

DIVERSE DISCIPLINES, ONE COMMUNITY

# BiomedicalComputation

REVIEW



## Human VERSUS Machine

Biomedical  
expertise meets  
computer  
automation

**PLUS:**  
COMPUTATIONAL  
BIOLOGY CATCHES  
THE FLU

Modeling the Bug,  
the Host, the World

Summer 2006



### FEATURES

## 8 Human Versus Machine:

Biomedical expertise meets computer automation

BY LOUISA DALTON

## 16 Computational Biology Catches the Flu:

Modeling the bug, the host, the world

BY KATHARINE MILLER

### DEPARTMENTS

- 1** GUEST EDITORIAL:  
SHARE AND SHARE ALIKE:  
A PROPOSED SET OF GUIDELINES  
FOR BOTH DATA AND SOFTWARE  
BY RUSS ALTMAN, MD, PhD

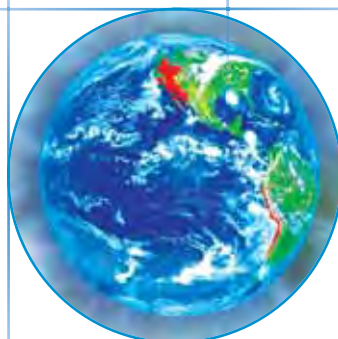
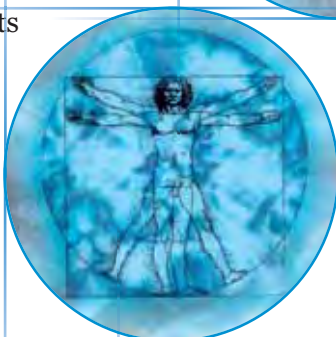
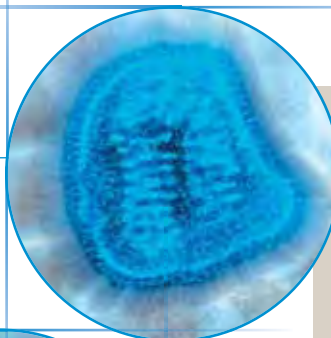
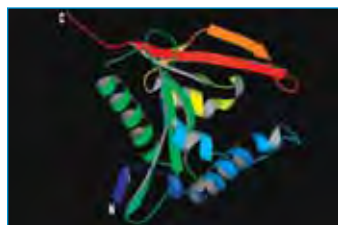
- 2** NEWS BYTES  
BY LOUIS BERGERON, MS,  
LINLEY ERIN HALL, AND KATHARINE MILLER
- Modeling Sex's (Evolutionary) Appeal
  - Modeling Whorls of Leaves
  - Finding the Best Molecule for the Job
  - Whole Virus Simulation
  - Predicting the Structure of Important Drug Receptors
  - Computation Competition

- 27** UNDER THE HOOD:  
COMPLEX STEP DERIVATIVES: HOW DID I MISS THIS?  
BY MICHAEL SHERMAN

- 28** FEATURED LAB:  
JANELIA FARM: CULTIVATING SCIENTISTS  
BY LOUISA DALTON

- 30** SEEING SCIENCE:  
MAKING DNA SMILE  
BY KATHARINE MILLER

ON THE COVER:  
COVER ART BY RACHEL C. JONES OF AFFILIATED DESIGN



### Summer 2006

Volume 2, Issue 3  
ISSN 1557-3192

**Executive Editor**  
David Paik, PhD

**Managing Editor**  
Katharine Miller

**Science Writers**  
Katharine Miller  
Louisa Dalton  
Linley Erin Hall  
Louis Bergeron, MS

**Community Contributors**  
Russ B. Altman, MD, PhD  
Michael Sherman

**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  
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](http://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)  
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)

RUSS ALTMAN, MD, PhD

## Share and Share Alike: A Proposed Set of Guidelines for Both Data and Software



A pair of challenges increasingly threaten the success of bioinformatics research: convincing biologists to share their data and convincing computational colleagues to share their code. Many of us learned to share in preschool, but we also learned that there are sometimes reasons to keep a strong grip on your own stuff. The barriers to academic sharing include protection of graduate student and post-doc publication priority, protection of intellectual property for institutional patents, protection of the patient confidentiality and privacy, and protection of rights for future publication and funding applications.

But these concerns can be overcome with appropriate guidelines. And I believe the parallel issues of data sharing and code sharing can be addressed together. Proven successes in both fields—the sharing of genomics data, and the open source movement—suggest the effort will be well worthwhile.

The ethic of data sharing permeates the genomics community, a fact that derives from the earliest examples of open biological databases: Genbank for storing DNA sequences and the Protein Data Bank for storing macromolecular three-dimensional structures. Through the efforts of visionary scientists, funding agency staff, and journal editors, the submission of data to databases simultaneous with publication became a standard in these fields. The genome sequencing project was successful in part because the organizers created rules very early on—the “Bermuda rules” required nightly release of sequence data. This led directly to the creation of databases to support the sharing of microarray gene expression data, such as the Stanford Microarray Database (SMD), and the Gene Expression Omnibus (GEO) at NCBI. But biomedical computation researchers who enter new fields with visions of Genbank, PDB, and GEO as the relevant precedents may be surprised at the degree of resistance to data sharing in other subdisciplines.

At the same time that biomedical computation researchers are struggling to convince their biology colleagues to share data, they are engaged in an intriguing debate about the merits of open-source code sharing. The open source movement points

to the emergence of Linux as a major precedent that shows the power of shared code. Closer to biomedicine, myriad examples of public domain software have energized certain fields, including EMBOSS for molecular biology, VTK for general visualization, and others. These tend to be larger projects with explicit dissemination goals. Of course, the NIH program that funds the seven National Centers for Biomedical Computation (NCBC) has software dissemination as a major goal, and has led to the creation of domain-specific portals such as Simtk.org. It is more difficult, however, to procure software created by an individual lab that is competing with other labs to create novel methods for biomedical computation.

These parallel problems merit a common solution. I would suggest that the community is converging on the included guidelines.

### PROPOSED GUIDELINES

- 1** Biological data sets and software for storing, analyzing, and visualizing biological data should be released to the public when the first-pass analysis and publication is substantially complete, and no later than one year after the appearance of the first full scale analysis.
- 2** Users should have no expectation of “support” for working with the data or software, beyond basic documentation sufficient for a motivated graduate student to understand and use it.
- 3** Citation of the original source paper, consistent with scholarly standards, should be mandated, and failure to cite should be considered scientific misconduct.
- 4** Downloads should be instrumented, and information about frequency of downloads and other measures of impact should be included in hiring and promotion materials. They should be routinely addressed and evaluated in letters of recommendation written by peers.
- 5** Funding agency staff and biomedical journal editors must be firm in enforcing the sharing of data and code. Manuscripts should have an identifier that lists the eventual location of the data or code, and a date when the data or code will be available.

The current climate of tight funding for biomedical research, with the end of the NIH budget doubling, could threaten the trend towards more open sharing as investigators become nervous about competitive advantage. However, it is critical to preserve the gains in this area achieved over the last decade, and to institutionalize the processes that guarantee continued sharing. □



# NewsBytes

## Modeling Sex's (Evolutionary) Appeal

Sex is a costly undertaking. Finding partners takes time and energy. Sexual contact can transmit disease. And if reproductive success is measured by how many genes you pass on, females would be better off reproducing asexually. But sex must be beneficial in some way—besides being fun—since so many plants and animals do it without going extinct. A new computational model described in the March 2, 2006, issue of *Nature* confirms one existing theory about why sex is advantageous on the genetic level.


“This is very difficult to measure in real organisms,” says **Ricardo Azevedo, PhD**, assistant professor of biology and biochemistry at the University of Houston. “But

things that take years or decades in the lab take only hours in the computer.”

Evolutionary biologists have posited several reasons for the success of sexual reproduction. The mutational deterministic hypothesis suggests that sex helps remove harmful mutations from a population because offspring receive genes from two parents. But the benefits of mutation purging can only overcome the costs of sex if the rate of harmful mutations is high. Multiple mutations must also be more harmful than would be expected from their individual effects, a condition known as negative epistasis. Azevedo’s model suggests that the mutational deterministic hypothesis may be true.

Azevedo along with **Christina Burch, PhD**, assistant professor of biology at the University of North Carolina, Chapel

Hill, and three graduate students created a model that treats each “organism” as a network of interacting genes. The network is expressed as a matrix of numbers (positive, negative or zero) representing the effect of each gene on the activity of every other gene in the organism. Large populations of sexually and asexually reproducing cyber-organisms (networks) were created with different rates of spontaneous mutation. In the first part of the simulation, each organism’s genes interact. Organisms that produce stable patterns of gene expression produce offspring in the second part of the simulation; unstable networks don’t—natural selection at work. When the populations reached equilibrium in their sensitivity to mutations, the sexual populations had become more insensitive to mutations than asexual popula-



“If the conditions in the model are real, then when sex evolves it creates conditions that help sustain itself over time,” Ricardo Azevedo says.

tions and had also evolved negative epistasis. Compared to asexual creatures, they more effectively purge negative mutations from the gene pool.

“If the conditions in the model are real, then when sex evolves it creates conditions that help sustain itself over time,” Azevedo says.

“The prevalence of sex begs to be studied,” comments **Andreas Wagner, PhD**, an associate professor of biology at the University of New Mexico. “To the extent that an abstract model can tell you anything about the evolution of sex, [Azevedo and Burch] have made an important contribution.” But, he says, he’d like to see the work confirmed in living systems.

Azevedo agrees this paper is a first step. He is trying to make the model more applicable to multicellular organisms while his collaborator, Burch, conducts experiments with viruses in order to confirm the model’s results.

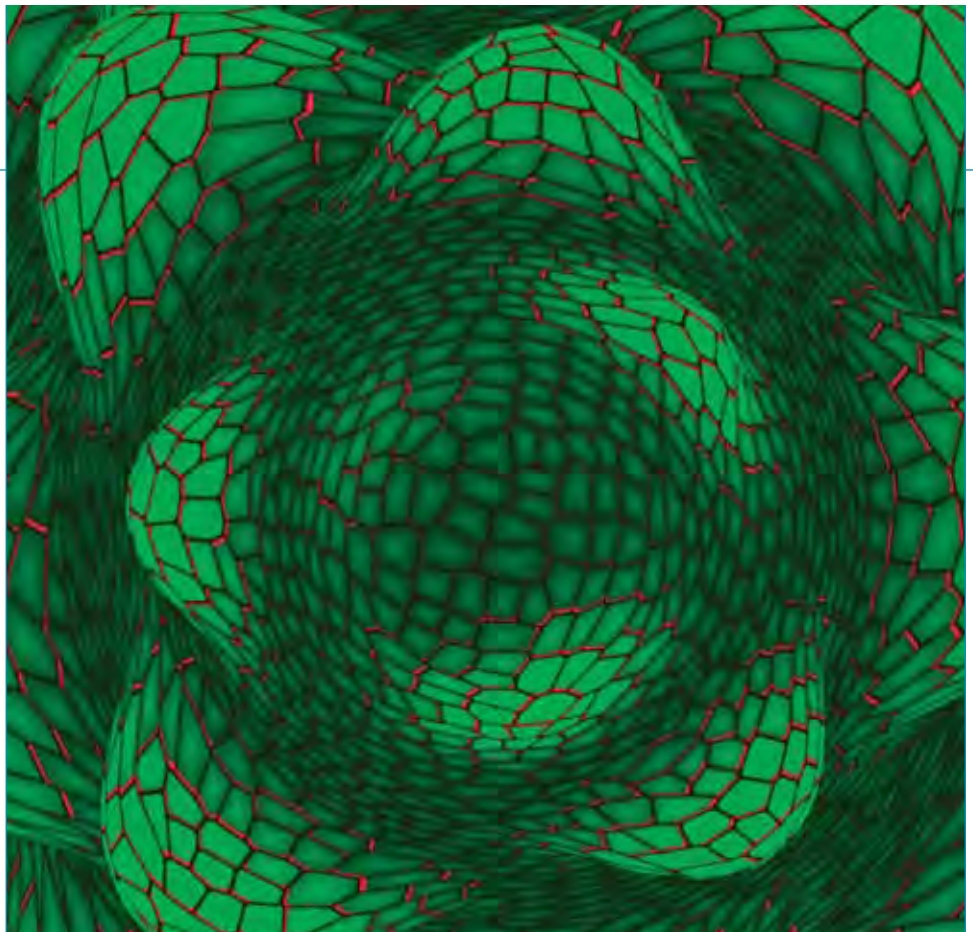
—Linley Erin Hall

## Modeling Whorls of Leaves

The petals of every flower and the leaves sprouting from every plant stalk have characteristic arrangements, a phenomenon called phyllotaxis. For two centuries, botanists have puzzled over the force driving such regularity.

“If you want to understand how plants acquire their form, this is one of the very key questions,” says **Przemyslaw Prusinkiewicz, PhD**, professor in computer science at the University of Calgary in Alberta, Canada. He and his colleagues recently presented a new cellular-level computer model of the process. The work appeared in the January 31, 2006, issue of the *Proceedings of the National Academy of Sciences*.

Previous experimental work by Prusinkiewicz’s Swiss collaborators had shown that a plant hormone, auxin, plays a crucial role in phyllotaxis, as does a protein called PIN1, which regulates the transport of auxin. The team hypothesized that there was a feedback mechanism in which the distribution of



*A computer simulation of the growing tip of a seedling of Arabidopsis thaliana, viewed from above. PIN1 proteins (red) facilitate transport of the plant hormone auxin (green), which in high concentrations promotes budding of leaves, seen here bulging out from the stalk. The feedback interaction of the protein and hormone produce the characteristic spiral pattern of leaves that form as the plant grows. Movies simulating the development of four different leaf arrangements can be seen in the paper’s supplemental material online at <http://www.pnas.org/cgi/content/full/0510457103/DC1#M1>.*

auxin determined the location of the PIN1 proteins, the position of which, in turn, governed the flow of auxin.

They devised a computer model to test the theory quantitatively by simulating the properties of individual cells during the growth of a small flowering plant of the mustard family, *Arabidopsis*.

The model assumes that the tip of

basal tissues of the stem and flowed up to the growing tip, they were only able to get the leaf patterns observed in nature when they altered the model to have auxin produced locally at the tip.

They also found that by varying the parameters of the model, they could produce the leaf patterns found in other plants, which, Prusinkiewicz says, “rein-

By varying the parameters of the model, the researchers produced the leaf patterns found in various plants.

forces our belief that what we have shown is actually true, and it is not just true in *Arabidopsis*, but also in other plants.”

Prusinkiewicz characterizes their model as part of a broader inquiry into how genes and molecular level processes determine the macroscopic forms of organisms, which he calls “one of the most fascinating questions in developmental biology right now.”

—Louis Bergeron, MS

Though the researchers initially assumed auxin was produced in the

forces our belief that what we have shown is actually true, and it is not just true in *Arabidopsis*, but also in other plants.”

Prusinkiewicz characterizes their model as part of a broader inquiry into how genes and molecular level processes determine the macroscopic forms of organisms, which he calls “one of the most fascinating questions in developmental biology right now.”

—Louis Bergeron, MS



## Finding the Best Molecule for the Job

Every pharmaceutical company wants to find the next blockbuster drug. Yet finding molecules with a complete set of desired properties is tricky because of the astronomical number of medium-sized organic molecules. Now researchers at Duke University have developed a novel way to design virtual molecules from scratch. The work was published in the February 17, 2006, online issue of the *Journal of the American Chemical Society*.

“The biggest challenge in chemistry is being able to design molecules for particular purposes,” says **Weitao Yang, PhD**, a professor of chemistry at Duke University. “You can only do experiments on real molecules, but virtual techniques let you use non-real molecules to explore the molecular space.”

