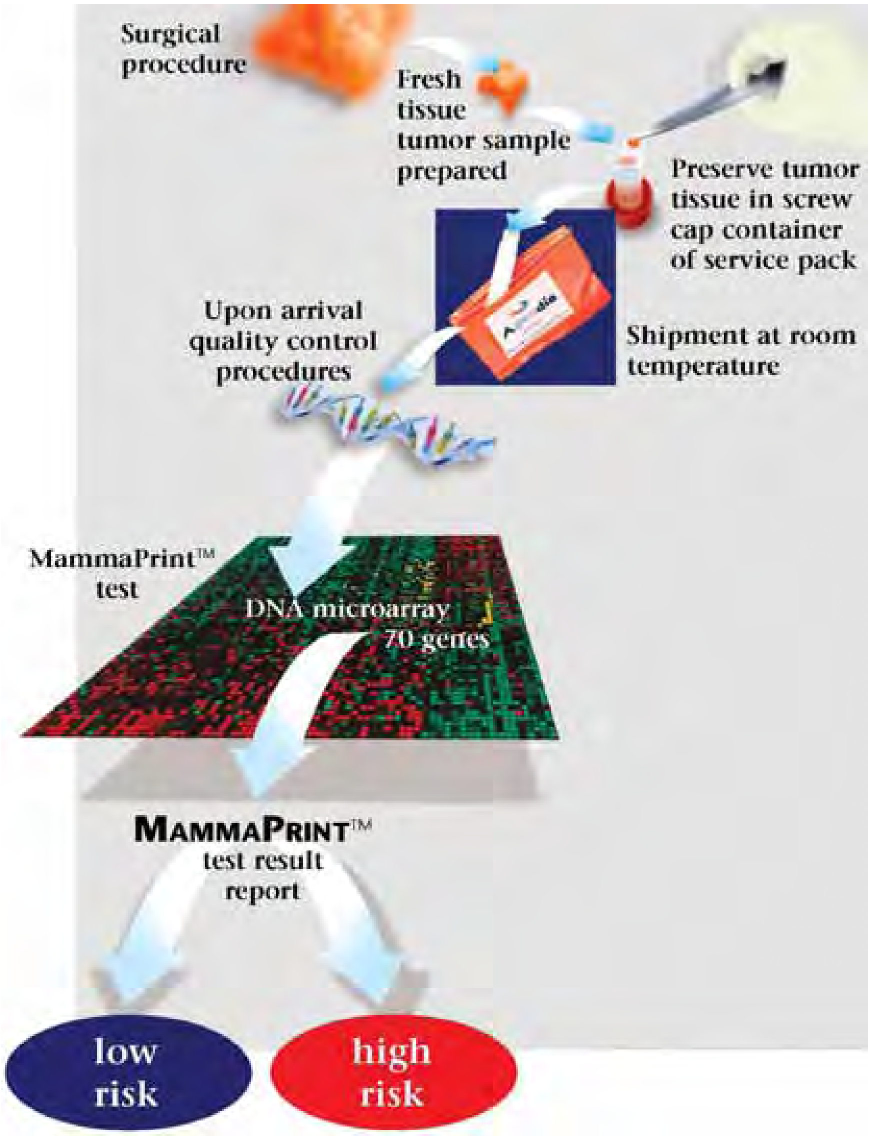
Biologists are also realizing the importance of having statisticians on their microarray teams, Wong says.

“I can clearly detect a changed perception about statisticians,” he says. “Before, the biologists wouldn’t even want to talk to you if you were a statistician. But now the biologists all realize that statistics has something to offer. It’s really raised the profile of our field.”

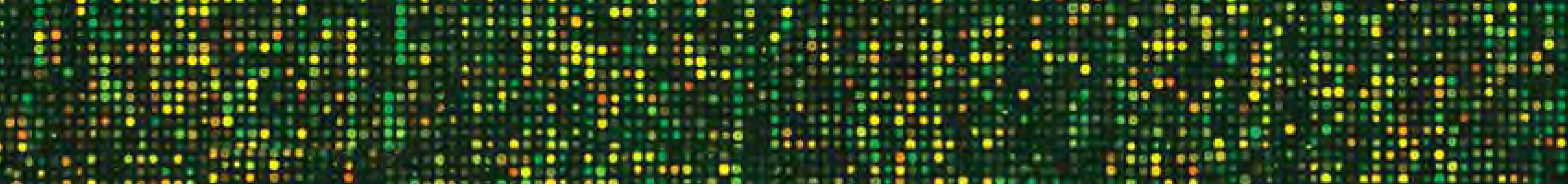
Back to Basics

Even if the statistical analyses are perfect, however, this does not guarantee a reproducible finding, Ransohoff cautions. Too often biologists and computational biologists overlook an even more basic problem: “Fancy math can’t undo biases that have been hard-wired into the data from fundamental errors in clinical study design,” he says.

