DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

PLUS

Tool Dissemination

Doing It Right

NCBC Update:

Shedding New Light On

BIOLOGICAL COMPLEXITY

Winter 2008/2009

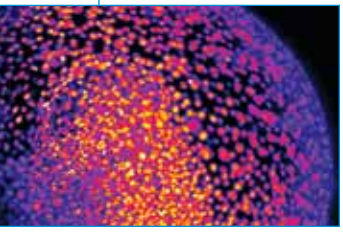
contents

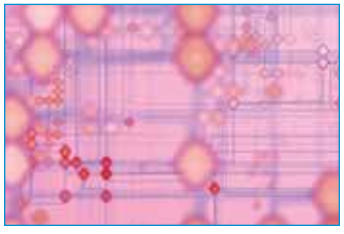
ContentsWinter 2008/2009

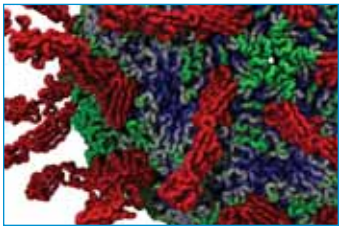
FEATURES

DEPARTMENTS

Cover Art: Created by Rachel Jones of Wink Design Studio using a collection of images from the NCBCs. Robotic arm is © Eraxion | Dreamstime.com







Winter 2008/2009
Volume 5, Issue 1
ISSN 1557-3192

Executive Editor David Paik, PhD

Managing Editor Katharine Miller

Associate Editor Joy Ku, PhD

Science Writers
Katharine Miller • Kristin Sainani, PhD
Cassandra Brooks • Emmanuel Romero
Hadley Leggett, MD • Kayvon Sharghi
Lisa Grossman • Lizzie Buchen
Michael M. Torrice • Michael Wall, PhD
Molly Davis • Stephanie Pappas

Community Contributors
Mark Musen, MD, PhD
Kartik Mani, PhD
Joy Ku, PhD

Layout and Design
Wink Design Studio

Printing
Advanced Printing

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

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

Biomedical Computation Review is published quarterly by:

simbios
An NIH National Center for Physics-Based Simulation of Biological Structures

Publication is made possible 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)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Yuan Liu, PhD (NINDS)
Richard Morris, PhD (NIAID)
Joseph Pancrazio, PhD (NINDS)
Grace Peng, PhD (NIBIB)
Nancy Shinowara, PhD (NCMRR)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)
Jane Ye, PhD (NLM)

GuestEditorial

BY MARK A. MUSEN, MD, PhD

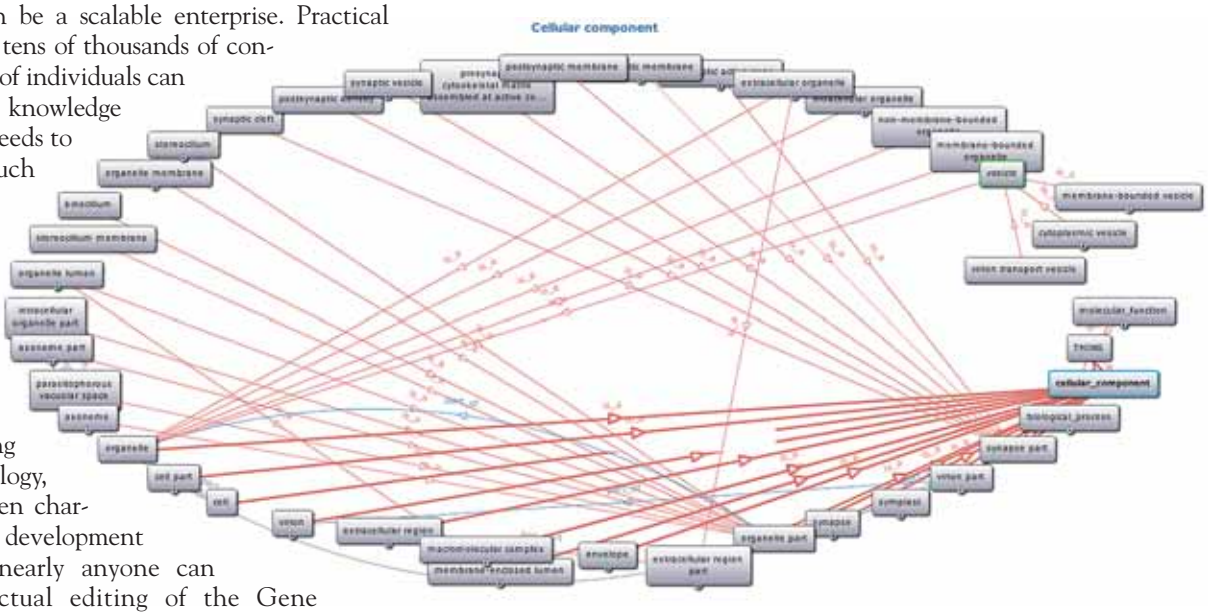
It Takes a Village: Building the Next Generation of Biomedical Ontologies

Although the notion of ontology has been around since Aristotle, the perceived need to develop ontologies in biomedicine has accelerated in recent years as investigators attempt to make sense of the terabytes of high-throughput data that are now finding their way into public repositories. While the number of biomedical terminologies and ontologies continues to increase as new areas of biomedical content become formalized, the creation and annotation of these resources can't quite keep up. The flood of information may necessitate a new approach involving vastly more ontology developers. It may, in fact, take a village.

The construction of biomedical ontologies has long been a cottage industry, with even vast systems such as SNOMED (the Systemized Nomenclature of Medicine) initially representing the handiwork of a very small group of dedicated individuals. Venerable ontologies such as the Foundational Model of Anatomy and the NCI Thesaurus represent the work of a surprisingly small set of developers. Nevertheless, as the demand for ever larger and more granular ontologies accelerates, and as large-scale systems such as the International Classification of Disease are being reengineered, the scientific community has increasingly raised concerns about whether ontology development ultimately can be a scalable enterprise. Practical ontologies comprise tens of thousands of concepts, and a handful of individuals can never have personal knowledge of everything that needs to be represented in such a system.

To address this problem, workers in biomedicine are attempting to democratize the development of large-scale ontologies. The engineering of the Gene Ontology, for example, has been characterized by an open development process to which nearly anyone can contribute. The actual editing of the Gene Ontology content, however, is still performed by only a handful of trusted curators. The National Cancer Institute is experimenting with an open process for extensions to the current content of the NCI Thesaurus via the BiomedGT initiative. Here, nearly anyone can annotate the ontology via a Web-based wiki and suggest changes and extensions, although again, the modification of the actual ontology content will be channeled through a set of trained individuals who understand principles of knowledge representation and the use of knowledge-editing tools.

Probably the best exemplar of an open, nearly democratic ontology-development initiative is the Open Directory Project (ODP). Founded more than 10 years ago, the ODP has enlisted more than 75,000 volunteers to flesh out the extensive open-content ontology of Web pages that has been adopted by Google, Yahoo!, Netscape, and a host of other companies. The ODP has generated an enormous ontology (commonly known as dmoz) that provides standard, categorized entrée to virtually all the content on the Web. All of us use dmoz, perhaps unknowingly, every time we browse the Web by categories in Google and Yahoo!, rather than searching the Web's free text for particular terms. Embracing everything imaginable that a user could search for, dmoz is a remarkable demonstration of how scalable ontology engineering can be, particularly when volunteers step forward to provide fine-grained descriptions of their particular areas of personal interest.



At bioportal.bioontology.org, the user is presented with a list of possible ontologies to explore and visualize. This screenshot of the Cellular Components ontology shows the relationship between major components and some of the minor components. Any registered user can comment on any ontology.

The dmoz ontology is very simple in its structure, and lacks the rich semantics of ontologies developed in formal knowledge representation systems such as the Web Ontology Language (OWL). When the developers of dmoz make modeling errors, the consequences are unlikely ever to impede the advancement of science or to threaten lives. Nevertheless, the dmoz ontology stands as a stunning example of how legions of volunteers can be mobilized to generate an enormous and undeniably useful ontology. Imagine if the lessons of dmoz could be applied to SNOMED or to BiomedGT!

At the National Center for Biomedical Ontology (NCBO), we are experimenting with ways in which the biomedical community can take an active part in contributing to the construction of scalable ontologies and controlled terminologies. Our BioPortal system allows any registered user to comment on any ontology in our distributed repository, to comment on the comments left by other users, and to demonstrate how the elements of one ontology may relate to those of another. We have used this capability extensively in the engineering of the Biomedical Resource Ontology used to describe the online software and data resources developed by the National Centers for Biomedical Computing and by the recipients of Clinical and Translational Science Awards. BioPortal, at present, does not play a role in completely open ontology editing, however.

There are very legitimate concerns about how we can maintain the quality of ontologies if the development process is democratized. Organizations such as the Open Biomedical Ontologies (OBO) Foundry have been established under the assumption that there must always be central management of ontology development to ensure the quality of the content. And yet there continue to be too much data, too many medical records, and too many experiments for the ontology-development community to keep up with existing needs.

It is clear that the ontology-development community needs at least to experiment with new methods of ontology engineering that can scale to future biomedical requirements.

I don't know whether the dmoz approach will really be practical in biomedicine, but it is clear that the ontology-development community needs at least to experiment with new methods of ontology engineering that can scale to future biomedical requirements. Surely there are ways to take advantage of the expertise distributed among all biomedical investigators in a way that will overcome many of the limitations of centralized ontology curation. Workers at NCBO are extremely excited about the possibilities that new technology might provide in enabling this more open approach to ontology engineering. Experimentation with community-based ontology development not only may accelerate the engineering of badly needed ontology content, but also can provide a laboratory for the study of new mechanisms for collaboration and interaction in biomedicine. □

DETAILS

BioPortal: http://bioportal.bioontology.org

The Open Director Project: http://www.dmoz.org

Mark A. Musen, M.D., Ph.D. is Professor of Medicine (Biomedical Informatics Research) and Computer Science at Stanford University. He is Director of the Stanford Center for Biomedical Informatics Research and principal investigator of the National Center for Biomedical Ontology (NCBO).

A Note from the Managing Editor:

Thanks to all who participated in the BCR survey. Your names were entered in a drawing for an iPod shuffle which went to Alan Villalobos from DNA2.0. The survey results are helping us to plan for the future.

If you didn't get a chance to answer the survey, you can still give us feedback on the magazine by visiting http://www.biomedicalcomputationreview.org and clicking on the "Feedback" link.