Yang along with colleague **David Beratan, PhD**, professor of chemistry, and post-doctoral fellows **Mingling Wang, PhD**, and **Xiangqian Hu, PhD**, developed an innovative approach. Rather than calculate properties of an enormous number of possible individual molecules, their framework approximates the properties over a continuous landscape in which the individual molecules lie. The model relies on knowledge of how atoms can be joined based on the

energy relationships between nuclei and electrons in atoms. This narrows down the possible combinations and smoothes out discrete characteristics, such as atomic number, and thus provides a continuous surface for optimization.

For their proof of concept, the researchers focused on the properties that determine the ability of an atom’s electron cloud to be distorted by external electric fields. So, for example, if six

“You can only do experiments on real molecules, but virtual techniques let you use non-real molecules to explore the molecular space,” says **Weitao Yang**.

groups of atoms could be located at each of two different sites, the model puts the different groups of atoms in the same spot simultaneously and then determines how well the different combinations fit. This repeats at a predetermined number of sites. Joining the best molecular groups or combinations—like snapping together Legos—yields a complete molecule with the best properties.

This approach quickly yields the molecular potential, but it doesn’t necessarily map back to a molecule that can be made. For example, the best group at a particular site might be a combination of 13 percent of one molecule and 87 percent of another. This is impossible, of course, since only one molecule can occupy a single location, so the preferred molecule would be used.

“I think it’s very elegant how Beratan and Yang approached the problem,” says **Ursula Rothlisberger, PhD**, an associate professor of computer-aided inorganic chemistry at the Swiss Federal

Institute of Technology in Lausanne, “But as a first step, it still has many limitations.” For example, it can only create simple molecules, as Yang would agree. He and his colleagues are now refining it to handle more complex systems such as designing optical materials for electronic devices. They plan to extend their work to drug design as well.

“We want to uncover many new materials that researchers didn’t know about before,” Yang says. “This method explores the design space much more efficiently.”

—**Linley Erin Hall**

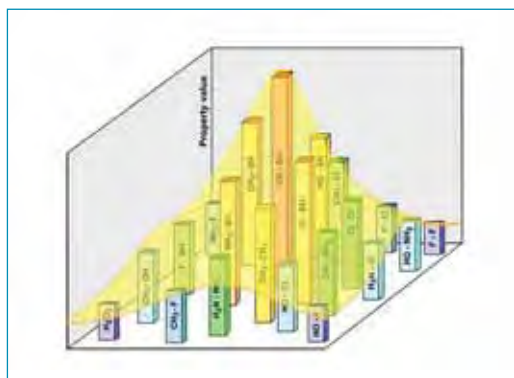
## Whole Virus Simulation

Giving new meaning to the phrase computer virus, researchers have created a computer simulation of an entire biological virus comprising approximately one million atoms.

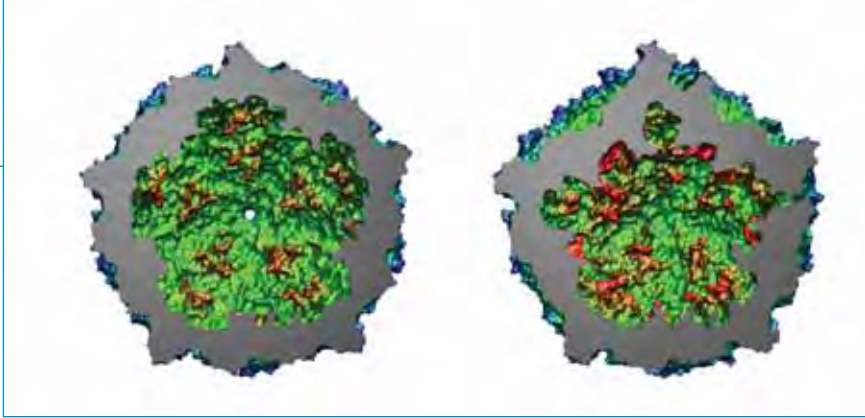
“It wasn’t clear before that one could do a simulation of such a large living system at an atomic level and learn something from it,” says **Klaus Schulten, PhD**, professor of physics at the University of Illinois at Urbana-Champaign. But when he and graduate students **Anton Arkhipov** and **Peter Freddolino** successfully simulated the satellite tobacco mosaic virus (STMV), they revealed some surprising features of the particle in the process. The work was published in the March 2006 issue of *Structure*, as a collaboration with virologists from University of California, Irvine.

Viruses must do two things: infect cells and transport their genetic material inside a stable container known as a capsid. In the case of the STMV, the capsid consists of 60 identical proteins produced by the virus’s genome. Crystallographers who had imaged the small virus believed all 60 pieces were arranged in complete icosahedral symmetry. The computer simulation, however, showed this to be an incomplete picture of the virus.

Schulten and his colleagues started with the crystallography image of STMV and then allowed the atoms to move according to their physical properties. For just over 10 nanoseconds (broken into 10 million time steps), “we let the laws of physics take over,” says Schulten. The



**Yang’s model allows researchers to find the best molecule for a desired property. In this graph, the bar heights represent the amount of a property that each candidate molecule possesses. The model finds the best molecule by evaluating different combinations of molecular groups along the smooth surface over the bars.**



**The collapse of the STMV capsid when simulated without the RNA core. The initial structure for this simulation (a) was the intact STMV capsid immersed in a drop of salty water (not shown). After only 5 nanoseconds of simulation, a prominent implosion of the capsid is observed (b). For both (a) and (b), a cut through the center of the capsid is shown. Courtesy of Klaus Schulten, Anton Arkhipov, and Peter Freddolino, University of Illinois at Urbana-Champaign.**

result: Although the capsid remained generally spherical, some of the symmetry was lost. “The virus developed a belt around an equator of the sphere, and that belt engaged in a back and forth motion,” Schulten says.

More important, simulation revealed that, unlike many other viruses, the STMV capsid is unstable without its RNA contents and depends on the RNA to assemble. “It seems that for this virus, the genomic material first aggregates into a sphere, and then recruits the 60 proteins to be a shell around itself,” Schulten says. “This is opposite to what one expected.”

Schulten and his colleagues hope that viral simulations of this type will help researchers understand how viral capsids shift from stable to unstable when they are infecting a cell. It’s possible that one might be able to interfere in an infection at the point when the capsid breaks apart, he suggests. “We want to use information gained from simulations to protect people from viral infections.”

In future projects, Schulten and his colleagues plan to simulate the poliovirus and other viral particles that are 4 to 10 times larger than STMV. Their success with STMV suggests that large scale simulations provide valuable, new information. “Had we done a partial simulation, we wouldn’t have learned as much,” he says.

—Katharine Miller

## Predicting the Structure of Important Drug Receptors

If you want to find a Tab ‘A’ that will fit into a Slot ‘B’, you’ll waste a lot of time if you don’t know the shape of the slot. For scientists trying to design new

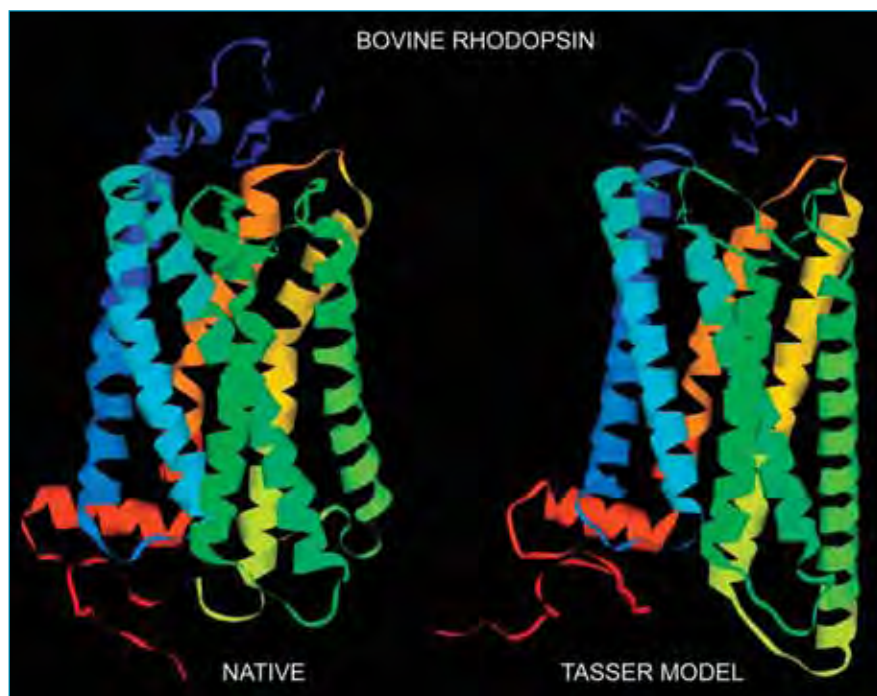
drugs, that is sometimes the precise problem: They seek a molecule that will snug itself into a nook whose shape is unknown, difficult to determine, and capable of changing as the fit is induced.

Now, a new computational tool promises to help rescue researchers from the task of fitting square pegs into undefined holes. It models the structures of the largest family of cell surface receptor proteins in the human body: G protein-coupled receptors (GPCRs). These receptors are encoded by about five percent of human genes and are the targets of about 45 percent of all modern medications. The 3D structures of most GPCRs are unknown because the molecules are extremely difficult to work with. Like all

proteins residing in cell membranes, they tend to fall apart when plucked from the membrane for analysis in a laboratory. Traditional approaches such as NMR and X-ray crystallography have only yielded a single GPCR 3D structure.

To sidestep the difficulties of the experimental approach, Jeffrey Skolnick, PhD, director of the Center for the Study of Systems Biology at the Georgia Institute of Technology in Atlanta, and his research team developed a structure prediction algorithm called TASSER. It takes whatever fragmentary information is known about a protein’s structure—or can be reasonably inferred from knowledge about related proteins—and feeds it into a structure assembly algorithm that combines the data in different ways, searching for the most energetically stable configuration.

“By looking closely at structures that are similar, you should be able to enhance drug discovery by not only designing towards what you want, but away from everything else,” says Skolnick, who estimates that of the 907 GPCRs in the human genome, TASSER has produced



**Bovine Rhodopsin is a GPCR whose structure is known from experimental work. Here, that known structure compares favorably with that predicted by TASSER.**

820 models that are likely to be correct. The work was published in *PLoS Computational Biology* in February 2006.

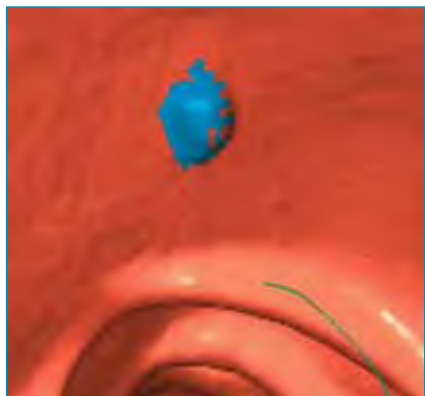
Because no one has determined the structure of these 900 proteins, an algorithm that can produce accurate predictive models should prove significant, comments **Harold Scheraga, PhD**, emeritus professor of chemistry and chemical biology at Cornell University.

Skolnick emphasizes that while he's confident most of the TASSER-generated models provide new insight into the GPCRs structures, he doesn't expect that many of the structures have been fully deciphered by this round of modeling. "What we're trying to do as best we can, is establish the plausibility of these [models] as hypothesis generators," he says, which should help guide drug development research away from dead ends and into productive avenues, where the tabs and slots of medication and receptor are most likely to mesh.

—**Louis Bergeron, MS**

## Computation Competitions Take Off!

From all parts of the computational spectrum, researchers are duking it out: They are throwing their algorithms into the ring to see which one will out-perform all others on a particular task. Contests that feature algorithms for protein structure prediction, natural language processing, and computer-aided disease detection



**Virtual colonoscopy image of a 0.8 cm polyp identified by CAD algorithm (shown in blue). Courtesy of R.M. Summers, MD, PhD, National Institutes of Health Clinical Center.**

are giving researchers a jolt of adrenalin and moving these fields forward.

"When you have a field with a quantitative basis and competing approaches in which high performance is one of the main outcomes, it seems like a natural setting for having a competition," says **Ron Summers, MD, PhD**, senior investigator and staff radiologist in the department of radiology at NIH. "It's also beneficial to the field. The spirit of competition encourages hard work to solve difficult problems."

Protein-structure prediction has been competitive since 1994 when the CASP

**"The spirit of competition encourages hard work to solve difficult problems," says Ron Summers.**

(Critical Assessment of Techniques for Protein Structure Prediction) contest drew 34 groups to register. Since then, the biennial event has steadily grown in popularity: 263 groups are registered for the 2006 bout, including several that will rely only on *in silico* tools, without help from human instinct (See Human vs. Machine feature story in this issue).

This year, competitive natural language processing (NLP) gets a boost from one of the National Centers for Biomedical Computation. In conjunction with the fall meeting of the American Medical Informatics Association, i2b2 (Informatics for Integrating Biology and the Bedside) is extending an open invitation to anyone who wants to challenge their own NLP tools using real clinical records.

"Clinical data is not easily accessible to a lot of people who want to work on this type of data," says **Ozlem Uzuner, PhD**, assistant professor of information studies at the State University of New York at Albany. "I2b2 and its partners have put together these data and that's what makes this a unique opportunity."

The competition is two-pronged. Researchers compete to effectively remove patients' identifying information from clinical data. (Note: I2b2 has already removed the real infor-

mation and replaced it with fictional data to protect patient privacy). In addition, they will parse hospital discharge summaries to accurately extract information on patients' smoking status. The work will help set the stage for researchers to work with clinical data without violating patient privacy.

A computer-assisted polyp detection "bake-off" is also on the horizon. In a traditional bake-off, says Ron Summers, the cooks are given the ingredients and they compete to produce the best cake. In the CAD polyp bake-off, the American College of Radiology Imaging Network

(ACRIN) provides researchers with a data set consisting of CT colonoscopy scans from about 200 patients. The researchers then run their CAD systems using these data. About a dozen academic and commercial researchers have expressed interest in participating.

"Various researchers have been producing systems and claiming outstanding performance on very small data sets," says Summers. "It was competitive but not fair. It was like everyone deciding the terms of their own race." Since the ultimate goal is to help patients, results need to be standardized, Summers says. "We need to know which approaches are better so everyone can move toward that and improve their systems." Hence the CAD competition, which Summers hopes will be underway by November.

—**Katharine Miller** □

### DETAILS

**CASP:**  
<http://predictioncenter.gc.ucdavis.edu/>  
**Challenges in Natural Language Processing for Clinical Data (sponsored by i2b2 in conjunction with AMIA):**  
<http://www.i2b2.org/NLP/Main.php>  
**Virtual Colonoscopy CAD Bake-Off:**  
For more information, contact **Ron Summers: rms@nih.gov.**



BY KATHARINE MILLER

## SimTK: Striving to Host the Best Bio-Simulation Tool Kit



**P**hysics-based simulation is a powerful new tool in the search for new therapies and surgical treatments. But many of the simulation tools now available were designed for planes, trains, and automobiles rather than biomedicine. Moreover, they aren't gathered into one integrated web site and aren't readily available to people who want to use them. With SimTK—an open-source, web-based tool kit—Simbios, a National Center for Biomedical Computing, is trying to change all that.

