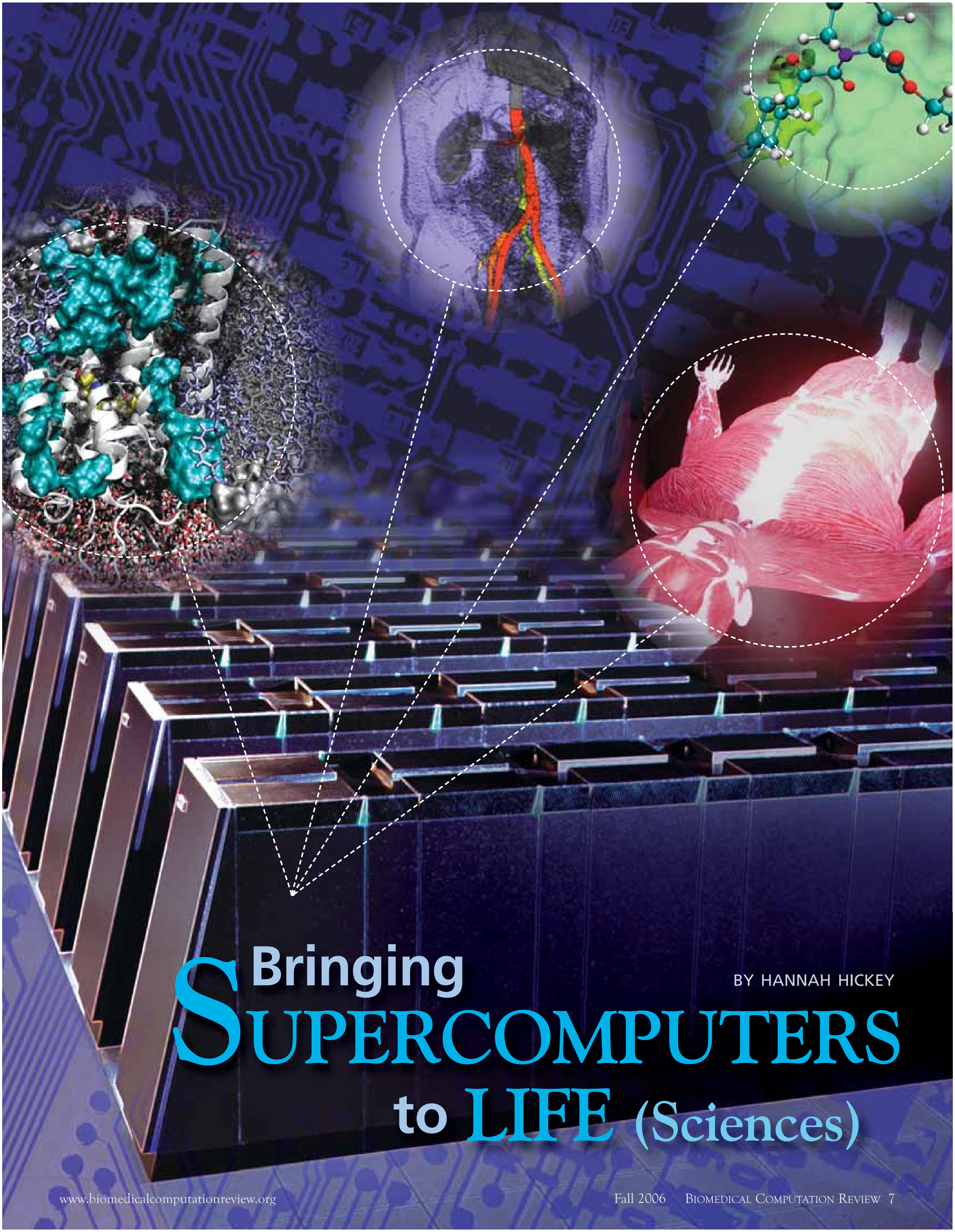