Starting in our next issue, we will launch a new "Debate" column, starting with the topic selected by the survey respondents: "To Mine or Not to Mine: Are clinical data repositories useful sources of untapped discoveries awaiting data-mining algorithms or are they too noisy and messy."

Best,

Kathy Miller MANAGING EDITOR

BY JOY P. KU, PhD

Stop Wheel Reinvention, Share Your Simulations!

Simbios has built a new publication repository that links publications to the research data and software behind them. The goal: to encourage and facilitate replication of published results and to foster use of what has already been accomplished rather than leaving others to reinvent the wheel. The repository is built upon Simtk.org—Simbios' web-based infrastructure that provides open access to simulation software tools and models—making it easy to use and accessible to all.

"The publication repository is more than just the collection of data, models, and software used in the publications," says **Jeanette Schmidt, PhD**, Executive Director of Simbios. "It provides the means for others to reproduce and build upon the results of your publication."

Giving Your Published Research a Future

Historically, when researchers have come across papers describing potentially useful software or data, their chances of actually getting their hands on that software or data were hit or miss. The student who did the research might have moved on, or the software developer might want to clean up the code first and take months (or longer) to do so. The Simbios publication repository for physics-based simulations of biological structures addresses this problem by providing a simple way to share and access the software, data, and other materials that support a particular research paper. It means that all the hard work behind the paper—the hours of coding, the repetitive experiments to get useable data—is captured and can easily enable future research.

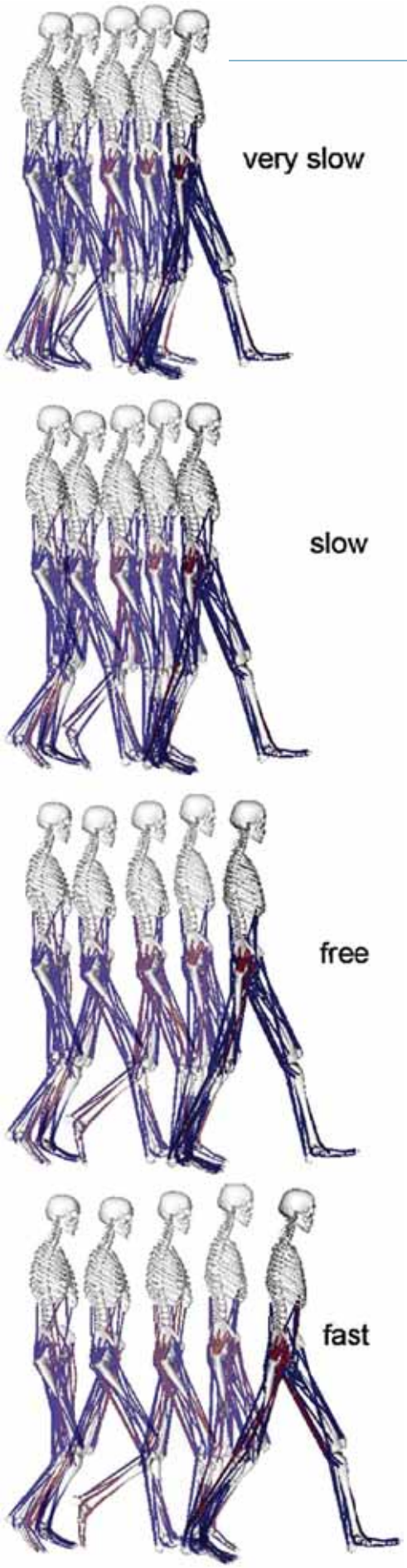
That's what motivated **May Liu, PhD**, a recent graduate from Stanford University's mechanical engineering department, to create a publication project on Simtk.org. She is sharing 32 walking simulations used to analyze how muscle functions change with walking speed in children. It's the largest number of simulations ever included in a muscle-driven simulation study, yet Liu sees it as just the beginning of further research rather than the end of the line. "The simulations themselves could become the starting point for a number of other studies," Liu says. "There's no reason why people should have to recreate simulations that already exist."

Dahlia Weiss, a doctoral student in structural biology and chemistry at Stanford University, has a similar perspective. She established a publication project for her article comparing her Climber software tool against four other tools for interpolating between two molecular structures as one morphs into the other. Climber, based on a non-linear interpolation method, turned out to be very good at producing intermediate structures for very large, complicated changes.

"Knowing that we have a really good tool and not making it publicly available just seems really pointless," says Weiss. She thinks Climber would be useful wherever high fidelity intermediate structures are required, not just for looking at structural movement.

Replicating Research Goes Beyond Software Sharing

But the Simtk.org publication repository is not just about software sharing. It supports and encourages sharing anything needed to replicate research results. For example, Stanford University researchers **Yuan Yao, PhD**, a post-doctoral fellow in the math department, and



On Simtk.org, May Liu created a publication project to share the three-dimensional simulation results from her latest publication analyzing eight subjects walking at four speeds (very slow, slow, free, and fast). Shown here are still images from simulations of a representative subject. The goal of the publication projects is to encourage and facilitate replication of published results. Courtesy of May Liu. Reprinted with permission from Liu, MQ, et al., Muscle contributions to support and progression over a range of walking speeds,* Journal of Biomechanics *(2008) 41:3243–3252.

Xuhui Huang, PhD, a research associate in the bioengineering department, and their colleagues developed Mapper, a tool that improves detection of low-density states within a massive amount of data. After creating a project for Mapper on Simtk.org, they submitted a paper showing how Mapper could be used to identify intermediate stable states during the RNA hairpin folding process, a difficult task when those states represent only two to three percent of the whole data set. In order for someone else to replicate that research, Yao and Huang posted (on Simtk.org) not only the Mapper software, but also the project's input data and instructions about how to use that data with Mapper.

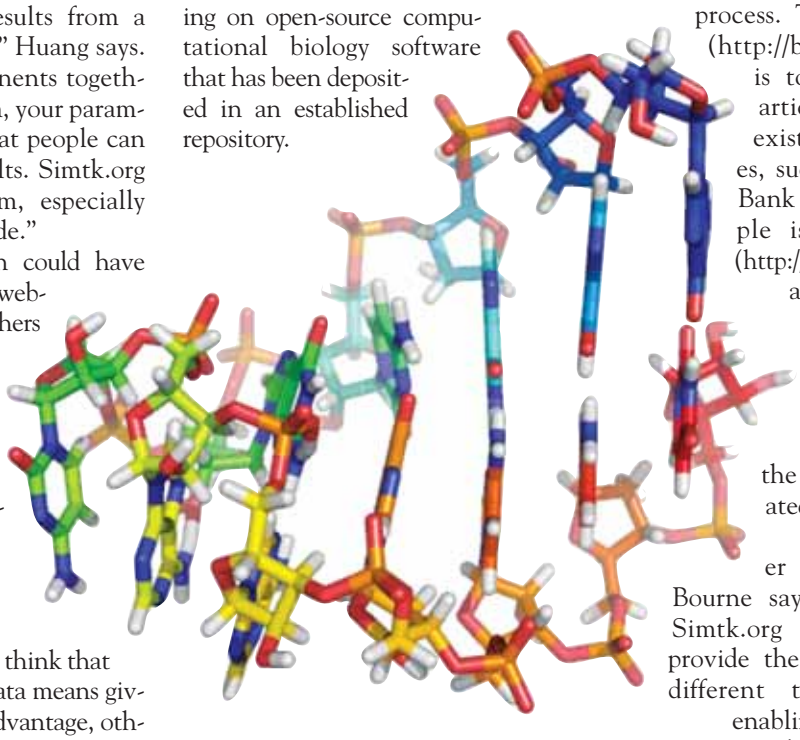
"To reproduce the results from a paper is not an easy task," Huang says. "You need all the components together—the data, the program, your parameters, instructions—so that people can easily reproduce the results. Simtk.org provides such a platform, especially with this publication mode."

While the information could have been posted on his own website, Yao says that researchers from other fields would not think to look there. For an interdisciplinary field, a common platform like the Simtk.org publication repository is particularly valuable.

Rewards for Sharing

While some researchers think that sharing their software or data means giving up their competitive advantage, others believe that it is a great way to build a successful career. "Careers often come from the application of software to make new discoveries in the life sciences," says **Philip Bourne, PhD**, a professor in pharmacology at the University of California at San Diego, and founding editor-in-chief of the journal *PLoS Computational Biology*. "So by making the software available, researchers open up that possibility to benefit from what other people do with the software as well."

Bourne acknowledges that the current system does not always reward the work involved in preparing and supporting open-source software: answering questions from software users and providing documentation, examples, and tutorials. All of that effort takes time that could be spent doing research that would generate more publications—the metric by which academics are primarily judged. To address this concern, Bourne says, *PLoS* is considering having a special section that only publishes articles reporting on open-source computational biology software that has been deposited in an established repository.



Molecular dynamics were used to simulate the folding of the RNA hairpin structure (above), generating hundreds of thousands of molecular structures. Mapper was then used to sort through all that data to identify the relatively infrequent intermediate states that occur during the folding process. Courtesy of Yuan Yao.

A First Step Toward the Publication of the Future

Bourne sees the Simbios publication repository as a very positive step: "It actually speaks to the dream that I have." He envisions all aspects of research being accessible, with the paper being an access point to the experiment. From the paper, a researcher could retrieve and manipulate the associated data, and possibly discover new links and relationships via the data and tools—not just the paper citations—enhancing the research process.

Bourne observes that there are an increasing number of efforts to capture this whole research work flow process. The goal of his BioLit (http://biolit.ucsd.edu) project is to connect open access articles with information in existing biological databases, such as the Protein Data Bank (PDB). Another example is the Insight Journal (http://www.insight-journal.org), an open access on-line publication focused on medical image processing and visualization where authors are encouraged to provide the data and software associated with their papers.

"Most people get together because of content," Bourne says. Efforts such as the Simtk.org publication repository provide the infrastructure to share different types of information, enabling a dialog between the people who are using and developing the content.

"My sense is that in the next ten years, scientific discourse is going to change very dramatically as a result of these kinds of things." Bourne says. □

DETAILS

Learn more about the projects mentioned in this article.