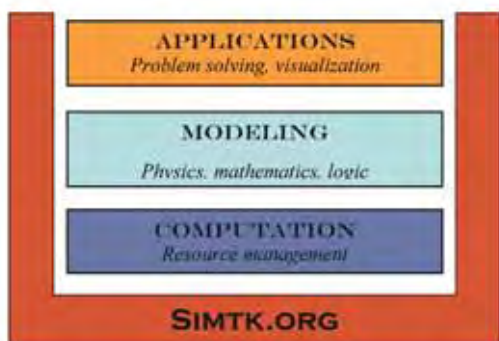
With the launch of [www.simtk.org](http://www.simtk.org), the infrastructure is now in place for people doing physics-based simulation to use SimTK's commercial-grade hosting tools. When they post their projects on the site, they receive reliable back-ups; state of the art version control; built in feature- and bug-tracking; individualized privacy options; and a library system to keep track of prior work. The hosting tools were built upon GForge and tailored to the specific needs of the biomedical community.

"We're striving for the highest quality, most reliable tools," Schmidt says. "We want SimTK to attract the best researchers and programmers."

For biologists, SimTK researchers are developing applications written in the language of biology rather than in computerese. For example, applications might refer to neuromuscular excitation or amount of energy expenditure rather than the linear algebra that underlies the program. So far, Schmidt says, SimTK has applications that scientists want, like a tool for computing the ion atmosphere around RNA molecules and an RNA visualization tool that will be augmented with dynamics later this year. But more are in the pipeline and Schmidt expects to post applications that biologists can download and use with a "whiz-bang" level of satisfaction.

Within Simbios itself, SimTK applications are being developed to simulate four diverse areas not previously combined: neuromuscular dynamics, cardiovascular dynamics, myosin dynamics, and RNA folding. Some are closer to fruition than others. For example, one team is developing tools for evaluating the optimal surgical strategy to help patients with cerebral palsy walk with greater ease. "We're not yet at the point where these tools are available for surgeons," says **Paul Mitiguy, PhD**, dissemination director for Simbios. Developing world class tools takes several years, he says. "The good thing is that we are well on the way to doing it." □

"We're building a coherent set of tools targeted to bio-simulation," says Jeanette Schmidt.



### Simtk.org functions on several levels:

**Simtk.org:** website, infrastructure, community

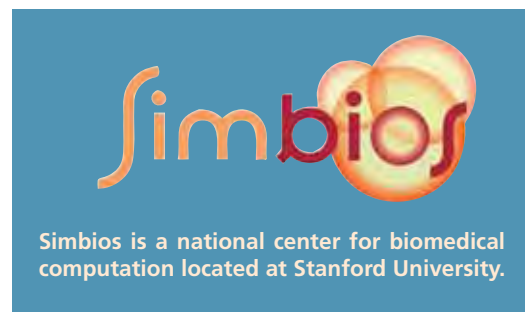
**Applications:** standalone, domain-specific application solving problems for particular users.

**Modeling Layer:** tools for expressing mathematical models of the physical world in a common framework

**Computation:** high-performance numerical methods, libraries, coded algorithms, management of resources needed for computational realization of models.

"We're building a coherent set of tools targeted to bio-simulation," says **Jeanette Schmidt, PhD**, executive director of Simbios. In this way, Simbios ensures that SimTK tools are open-source resources. "We are making these tools available to everyone," Schmidt says, "The NIH deserves a lot of credit for mandating that."

For SimTK to be effective, its tools must be usable not only by computer scientists who design physics-based simulation tools, but also by biologists who want efficient, easy-to-use downloadable applications. Progress has been made on both of these fronts.



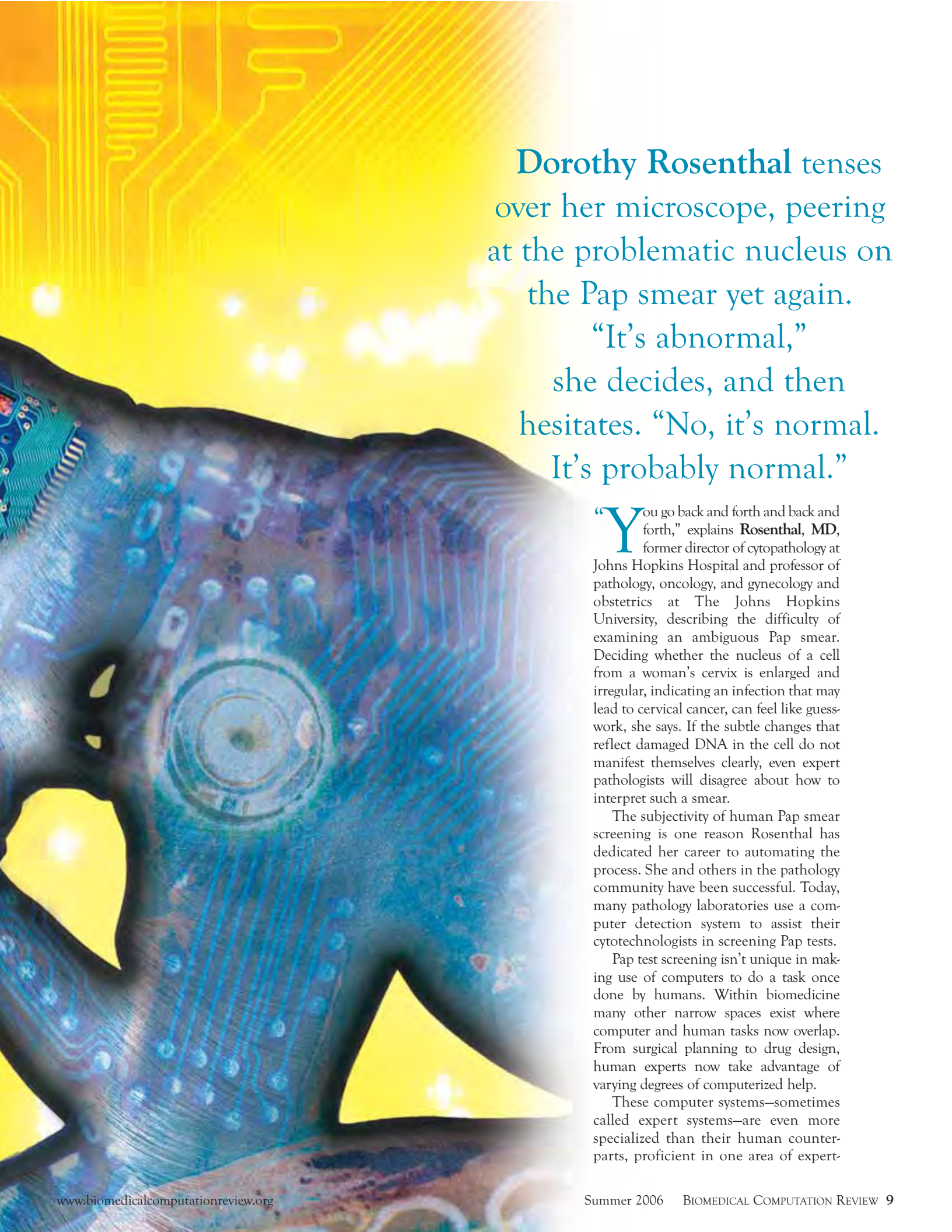
Biomedical  
expertise  
meets  
computer  
automation

BY LOUISA DALTON

# Human vs Machine







**Dorothy Rosenthal** tenses over her microscope, peering at the problematic nucleus on the Pap smear yet again.

“It’s abnormal,” she decides, and then hesitates. “No, it’s normal. It’s probably normal.”

**Y**ou go back and forth and back and forth,” explains **Rosenthal, MD**, former director of cytopathology at Johns Hopkins Hospital and professor of pathology, oncology, and gynecology and obstetrics at The Johns Hopkins University, describing the difficulty of examining an ambiguous Pap smear. Deciding whether the nucleus of a cell from a woman’s cervix is enlarged and irregular, indicating an infection that may lead to cervical cancer, can feel like guesswork, she says. If the subtle changes that reflect damaged DNA in the cell do not manifest themselves clearly, even expert pathologists will disagree about how to interpret such a smear.

The subjectivity of human Pap smear screening is one reason Rosenthal has dedicated her career to automating the process. She and others in the pathology community have been successful. Today, many pathology laboratories use a computer detection system to assist their cytotechnologists in screening Pap tests.

Pap test screening isn’t unique in making use of computers to do a task once done by humans. Within biomedicine many other narrow spaces exist where computer and human tasks now overlap. From surgical planning to drug design, human experts now take advantage of varying degrees of computerized help.

These computer systems—sometimes called expert systems—are even more specialized than their human counterparts, proficient in one area of expert-

ise, at sea in all others. The IBM super-computer Deep Blue, which played and beat chess grandmaster Garry Kasparov in 1997, is such a system.

Some have wondered whether Deep Blue launched a computer revolution that would extend into all spheres. Are Pap test screening and chess-playing just the first of many arenas in which computers will one day outperform humans?

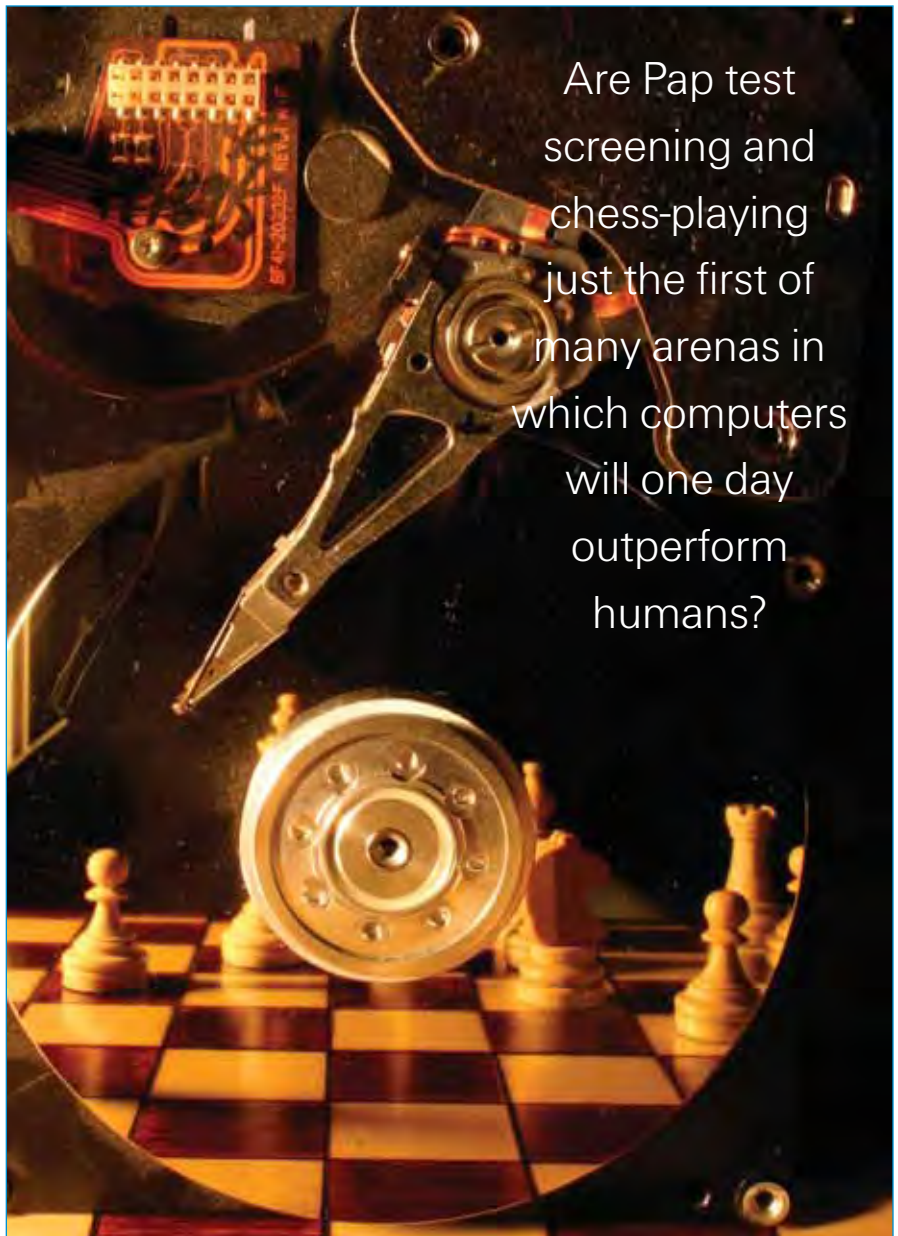
If so, it's a slow-moving revolution, at least in biomedicine. Although computers now rival, even surpass, human performance in some biomedical specialties, the challenges to widespread acceptance and use are still great.

There's the often sticky issue of changing a human's routine so that a computer can help. What the computer provides has to seem worth the trouble. Plus, people desire to stay sharp at their own specialty and want to safeguard against machine error. As a result, introducing computer automation in biomedicine has been extremely challenging on many levels.

Nevertheless, three examples—clinical diagnosis, image interpretation and protein structure prediction—illustrate the promise of developing computerized expertise. Computers can complement and augment native human capabilities and in some cases, even replace humans at the most mundane and repetitive tasks freeing scientists and doctors to pursue more interesting work.

### MEDICAL DIAGNOSIS

Next to every bed at LDS Hospital in Salt Lake City is a computer terminal that gathers patient statistics: blood pressure, medications, ventilator activity and other key bits of information. The



Are Pap test screening and chess-playing just the first of many arenas in which computers will one day outperform humans?

hospital laboratory, the front desk, radiology, and physicians themselves.

Every time new data enter the system, the computer reevaluates patient status and decides, for example, whether or not to alert a doctor or recommend a medication adjustment. Physicians also use

chair of the medical informatics department at the University of Utah, says it has been working smoothly for years. The HELP system started operating in 1967 and is one of the pioneers of hospital decision support. In some of its specialized functions, it “provides more

Although computers that could perform a medical diagnosis and recommend treatment were among the earliest expert systems, the medical community has far from embraced them.

data are collected and managed by a hospital-wide information and decision support system called HELP—Health Evaluation through Logical Processing. The system also collects data from the

the system interactively for help with diagnosis, data interpretation, patient management, and clinical protocols.

**Reed Gardner, PhD**, one of the designers of the system, and former

consistent, uniform care to people than physicians do,” he says. Yet Gardner thinks he could count on one hand the number of other hospitals with a similar system. “It is really not as widespread as





The computer room at LDS hospital in 1965 housed the working hardware and backup hardware for its hospital information system, HELP. HELP's creators include, from left, Homer R. Warner, MD, PhD; T. Allan Pryor, PhD; and Reed M. Gardner, PhD. Courtesy of LDS Hospital, Salt Lake City, Utah.

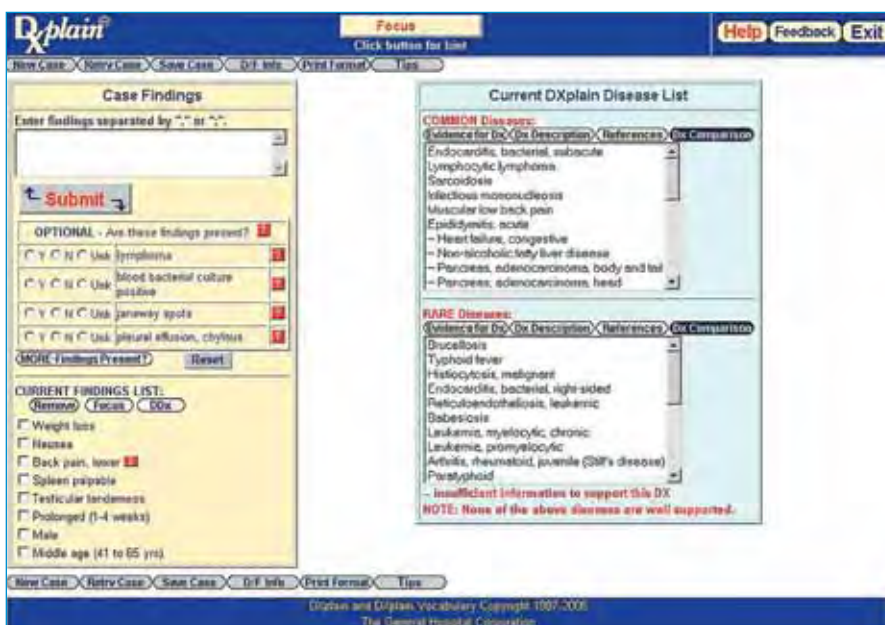
I would imagine 40 years later," Gardner says. Although computers that could perform a medical diagnosis and recommend treatment were among the earliest expert systems, the medical community has far from embraced them.