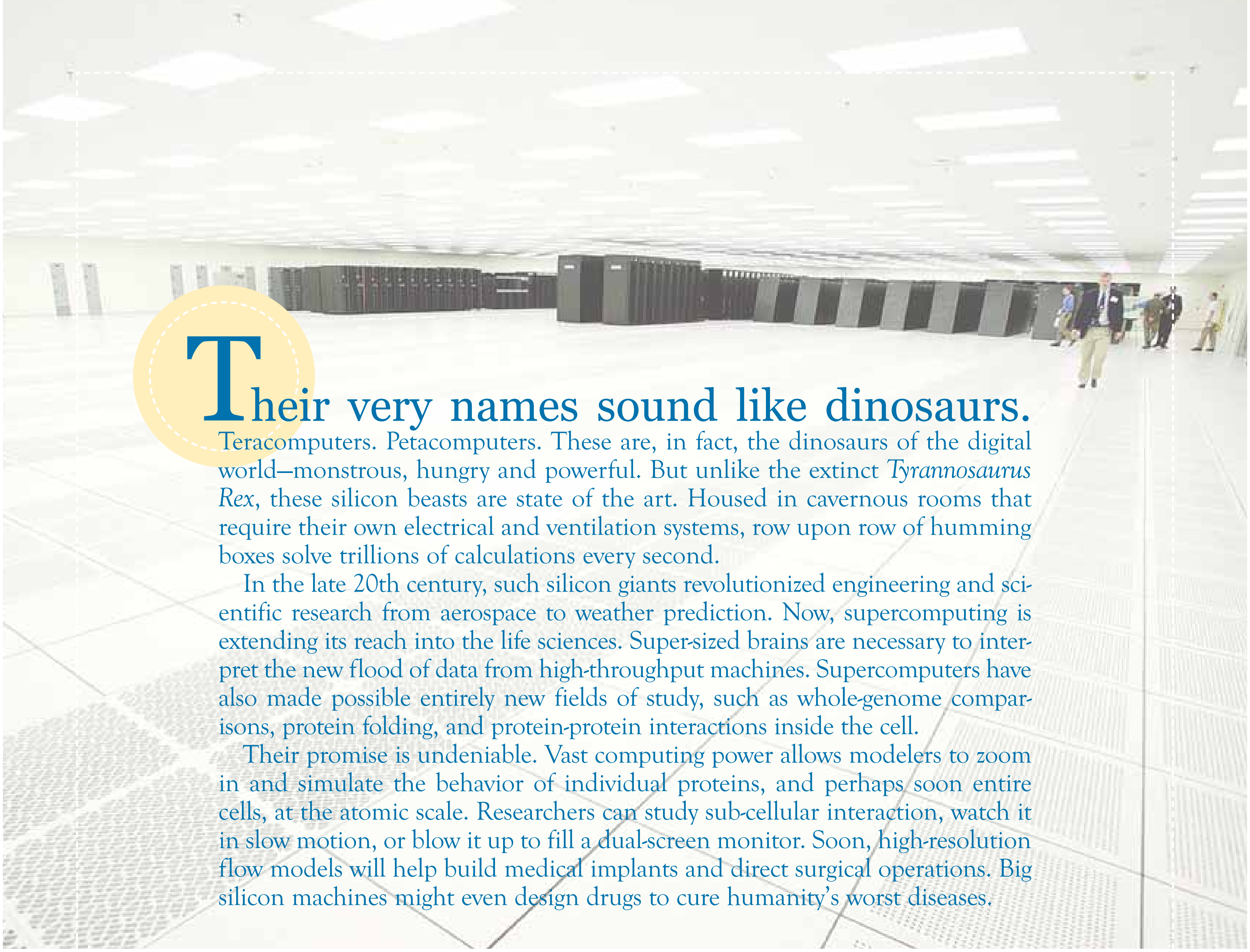