Walking simulations: http://simtk.org/home/mspeedwalksims
Mapper: http://simtk.org/home/mapper
Climber: http://simtk.org/home/climber

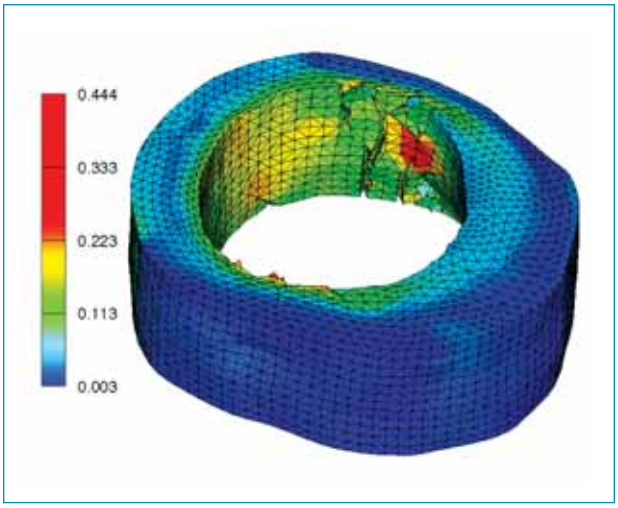
Simbios (http://simbios.stanford.edu) is a National Center for Biomedical Computing located at Stanford University.

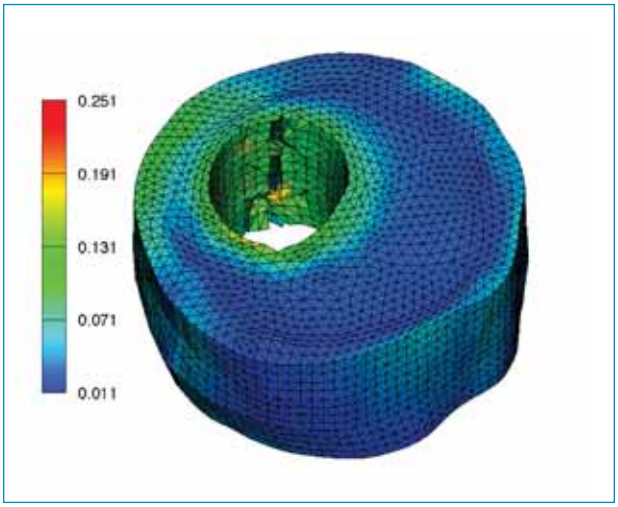
NewsBytes

Modeling Cracks in Clogged Arteries

Every year, doctors in the United States perform more than a million angioplasties: By inflating a tiny balloon inside a clogged artery, cardiologists can compress fatty plaques and restore blood flow. But the balloon also applies high pressure that can crack the wall of hardened, fat-lined arteries—sometimes with disastrous results. Now, structural engineers have created the first fully three-dimensional model to predict how arteries fracture under such stress.

"Once you have the true geometry [of the artery], this model applies pressure to simulate the presence of a balloon and evaluate the possibility of breaking the plaque or rupturing the artery walls," says author **Anna Pandolfi, PhD**, an associate professor of structural mechanics at the Politecnico di Milano in Italy. The research appears in the October 2008 issue of *Computer Methods in Biomechanics and Biomedical Engineering*.





Evolution of cracks in a clogged human artery depends on the geometry of the arterial wall and the pressure inside the artery. In the first simulation (left), a 40-percent-narrowed artery fractures at a blood pressure of 260 mmHg. In the second simulation (right), an 80-percent-narrowed artery fractures at a blood pressure of 380 mmHg. Colors show the distribution of stress on the arterial wall, measured in megapascals. Courtesy of Anna Pandolfi. Reprinted from Pandolfi A and Ferrara A, Numerical modeling of fracture in human arteries, in* Computer Methods in Biomechanics and Biomedical Engineering *(2008) 11(5):563.

In lab experiments, arteries tend to break when exposed to pressures of 0.3 megapascals or more—about 20 times the average human blood pressure. But angioplasty can easily generate such forces, and some areas of diseased arteries are particularly fragile.

To better understand how arteries fracture, Pandolfi and her colleague **Anna Ferrara, PhD**, of the Politecnico di Milano, combined high-resolution magnetic resonance imaging (MRI) of a patient's arteries with a model they previously developed to describe fracture in brittle solids, such as glass. Using a technique called finite element analysis, they divided the artery wall into small volumes and assumed each chunk had a uniform behavior. Then they simulated several high-pressure scenarios and monitored the evolution of arterial cracks.

"What we got was an interesting correspondence with the medical data," Pandolfi says: As others had seen in a clinical setting, cracks usually began at the edge, or "shoulder," of a fatty plaque.

But, Pandolfi says, the model has limitations: An MRI scan can only describe an artery's shape, not its mechanical properties, such as resistance. And these parameters vary from patient to patient, depending on the extent of arterial disease. To get individualized data, Pandolfi says, one must test a piece of artery outside the body or do an *in situ* experiment—dangerous procedures in a patient with unstable arteries.

"The key thing is to get more data and do more tests on human tissue," says **Gerhard Holzapfel, PhD**, professor of biomechanics at Graz University in Austria who published his own model of arterial fracture last year. "When we throw in more data," he says, "I am very certain we can actually define a more optimal stent, on a computer, for a specific lesion."

—By Hadley Leggett, MD

Modeling Muscles From the Inside Out

A new model of skeletal muscle starts from the micro-mechanical properties of the smallest possible unit—the sarcomere—and builds up to the muscle fibers and then to the muscles themselves. In addition, it places the fibers in their natural context—within surrounding soft tissue. The effort brings a new degree of flexibility and realism to muscle simulation.

"The idea behind micromechanical modeling is to imitate the behavior of the material as well as possible," says lead researcher **Markus Böl, PhD**, professor of mechanics of polymers and biomaterials at the Braunschweig University of Technology in Germany. "We're trying to include all the micro-parameters we can. In this way we do not have to fit the material behavior to the experimental data." His work appears in the October 2008 issue of *Computer Methods in Biomechanics and Biomedical Engineering*.

Scientists started making mathematical models of muscles in the 1920s. Most attempts to date were one-dimensional, and they ignored the soft tissue surrounding muscle fibers, Böl says. Also, they usually were built from the outside in: Scientists would look at the way a muscle behaved and tweak their

model's parameters (such as the number of contractions per second) until it matched the behavior. This led to some accurate but limited simulations.

Böl's work builds muscles from the inside out. He uses the finite element method, originally developed by aerospace engineers to design planes, to divide a muscle into discrete parts that each behave differently. Previous finite element muscle models used a continuum-based approach, which lumped all muscle fibers together and treated them as a single unit. But Böl gets into the nitty-gritty of each tiny fiber. In essence, his modeled muscles behave like a bunch of ropes of different thicknesses attached at the same point. Because the model describes each rope independently, Böl can plug in any parameters he wants and get realistic behavior back out.

In his model, Böl splits the muscle into an active element (the contractile muscle fibers) and a passive one (the incompressible tissue that surrounds them). Putting the "ropes" into the realistic environment of soft tissue yields a more complete picture, he says.

The model has both experimental and clinical value, Böl says. Scientists will use it to test the properties of living muscle, or to help doctors design unique treatments for patients, he believes. He is now working with sports doctors to refine and implement his approach. "But I have to say, these are first trials and work is still in progress," he cautions.

The new model can simulate any biological tissue that contracts, not just skeletal muscles, says **Ellen Kuhl, PhD**, professor of mechanical engineering at Stanford University. Kuhl was so impressed that she is now working with Böl to model heart tissue, with the goal of helping researchers develop a patch to replace dead tissue after a heart attack. "I think the cardiac application is even more sexy, because many more people could benefit from it," she says.

—By Lisa Grossman

"Digital Embryo" Created

How does a humble zygote grow into a fully functioning animal, billions or trillions of cells strong? This question has intrigued biologists for centuries. Now scientists have generated the first complete developmental blueprint of a vertebrate—a "digital embryo" mapping the positions, divisions, and movements of every cell during the first 24 hours of a zebrafish's life.

"Such reconstruction of a complex vertebrate embryo had not been achieved before," says **Philipp Keller**, a PhD candidate at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Keller is lead author of the paper, which appeared in the October 9, 2008 issue of *Science*.

Developmental biologists have long coveted such a tool, but imaging a complex organism's growth presents a serious hurdle. After just one day, for example, a zebrafish already has 20,000 cells and a beating heart. To meet that challenge, Keller and his colleagues developed a new technique called digital scanned laser light sheet fluorescence microscopy (DSLM).

DSLM generated a three-dimensional image of the embryo by combining about 400 pictures taken along slightly different planes. The team repeated this process every 60 to 90 seconds, tracking changes as the zebrafish developed. In 24 hours, this amounted to about 400,000 images for each embryo.

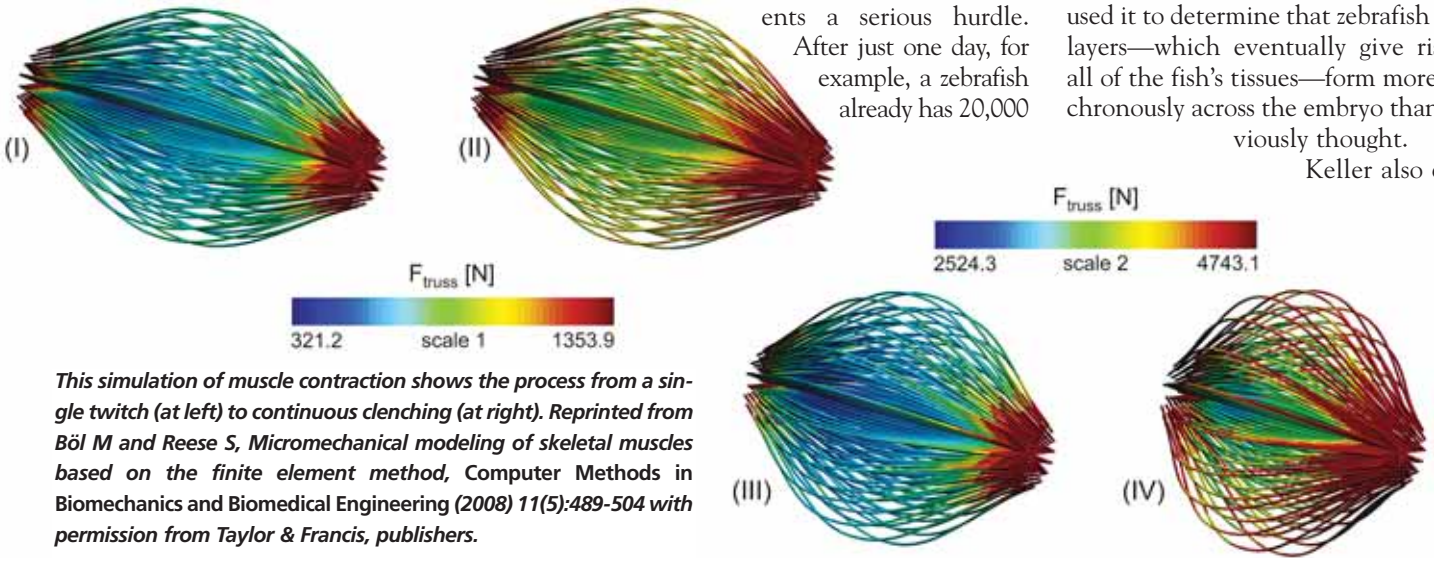
To deal with this deluge of data—three terabytes per embryo—the researchers developed a computational pipeline. They wrote algorithms defining the structure of cell nuclei, then ran the microscopy data through a network of more than 1000 computers at EMBL and the Karlsruhe Institute of Technology in Germany.