One of the sticking points relates to those data-gathering bed monitors at LDS Hospital. Gardner went to a great deal of trouble to create the terminals—even building many of the first ones with his own hands. He knew that if the

bedside computer wasn't automatically collecting data and sending the information to HELP, a doctor or nurse would have to type it in. That would have required a change in workflow. When Gardner first tested HELP, he actually hired computer technicians to do the data input for the doctors.

"Physicians are intensely practical," says Octo Barnett, MD, a developer of DXplain, a decision support system primarily for diagnosis that has been operating at Massachusetts General Hospital for more than 18 years. "They won't do something that takes a lot of time and effort to do and doesn't have a lot of payoff for them."

On top of the inconvenience of data entry, the perceived benefit of using



*DXplain, a decision support system primarily for diagnosis, was developed at the Massachusetts General Hospital in 1986 and has been available internationally over the Internet for the past 10 years. Physicians can submit symptoms (left) and the system suggests possible diagnoses (right). Courtesy of Octo Barnett, MD, Massachusetts General Hospital.*

“We feel that [the machine] helps us look harder at the cases that are most likely to be abnormal,” Nance says of the Pap test screening system used by Rex Healthcare.

such a system is low, according to **Eta Berner, EdD**, professor in the health informatics program at the University of Alabama at Birmingham. Computerized diagnosis, she says, solves a problem that some don't think needs a solution. Berner studies diagnostic errors and believes that many such errors go undetected by physicians. If your doctor doesn't get a diagnosis right the first time, she says, either you will go to another doctor, to the hospital, or back to your doctor who will try something else until you get sick enough that the correct diagnosis is obvious. None of those scenarios conveys to the physician that an error was made.

However, there is a growing awareness that medical errors, including misdiagnosis, are indeed a problem. In 1999, the Institute of Medicine (IOM) of the National Academies of Medicine released a startling report stating that medical errors cause an estimated 44,000 to 98,000 deaths a year in U.S. hospitals. The types of errors include misdiagnosis, incorrect drug dosing, equipment failure, infections, blood transfusion related injuries, and misinterpretation of a medical order.

Since then, the Institute of Medicine

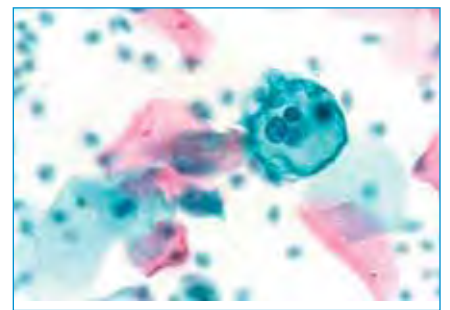
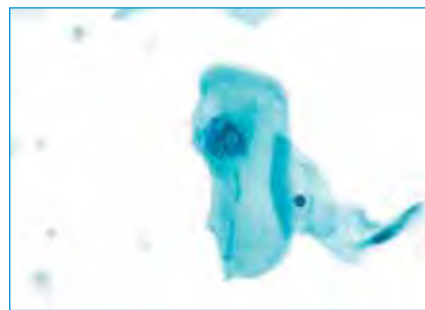
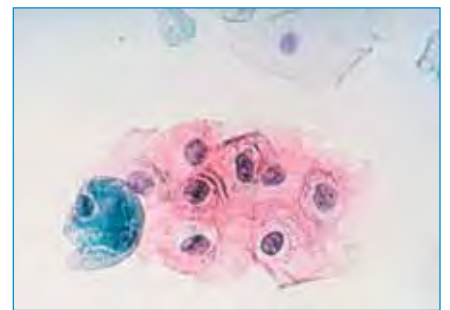
has called for computerized healthcare tools that effectively capture patient information and offer decision and diagnosis support aids.

Hospitals, nursing homes, and doctor's offices have begun to respond. They have preferred systems that, like HELP, primarily alert, remind, inform, and suggest, not just diagnose. Such activities fit into clinical practice better partially because they augment the staff's efforts, rather than replacing them. Berner calls these systems the low-hanging fruit: recommending the most cost-effective antibiotic, advising the pharmacist on drug-drug interactions, double-checking blood types before a transfusion, or carefully guiding and monitoring as a patient is weaned from a ventilator.

The doctors are generally positive if the system works well, Gardner adds, but they don't like it if one of the functions over-alerts, a server is down, or data doesn't get properly entered and lab results get delayed.

### INTERPRETING MEDICAL IMAGES

When pathologist **Keith Nance, MD**, and his coworkers at Rex Healthcare were



*Liquid-based Pap tests can vary from slides that clearly indicate a precancerous state (many abnormal cells with enlarged nuclei and irregular contours), as shown at top, to slides that are much more ambiguous because they contain only one or two atypical, abnormal-looking cells, as shown at bottom. Courtesy of Dorothy Rosenthal, Johns Hopkins Hospital.*



told that the Pap test screening system they just bought could detect abnormal cervical cells better than humans could, it wasn't that they didn't believe it. They simply wanted to make sure for themselves. So for a few years, they double-checked everything the computer analyzed. Out of more than 100,000 cases, the machine missed only one case of the type of abnormal cell called high-grade dysplasia, less than 0.00001 percent of occurrences.

"Now that's good," Nance says, "because the human miss-rate is considered to be five to 10 percent." The computer wasn't quite as good at picking up low-grade dysplasia; it missed about three percent of them. Still, humans miss roughly five percent, Nance says. "Basically, we proved that the machine is better than humans."



That's why the machine that Rex Healthcare purchased, the FocalPoint slide profiler sold by TriPath Imaging, is FDA-approved to independently sign off 25 percent of the slides it sees. It dubs them as requiring "no further review" by human or machine. In addition, it ranks the remaining 75 percent from most likely to be abnormal down to least likely. Nance and his coworkers are pleased with the machine. "We feel that it helps us look harder at the cases that are most likely to be abnormal," he says, and rescreen the cases that really should be rescreened.

Pap test screening is a bright success story of computer assistance. Computer automation of the task is well accepted and cost-effective. It tackles a task for which there aren't enough people to do the work. And because routine Pap test screening "consists of long, tedious intervals between interesting cases," says Rosenthal, most humans gladly welcome help.

Effective, automated Pap test screening is now a reality, but getting to this point wasn't easy. It has taken patience and consistent funding. Back in 1979, the National Cancer Institute sent out a call for proposals to develop automated Pap smear screening systems. At that time, however, "the computers weren't capable of doing the kind of number-crunching we needed them to do," Rosenthal says. In 1987, an exposé in *The Wall Street Journal* about widespread

poor practices for Pap smear screening led to a public outcry and an even greater interest in automation. The U.S. government funded many groups from the late 1970s to the late 1980s. From 1990 to now, private money has developed the devices, Rosenthal says.

Despite the long incubation period and the cautious mistrust displayed by groups like Nance's, computer automation of cervical cancer screening survived. Machines offered by both TriPath Imaging and Cytoc Corporation now increase the productivity of the cytotechnologists and pathologists.

A radiologist's search for telltale signs of breast cancer (such as breast calcifications, tumors, or other lesions) on a film from an x-ray mammogram shares some

miss something obvious. The problem is very appropriate for a computer because computers search "pixel by pixel, area by area, without getting phone calls, without getting tired," says **Maryellen Giger, PhD**, professor of radiology at the University of Chicago.

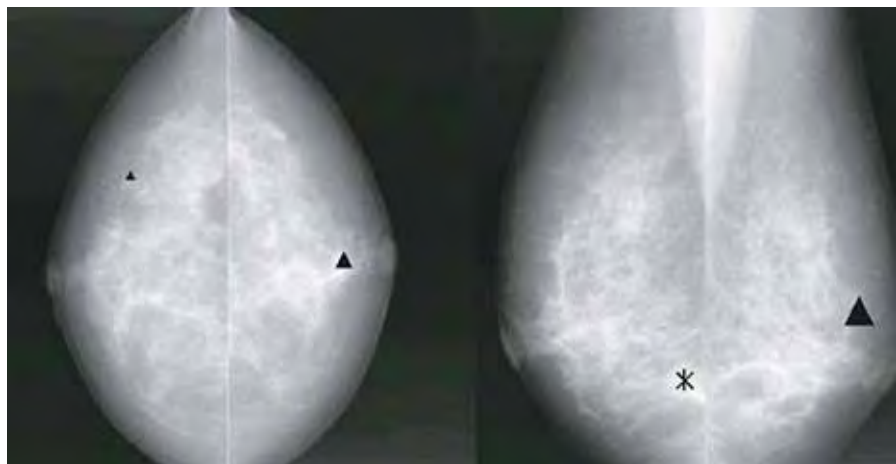
Giger develops computer algorithms and software to aid the radiologist. For more than 20 years, she has worked on computer-aided detection and diagnosis (CAD) of radiological images.

The first computer system that could search a mammogram for breast cancer was FDA-approved in 1998, and Giger estimates that a computer reads approximately a quarter of the screening mammograms performed in the United States nowadays. Unlike cervical cancer screening, however, the physician always has the first look at a mammogram. After the unaided radiologist searches carefully for cancer on a film,

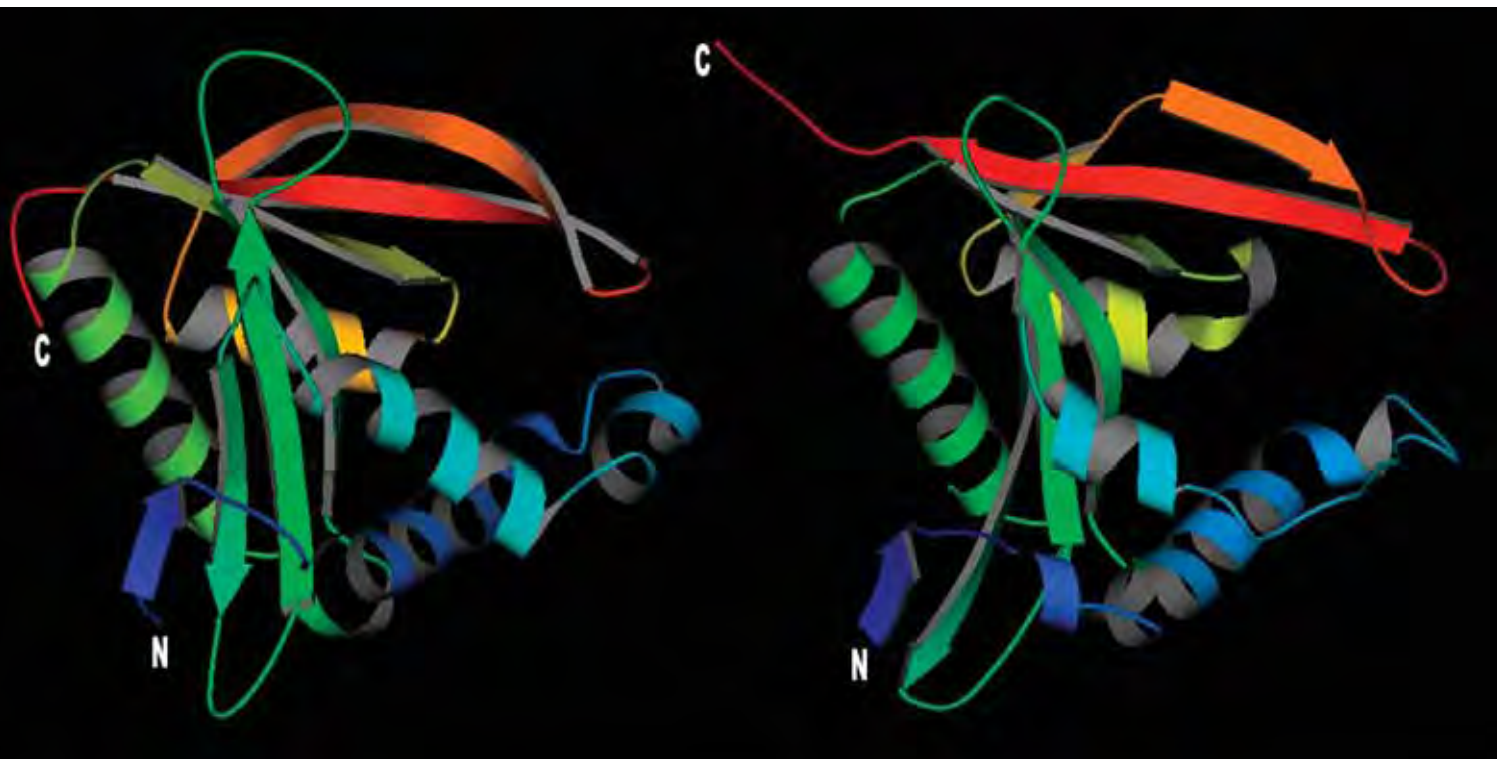
"If you let the computer do it first, there is the possibility that the radiologist gets lazy," says Sandy Napel. For the present, "it is thought best to put the radiologist in competition with the machine."

challenges with the pathologist's search for abnormal nuclei on Pap tests. With such long intervals between abnormal cases, it is easy for a human to get distracted, give up too quickly, or simply

the computer outputs the analysis of the digitized film, designating possible cancers. The radiologist then rechecks the image and either accepts or rejects the computer's suggestions. Most studies show



*After reviewing a mammogram, a radiologist may activate a system such as R2 Technology's ImageChecker, which places asterisks at possible cancer masses and triangles at calcifications indicative of cancer. The larger the mark, the greater the likelihood of cancer. Courtesy of R2 Technology.*



*A predicted structure of an acetyltransferase generated by an automated server in Daniel Fischer's laboratory was among the best of the entries in CAFASP. The predicted structure (right) differs from the experimental structure (left) by just a few angstroms. Courtesy of Daniel Fischer, University at Buffalo (The State University of New York) and Ben Gurion University in Israel.*

“Biologists don’t want to write to the winner of CASP and ask him to spend weeks modeling their particular protein,” Fischer says. “They want to go to the winner of CAFASP on the Internet and push a button.”

that by using this system, radiologists potentially catch more cancers, catch them earlier, and catch them in younger women.

The computer’s second-reader status in mammography actually adds some time to the radiologist’s review. And even though the computer catches calcifications better than most radiologists, for the moment, radiologists have resisted the idea of allowing the computer the first shot. “If you let the computer do it first, there is the possibility that the radiologist gets lazy,” says **Sandy Napel, PhD**, professor of radiology at Stanford University School of Medicine. “The radiologist must take a very close look at the images,” he adds, and being told where to look might keep the radiologist from looking closely everywhere. Even with the current system, “there is concern that as radiologists get more comfortable with the technology and see how effective it is at finding lesions, they may press the second reader button sooner.” For the present, Napel says, “it is thought best to put the radiologist in competition with the machine.”

### PROTEIN STRUCTURE PREDICTION

For years, the ongoing joke among computational biologists was that the

protein-folding problem had again been solved that year. The long-standing problem consists of predicting the final, balled-up form of a protein given only its linear, amino acid sequence.