# Bringing Supercomputers to LIFE (Sciences)

BY HANNAH HICKEY

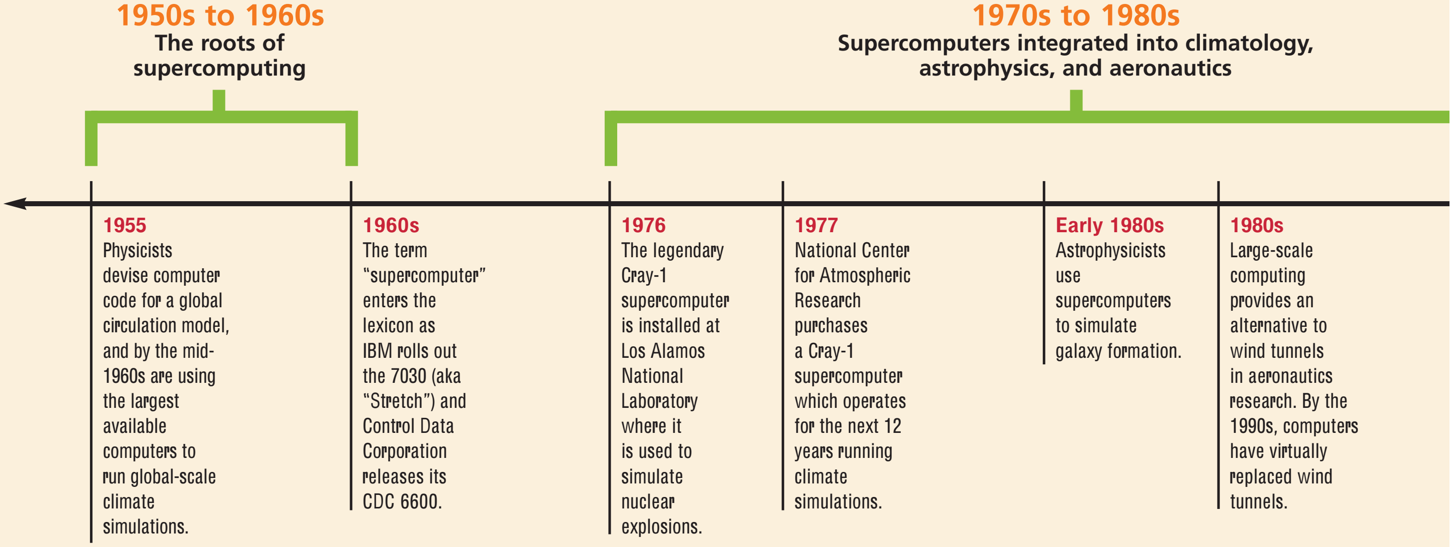


Their very names sound like dinosaurs. Teracomputers. Petacomputers. These are, in fact, the dinosaurs of the digital world—monstrous, hungry and powerful. But unlike the extinct *Tyrannosaurus Rex*, these silicon beasts are state of the art. Housed in cavernous rooms that require their own electrical 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

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

**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.

**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.

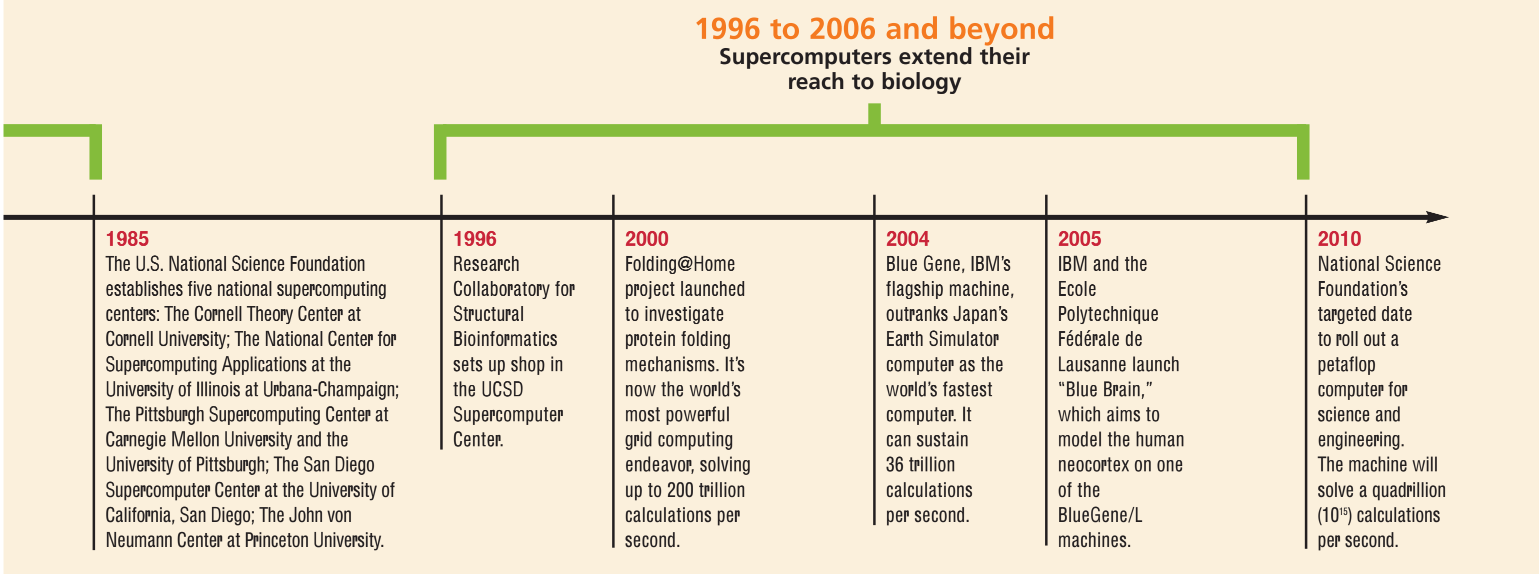
"It used to be that the power of those machines [at supercomputing centers] was many orders of magnitude more than what anybody had access to," says **Philip Bourne**, **PhD**, professor of pharmacology at the University of California in San Diego and editor-in-chief of *PLoS Computational Biology*. "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.