The computational analysis picked out every nucleus. Keller's team then processed this information into comprehensive databases of cell positions, divisions, and migratory tracks. In all, they catalogued 55 million nucleus entries.

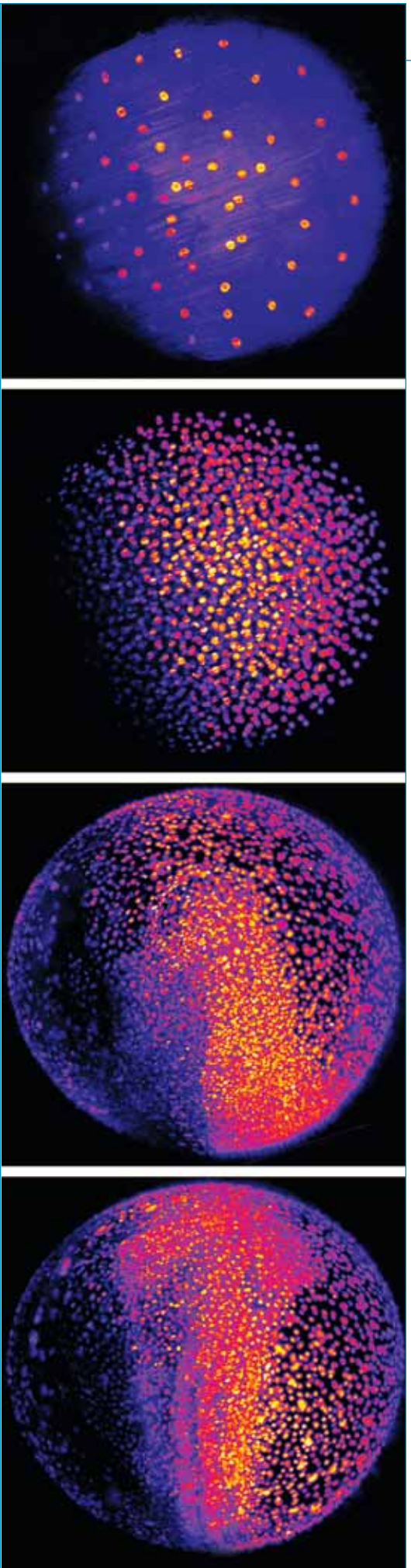
Digitizing the data was key. "Microscopy tells you about phenomena from a qualitative point of view," Keller says. "But with digital embryos, we can count the number of cells that are involved in a process and see what they do."

The digital embryo has many potential uses. For example, the researchers used it to determine that zebrafish germ layers—which eventually give rise to all of the fish's tissues—form more synchronously across the embryo than previously thought.

Keller also envi-



This simulation of muscle contraction shows the process from a single twitch (at left) to continuous clenching (at right). Reprinted from Böl M and Reese S, Micromechanical modeling of skeletal muscles based on the finite element method,* Computer Methods in Biomechanics and Biomedical Engineering *(2008) 11(5):489-504 with permission from Taylor & Francis, publishers.



sions applications in tissue engineering and the study of tumor growth. Overlaying the digital embryo with genomic data also could be powerful, he adds. Researchers could learn which genes regulate vital developmental processes, such as organ formation. To encourage such progress in multiple fields, the researchers made their data public.

"Microscopy tells you about phenomena from a qualitative point of view," Philipp Keller says. "But with digital embryos, we can count the number of cells that are involved in a process and see what they do."

DSLM images of a zebrafish embryo at four different time periods, between 1.5 and 20 hours post-fertilization. Different colors indicate different densities of nuclei (blue and purple are least dense, while yellow is most dense). Courtesy of Philipp Keller.

"This paper is groundbreaking," said **Kees Weijer, PhD**, professor of developmental physiology at the University of Dundee in Scotland. "And making all the data available is very helpful since these coordinates will be used to compare the development of mutants."
—By Michael Wall, PhD

The Circuitry of Yeast

For centuries, yeast has helped scientists understand how cells work. Now, two inventive teams have applied an engineering approach coupled with computer modeling to reveal new details about key biological pathways by which yeast cells regulate themselves in a changing environment, as reported in the January 25, 2008 issue of *Science* and the August 28, 2008 issue of *Nature*.

"What's interesting to me was looking at this biological system from an information-processing perspective," says **Jerome Mettetal, PhD**, a physicist at the Massachusetts Institute of Technology and lead author of the Science paper. "By applying temporally varying inputs, you can find out a lot about the system that you wouldn't be able to see otherwise."

Traditionally, biologists measure how cells respond by adding or taking something away in a steady-state context. But in real cells, inputs from the environment vary constantly. To understand the mechanisms by which cells respond to changes, the two teams created microfluidic arrays that confine yeast cells in a chamber and feed them in regular cycles, controlled by software. Based on the output, each team generated a model of the inner workings of the cells.

Mettetal's team added bursts of salt to the microfluidic array in order to tease out how yeast responds to changes in osmotic pressure—the salt level in the surrounding medium. They then built a model based on the response generated by the yeast. When they compared their model to known cell responses to osmotic changes, they discovered new roles for three different negative feedback loops—the processes by which a biological system reestablishes equilibrium.

The research team on the *Nature*

paper applied a similar engineering approach to better understand how yeast cells respond to fluctuations in nutrient levels. If yeast is deprived of its favorite sugar (glucose), it will consume an alternative and less nutritious sugar (galactose). The researchers created a sinusoidal input by alternately feeding and starving yeast of glucose on different time scales while galactose was constantly present in the environment. The cells responded to long-term changes in glucose, but not to faster fluctuations.

The researchers then made a model based on the well-known metabolism of galactose. But the experimental yeast was responding much faster to the glucose fluctuations than the model predicted. "This suggested something was crucially missing from the model," says co-author **Jeff Hasty, PhD**, associate professor of bioengineering at the University of California, San Diego. Studying live yeast provided the answer: The messenger RNA necessary for the galactose metabolic pathway was degraded when glucose was present. "The most exciting thing is that without the model, none of this would have happened," said Hasty.

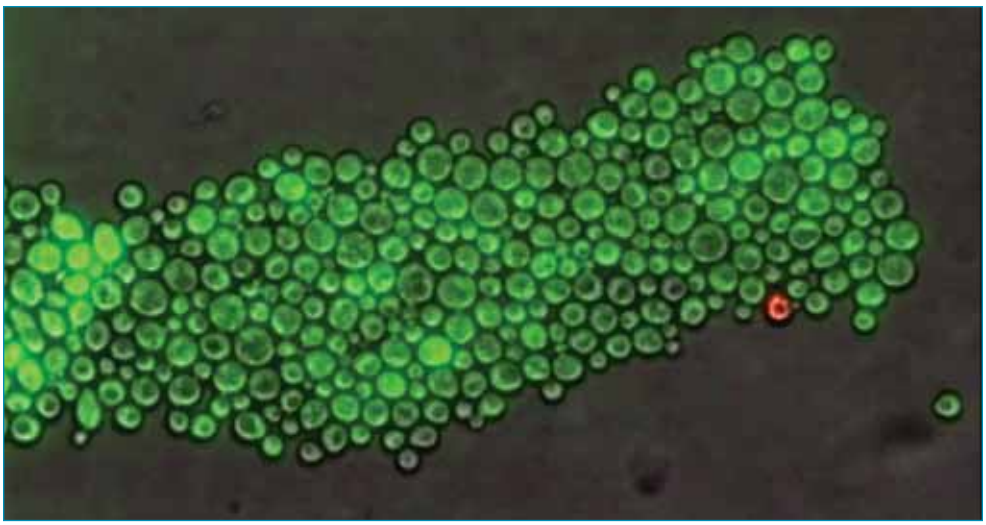
"The broader contribution of each of these pieces will be to point to the value of using periodic input signals as a means to tease out the structure and function of the underlying system," says **James Collins, PhD**, professor of biomedical engineering at Boston University. "I am already beginning to think about how these might be interesting tools to use to look at other systems, bacteria in particular."

—By Cassandra Brooks

Watching a Molecule Bind

Like a paper clip being pulled to a magnet, a small molecule called ADP gets pulled into its port in a new simulation. Because of a simple case of opposites attract, it's the first time computational biochemists have successfully simulated a molecule—or ligand—being drawn into its binding site in an unbiased simulation.

"Nobody has been able to capture and describe the full process of ligand binding to a binding site while permitting natural motion of the ligand," says **Emad Tajkhorshid, PhD**, assistant professor of biochemistry, pharmacology and biophysics at the University of Illinois at Urbana-Champaign. "We think we are getting the most faithful representation of the binding site, because in our simulations, the protein is dynamic and allowed to freely react to and establish new interactions with the ligand as it binds."



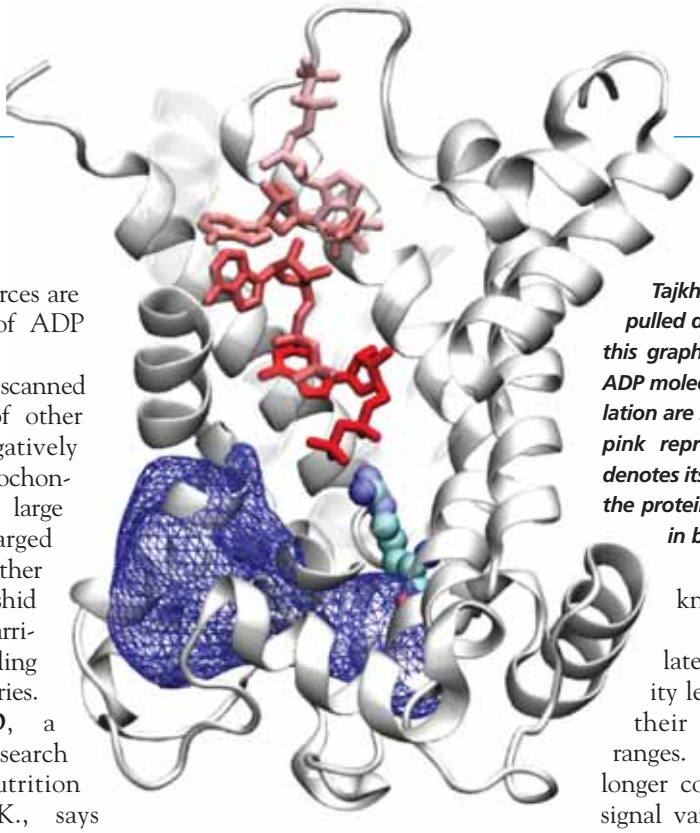
Yeast grows in a microfluidic chamber designed at the University of California, San Diego. Regular nutritional inputs, generated in a wave-like pattern, reveal aspects of how the cells regulate their metabolism and internal environments. The green background color signals that it is a galactose rich environment. Photo credit: UC San Diego Jacobs School of Engineering.