The problem of predicting protein structure is now a bottleneck to progress. Plenty of amino acid sequence data are being generated by genome projects, but computers can’t yet use that information to predict protein 3D structure—a valuable piece of information for rational drug design. Although researchers can get to final protein structures experimentally with x-ray crystallography, trying to crystallize and determine structures for all proteins (even just the hundreds of thousands of human proteins) will simply take too long given current technologies. Many scientists believe that we need a computational solution.

During the 1980s, groups regularly developed models that worked well on one protein only to find that the model didn’t work for every protein. Finally, in 1993, a pair of researchers got frustrated enough to declare a competition. **John Moult, PhD**, at the Center for Advanced Research in Biotechnology at the University of Maryland, and **Krzysztof**



**Fidelis, PhD**, director of the Protein Structure Prediction Center at the University of California, Davis, set up the Critical Assessment of Techniques for Protein Structure Prediction (CASP): an open competition held every other year where prediction groups compare their strategies head-to-head on new proteins. The experimental structures are published at the end of each competition, clearly revealing which groups performed well and which did not.

Most submissions come from research groups that use a combination of modeling

They've gotten better—so much better that in the latest CASP/CAFASP competition, “only a handful of human predictors did better than the best of the servers” in one of the prediction categories, Fischer says. The difference between a prediction made by an automated server and that of a human expert who chooses which programs to run and improves the results manually is getting “smaller and smaller,” Fidelis says.

Still, no predictions, whether submitted by human or machine, yet reach the quality of an experimentally deter-

puter take over a task so that humans can work on other problems. Losing a human skill because of technological advances is something humankind has been doing for a long time: from the loss of hand spinning with the invention of the spinning jenny to the loss of slide rule proficiency with the invention of the calculator.

Yet some, even some of those same researchers who support computer automation in biomedicine, also worry about the loss of a human art. If the computers at the pathology laboratory

## If the computers at the pathology laboratory were to break down for two days, asks Napel, would we have enough cytotechnologists and pathologists to screen all the Pap smears?

programs, prediction algorithms, human familiarity with protein families, and gut instinct to come up with their predictions.

Yet increasingly, predictions are also coming from computers alone—automated servers that, except for help in setting the initial parameters, receive no human input at all. **Daniel Fischer, PhD**, a professor of bioinformatics at the University at Buffalo and the Ben Gurion University in Israel set up a parallel competition with CASP solely for automated servers. CAFASP (Critical Assessment of Fully Automated Structure Prediction) runs at the same time as CASP and uses the same data, making it “not only a competition of who is the best server, but also a competition of humans versus machines.”

Initially, not everyone liked the idea of automated servers competing with the human predictors. Yet Fischer felt strongly that the automated servers needed their own place in the competition because automation has to be the ultimate goal of the field. “Biologists don’t want to write to the winner of CASP and ask him to spend weeks modeling their particular protein,” he says. “They want to go to the winner of CAFASP on the Internet and push a button.”

The first year that CAFASP ran alongside CASP, the server predictions were downright lousy, Fischer says.

mined structure. The best predictions of both humans and computers position the backbone at least 1 to 1.5 angstroms away from where it ought to be. That’s good enough for some tasks, Fischer says, such as predicting how a protein assembles in a complex, but not yet good enough to create a drug that will act on the protein. “We still hope that someone will figure out how to do the last bit,” Fischer says.

Eventually, it is inevitable that automated servers will take over the task of protein structure prediction, and CASP competitors will have worked themselves out of a job.

Fischer is not very nostalgic about it. He says that both computationalists and biologists will then be freed up to work on weightier issues. We don’t worry about letting a calculator compute cubed roots for us, he says. “Structure prediction itself is a wonderful problem, I love it. But it is not the big picture. The big picture is, do you know what the protein does? Do you know how to suggest a drug to interact with it? I’ll be very happy if no human ever does that again by hand, and researchers concentrate on the more interesting and challenging problems of the 21st century. Protein structure prediction is a 20th century problem.”

Many researchers echo, to some degree, Fischer’s willingness to let a com-

were to break down for two days, asks Napel, would we have enough cytotechnologists and pathologists to screen all the Pap smears? Will we become too dependent on computers? Napel’s concern is a common one, and one reason that the automation of biomedical tasks is a slow-moving trend.

Rosenthal says that Napel’s concern is valid; we probably would not have enough humans at hand to screen the Pap tests if the machines broke down. But the benefits outweigh the loss, she says. The cytotechnologists who used to do mass screening are focusing more on the interesting, abnormal cases, she adds. Novel molecular and genetic tests are being applied to cytology samples, including Pap tests. Salaries will go up as cytotechnologists learn more and their skills become more valuable. Though they lose a task, their skill improves elsewhere, and the field advances.

**Chung-Jen Tan, PhD**, senior manager of the Deep Blue development team at IBM, referred to a similar effect in the world of chess shortly after the 1997 match between Deep Blue and Kasparov. He pointed out that there was more to the victory than just a game of chess. “This will benefit everyone,” he said, “from the audience to school children, to businesses everywhere, even to Garry Kasparov.” □

# Modeling the bug, the host, the world



*The Bug: This negative-stained transmission electron micrograph (TEM) depicts the ultrastructural details of an influenza virus particle. Credit: Cynthia Goldsmith, CDC Public Health Image library. The Host: Credit: Leonardo DaVinci. The World: Credit: JupiterImages.*



# Computational Biology

## CATCHES THE

# FLU

BY KATHARINE MILLER

**I**n 1918, the so-called Spanish flu killed more than 20 million people worldwide. Almost ninety years later, we're faced with the possibility of a flu pandemic that could spread even faster in this globally-connected world.

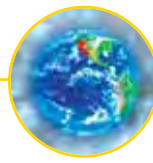
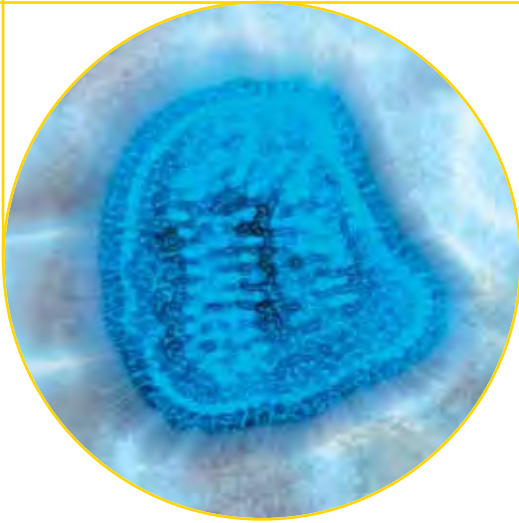
In preparation, researchers are racing to understand what makes one flu bug more infectious or more deadly than another; how best to prevent or treat influenza; and how to control a worldwide outbreak of a deadly strain. Some of these researchers are turning to an approach that was not available in 1918: computational modeling. From the molecular scale to the cellular, organismal and epidemiological, computational biologists are teaming with experimentalists to tackle tough questions about influenza and other infectious diseases.

At all scales, complexity rules the day—making influenza an appealing target for computational research. The intricacy of viruses themselves, the many interactive components of the human immune system, and the complicated networks of human interactions that lead to disease spread, all call out for integrated models that require powerful computation.

"Complexity really does matter," says **Ira Longini, Jr, PhD**, a professor of biostatistics at the University of Washington School of Public Health with reference to his model of pandemic flu spread in the United States, "And we have the computational ability to handle it now."







# The Bug:

## Modeling Shape-Shifting Viruses

The flu virus is an evolutionary marvel. Teams of experts design an appropriate flu vaccine annually just to keep up with the microbe's ability to evade the human immune system. Multiple strains circulate, and no one can predict when a new strain will emerge by mutation or recombination with another strain so that it can jump from another species to humans.

Computational biologists approach

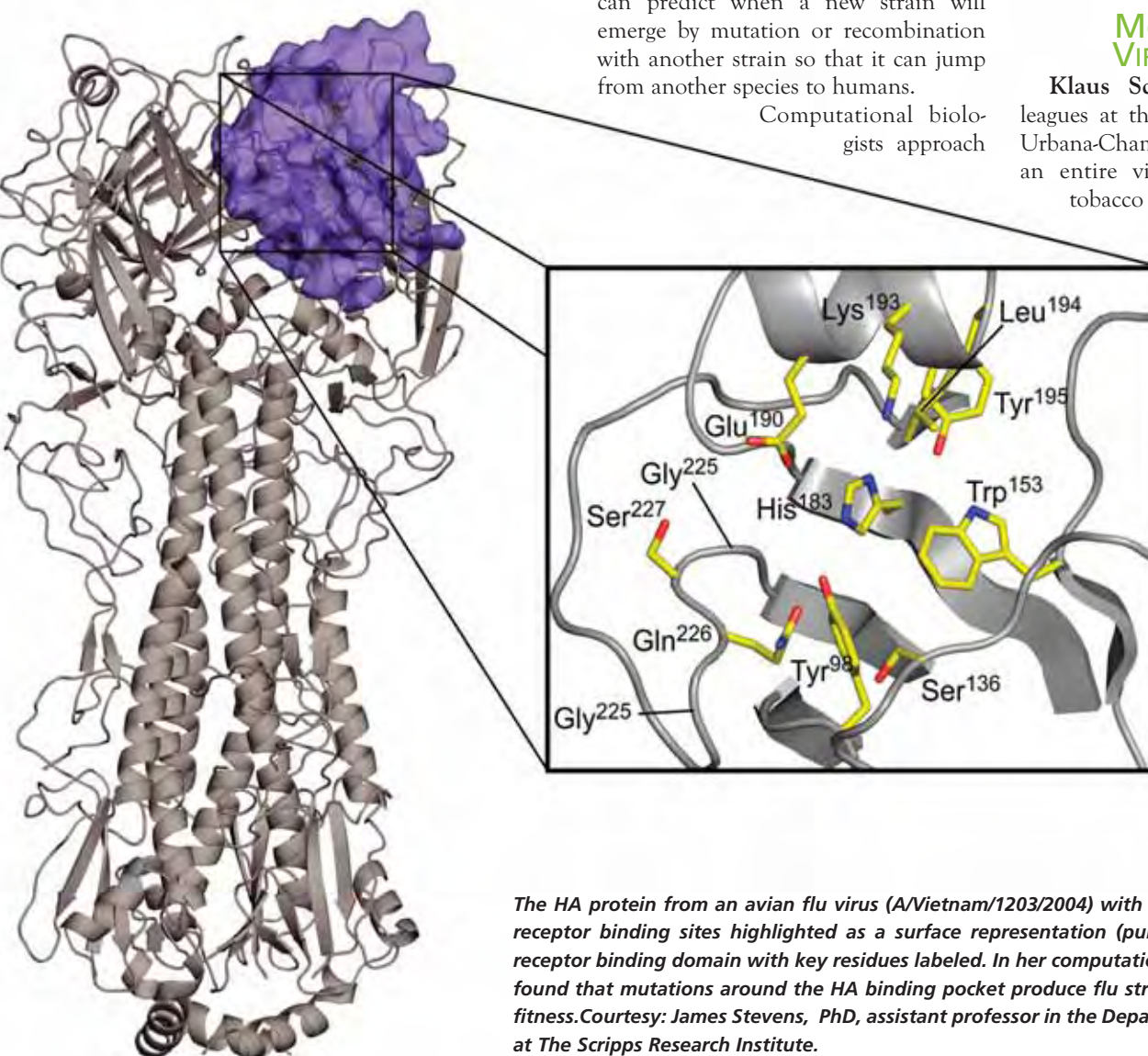
this ever-changing bug from several angles: some simulate entire virus particles to detect their vulnerabilities; others model viral evolution to predict future strains; still others use bioinformatics approaches to design better vaccines.

### MODELING THE VIRUS PARTICLE

Klaus Schulten, PhD, and colleagues at the University of Illinois at Urbana-Champaign recently simulated an entire virus particle—the satellite tobacco mosaic virus (STMV), one

of the smallest known viruses (see the News Bytes section of this issue). Allowing all of the virus's one million atoms to move for 10 nanoseconds showed surprising features of the tiny particle—and hinted at possible interventions to prevent infection.

Whole-virus simulations for flu—1000 times bigger than STMV—are still a ways off. But researchers could simulate pieces of the viral capsid—the exterior casing that holds a virus's genetic material—or they could model some parts of



*The HA protein from an avian flu virus (A/Vietnam/1203/2004) with one of the molecule's three receptor binding sites highlighted as a surface representation (purple) and a close-up of the receptor binding domain with key residues labeled. In her computational models, Robin Bush has found that mutations around the HA binding pocket produce flu strains with greater long-term fitness. Courtesy: James Stevens, PhD, assistant professor in the Department of Molecular Biology at The Scripps Research Institute.*



the virus in atomic-level detail while leaving other parts imprecise.

## PREDICTING FLU STRAIN FITNESS

**Robin Bush, PhD**, associate professor of ecology and evolutionary biology at the University of California, Irvine, is modeling how specific flu virus surface proteins evolve. For flu, evolutionary fitness is largely determined by the virus's ability to evade the host's immune system.

In a 1999 paper in *Science*, Bush proposed a way to predict which of the then-current lineages of influenza A was evolutionarily most fit—that is, likely to have the most descendants.

"In H3N2 [a common strain of influenza A], we have a long skinny family tree with many lineages that quickly go extinct," she says. "Why is this?" To answer that question, Bush focused on the gene for haemagglutinin (HA), a flu virus surface protein that provokes a strong immune system response. She found that the fit strains exhibited changes in amino acids in the HA binding pocket—the place where antibodies of the immune system latch onto the flu virus.

"It doesn't take much in the way of amino acid changes to keep an antibody from binding again," she says. "Antibodies are very specific. So it's not surprising that changes around the binding pocket affect the fitness of the virus."

Bush then attempted to computationally model which mutations around the HA binding pocket would lead to long-term fitness. In 9 of 11 simulations, she found that mutations in any of 18 specific amino acids predicted that a strain's descendants would continue to infect

to pick one strain over another, she says, "you *might* pick the one that had the predicted binding pocket changes."

## FILTERING THE VIRAL GENOME TO DESIGN VACCINES

**Anne De Groot, MD**, associate professor of medicine at Brown University, is tackling vaccine development head on. She's using bioinformatics to rationally design vaccines.

Annual flu vaccines are produced by

EpiVax fishes for antigens using epitopes as bait.

"It's a way of filtering genome information to find what's immunologically relevant," says Anne De Groot.

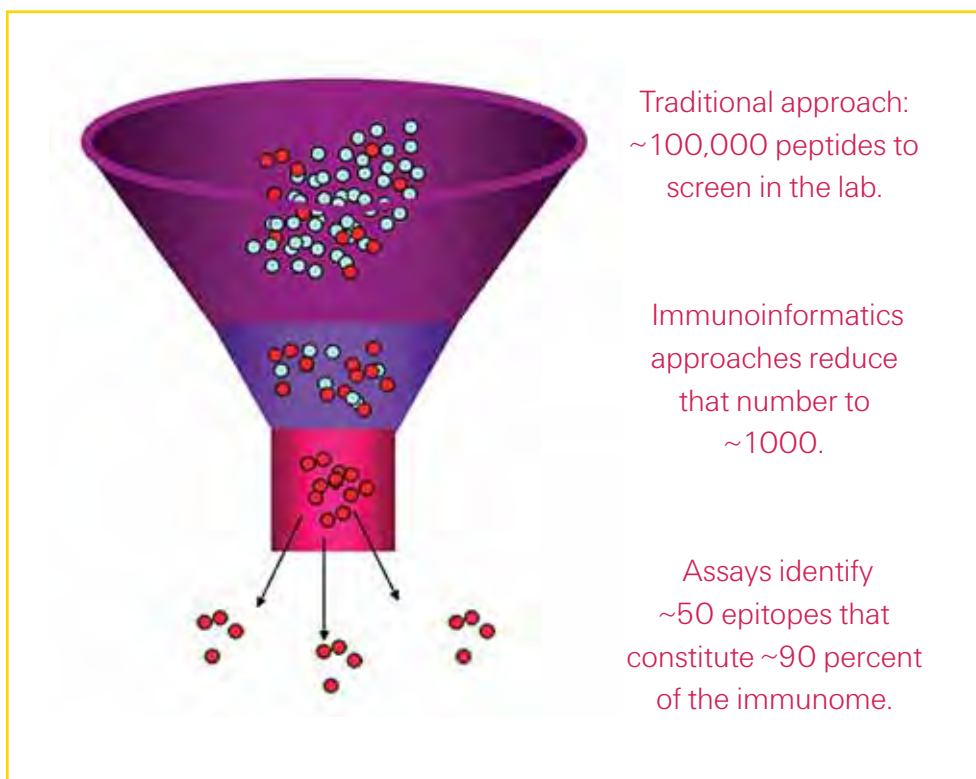
humans in ensuing years. But, Bush says, she is unable to predict if or when those expected descendants would appear.