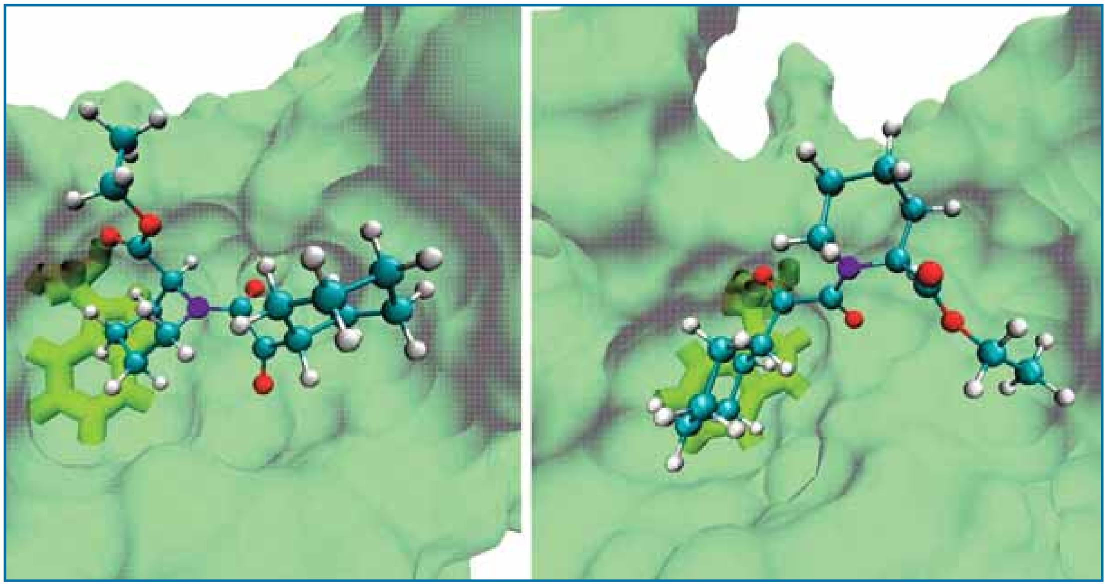
biology researcher at the San Diego Supercomputing Center.

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.



*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.*

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."