Until now, says Emad Tajkhorshid, "nobody has been able to capture and describe the full process of ligand binding to a binding site while permitting natural motion of the ligand."

Tajkhorshid and graduate student **Yi Wang** describe their simulations in the July 15, 2008 issue of the *Proceedings of the National Academy of Sciences*.

Tajkhorshid and Wang simulated the binding of adenosine disphosphate (ADP), a molecule involved in fueling the cell, to the ADP/ATP carrier protein (AAC) located in the membrane of mitochondria—the cell's power generation plants. For ADP to be shuttled into the mitochondria, it must first float into a cavity inside AAC and bind to it—an event that lasted 100 nanoseconds in the simulations.

Previously, simulations of molecular binding have required an active force to produce the attachment. But placing the ligand (in this case ADP) at the mouth of the ligand binding site (here, the AAC cavity) in molecular dynamics simulations is more faithful to biological reality. Initially, Tajkhorshid thought that the ADP would just float away. Instead it moved right into place. He and his colleagues found that AAC uses a special bait to lure ADP to its binding site: Positively charged amino acids line the sides and bottom of the AAC cavity, creating a surprisingly strong electrostatic potential that attracts the negatively charged ADP. They called this process "electrostatic funneling." And



Tajkhorshid and Wang watched as ADP was pulled down into the cavity of the AAC protein. In this graphic, the AAC structure is outlined in black. ADP molecules at different stages of the 100 ns simulation are shown in colors ranging from pink to red—pink represents ADP's starting position and red denotes its final binding state. The strongest region of the protein's positive electrostatic potential is shown in blue mesh. Courtesy of Emad Tajkhorshid.

because of it, no additional forces are needed in the simulations of ADP binding to AAC.

In addition, when the team scanned the amino-acid sequences of other molecules that shuttle negatively charged molecules across mitochondrial membranes, they found large numbers of positively charged amino acids not present in other membrane proteins, Tajkhorshid says. He suspects these other carriers also use electrostatic funneling to pull in their molecular quarries.

Alan Robinson, PhD, a researcher at the Medical Research Council Dunn Human Nutrition Unit in Cambridge, U.K., says Tajkhorshid has "published what looks like the most reasonable structure of ADP bound to the carrier." This structure may serve as the starting point for more detailed studies of how ADP binds to AAC and how it triggers the protein to open, he says.

— By Michael M. Torrice, PhD

Identifying a Cell's Weakest Link

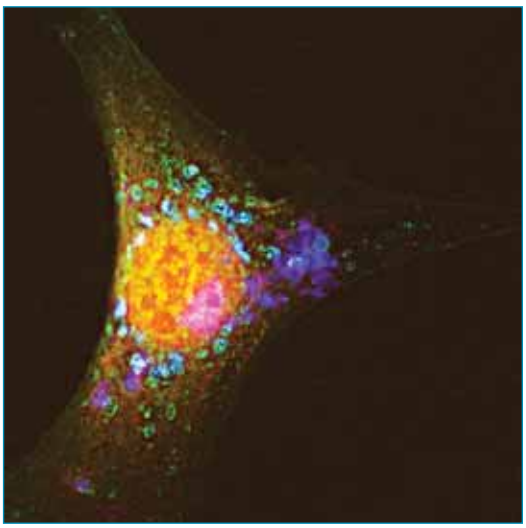
To understand why bridges collapse or computers fail, engineers might create models of these systems and push them beyond their limits. Now, computational biologists are using a similar approach to understand the causes of cell death. By driving their model of the cell beyond experimentally observed values of certain important cellular ingredients, they push it to the "breaking point"—uncovering the weakest links. The process revealed some new biological roles for several key signaling molecules—the kinases ERK, Akt, and MK2.

"It showed us things that, in retrospect, we couldn't see looking by inspection of the original model," says co-author **Michael Yaffe, PhD**, associate professor of biology and biological engineering at the Massachusetts Institute of Technology (MIT). The work was published in the October 17, 2008 issue of *Cell*.

The mechanisms by which proteins influence cytokine-induced apoptosis, or cell death, are poorly understood. So Yaffe and colleagues **Kevin Janes, PhD**, a recent MIT graduate, and **H. Christian Reinhardt, PhD**, a postdoctoral associate at MIT, built a model of the cell using carefully collected data. Included in the model were nearly 8,000 measurements of protein signals in response to combinations of three cytokines that help dictate the fates of cells: tumor necrosis factor (TNF), known as the "death stimulus," and epidermal growth factor (EGF) and insulin, known as "survival stimuli."

The researchers then manipulated the model to drive the activity levels of the proteins outside of their experimentally observed ranges. When the model could no longer computationally fit one of the signal variables, it would stop making predictions. This "breaking point" highlighted the protein that caused the failure. Thus, the technique acts as a sort of high-throughput screen, revealing new hypotheses about proteins previously thought to have well-defined roles within the cell. The team then verified these hypotheses experimentally, leading to surprising new insights about how the signaling proteins communi-

"Signaling networks are so complicated right now that common sense doesn't always hold true," Michael Yaffe says.



Breakpoint model analysis pushes cellular ingredients beyond their normal ranges to see which ones are critical to a particular cellular process. Here we see fluorescent proteins highlighting the subcellular location of several different key signaling molecules (phosphoinositide-binding domains), which function together with lipid and protein kinases and phosphoserine/threonine-binding domains, to control a wide variety of cellular events. These are the kinds of molecular interactions that could be studied using breakpoint model analysis. Courtesy of Seth J. Field and Michael Yaffe.

cate. "Signaling networks are so complicated right now that common sense doesn't always hold true," Yaffe says.

"The thing that makes me really stop and pay attention is the methodology, which I found of special note," says **Raphael Levine**, **PhD**, distinguished professor of chemistry at the University of California, Los Angeles. "Instead of trying to see if the model can predict something new, they tried to drive it to say something which they know it shouldn't say. As a result, they were successful in finding some new biology."
—By Kayvon Sharghi

Diagnosing Cell Circuitry

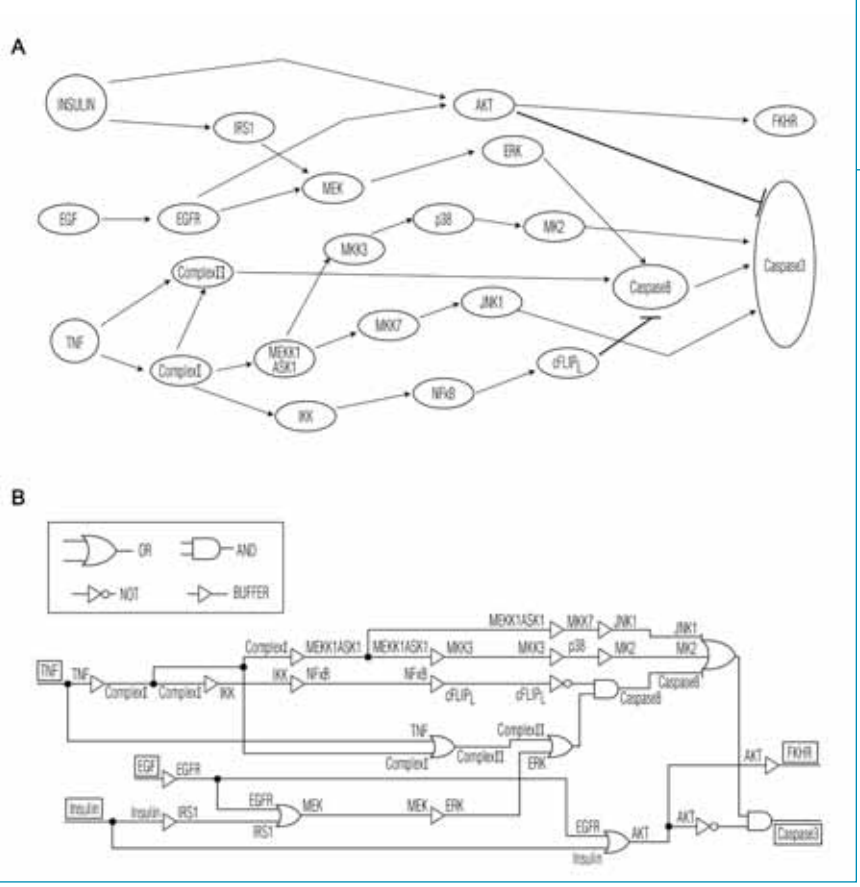
To biologists, a computer's motherboard may just look like highways of circuitry connecting various chips. But if they focus harder, they might see a model for disease, according to new research.

Just as a single corrupt circuit can foul a computer's operation, a faulty molecule can upset a healthy body. "If your body is not functioning correctly, then the molecules inside your cells are causing the problem," says **Effat Emamian**, **MD**, president and CEO of Advanced Technologies for Novel Therapeutics in New Jersey.

The parallels between signal transduction pathways in a cell and circuit networking in a motherboard inspired Emamian's team to identify defective cell pathways in the same way that engineers inspect faulty circuits. This technique, known as fault diagnosis, can pinpoint the molecules that are most critical to a cell's function.

Such an accurate assessment may lead to more precise medicines. Most new drugs in trial are toxic, Emamian says, because they often target molecules essential for cell function. Fault diagnosis can reveal safer molecules to target. The work appears in the October 21, 2008 issue of *Science Signaling*.

Lead author **Ali Abdi**, **PhD**, associate professor of electrical and computer engineering at the New Jersey Institute of Technology, helped test Emamian's theory. Abdi re-envisioned three previously studied cell pathways as electronic circuits: tumor suppressor p53, cell death regulator caspase3, and a nerve-cell network called CREB. His reconstructions used binary language to characterize a molecule's state in its pathway as "active" or "inactive." Relationships between molecules were organized into decision-making operations using Boolean logic where each relationship contains only two possible values—on or off. This allowed the researchers to write algorithms predicting which molecules were critical to a pathway's smooth functioning. The algorithms confirmed what was known about p53 and caspase3, but they also revealed new critical molecules in the CREB network.