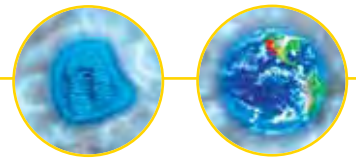
Bush also cautions that her work is not likely to contribute greatly to annual flu vaccine design. Such vaccines contain three different flu viruses, and decisions about which lineage of each to include are made by experts based on many factors. If there was no other way

growing viruses in eggs, killing them, and then combining the dead viruses with other ingredients known as adjuvant. The process is slow, so vaccines must be designed several months before the flu season begins, with strains from the prior flu season. Vaccines containing the entire contents of dead viruses—including tens of thousands of proteins with unknown side effects—can also be

risky. "You have to be very careful about what you put in a vaccine," says De Groot, pointing to the Lyme disease vaccine that appears to have caused arthritis in some patients. "We've been lucky with some other whole-virus vaccines such as polio and cholera that have not produced deleterious effects, but all of the proteins produced by a virus have potential cross-reactivity. When you create an immune response to them, you could be

*The traditional approach to epitope-mapping a typical pathogen genome could involve synthesizing 100,000 overlapping peptides. An immunoinformatics approach remarkably reduces that figure to ~1000, accelerating the discovery of ~50 epitopes that comprise over 90 percent of the immunome in an actual infection. Courtesy of Anne S. De Groot.*





creating auto-immunity or pre-setting an immune response that you don't want."

A different approach is to put the vaccine together one piece at a time, so you know exactly what's going on, De Groot says. In addition to being safer, this approach should allow development of a vaccine in response to the current flu strain (rather than last year's) because individual pieces (peptides) can be rapidly manufactured.

EpiVax, a Rhode Island biotech company founded by De Groot in 1998, uses computational tools to design peptide-based vaccines. They use a process called fishing for antigens using epitopes as bait. "It's a way of filtering genome information to find what's immunologically relevant," says De Groot. In the 1990s, researchers developed algorithms that can pick out gene motifs that are likely to stimulate the immune system. Using their own version of such algorithms, known as EpiMatrix, EpiVax filters a particular pathogen's genome to pick out snippets likely to produce immuno-stimulatory peptides known as epitopes. These peptides can then be synthesized in a lab and mixed with blood from people who have been previously exposed to the particular pathogen. The epitopes that successfully "fish out" responses in the blood are presumably part of an antigen—one of the viral or bacterial proteins to which the person's immune system responded during the earlier infection. Such antigens and/or their epitopes are potential ingredients in a vaccine, since they produce valuable immune responses. This approach has led to potential vaccines for HIV and meningitis that are now in clinical trials.

A bioinformatics approach might also contribute to development of a universal flu vaccine, De Groot says. Algorithms can screen the genomes of all the various flu strains to look for genomic sections that are pretty short and don't change. "They're kind of like the flu thumb or index finger: they are critically important to the function of the virus," De Groot says. Running these regions through another algorithm will reveal whether they stimulate the immune system. If they do, then a flu vaccine containing these proteins might induce immunity to a group of flu strains rather than just one.

## The Host: Modeling the Immune System

Using computation to understand the flu virus and its proteins only covers half the story. The T-cell mapping interface used by De Groot hints at the other half: the host immune system.

When a virus or bacterium invades the human body, it stimulates a cascade of immune system events to fend off the intruder. Over the last hundred years, experimentalists have cleverly studied these events in contexts where only one component changes at a time. It's work

tional immunology comes from HIV work published in 1995 by **Alan Perelson, PhD**, and **David Ho, MD**. It led directly to the realization that HIV could be treated with cocktails of drugs—an approach that has greatly reduced the number of deaths due to AIDS.

This research demonstrated that it's not too soon to take computational immunology seriously, Kepler says. Moreover, he adds, "the rate of accumulation of new information is so fast, that if

"Computation is a way to take all these objects [the pieces of the immune system] and put them back together into a form where the goal is not to minimize variation but to keep track of it," says Thomas Kepler.

that has generated huge amounts of data about more than 20 different types of immune cells and a few thousand participating molecules. But what's missing, say computational immunologists, is an integrated view of the puzzle.

"Computation is a way to take all these objects and put them back together into a form where the goal is not to minimize variation but to keep track of it," says **Thomas Kepler, PhD**, professor of biostatistics and bioinformatics at Duke University.

The current poster-child of computa-

we don't start now, we'll never catch up."

It's a view shared by leaders at The National Institute of Allergy and Infectious Diseases (NIAID) who, in 2004 and 2005, funded four computational immunology projects. Three of these are using flu as a model pathogen.

### THE IMMUNE SYSTEM AS A BLACK BOX

Under an NIAID grant to **Penelope Morel, MD**, associate professor of immunology at the University of Pittsburgh, researchers are modeling



how respiratory infections (influenza, tuberculosis and tularemia) affect the local immune response in the lungs. The group will be gathering data about how macrophages in the lung respond to each virus by measuring such things as secretions (cytokines) and cell surface-markers as they change through time. But the goal is to take the experimental measurements and plug them into computational models. “If your model doesn’t match the data, then you know something’s missing,” says Morel. “That exercise is a highly valuable one.”

The project’s flu modeler, **Shlomo Ta’asan, PhD**, professor of mathematical sciences at Carnegie Mellon University, is taking a highly mathematical approach: He will look at the immune system as a black box, without

“The model is easy to write out,” says Hulin Wu, “But there’s no validation without data.”

making assumptions about the biology.

“We don’t put anything into the model except the data that come out of the experiments,” he says. “Our algorithm will spit out something that might be intuitive for biologists, and it might not.” He hopes to find out if math can cut through biological intuition to gain

some new truth. After creating a model that seems to reproduce the experimental results for macrophage responses, Ta’asan says, “Then we want to see how to manipulate it with various drugs.”

The biggest challenges to Ta’asan’s model are practical ones. One mouse doesn’t have enough blood to cover all the necessary tests and must be sacrificed to get certain measurements. In addition, microarray data are highly variable and there is fuzziness in the measurements. Ta’asan says some people simply ignore that variability, but he thinks it says a lot about the system and should be accounted for mathematically. “We’re thinking about using some fuzzy logic ideas or probabilistic approaches,” he says. “We don’t want to pretend it’s not a problem.”



*This 1976 photograph shows an elderly female receiving a vaccination by a public health clinician during the nationwide swine flu vaccination campaign, which began October 1, 1976. Courtesy: Centers for Disease Control and Prevention.*

## MODELING THE IMMUNE SYSTEM USING EXPERTISE

**Hulin Wu, PhD**, professor of biostatistics and computational biology at the University of Rochester shares these concerns. He, like Ta'asan, received an NIAID grant and is modeling the immune system response to flu. But Wu is taking a more traditional approach: He develops his models based on immunologists' and virologists' current theories about flu infection. And he needs lots of data on the kinetics of the virus and the cells with which it interacts. For example, he needs to know how fast the flu virus proliferates and dies; the infection rate for various cell types; and the rate of production of T-cells, antibodies, CD4 and CD8 cells, and lymphocytes. On top of that, he needs this data from several different locations (lung and lymph nodes, for example) at various time points so that he can model the host

*"It's difficult to understand parallel events without the benefit of computational approaches," says Stuart Sealfon.*

reaction as the virus and immune cells migrate between compartments.

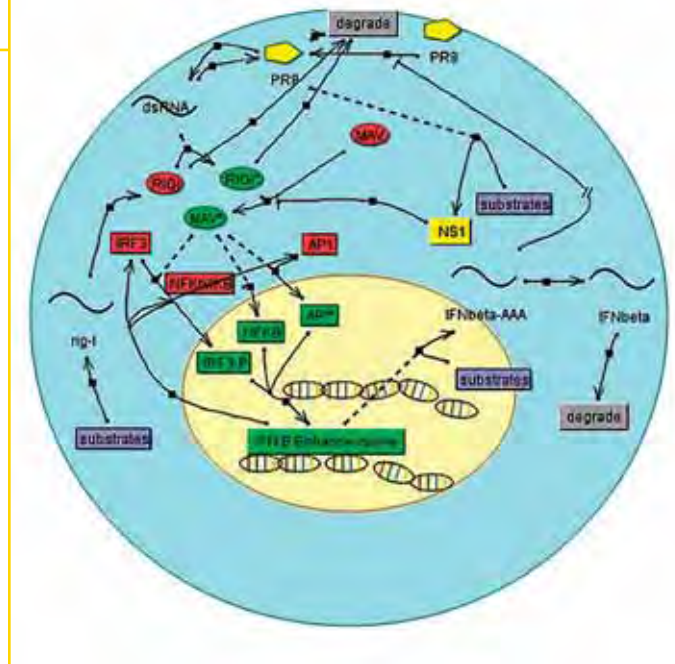
But he's finding that such data just doesn't exist for flu. Coming from HIV modeling, this can be frustrating. "HIV is a long-term infection. You can measure the immune response over many years," he says. "Flu lasts only one week, and then everything's gone." Getting enough measurements in a short time span is challenging but essential. "The model is easy to write out—to describe the interactions between the virus and the immune system in the lung, the spleen and the lymph nodes. But there's no validation without data."

The data-gathering problem would be even worse in the event of a bioterrorist event, he says. "How can we collect enough information quickly to deal with a new engineered virus?" That's when an immune system model would prove valuable. If there's a model in place for an existing flu virus, it can be quickly adjusted to a new one, he says.

## MODELING MOLECULAR LEVEL IMMUNE RESPONSES

Another NIAID group led by **Stuart Sealfon, MD**, professor of neurology at Mount Sinai School of Medicine in New York City, is using computation to get a handle on the immune system's response to flu at the molecular level. They are modeling the ways that flu viruses evade or undercut the immune system's efforts, specifically focused on the dendritic cell—the transitional cell between the innate and adaptive immune systems.

The team starts with experimental work: They infect dendritic cells with non-pathogenic viruses containing specific components of the flu virus such as NS1 (a protein that shuts down some parts of

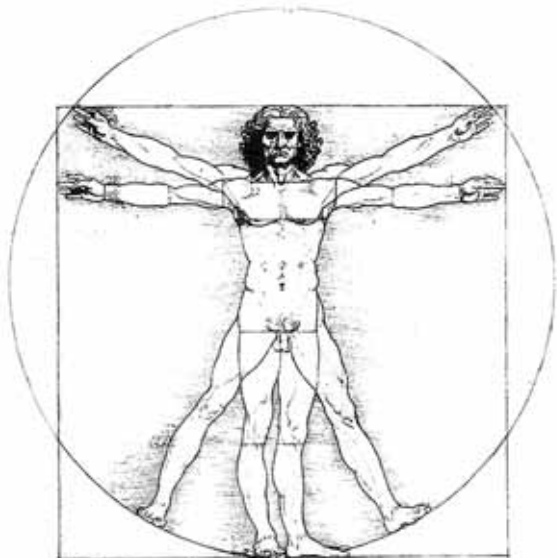


*This is the graph of a preliminary 18-equation model of flu virus antagonist effects on interferon production in dendritic cells. Both the graph and model were developed by Mount Sinai researchers using BioPathwise, a signaling simulation program developed by BioAnalytics Group LLC. Courtesy of Stuart Sealfon.*

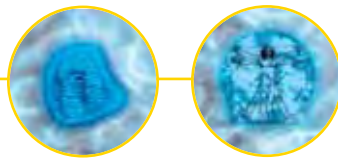
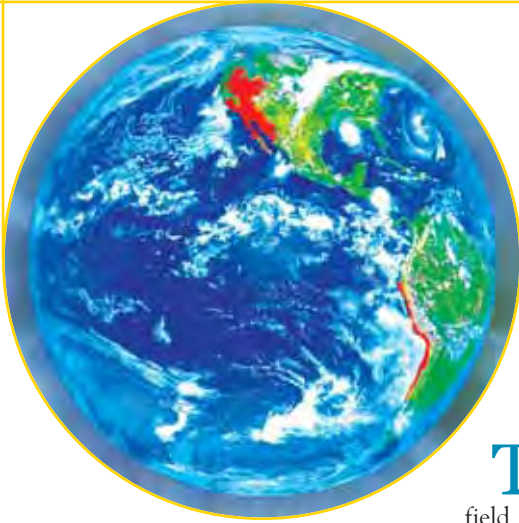
the normal signaling in such cells). This generates large amounts of data on gene and protein changes. The computer model then tracks all of these changes at once. "It's difficult to understand parallel events without the benefit of computational approaches," Sealfon says.

One of the modeling challenges, Sealfon says, is dealing with events that occur on different time scales. Signaling events take place over minutes, gene induction occurs over hours or a few days, and secretion and stimulation occur throughout the infection period. These multi-scale modeling problems still need to be addressed, he says. But if the challenges can be overcome, "ultimately, this work can help us to develop strategies to circumvent the virus's actions." And in the event of a new strain, the model can help identify the evasive tactics used by the new flu bug, which might lead to an appropriate therapy or vaccine.

Computational immunology still has a long way to go before it will fulfill its promise, Kepler concedes. But the field is really opening up, as technology provides more and more ways to measure the many complex interactions of the immune system. "There has already been a lot of good work in computational immunology," he says, "but it will have a very different flavor in the next few years."