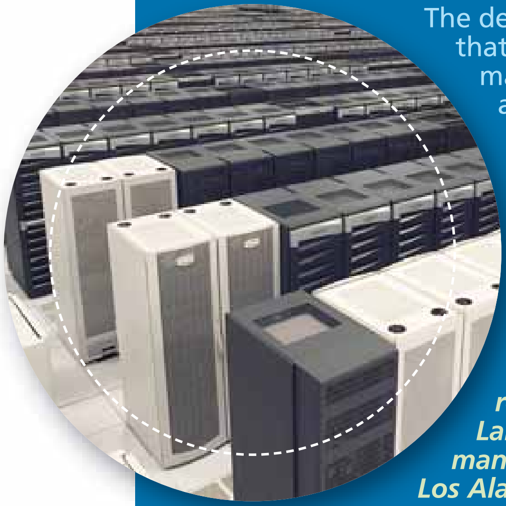
The Human Proteome Project recently finished rough predictions for all the proteins in the human genome in a single year—a job that would have taken a century on the available laboratory cluster.

"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

# What's a supercomputer?

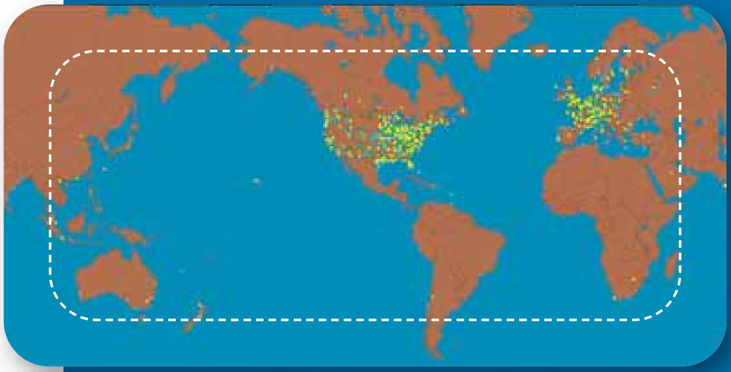
The definition of "supercomputer" is fluid—it just means a machine that's among the world's fastest. Not only is the world's fastest machine always changing, but so is the architecture for creating a supersized computer.



◀ **"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-

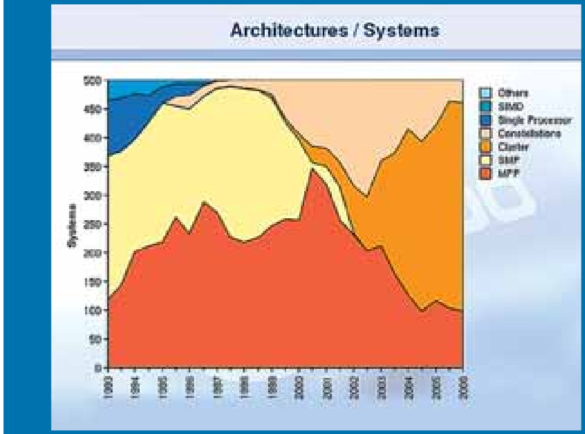
## Top of the FLOPS

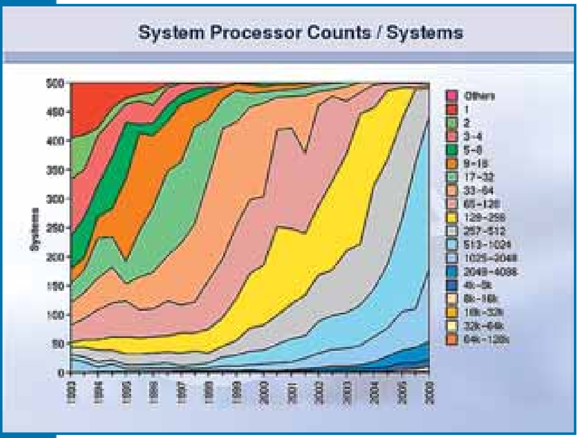
The widely quoted Moore's Law predicts that processing power will double every 18 months. So far the trend, attributed to Intel cofounder **Gordon Moore**, has held true. Processors continually speed up and supercomputers combine them in ever larger numbers. Today's fastest computers, including the Blue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