A simple model of the caspase3 network (top) shows the various regulatory molecules and their relationships to each other. Depending on which regulatory molecules are active or inactive, caspase3 will induce cell death. This network can be re-envisioned (below) as an electronic circuit after organizing previous knowledge of the molecules' relationships using Boolean logic. Algorithms applied to this circuit can predict molecules to which a pathway's signal is most vulnerable. Reprinted with permission from Abdi A, et al., Fault Diagnosis Engineering of Digital Circuits Can Identify Vulnerable Molecules in Complex Cellular Pathways,* Science Signaling*, (2008) 1(42):ra10.

The approach is a good start for quickly identifying essential points in cell networks, says **Kevin Janes**, **PhD**, assistant professor of biomedical engineering at the University of Virginia. But while Boolean logic can make good approximations, it may oversimplify the relationships for some networks, he says. For example, Emamian's approach doesn't allow consideration for graded responses between "active" and "inactive." "But it's not a fundamental flaw," Janes adds.

The team acknowledges these limitations in its *Science Signaling* paper. The next step, Emamian says, is to focus on larger networks, and not necessarily just signaling pathways. "We can analyze metabolic pathways, or pathways that also have several critical enzymes playing in the whole game."
—By Emmanuel Romero

Cancer's Signature—Written in Blood

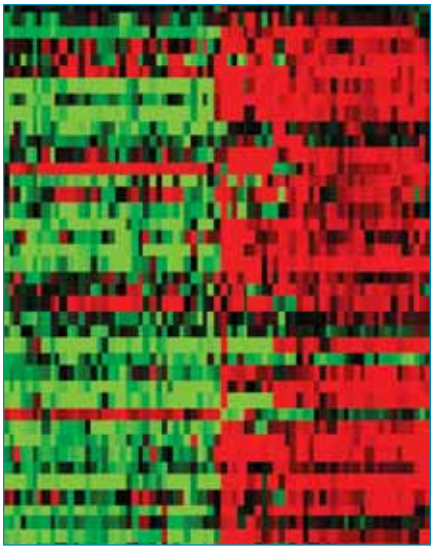
When it comes to deciphering the health of the body, the blood carries a potential mother lode of protein clues. Given the ease of extracting blood, such proteins could serve as efficient health barometers. But it's tough to distinguish between the multitude of proteins naturally found in blood and those that are secreted into the blood—including those secreted by diseased tissue such as cancer. Their signal may get swamped by the many other proteins present in blood, thwarting efforts to discover useful infor-

"Figuring out which proteins are secreted into the blood is like searching for a needle in a big, big haystack," says Ying Xu, PhD. "This [algorithm] sorts through all that hay."

mation. Now, scientists have developed an algorithm that sorts through the multitude, expediting the search for blood-based cancer biomarkers.

"Figuring out which proteins are secreted into the blood is like searching for a needle in a big, big haystack," says **Ying Xu**, **PhD**, professor of bioinformatics and computational biology at the University of Georgia. "This [algorithm] sorts through all that hay."

To develop their algorithm, Xu and his colleagues began by scouring the literature for all proteins known to be secreted into the blood, regardless of their origins. They then analyzed the amino-acid sequences of these proteins to identify common features, such as signal peptides, transmembrane domains, solubility, and secondary structure. They discovered 18 features that were powerful predictors of blood secretion, and used them to train a computerized classifier.



This microarray shows genes that differ in regulation between cancerous and non-cancerous lung tissue. Ying Xu's classifier can predict which of the proteins made by these genes may be useful as blood-based biomarkers. Courtesy of Ying Xu.

When the researchers applied the classifier to other data sets, it could distinguish proteins secreted into the blood from all other proteins in the blood with more than 80 percent accuracy. The results appear in the October 2008 issue of *Bioinformatics*.

Xu and his colleagues are now using microarrays to identify differences in gene expression levels between cancerous and non-cancerous stomach tissue. Using their classifier, they can then sift through the data to zero in on genes that produce proteins that are most likely to be secreted into the blood, followed by validation with mass spectrometry.

"We've already identified proteins that are elevated during different stages of stomach cancer," Xu says. "Typically, in order to find out what stage it's in, you'd have to actually cut the patients open and do a biopsy. Our markers could be the first markers to provide information about cancer stage."

By applying his biomarker discovery pipeline to a range of cancers, Xu ultimately hopes to identify general biomarkers that apply to any cancer. He envisions doctors detecting various cancers at early stages with a simple blood test.

Bo Huang, **PhD**, a post-doctoral fellow at Vanderbilt University, hopes to use Xu's classifier to find biomarkers for breast cancer. "These results provide a powerful method to discover potential biomarkers, not only for cancers but also for many other diseases," Huang says.

—By Lizzie Buchen

Blurring Data for Privacy and Usefulness

Hospitals with research agendas share a common problem: how to use medical records for research while protecting patient privacy. One approach—the data-protection equivalent of blurring the face of an anonymous source on television—has now been tested using real-world data. The results, which show promise for protecting privacy without rendering the data set useless, appear in the September/October 2008 issue of the *Journal of the American Medical Informatics Association*.

"It's not a theoretical problem," says **Khaled El Emam**, **PhD**, associate professor at the University of Ottawa and Canada Research Chair in electronic health information, who collaborated with **Fida Kamal Dankar**, **PhD**, on the paper. "We're trying to protect privacy, but we need the tools."

Just as the nightly news renders the faces of anonymous sources unrecognizable, the approach known as k-anonymity blurs distinctive variables to reduce the risk that someone could trace patients with distinctive characteristics. For example, the approach might cut birthdates down to birth years. And easily identifiable outliers—the octogenarian in a college town, the teenager in a retirement community—are omitted. The remaining information contains at least k data points that look identical, where $1/k$ is deemed an acceptable level of risk.

That works in theory, but the actual risk depends on the type of data set and what an intruder wants from it. A prosecutor digging up dirt on a defendant would try to re-identify a specific person in the database. A journalist trying to discredit an organization's data-security procedures would also only need to re-identify one person, but it wouldn't matter who. El Emam set out to test whether k-anonymity works in both circumstances. His findings: k-anonymity correctly predicts the risk of re-identifying one specific individual with minimal harm to the value of the database (the prosecutor example). But using k-anonymity to protect against re-identifying an arbitrary person (the journalism example) is unnecessarily strict and compromises the research quality of the data.

Since researchers choose *k* based on statistical theory, El Emam suggests data custodians run test cases to verify if the *k* is sufficient, or if it's overprotective, as in the journalism example, before making the data available to researchers. If needed, the number of groupings of *k* identical data points could then be adjusted to ensure that the actual risk approximates the theoretical risk of 1/*k* and, in this way, keep the risk acceptably low while preserving data.

"What is needed are the steps to turn this article into a practical tool that custodians can use in conjunction with researchers," says **Joan Roch**, chief privacy officer for Canada Health Infoway in Montreal, Quebec.

El Emam says he plans to continue exploring actual risks in various data-security scenarios: "It's a big problem, and we've solved part of it."

—By Stephanie Pappas

Modular Modeling

Biological models can quickly become as complex as the systems they represent. And minor changes can necessitate a complete rewrite of the model. But researchers may soon snap their models together like LEGOs, using a new programming language called Little b, which uses modularity to simplify biological modeling. Eventually, the authors hope to turn Little b into an easy-to-use tool for biology labs.

"I think that as an everyday tool, it [Little b] is going to be kind of like the microscope," says **Aneil Mallavarapu, PhD**, lead developer of Little b and a senior research scientist in systems biology at Harvard Medical School. "We're essentially building a new kind of gel, a new type of microscope for the lab." The work appears in the June 2008 issue of the *Journal of the Royal Society Interface*.

Biologists traditionally create models to describe unique systems, such as the development of fruit fly embryos or the actions of a phosphorylation cascade on gene transcription. Such computational models are usually based on lists of the system's properties, which detail every molecular interaction in the system. This allows researchers to tailor models to the precise questions being asked, but it also constrains the model's usefulness, because it can only probe into one area.

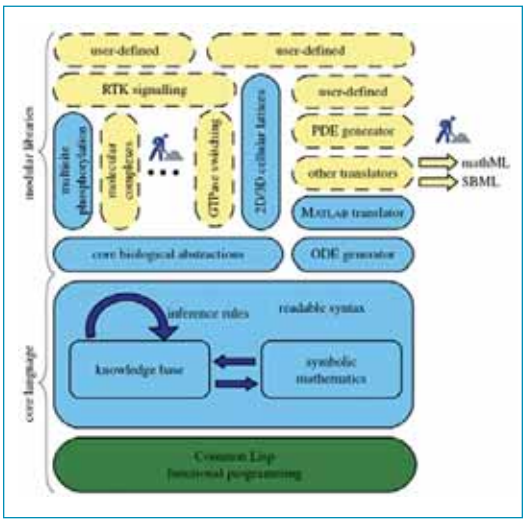
Little b strives to break down biological systems into modules that can be used regardless of the specific context, such as "nuclear export" or "membrane localization." It then defines those parts in a mathematical language. Researchers can use Little b to put together assorted modules to describe their system; Little b then uses those symbolic modules to write out executable code that a scientist could use in a simulation program like MATLAB. "I've given Little b the power to reason about biological objects," Mallavarapu says.

Mallvarapu is excited about the possible use biologists might make of Little b. He would like to see the language help uncover the complex pathways involved in diseases. He hopes that researchers will eventually build entire virtual cells or virtual plants collaboratively, increasing their ability to study their projects *in silico*.

> "I've given Little b the power to reason about biological objects," Aneil Mallavarapu says.

While the idea of breaking down biological systems into modular chunks may seem logical, Little b may not arrive in the lab immediately, says **Birgit Schoeberl, PhD**, a senior director of research at Merrimack Pharmaceuticals, Inc, in Cambridge, Massachusetts. "I'm excited about the concept and what I see, but in my own experience, it isn't straightforward," Schoeberl says. "I think it's not quite ready for non-developers. I hope he keeps developing it, or someone takes it on to keep working on the idea."