# The World: Modeling Flu Spread

The field of computational epidemiology is a much more mature field than computational immunology, Kepler says. Because epidemiologists have always dealt with disease spread across large populations, it's not as big a leap to computation on a national and global

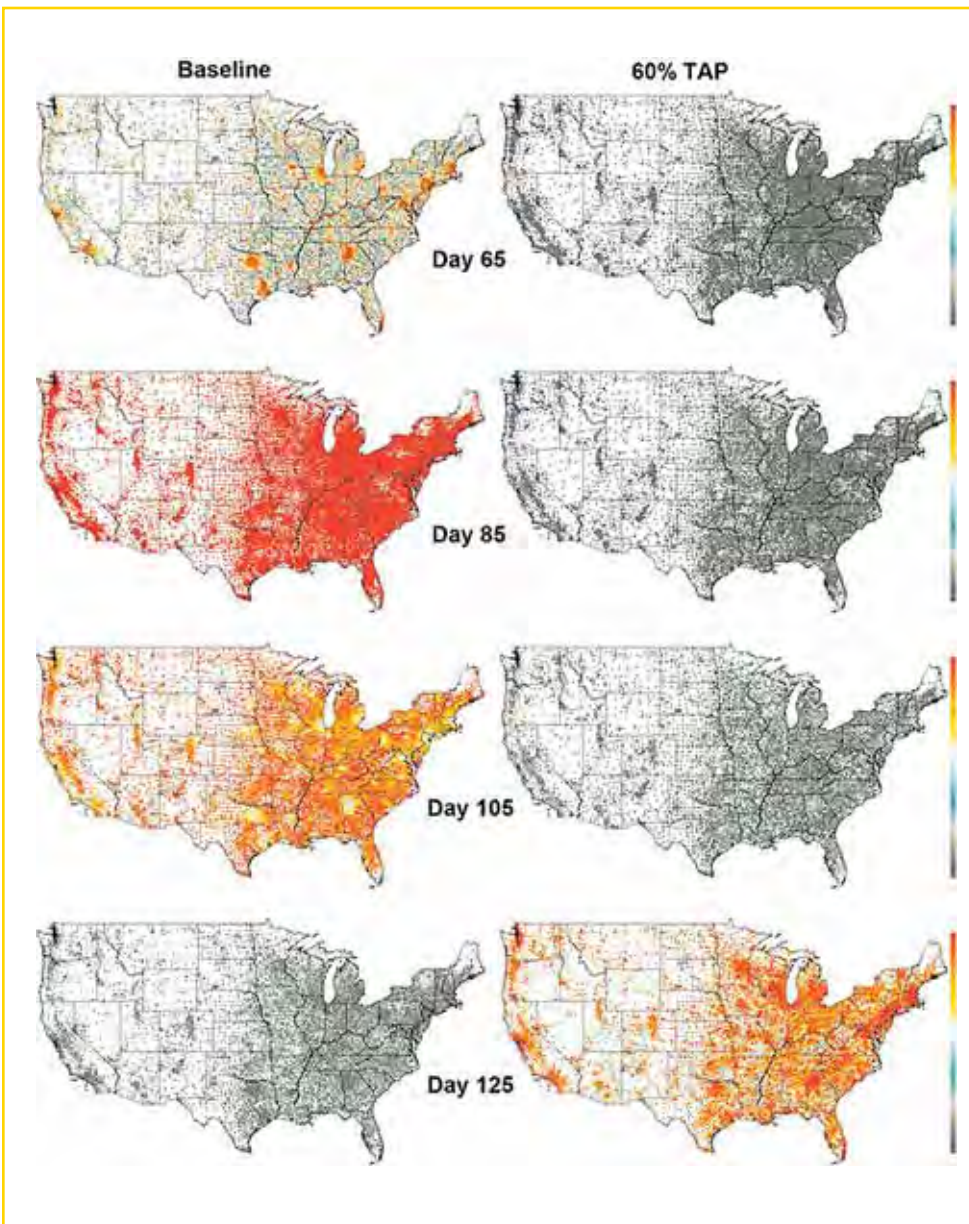
level. And the main computational approaches to epidemiological problems—agent-based modeling and graph theoretical methods—are well-established.

What is new, however, is the current United States effort to bring infectious disease modeling under one umbrella. In

2004, the National Institute of General Medical Sciences (NIGMS) within the NIH created the Modeling of Infectious Disease Agent Study (MIDAS), a program that funds several epidemiologic modeling efforts, gives them access to supercomputers, and also coordinates them in hopes of producing results that will be useful to policymakers.

MIDAS literally gets everybody in the room—programmers, data collectors, database designers, biologists, epidemiologists and statisticians—to try to iron out all the potential areas of disagreement. In particular, they try to reach consensus about what parameters should be part of the model. Telling policymakers that one program

*On day zero in these simulated pandemic outbreaks ( $R_0$  of 1.9), infected individuals arrived at 14 major international airports in the continental United States. The number of ill people at a given point in time is indicated by the color scale from green (0.3 percent) to red (3 percent) of the population. More than 40 percent of the entire U.S. population end up getting ill with no interventions (left). The use of antivirals slows the spread (right) until the stockpile of 20 million courses runs out, at which point there's a delayed nationwide pandemic. Courtesy: PNAS.*





gets one result and another program gets a different result simply won't do, says **Irene Eckstrand, PhD**, scientific director for the MIDAS program. "So we try to work all those things out in-house."

The end goal is for MIDAS to be able to tell policy makers: Based on our models, a specific intervention in a specific type of epidemic will likely have a specific effect. But, Eckstrand cautions, the models are all stochastic—they don't give the exact same answer back twice. Uncertainties are built into the models because many parameters are probabilistic. For example, the likelihood that a person will stay home on a given date rather than spread the disease to one or more people can be assigned a specific probability so that the outcome will vary each time the model runs. So each computer model must be run multiple times on a given set of parameters in order to produce a distribution of results that express the range of possible outcomes as well as the most likely outcomes.

Although the MIDAS approach could be applied to any infectious disease, the researchers decided early on—before the current concern over avian flu—that it would be interesting to model pandemic influenza. "The timing was pretty remarkable," says Eckstrand.

Because of this fortuity, MIDAS models published in *Science* and *Nature* in August 2005 and in *Proceedings of the National Academy of Sciences (PNAS)* and *Nature* in April 2006 were front page news.

MIDAS grantee **Ira Longini** co-authored two of these high-profile papers. His August 2005 paper in *Science* looked at ways to stop an outbreak of flu in an imaginary population of 500,000 people in Southeast Asia. He and his colleagues found that an outbreak could be contained if a sufficient stockpile of antiviral drugs could be delivered rapidly enough—within three weeks of the first human-to-human transmissions. In practice such an

approach would be difficult to implement in Southeast Asia but the model will help policymakers plan for an effective response.

### MODELING FLU ACROSS THE UNITED STATES

Longini's April, 2006 model in *PNAS* focused closer to home: What interventions would help contain a flu pandemic in the United States? Instead of an imaginary population, this model was built on census tract data for 281 million people and relied on extensive knowledge about peoples' travel and activity patterns. "We're all pretty predictable, really," he says. "We all get up, go to work, go shopping, and get together

"These models aren't meant to be predictive tools," Longini says. "They are meant to evaluate strategies for intervention."

with our neighbors." So Longini's model breaks down social contacts into 12-hour time periods (day and night) in seven different contexts ("mixing groups"). In some contexts, close contact occurs (home, work, schools); in others, it's more occasional (shopping malls).

The key variable for Longini's model is a number called  $R_0$ , which represents how transmissible a strain will be. Specifically:  $R_0$  is the number of people, on average, that a typical infectious person infects during the infectious period in a fully susceptible population. If that number is bigger than one, then the disease will spread. Less than one and it will disappear.

No one really knows what the  $R_0$  for a new pandemic flu strain would be. It's thought that newly emerging strains that haven't had a chance to adapt to humans might have a low  $R_0$  and there-

fore may die out. But no one has observed an emerging infectious disease before it becomes well adapted. "We kind of missed HIV and SARS," Longini says. But now, with better surveillance, virology and field epidemiology, "Flu might be the first emerging disease where we really have an opportunity to watch what happens."

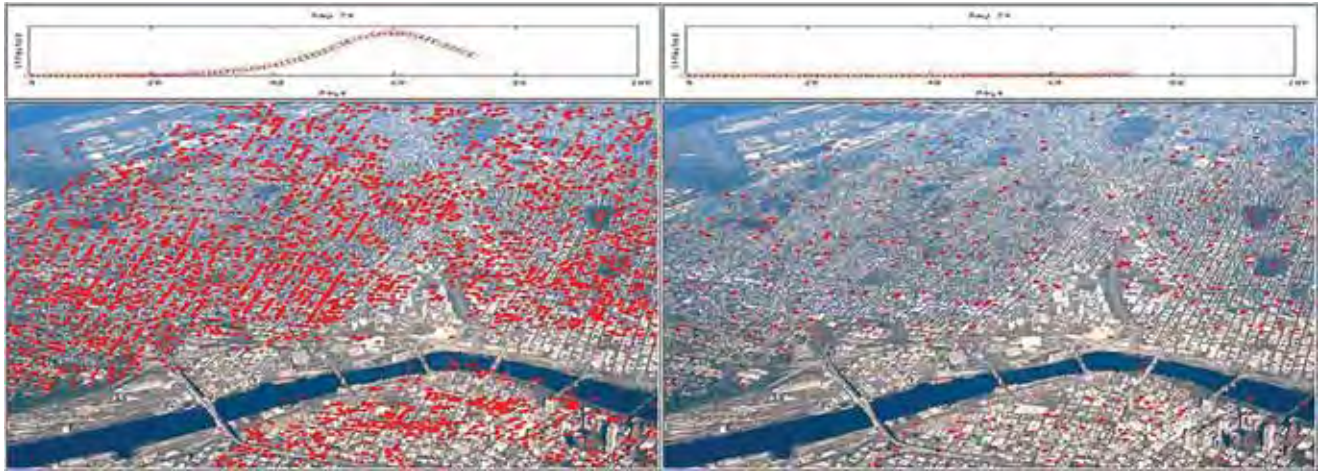
In Longini's computer model of the United States, he's assuming a well-adapted virus, so he starts with a pretty high  $R_0$  of 1.6 to 3.0 (the  $R_0$  for Smallpox is 5; for the 1918 flu, about 2). But  $R_0$  is only the starting point for the model. As different intervention strategies are tried, the  $R$  value changes. "These models aren't meant to be predictive tools," he says.

"They are meant to evaluate strategies for intervention."

For a flu pandemic with an  $R_0$  of 1.6, Longini and his colleagues found that any of several individual strategies such as antiviral drugs, child-first vaccinations or school closures could be fairly effective in reducing the incidence of flu below ten percent (the rate for a typical annual flu season). If the  $R_0$  is higher than 1.9, however, only vigorous application of multiple strategies would reduce the outbreak's impact.

Longini has been working directly with the government on intervention scenarios. For example, he can compare the impact of stockpiling 10 million versus 100 million courses of Tamiflu. And he can say closing schools is more effective than other social distancing measures while travel restrictions appear to have little impact (findings confirmed by the other MIDAS model for the United States published in *Nature*). But he's quick to point out that the model cannot predict what will actually happen. "We can say one strategy might be better than another or one might be totally ineffective and another has a good chance of being effective. So we can make those sorts of statements, and that's about as far as we can go."





An aerial view of Portland, Oregon, in a simulated epidemic performed by Eubank and colleagues. In each image, red dots represent a location with at least one infected person present at a set time on a given day. On the left is the baseline spread after 60 days. At right is what happens if 75% of households make the decision to stay home during those 60 days. These simulations are not meant to represent what would happen at a particular address, but to indicate the general consequences of such extreme behavioral modification. Courtesy: Stephen Eubank.

## GETTING DOWN IN THE WEEDS

Despite the unifying influence of MIDAS, its grantees still have their own particular approaches to modeling epidemics, says Eckstrand. “There’s an interesting discussion about how much detail you need to know in order to build higher level estimates,” she says. **Stephen Eubank, PhD**, is a MIDAS grantee who works with “down in the weeds” information, Eckstrand says.

Eubank, project director of the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute, believes models need detail in order to best address policymakers’ needs. To evaluate the relative effectiveness of strategies such as telecommuting, limiting meeting sizes or setting quotas on the number of people in a grocery store, a model must contain sufficient details about what individuals are actually doing and where. Eubank’s model consists of individual agents that each represent a single person assigned a set of activities at reasonable locations given where they live. “So we don’t have a knob in our model that says ‘reduce contact rates by thirty percent,’” he says. “Instead we have knobs that say, keep some people home from work; or don’t let more than ten people in this room.”

Right now, Eubank’s models can only be applied to one city at a time. “It’s hard to support both the amount of detail that we’re talking about in our model and the scale of the whole country. It becomes a question of computer

resources.” For a city of a million or so, each engaged in five to ten activities, a simulation covering 60 days can take an hour or two on 30 CPUs. But expanding such detailed models nationwide would require very large clusters of computers and large quantities of data. Eubank’s hope is to develop grid-based platforms. “It’s unlikely that any one person or organization would want to model this much detail for the whole United States,” he says. “But at the local level, there are good arguments for why an urban area should have such a model of itself.” It would be useful not only in the event of an epidemic, but for other kinds of urban planning.

If cities participate in Eubank’s plan, they could then tie their models together in a grid to create a nationwide, detailed model. “So we’d have this loose federation of urban or regional models interacting across the grid, each maintained by someone with a vested interest in having a good model of their area.”

## OPTIMIZING INTERVENTIONS

**Catherine Dibble, PhD**, assistant professor in the department of geography at the University of Maryland, College Park, also offers a different perspective within MIDAS. As a collaborator on the MIDAS grant headed up by **Donald Burke, PhD**, at Johns Hopkins University, she has developed tools for

doing two things many others haven’t done: risk analysis and optimization. “Most pandemic modelers decide the interventions and settings by hand and run them through the simulations,” she says. “We do that too, but we can also optimize interventions and evaluate their risks.”

So, while Longini and other MIDAS

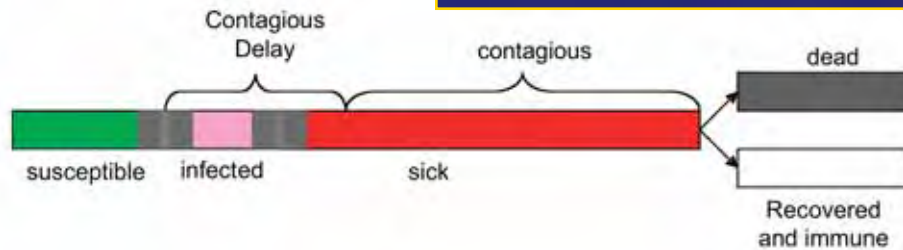
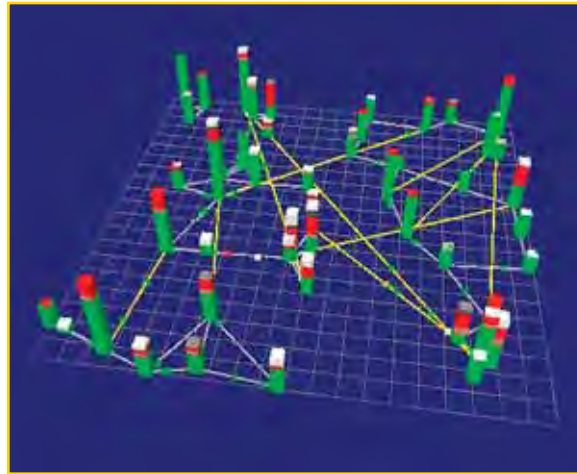
Telling policymakers that one program gets one result and another program gets a different result simply won’t do, says Irene Eckstrand, “So we try to work all those things out in-house.”

modelers (such as **Neil Ferguson, PhD** and **Mark Lipsitch, PhD**; see [www.epi-models.org](http://www.epi-models.org)) recommend which local interventions and combinations of interventions could be most effective, Dibble has the capacity to evaluate the optimal geographic deployment of those recommended interventions and associated scarce resources such as Tamiflu and vaccine supplies.

In addition, Dibble’s risk analysis tools can evaluate the optimal strategies



*Catherine Dibble's approach to flu modeling is similar to her earlier work with SARS (published in 2003). In this simulated landscape of SARS spread, each tower represents a population center. Colors represent the health status of individuals at each center:*  
 green = healthy  
 pink = infectious  
 red = symptomatic  
 gray = deceased  
 white = recovered (now immune).



**“If these models can be useful at all,” Catherine Dibble says, “people need to be comfortable with them and understand how a particular intervention can help.”**

to see how well they deal with events that don’t go as planned. As Dibble explains, “some interventions might give a good outcome under some conditions, but, compared to other possible interventions, might be more sensitive to chance events that could work against them.”

But optimization and risk analysis would require huge amounts of computer resources if applied to the fully detailed national models, Dibble says. “Effective optimization requires a model that represents key aspects of geographic structure and travel behavior, yet is simple enough to run hundreds of thousands of times to fully explore alternative geographic deployments and to explore uncertainties, sensitivities and risks.”