“This is not fancy molecular stuff, its basic study design that goes back to the 19th century. If case and control samples are not maintained the same way, then we might develop molecular signatures that simply tell us what refrigerator the samples were stored in,” Ioannidis adds.


Surgical procedure
Fresh tissue tumor sample prepared
Preserve tumor tissue in screw cap container of service pack
Shipment at room temperature
Upon arrival quality control procedures
MammaPrint™ test
DNA microarray 70 genes
MammaPrint™ test result report
low risk
high risk

Commercial microarray products such as the Mammaprint and Oncotype DX (not shown) gene expression tests determine whether breast cancer is likely to recur and thus should be treated aggressively. Courtesy of Agendia, Inc.



For example, a 2002 *Lancet* paper (by **Lance Liotta** and **Emanuel Petricoin**) announced the development of a highly accurate blood test for early stage ovarian cancer.[11] Ovarian cancer is usually fatal because it is diagnosed too late, so accurate early detection would be a huge leap forward—exactly the incredible payoff that the "-omics" technologies have long promised to deliver. The test was based on proteomics—patterns from mass spectrometry, rather than microarrays—but the study design issue is the same.

Dismissing microarray technology now would be like stopping flight travel because the first few planes crashed, Horvath says.

The finding launched a commercial test (OvaCheck, Correlogic Systems, Inc.); prompted an unprecedented congressional resolution granting more funding; and was deemed one of the top ten medical breakthroughs of 2002 by *Health* magazine.

But soon after the initial paper, other scientists began questioning the results. Many now believe that Liotta and Petricoin's findings were actually an unintentional artifact of differences in the way the cancer and non-cancer samples were processed. The authors had found a real statistical pattern that separated the groups, but it wasn't a signature of the ovarian cancer.

To avoid such errors, Chu always processes a patient sample at the same time as its control. Some people might consider his attention to detail obsessive-compulsive, he says. "But actually you almost have to be more obsessive with microarray data than with almost any conventional biological experiment."

Microarray teams should also include clinical epidemiologists to address these basic study design issues, Ransohoff concludes.

A More Mature Field

The result of high-profile failures has been an unwarranted backlash against microarray technology, reflects **Steve Horvath, PhD, ScD**, associate professor of biostatistics and human genetics at the University of California, Los Angeles.

"Five years ago there was wide enthusiasm about microarrays, so people were probably a little bit too naive about the challenges that lay ahead," he says. "Now the pendulum appears to have swung back in the opposite direction, where people are much too negative about the promise of microarray data."

Indeed, the backlash has overshadowed some exciting successes. In 2005, the FDA approved the first microarray-based clinical test, AmpliChip (from Roche and Affymetrix). The test identifies genetic variations in the gene for cytochrome P450—an enzyme that metabolizes common drugs—and allows doctors to personalize drug choice and dosing accordingly.

A 21-gene expression test for breast cancer, Oncotype DX (Genomic Health), has been validated in large, independent studies. By distinguishing lower and higher risk tumors, Oncotype DX may spare up to half of women with a common type of early-stage breast cancer from unnecessary chemotherapy. A 2005 analysis showed the $3000 test to be cost-effective because of the averted chemotherapy.[12] Oncotype DX is now being tested in a major prospective clinical trial sponsored by the National Cancer Institute.

A 70-gene breast cancer test developed in the Netherlands, MammaPrint (Agendia), is undergoing a second round of validation studies. The jury is still out, but it is already being used in some clinical settings. A recent study in the *New England Journal of Medicine* found that though MammaPrint and Oncotype DX only overlap in one gene, they give similar results—they agreed about whether tumors were "high" or "low" risk in 81% of cases.[13]

While these examples fall far short of a finger-prick test that instantly sizes up your current and future health, they show that microarray data are not an empty wasteland. Dismissing microarray technology now would be like stopping flight travel because the first few planes crashed, Horvath says. These early crashes led to strict and effective safety procedures for flight, and, similarly, early failures in the microarray field have led to stricter standards to ensure reproducibility, he says.

As a more mature field faces its second decade, it is also adopting a more realistic outlook. Microarray users acknowledge that an all-inclusive finger-prick test is unlikely to materialize anytime soon, but they have a more modest goal for their next decade: to streamline their search for meaning in a vast sea of data.

Footnotes

[1] Lander ES, *Nat Genet*, Jan 1999.

[2] Frantz S, *Nat Rev Drug Discov*, May 2005.

[3] Ruschhaupt M, et al., *Stat Appl Genet Mol Biol*, Jan 2004.

[4] Knudtson KL, et al., *J Biomol Tech*, Apr 2006.

[5] Golub TR, et al., *Science*, Oct 1999.

[6] Sorlie T, et al., *Proc Natl Acad Sci*, Sep 2001.