—By Molly Davis □



Little b is based on a core language, which includes the Lisp language it was created in (green) and the knowledge base, symbolic mathematics and syntax modules that allow Little b to reason about biological systems. It also includes modular libraries that describe specific biological interactions, and translators that can generate code used in simulations. Blue areas exist within the current framework; yellow areas are currently under development or are envisioned for future work. Reprinted with permission from Mallavarapu, A, et al., Programming with models: modularity and abstraction provide powerful capabilities for systems biology,* Journal of the Royal Society Interface, *online publication, July 23, 2008.

After four years, the seven National Centers for Biomedical Computing (NCBCs)—established largely to build a national biocomputing infrastructure—have, as one might expect, produced an impressive array of computer tools. >

NCBC Update: Shedding New Light On BIOLOGICAL COMPLEXITY

By Katharine Miller

But it's the Centers' wide-ranging impact on biomedicine that takes center stage. From AIDS to diabetes, prostate cancer or schizophrenia, the NCBCs are changing the landscape of disease research by shedding new light on biological complexity.

"The impact on biology and medicine happened faster than anyone expected," says **Russ Altman, MD, PhD**, co-principal investigator for **Simbios, the National Center for Physics-based Simulation of Biological Structures**, an NCBC grantee at Stanford University.

And that impact springs from the way the NCBCs function, says **Andrea Califano, PhD**, who heads the **National Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet)** at Columbia University. "Developing new tools in the context of solving specific scientific, biological or medical problems is what I think has allowed the NCBCs to successfully penetrate the broader community with tools, techniques and methodology," he says. "We've shown what can be accomplished by applying these tools to biological problems."

And while the specific breakthroughs enabled by NCBC tools varies with the tool being used or the disease being studied, it is clear that they are all helping researchers approach the complex system that is the human body. "Dealing with complexity is the essential challenge of this century in biology," says **Scott Delp, PhD**, co-PI for Simbios. "And you can't do it without computers."

"Developing new tools in the context of solving specific scientific, bicalifano, biological or medical problems is what I think has allowed the NCBCs to successfully penetrate the broader community with tools, techniques and methodology." says Andrea Califano.

NCBC 10 YEAR GOALS

Califano: **"One critical thing we hope to accomplish is to create a new breed of biologist trained both in computational and experimental sciences. You already see evidence of this in some labs. Now, as never before, some of the projects enabled by the NCBCs have computation and experimental biology playing hand in hand rather than in a pipeline fashion. That is also reflected in the tools that we generate. Unlike other platforms, geWorkbench was created for an experimental biologist who wants to learn enough computational biology to be able to analyze data. It's easy and intuitive to use, and the researcher doesn't have to learn complex scripting languages. The emphasis has been on enabling experimental labs to use more and more computational tools. Across the entire set of activities at MAGNet, the real aim is to fuse the two disciplines and to create a really interdigitated boundary between the computational and experimental life sciences."**

Andrea Califano, PhD, is the principal investigator for the National Center for Multiscale Analysis of Genomic and Cellular Genomics (MAGNet) and professor of biomedical informatics at Columbia University.

MAGNet

Here, following a few years of hard work, the NCBC PIs reflect on what they've accomplished so far, how they've gained traction in the research community, and what their goals are going forward.

NCBC Tools: Enabling Discovery Across The Disease Spectrum

From the start, each NCBC's tool and infrastructure development goals were driven by a cluster of specific biological problems—commonly referred to in NCBC parlance as the "driving biological problems" or DBPs. After a few years, these DBPs were replaced by a new set of DBPs, ensuring that the tools would be suitable for multiple purposes. That strategy has worked.

"To a certain degree, the tools and biology are push-pull kinds of associations," says **Art Toga, PhD**, principal investigator for the **Center for Computational Biology (CCB)** based at the University of California, Los Angeles. "The tools get developed because you couldn't do something without them. And vice versa, you get this tool and you decide to pose new questions. You end up pushing and pulling so that both are advanced."

Thus, NCBC tools that were developed to address one biomedical problem have proven to be broadly useful. For example, at **i2b2—Informatics for Integrating Biology and the Bedside**—an NCBC based at Harvard, tools developed to allow the use of medical record systems for clinical research initially focused on diseases such as asthma, obesity and depression. Now, however, these tools have been adopted at 18 large academic health centers with no apparent limit on the number of diseases that can be studied, says i2b2 principal investigator **Zak Kohane, MD, PhD**.

And imaging tools originally developed by CCB and the **National Alliance for Medical Image Computing (NA-MIC)** to study schizophrenia in the brain are now proving useful in studying many other brain diseases, as well in prostate cancer (at NA-MIC) and cardiovascular disease (in CCB's case). "You can begin to see how the shape-modeling approach [we've developed] is applicable to a whole range of biological problems," says Toga of CCB.

Similarly, OpenSim, a software program developed by Simbios to study human movement and movement disorders, was first used to conduct research into one of the Simbios DBPs, cerebral palsy, but is now being used more broadly. Indeed, it has been adopted by more than one thousand individuals working on any number of problems including osteoarthritis, Parkinson's disease and stroke.

This is the vision of the NCBCs—to provide the underlying computational tools that will advance the field of medicine and biology, across a spectrum of diseases. As **Mark Musen, PhD**, of the **National Center for Biomedical Ontologies (NCBO)** at Stanford, says, "We're enablers. We are providing the foundation by which

"Dealing with complexity is the essential challenge of this century in biology," says Scott Delp. "And you can't do it without computers."

Kikinis: **"We are developing algorithms and a platform—the NAMIC kit—for analysis of diagnostic images. I think that platform will be one of our major accomplishments. It is free and open source with a very liberal license, and it will continue to be developed. That will continue for a long time. So our goal is to develop enabling technologies and make them accessible. That will be one of the legacies of the center."**

NCBC 10 YEAR GOALS

NA MIC

Ron Kikinis, PhD, is the principal investigator for the National Center for Medical Image Computing (NA-MIC) as well as director of the Surgical Planning Laboratory of the Department of Radiology, Brigham and Women's Hospital and Harvard Medical School, and professor of radiology at Harvard Medical School.

investigators can do research that will impact human health. Our goal is to create the kinds of tools that would be valuable to everybody."

Ron Kikinis, PhD, head of NA-MIC, concurs. "We will not solve cancer but we will provide the people who are fighting cancer with better tools to fight their fight," he says. "And the DBPs will use these tools and promote those tools into their communities—so that makes it possible for lots of different diseases to be addressed."

Brian Athey, PhD, co-PI for the National Center for Integrative Biomedical Informatics, centered at the University of Michigan, agrees. While his center's tools have contributed to a better understanding of type 2 diabetes and prostate cancer progression, the tools' reach extends much farther: "We're opening doors to new research," he says.

"The tools get developed because you couldn't do something without them. And vice versa, you get this tool and you decide to pose new questions. You end up pushing and pulling so that both are advanced," Art Toga says.

NCBC Challenge: Putting It All Together

For the last thirty years, biology has been about breaking things down into their fundamental parts to understand them. "But things don't work as independent parts," says Delp. "Theoretical and computational biology let you put things back together to understand the whole system."

Several of the NCBC PIs cite the reassembling of biological pieces as a major focus of their efforts. For example, literally thousands of experiments have looked at how elements of the neuromuscular system (muscles, joints, connective tissue) operate independently. But, Delp says, looking at those elements separately doesn't tell you how people move. OpenSim lets researchers put the pieces together. "When you can code the details accurately in a computer framework, then you can understand how the system works," Delp says.

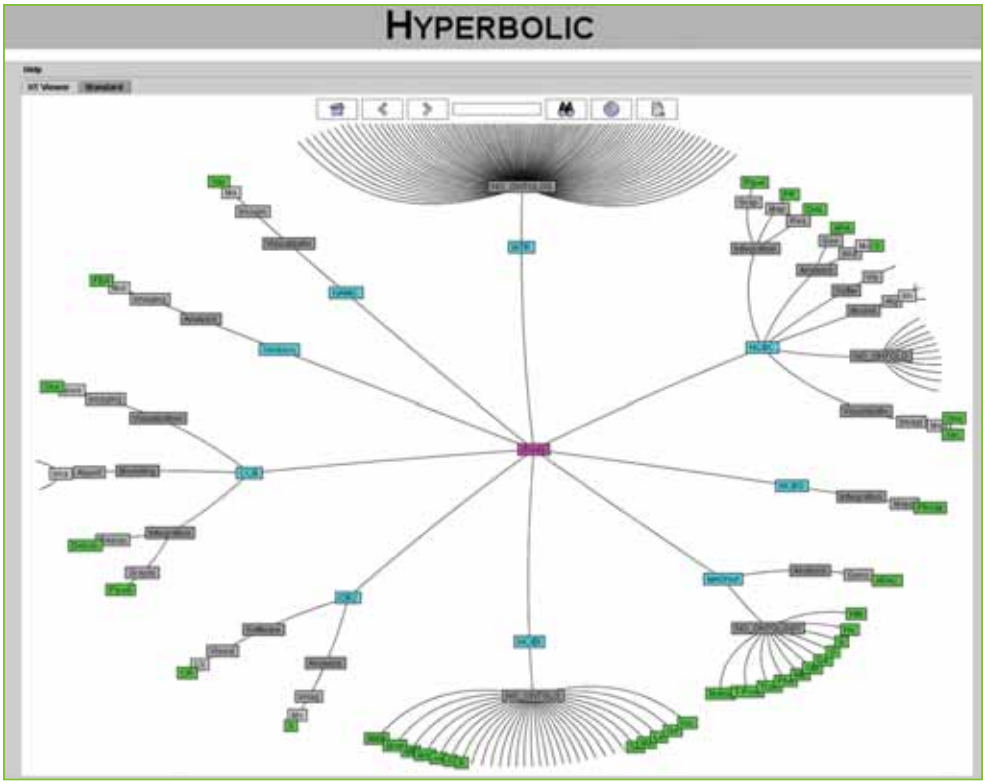
Likewise for the brain, says CCB's Toga. Brain researchers have typically focused on only one variable at a time—for example, electrical activity, blood flow, distribution of receptors, gene expression patterns, or cortex morphology. But, Toga says. "All of these brain changes are happening in concert." To understand the brain requires re-integration of these events. CCB, Toga says, is providing the tools, mechanisms, and strategies to put things

Toga: **"Our hope is to continue to integrate what we know about the brain in a way that allows us to ask questions such as: 'How does the brain change throughout a person's life?' These sorts of emerging questions are provocative. And we can only ask them because of computation. So by the end of our ten years, we really hope that our center produces new research programs that can continue to evolve in accord with the basic thrust of the NCBCs. Because you know, it doesn't finish. They haven't finished mapping the earth yet and there's only one of those! How can anyone possibly suggest we will ever finish mapping the human brain when there are billions of them? So we'll continue to layer on what we already know without throwing away our previous efforts."**

NCBC 10 YEAR GOALS

Center for Computational Biology (CCB)

Arthur Toga, PhD, is the principal investigator for the Center for Computational Biology (CCB), and a professor of Neurology and Director of the Laboratory of Neuro Imaging at the University of California, Los Angeles.