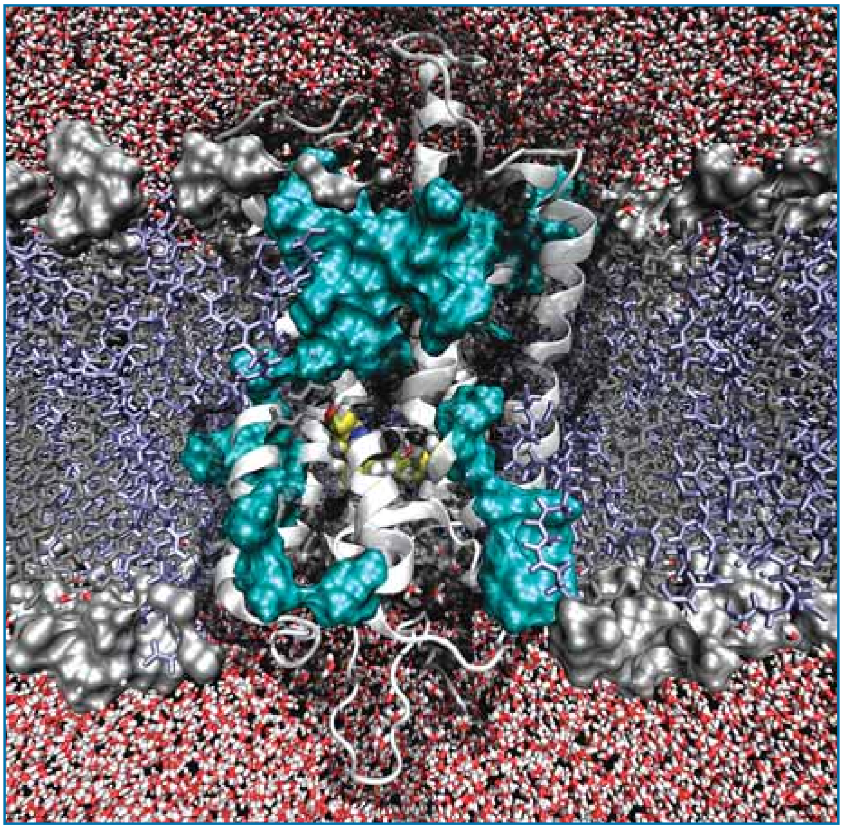
"I think we will model larger and larger biological systems," Germain predicts. He also sees the models themselves improving. While simulating a living thing is not inherently different from recreating a physical event—exploding galaxies, say, or air flowing over an airplane wing—biology has more complex 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.

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.

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.

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.”

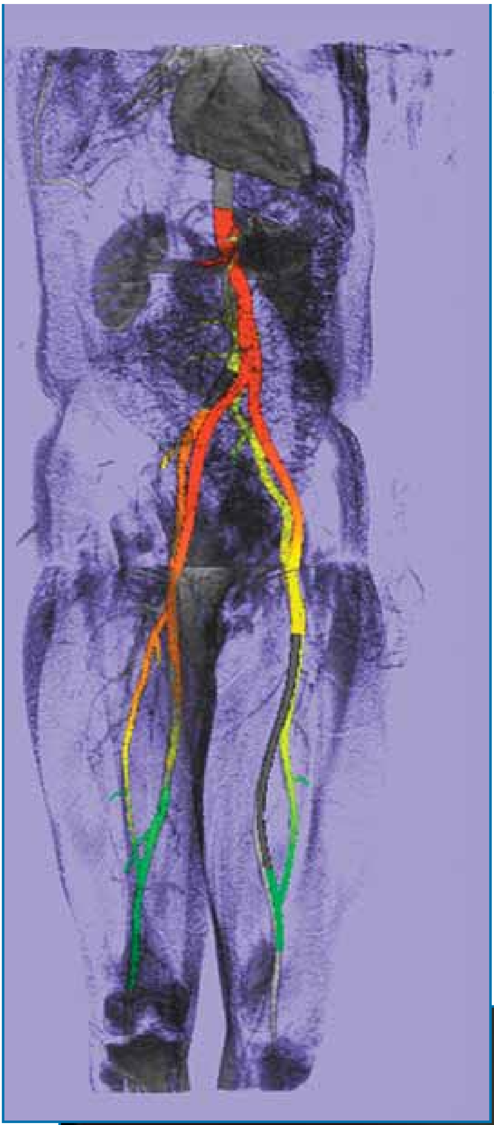
## Genetics’ Information Overload

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.