[7] Petty RD, et al., *J Clin Oncol*, Apr 2006.

[8] Mehta T, et al., *Nat Genet*, Sep 2004.

[9] Dave SS, et al., *N Engl J Med*, Nov 2004.

[10] Michiels S, et al., *Lancet*, Feb 2005.

[11] Petricoin EF, et al., *Lancet*, Feb 2002.

[12] Hornberger J, et al., *Am J Manag Care*, May 2005.

[13] Fan C, et al., *N Engl J Med*, Aug 2006. □

BY LOUISA DALTON

Santa Fe Institute: Addressing Complexity

In the summer, it can be hard to find a place to sit at the Santa Fe Institute. Much of the year, only about a dozen researchers make their home at the multidisciplinary research organization. But in the summer, they double-up in work spaces with many of the roughly 70 external faculty who have primary appointments elsewhere. Add to that flocks of scientists and students moving in and out for workshops and summer school, and it's easy to see how all the chairs get filled in the one-story, pueblo-style research institute.

The visitors and faculty come with widely diverse backgrounds—from the hard physical sciences to the natural sciences, biomedical sciences and social sciences, "sort of ending up somewhere in the fringes of archaeology," says SFI president **Geoffrey West, PhD**. Since SFI's founding more than 20 years ago, the stand-alone organization, funded primarily by private money, has been an evangelist of multidisciplinary research. The institute is a safe haven for such research, West says, a place where investigators can ask fundamental, wide-ranging questions, even if they are not sure the questions will bear fruit.

The types of questions SFI researchers ask—big picture questions midway between multiple fields—tend to fall into a bucket now referred to as complexity science. Answering these questions frequently uncovers commonalities (often mathematics-based) between vastly different systems, whether they are social, natural, or artificial systems. West is now helping to organize a workshop centered on the question, "How are huge databases going to change the way we do science?" Although an astronomer posed the question, huge datasets are also a challenge in high-energy physics, biology, even the social sciences. The workshop will bring together 30 to 40 scientists from informatics, computational biology, astronomy, population dynamics, genomics, high-energy physics, and more. West hopes that some disciplinary cross-fertilization will occur, and new ideas will emerge.

Biocomputation is one of the well-fertilized fields pursued by SFI researchers. For example, **Tanmoy Bhattacharya, PhD**, one of SFI's relatively permanent researchers, was trained as a physicist but has branched out into computational biology. He studies how the rapid evolution of HIV makes it difficult to fight. So many strains now exist; he tries to take a step back and understand the virus from a global evolutionary perspective. His research will likely help others choose the best strains to use for HIV vaccine development.

Another faculty member **David Krakauer, PhD**, studies the stability and evolution of biological organizations with the help of computational principles. He believes that the science of adaptive systems could have an important impact on the design of robust, distributed computing systems.

For the most part, research done at SFI isn't driven by the promise of immediate applications. "That's what makes this place different," Bhattacharya says. "Here you can really ask important, long-term questions, ones that don't have immediate relevance ... Here you can actually understand things. And history tells us that understanding never goes to waste." □

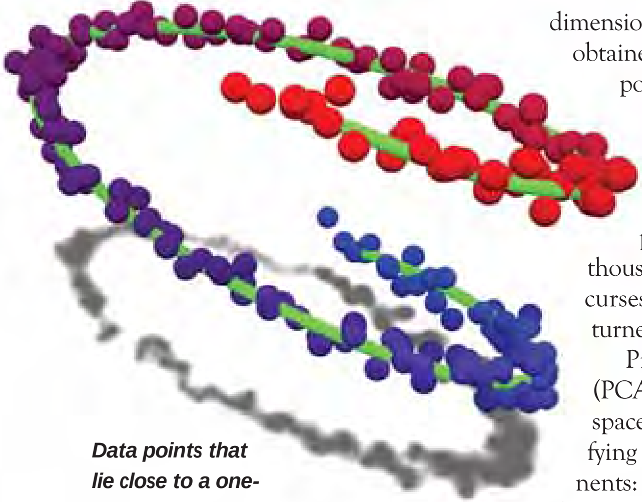
BY ZACHARY PINCUS

Dimension Reduction and Manifold Learning

When Less Is More

The Fall 2005 “Under the Hood” column discussed the *curse of dimensionality*—too many numerical components for each data point—and the *curse of dataset sparsity*—too few data points. One way to treat these problems in concert is to examine the geometric relationships between the data points, and represent the data with fewer descriptors that retain the salient structure.

This illustration demonstrates data in three dimensions that has such structure. Though each point is described with an (x,y,z) triplet, a single number—the parametric position along the spiral—may be sufficient to characterize that point for many applications. In technical terms, the points lie along a one-dimensional *manifold* (plus noise) that has been *embedded* in a three-dimensional space. Methods typically referred to as *manifold learning* or relating to *embeddings* seek to find such simpler parameterizations.