Dibble’s model is designed to evaluate the effect of travel restrictions between transportation hubs in the

event of pandemic flu. She created a network with healthy individual agents (green) distributed at each transportation hub all across the continental United States. Each population center can be visualized as a tower with its height determined by its relative population. “Then we drop one or more infected individuals into the landscape,” says Dibble. “They are pink.”

As time goes by, different people make different travel decisions (modeled using actual airline routes and travel data) and the infected agents start “sneezing” on people (infecting them) at a rate consistent with a particular  $R_0$  and whichever interventions may be imposed. Sometimes the epidemic fizzles out—the equivalent of the infected person going home and not giving the disease to anyone. When it doesn’t fizzle out, pink

(infectious), red (sick), gray (dead), and white (recovered) people appear on the landscape, with travel decisions leading to diffusion among cities. “We focus on evaluating the relative pandemic risks across cities: Which cities in the United States are likely to be hit soonest or more often,” she explains.

In the event of a pandemic, her model can suggest how to allocate the available (and limited) resources effectively, Dibble says. Spreading interventions uniformly over the population might seem fair, but it might not control the pandemic as effectively as targeting the resources to particular cities.

Convincing policymakers to focus resources geographically could be a big challenge, she says.

“If these models can be useful at all, people need to be comfortable with them and understand how a particular intervention can help.” That kind of public awareness, she says, will be key. According to her, “Communication may turn out to be more important than any particular model, vaccine or resource.”

### BRINGING BUG, HOST AND WORLD TOGETHER

As with many modeling endeavors, the question arises: What if the models could be integrated across the scales? Will we eventually model an evolving flu virus interacting with the host immune system in such a way as to predict, with reasonable reliability, its effect on a population? If so, that day is not near. But even now, efforts by MIDAS researchers might help stem the spread of a flu pandemic, potentially saving millions of lives. Even if the models don’t help, and a pandemic rages uncontrollably, the work will help prepare us for the next time, and the one after that. “Pandemic flu is a big threat,” says Longini, “but it’s also a really important scientific opportunity.” □



BY MICHAEL SHERMAN

### Complex Step Derivatives: How Did I Miss This?



One of the tasks faced by every scientific programmer sooner or later is the need to compute the derivative  $f'(x)$  from code for the original function  $f(x)$ . This need arises in design and minimization problems, for example. In practice  $f$  is often an enor-

sober perspective is that we lost *nine orders of magnitude!*

We can improve this somewhat by calculating more terms in the Taylor expansion of  $f'$ ; for example,  $(f(x+h)-f(x-h))/2h$  (central difference) gives a few more digits at twice the

have to modify  $f$  to accept a complex argument. That's easy in languages like C++ and FORTRAN with built-in com-

This result is so surprising you have to see it to believe it.

mous, messy, "legacy" numerical computation, and  $x$  is a vector of many arguments. We all know the standard finite difference trick:

$$f'(x) \approx (f(x+h)-f(x)) / h$$

which converges exactly on  $f'$  in the limit  $h \rightarrow 0$  ... except *that* can only hap-

pen in a calculus textbook! In practice, roundoff error ruins the accuracy of the difference  $f(x+h)-f(x)$  as the arguments get closer together. So we try to balance roundoff error caused by  $h$  being too small against "truncation" error from  $h$  too large. Optimal balance is *usually* found near  $h=\sqrt{\epsilon}$  where  $\epsilon$  is the precision with which  $f$  can be calculated, although exceptions abound. On a good day, this yields seven correct digits of  $f'$  when  $f$  has sixteen. Most of us think of that as "about half the accuracy" but a more

expense. We're still down *six* orders of magnitude (assuming we picked  $h$  well). William Press expressed it best:

It is disappointing, certainly, that no simple finite-difference formula ... gives accuracy comparable to the machine accuracy.—*Numerical Recipes in C++* (2003)

plex numbers, and some automated tools have also been developed.

This result is so surprising you have to see it to believe it. The inset shows a complete C++ program that differentiates  $f(x)=\sin(3x)\log(x)$  by finite, central, and complex step differencing (in yellow), and analytically to check the answer. Here is

```
typedef double      Real;
typedef complex<Real> Complex;
const Complex i(0,1); // sqrt(-1)

Real f(Real x) {return sin(3.*x)*log(x);}
Complex fc(Complex x) {return sin(3.*x)*log(x);}

int main() {
    Real h2=1e-7, h3=1e-5, hc=1e-20, x=0.7;
    Real dfdxFinite = (f(x+h2) - f(x)) / h2,
          dfdxCentral = (f(x+h3) - f(x-h3))/(2*h3),
          dfdxComplex = fc(x*hc*i).imag() / hc,
          dfdxAnalytic = sin(3*x)/x
                        + 3*cos(3*x)*log(x);

    printf("f' finite = %.16f\n", dfdxFinite);
    // ...
}
```

```
f' finite = 1.7733539392494890
f' central = 1.7733541058356781
f' complex = 1.7733541062373444
f' analytic = 1.7733541062373446
```

the output, with correct digits highlighted: Complex step matched to *sixteen* decimal places, full machine precision. I chose  $10^{-20}$  as the complex step size, but  $10^{-100}$  works just as well!

Simbios Center faculty member Michael Levitt recently reported a breakthrough in coarse grained molecular modeling of myosin. He replaced a numerical difference calculation of a large matrix with the complex step method, and can now closely match all-atom normal modes with a simplified model. Perhaps more importantly, he now has a reason to gloat about being a FORTRAN programmer!

There is much more to learn about this fascinating idea, including some practical issues to consider, a deep relationship with automatic differentiation theory, and historical roots in work done in the 1960s by Simbios Scientific Advisor Cleve Moler. For more information, see the referenced paper. Then give it a try yourself and let us know what happens. □

But maybe we all did too much science and not enough math. I recently stumbled on a paper<sup>1</sup> reporting an amazing result from complex analysis:

$$f'(x) \approx \text{Im}[f(x+hi)] / h$$

This says to perturb  $f$  along the *imaginary* axis, and then take the imaginary part of the result. Otherwise it looks deceptively like the usual finite difference formula. But look again—this one contains no subtraction and hence no roundoff error. So we could hope to make  $h$  smaller to reduce the truncation error as well. Here is the amazing part: you can make  $h$  as small as you like, and as long as  $h < \sqrt{\epsilon}$  you'll get the derivative to *machine accuracy*. Of course you do

#### DETAILS

Michael Sherman is Chief Software Architect for the Simbios Center.

#### FOOTNOTES

1 Martins, J. R., Sturdza, P., and Alonso, J. J. 2003. The complex-step derivative approximation. *ACM Trans. Math. Softw.* 29(3) (2003).

BY LOUISA DALTON

## JANELIA FARM: Cultivating Scientists

The folks at Howard Hughes Medical Institute who dreamed up Janelia Farm say it is as much a social innovation as a scientific one. “We are creating a different culture here,” says **Gerald Rubin, PhD**, director of HHMI’s first freestanding research institute under construction in Loudoun County, Virginia. “Most professors don’t do lab work anymore. They spend time on committees, write grant proposals, and teach. We want to be on the much more adventuresome end of things.”

Researchers at Janelia Farm, Rubin says, will above all do research with their own hands. They will have small, easy-to-manage laboratories and no teaching, grant-writing, or administrative responsibilities. They will work on a campus designed to promote run-ins with other researchers, especially those from vastly different backgrounds. And they will self-assemble into novel, cross-disciplinary collaborations to work on long-term, unwieldy scientific problems difficult to tackle in a single laboratory.

That’s the idea behind the social experiment of “The Farm.” Built on a 689-acre tract of land 30 miles outside of Washington, D.C., Janelia Farm is due to start operating this summer and will have its grand opening in early October 2006.

When **Sean Eddy, PhD**, associate professor and a computational biologist at Washington University School of Medicine in St. Louis, heard Rubin speak in 2001 about the creation of Janelia Farm, he wanted in. “Gerry said, ‘We don’t know what we are going to do yet.’ I said, ‘I don’t care. Keep me posted. The culture by itself is an attractive thing for me. Hopefully scientifically, it’ll be a good fit.’”

HHMI eventually settled on two broad initial goals for Janelia Farm: first, develop computational tools for image analysis and second, identify how neuronal circuits process information. HHMI deliberately chose ambitious goals that are best suited to a 50-year multidisciplinary collaboration rather than goals that could be addressed with a five-year federal grant.

Eddy was chosen as one of the Janelia Farm group leaders, even though he specializes in computational genome sequence analysis rather than neuroscience or image analysis. He’s thrilled with the challenge of working his way over to neurobiology. He lists some of the research ideas he has been throwing around in rank



**Top: Gerald M. Rubin, PhD, HHMI vice president and director of the Janelia Farm Research Campus. Above: Janelia Farm buildings are designed to foster impromptu conversations. Photos by Paul Fetters.**

order from most to least sane. Perhaps he’ll take software he has already developed for identifying mRNA secondary structures and apply it to the study of neuronal mRNA localization. Or, in collaboration with others, he might use computational techniques to build synthetic promoters for specific neurons in the fly, the worm, and the mouse. For that, he’d want to work with other group leaders such as **Julie Simpson, PhD**, who just finished a postdoctoral fellowship at the University of Wisconsin-Madison and has been mapping the brain of the fruit fly; **Karel Svoboda, PhD**, a neuroscientist at Cold Spring Harbor who has found a way of monitoring individual synapses in the mouse brain; and Rubin, who led the effort to sequence the fruit fly genome. One of Eddy’s





HHMI's new research campus is located on a 689-acre tract in northern Virginia. Photo by Paul Fetters.

more crazy ideas, he says, is to try to simulate the behavior of the *C. elegans* worm from its wiring diagram. He's organizing one of the first onsite scientific meetings at Janelia Farm to discuss this challenge.

Although only eight of what will eventually be 24 Janelia Farm group leaders have been picked so far, Eddy's already collaborating with most of them. That's exactly

important than the 1200 CPUs and the 10,000 mouse-cage vivarium, Peterson says, are the large round tables in the dining room, the housing for visiting scientists, and the pub, where "productive collisions" between researchers are sure to occur.

Eddy is counting on at least one of those productive collisions happening to him. "From osmosis, from hang-

Researchers at Janelia Farm will self-assemble into novel, cross-disciplinary collaborations to work on long-term, unwieldy scientific problems difficult to tackle in a single laboratory.

what Rubin hoped would happen. Whatever their current expertise, all of the group leaders are extremely creative thinkers. Most have heavily quantitative backgrounds in areas such as computer science, physics, or mathematics. And most make a habit of inventing things—physical tools, gene lines, or analysis techniques.

Even the director of information technology at Janelia Farm, **Marshall Peterson**, who worked as vice president of IT at Celera Genomics, will invent tools to help the researchers as they need them. Peterson says that the goal for IT at Janelia Farm "is to build a very flexible shared computing environment that we can scale and adapt as needed when people come to us with computational challenges." He is starting off with 1200 CPUs and 150 terabytes of storage and leaving room for whatever else they might need. "The trick is to not paint yourself into a corner," he says.

Other tools at Janelia Farm include equipment for electron microscopy, light microscopy, genomic sequencing, instrument fabrication, transgenic animal studies, and more. Peterson emphasizes, however, that at Janelia Farm, "it's not all about tools. It is about people getting together, talking, interacting, exploring, giving full reign to their imaginations." Almost more

ing out with all of these neurobiologists," he says, "I'm hoping to eventually have a smart idea." If Janelia Farm works the way it has been designed to work, Eddy and his colleagues will have many. □

#### JANELIA FARM QUICK FACTS

**WHAT:** Howard Hughes Medical Institute's first freestanding research institute

**WHERE:** Ashburn, Virginia—roughly 30 miles northwest of Washington D.C.

**STAFF:** Will have up to 300 resident research staff (including group leaders, postdocs, and graduate students) and 80 support staff, plus up to 100 visiting scientists

**FACILITY:** \$500 million research campus (including a 900-foot-long laboratory building, conference facilities and hotel, and visiting scientist housing) designed by Rafael Viñoly to foster collaborative science and adapt to changing needs

**FUNDING SOURCE:** Howard Hughes Medical Institute

**HISTORICAL MODELS:** Medical Research Council Laboratory of Molecular Biology (MRC LMB) in Cambridge, England, and AT&T's Bell Laboratories in New Jersey

**SCIENTIFIC CONFERENCES:** Will host at least twelve per year

**WEB PAGE:** <http://www.hhmi.org/janelia/>

*Biomedical Computation Review*

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

seeing science

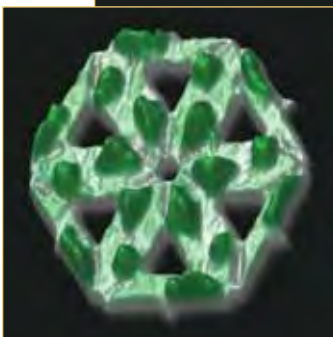
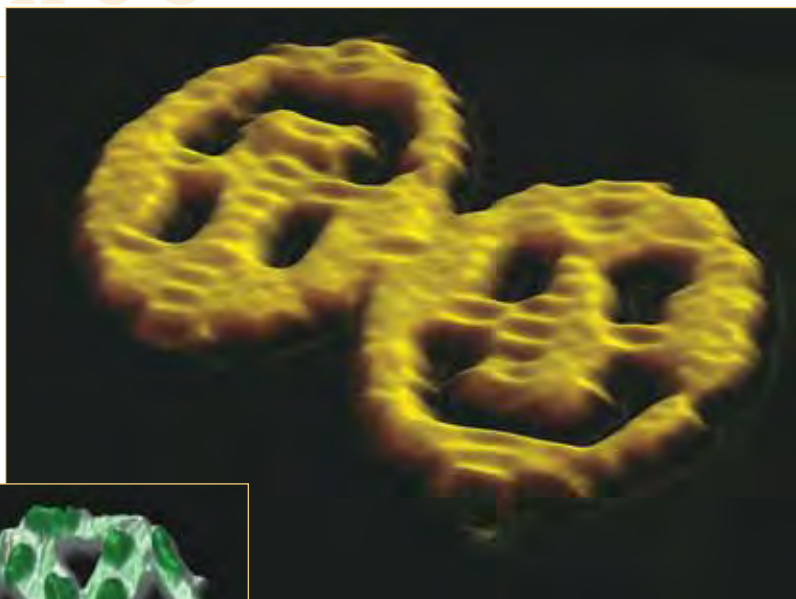
## SeeingScience

BY KATHARINE MILLER, MANAGING EDITOR

### Making DNA Smile

**D**esigning nanostructures of DNA just got easier. **Paul Rothmund, PhD**, a senior research fellow at Caltech has found a way to coax a long strand of DNA into a pre-determined geometric shape by mixing it together with some well-designed “staples” (oligonucleotides). After designing the shape and the staples on a computer, Rothmund has produced smiling faces, a map of the western hemisphere, stars, and a wide range of geometric shapes with good to excellent yields: between 60 and 90 percent of the time, the intended shape

comes out perfectly. The work, which could have ramifications for the design of nanodevices, was published in the March 16, 2006 issue of *Nature*; more shapes can be seen at <http://www.dna.caltech.edu/~pwkr/>.



*Each of these two smiling friends is actually a giant molecular DNA complex, 100 nanometers across and 5 megadaltons in mass. They are created by self-assembly, in a single reaction step, in which a 7000 base long single strand of DNA is folded by about 250 short DNA strands, each about the length of a PCR primer. Roughly 50 billion smileys are made in a single drop of water at once. The hexagon is about 250 nanometers across and is composed of 6 origami triangles linked together. Courtesy: Paul Rothmund and Nick Papadakis.*