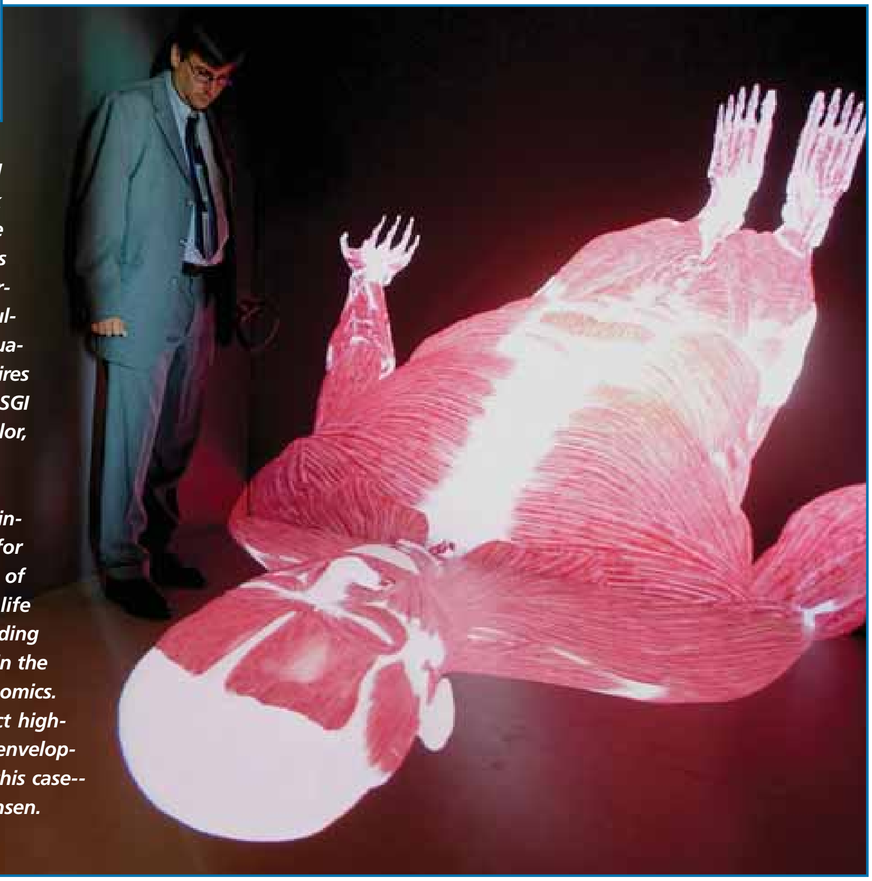
And this is only the beginning. Every year it gets cheaper to sequence more genomes.

“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.*

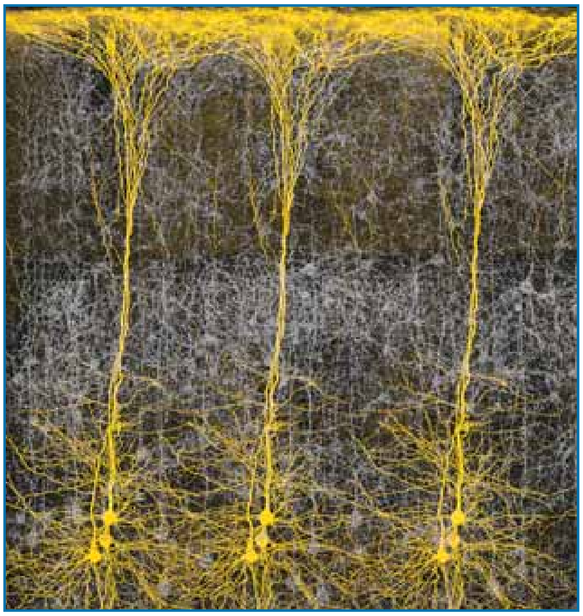


**Christoph Sensen, PhD**, *professor of bioinformatics and director of the Centre for Advanced Technologies at the University of Calgary, looks down on a larger-than-life image of muscle structures. He is standing inside the CAVE, a 4D virtual environment in the Sun Center of Excellence for Visual Genomics. 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.

“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.”

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



*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.*

## 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.

“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.” □