Data points that lie close to a one-dimensional manifold (green line).

The simplest methods assume that the manifold is linear. Vast decreases in the dimensionality of the data points can be obtained by simply noting that n data points cannot span a linear space of dimension greater than n–1. For example, three points form no more than a two-dimensional plane, regardless of whether the points are vectors in ten or ten thousand dimensions. Thus, the two curses of biomedical data can be turned against one another.

Principal component analysis (PCA), for example, finds the subspace spanned by the data by identifying the data set’s principal components: The first such component is the direction along which the data has the most variance; the second component lies orthogonal to the first and best accounts for the remaining variation; and so on. The data in the illustration have three principal components: first, the central axis of the spiral; then, the long axis of the ellipse that remains when the spiral axis has been projected away; and finally the ellipse’s short axis. In many cases even highly nonlinear data can be described well with a small subset of the leading principal components. For example, if the illustrated data were ten-dimensional but had the same spiral structure, three principal components would describe the data exactly. Alternatively, if the data were described only with the first one or two principal components, much meaningful structure would still be retained.

A related technique called multi-dimensional scaling (MDS) operates over the distances between data points. MDS finds positions for the points in low dimensions such that the inter-point distances are changed as little as possible. If the Euclidian distances are provided, MDS and PCA are identical. However, other distance measures can also be constructed. The Isomap method finds the distance between points by measuring the length of a path that is constrained to “hop” from point to point along the data cloud. This distance approximates the *geodesic* distance (e.g., distance along the spiral). Application of MDS to these distances can easily recover many nonlinear structures.

The two curses of biomedical data can be turned against one another.

In contrast to MDS and Isomap, which consider the distances from every point to every other, several methods, such as Laplacian eigenmaps and locally linear embedding (LLE), deal only with distances between points and their close neighbors.

It is my hope that this extremely brief column has interested you in this fertile field of research. For a full-fledged introduction, refer to Christopher Burge’s excellent tutorial.[1] □

DETAILS

Zach Pincus is a PhD Candidate in the Biomedical Informatics program at Stanford University. He works in the lab of Julie A. Theriot, developing methods for automatic and quantitative interpretation of images from microscopy, and cellular structures therein.

FOOTNOTE

[1] Christopher Burge. “Geometric Methods for Feature Extraction and Dimensional Reduction: A Guided Tour,” in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Eds. L. Rokach and O. Maimon, Kluwer Academic Publishers.

ACKNOWLEDGEMENT: Thanks to Brian Naughton for many discussions of manifold learning.

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation
REVIEW

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

seeing science

SeeingScience

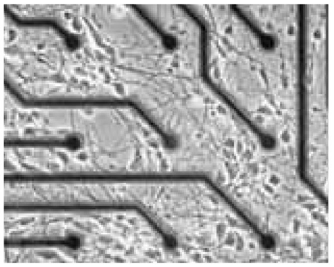
BY KATHARINE MILLER

MEART: The Semi-Living Artist

MEART's creators link the basic components of the brain (isolated neurons) to a mechanical body (robotic arms) through the mediation of a digital processing engine across the Internet. The goal: to create a semi-living entity that will seemingly learn, evolve, and become conditioned to express itself through art. The project was originally created in 2001, but has changed and evolved in seven art galleries since then, most recently in Shanghai, China this year.

How it works: Neuroscientist Steve Potter, PhD, who runs the Laboratory for Neuroengineering at Georgia Institute of Technology cultures neurons from the cortex of embryonic rats in a Petri dish with 60 microelectrodes (a multi-electrode array or MEA) embedded in it. Through these electrodes, the researchers/artists can send sensory inputs (electrical stimuli) to generate responses (action potentials) that can be converted into movement (drawing) by robotic arms located remotely (over the Internet). For each artwork, a single camera shot of a viewer in the art gallery is converted into a 60 pixel image (corresponding to the 60 microelectrodes) and is sent to the neurons as an electrical stimulus. Computer software processes the resulting neural signals to create a vector that represents the relationship between the current position of the drawing arm and the position on the culture plate of the highest neural activity. This vector is then used to move the robotic arms holding colored pens.

By closing the loop from neural activity to behavior, sensing, and stimulation, the researchers hope that the semi-living artist will learn something about itself and its environment. At the same time, MEART provokes humans to explore questions such as: "What is creativity?" and "How does the evolution of biotechnology alter our definition of creativity?"



Here we see the MEART brain (neurons cultured on a multi-electrode array), the artist's body at work (moving robotic arms), and one of the artist's drawings. Courtesy of Symbiotica Research Group at the University of Western Australia in collaboration with Steve Potter's Lab. Photos by Philip Gambien (robotic arm; drawing) and Steve Potter (multi-electrode array).