Working together NCBC researchers created iTools—a way to manage the description of computational biology data, tools, and services. Using the iTools hyperbolic viewer a researcher can displays all of the activities of the NCBCs organized by Center (as shown here) or by activity. iTools also lays the groundwork for interoperability among diverse biomedical computing tools. Reprinted from Dinov, ID, et al., 2008 iTools: A Framework for Classification, Categorization and Integration of Computational Biology Resources. **PLoS ONE *(2008) 3:(5):e2265.***

back together. "Observations from one project in 2007 can be combined with other observations in another laboratory using different subjects and techniques in 2008," Toga says. "That transition in science is revolutionary, and the computational strategies that enable it are only now beginning to emerge."

MAGNet hopes to provide a similar service at the genetic and cellular level. Very few diseases are caused by a single gene, Califano says. Usually a complex interplay of genetic and epigenetic factors is involved. "But what has been lacking is a framework for integrating genetic,

"We're enablers." Mark Musen says. "We are providing the foundation by which investigators can do research that will impact human health. Our goal is to create the kinds of tools that would be valuable to everybody."

Musen: **"We are thinking about what it would mean to be able to move biomedical knowledge from prose to machine-processable format. The long-term vision is to create the infrastructure and tools so that biomedical literature could be intelligible to both people and machines. Ultimately this could allow intelligent computer-based agents to read the literature, to make associations between scientific contributions, and to synthesize ideas from the literature. That would obviously change the way we do science in a very profound way. But there are lots of baby steps until we can do that."**

NCBC 10 YEAR GOALS

Mark Musen, MD, PhD, is the principal investigator for the National Center for Biomedical Ontologies (NCBO) and professor of medicine at Stanford University School of Medicine.

NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

epigenetic, functional and structural data—and getting an answer that can really dissect disease," he says. MAGNet's goal is to establish such a framework and to show that the framework can integrate data in meaningful ways for several diseases. "We already have proof of concept for glioblastoma multiforme—a cancer that produces the worst possible prognosis in patients," Califano says. The results for that work will be published in the next few months. "This kind of proof of concept in a disease is of course important, but at the same time the methodology becomes universal."

NCBCs: More than the sum of their parts

The NCBCs are also working together in various ways to ensure that they have a broad impact. In some ways this is a surprise, say the NCBC PIs, because the NIH cast such a wide net—with centers that cover ontologies, simulations, clinical systems, systems biology and imaging. "Given the breadth of the needs and the solutions to biomedical computing problems," says Kohane, "it wouldn't have been surprising if there had been no overlap and the synergies had been fewer."

"What bioinformatics was five years ago is frankly just a glimmer of what it is today," Brian Athey says. "It's exploding into something much more robust. And that's going to continue for a while."

NCIBI is also integrating many different high-throughput data types to better understand complexity. "We do not yet understood the full complexity of the architecture of the human genome," Athey says. "Only 2 percent of the genome are 'genes' and we're learning more and more that the other 98 percent are doing things." To tackle that problem, he says, computational biology is making huge strides. "What bioinformatics was five years ago is frankly just a glimmer of what it is today," Athey says. "It's exploding into something much more robust. And that's going to continue for a while."

Yet the NCBCs have found overlap and have helped each other. For example, the i2b2 center collaborated with NCIBI around Type 2 diabetes, Kohane says. And NA-MIC nicely complemented i2b2's major depression DBP by correlating patient imaging with what was being seen genetically. Similarly, ontologies from NCBO have been helpful to CCB in constructing their brain atlas; and CCB and Simbios have used some of NA-MIC's visualization tools.

Even though the NCBCs might be developing different tools, Califano says, "when you

Athey: "There's much more work to do to figure out how to use systems biology more effectively to understand disease and its complications. The daunting complexity of biological systems is becoming more and more clear. To gain an understanding of that complexity, we need an integrative approach that's iterative and that allows the integration of many different kinds of data types around hypotheses and models. The abundance of high throughput data we're presented with from next generation sequencing, and what that's revealing about the transcriptome and alternative splicing, and all the components we haven't yet annotated—it's just astounding. It's literally changing our basic understanding of cells and their complexity and function. And, frankly, it's changing what our understanding of a gene is. So there's a lot of work to do. I think that's the theme. And each success brings on new challenges."

Brian Athey, PhD, is the principal investigator for the National Center for Integrative Biomedical Informatics (NCIBI), associate professor of biomedical informatics at the University of Michigan, and director of the Michigan Center for Biological Information.

NCBC 10 YEAR GOALS

NCIBI
National Center for Integrative Biomedical Informatics

tackle a biological problem you must tackle it from several angles." So for example, MAGNet and NCIBI have several DBPs that focus on analyzing genomic data as a way of studying neurodegenerative diseases, diabetes or cancer. But these same diseases also need to be studied using data from large cohorts, which ties in to what i2b2 does at Harvard to use medical records to study large populations. It also ties in to the ontology work of Mark Musen, Califano says, because ontologies provide an essential foundation for other work. And, he says, when you look at the actual problem you're trying to understand, all sorts of issues related to physical modeling also come up. Indeed, according to Altman, eventually cellular physics will become an essential piece of systems biology.

"The reality of why all the centers come together is precisely around the biology, Califano says. "We develop all the different techniques and infrastructure to tackle biology problems, but when you actually want to tackle one of these problems, you require all of these approaches."

And those multiple tools also need to be kept organized. So one key activity that has united all the centers, says Musen, is the creation of an online tool that allows biomedical software resources to be easily identified and searched online. Called Biositemaps, the tool, seeded with information about the NCBC tools, can inform search engines about software available from any organization that creates a simple Biositemap file as described on the site (http://www.biositemaps.org). NCBO is providing the ontology behind the tool but, Musen says, "It's a product of all the NCBCs that would not have been possible without the cooperative involvement of all the different centers."

"If we actually successfully did a big population study and discovered something important or successfully calculated how to design a vaccine or predicted a new drug for a specific disease, then we'd be bringing ourselves to the next level," says Kohane. "We'd be solving a biomedical problem of true health relevance. In fairness, I think we're all trying to get there, but we're not there yet."

NCBC 10 YEAR GOALS

Kohane: "Within the ten-year time frame, the goal would be to establish a kind of scientific ecosystem around the country where we can use entire healthcare systems as a unit of study. We'll be able to look at reproducibility across multiple academic health centers to see if we're seeing, for example, the same adverse drug events (so that we can push early warnings to prevent such events); or compare efficacious therapies; or compare whether we have reproducible findings in genomics or proteomics across populations. This approach will allow us to do research in a more cost-effective way. And although it sounds venal to talk about cost, cost is a key rate-limiting factor in large population studies. So if we can do clinical research, including genomic measurements in populations of 10,000 to 100,000, that's really a game-changer."

Isaac Kohane, MD, PhD, is the principal investigator for Informatics for Integrating Biology and the Bedside (i2b2), as well as Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School, and Chair of the Informatics Program at Children's Hospital, Boston.

i2b2

NCBCs: Bench to bedside

Whether casting a wide net to enable research in lots of areas is enough to render the NCBCs successful remains to be seen. Curing a disease would be better. "If we actually successfully did a big population study and discovered something important or successfully calculated how to design a vaccine or predicted a new drug for a specific disease, then we'd be bringing ourselves to the next level," says Kohane. "We'd be solving a biomedical problem of true health relevance. In fairness, I think we're all trying to get there, but we're not there yet."

"The challenge is," says Delp, "that it takes time to build the tool, teach people how to use it, get it adopted, make a discovery and then translate that into clinical care." Currently, says Delp, "OpenSim is only halfway down that pipeline and is just beginning to see the first examples where new discoveries will enhance human health."

"Adoption by companies is one indication that what we're doing will eventually make a difference to clinical practice," Ron Kikinis says. "We are not yet at that point, but I have these early indicators."

Kikinis says NA-MIC's tool kit is similarly poised for bedside use. He's beginning to see the first signs—such as questions at seminars, and email inquiries—that companies are interested in it. "Adoption by companies is one indication that what we're doing will eventually make a difference to clinical practice," he says. "We are not yet at that point, but I have these early indicators."

Migrating computational biology from the bench to the bedside remains a challenging goal for all the centers. But, as Toga sees it, "I think these computational strategies, which are the hallmark of this program, are having a great effect on accelerating that." CCB is modeling the effect that HIV and Alzheimers have on the brain. These are diseases that will strike people we all know, Toga notes. "So our work immediately transforms a mathematical problem [shape modeling] into something with obvious and immediate clinical value," he says. "And the time frame for doing that is getting shorter and shorter and shorter."

Kikinis summed it up succinctly: "What are the NCBCs doing for biology? Everything. That's by design, but now you can say that they're actually delivering, and there's a sense of excitement. It's clear that things are moving." □

Delp: "The goal is twofold, really. One, that we'll produce a set of tools that are ubiquitous in biomedical research so that every investigator who is interested in how physics affects biological function will have SimTK-based tools as part of their laboratory. The second objective is that we and others will use those tools to make new discoveries that enhance human health."

NCBC 10 YEAR GOALS

simbios

Scott Delp is co-principal investigator for the National Center for Physics Based Simulation of Biological Structures (Simbios) and a professor of bioengineering and mechanical engineering at Stanford University.

By Kristin Sainani, PhD

Tool Dissemination

DOING IT RIGHT

Biomedical computing at academic research centers has been compared to a cottage industry. Lots of individuals work away on their focused research projects, generating useful algorithms. But quite often, the knowledge gained is lost when researchers move on to new projects. Yes, they might post their code on Web sites. But is it useful to anyone else without support and documentation? And how can people find it in the first place?

To overcome the cottage industry mentality, the National Institutes of Health (NIH) is placing a greater emphasis on dissemination as a piece of the National Centers for Biomedical Computing (NCBCs) as well as for other grantees.

But what does it really take to turn an impressive algorithm into a widely disseminated, prolific computational tool? The transition might be harder than you think.

"Today, our software is very widely used, but it didn't take off right away. It took years," says **Klaus Schulten**, **PhD**, speaking about the molecular dynamics simulator NAMD (http://www.ks.uiuc.edu/Research/namd/) and the molecular graphics viewer VMD (http://www.ks.uiuc.edu/Research/vmd/), which together have more than 100,000 users. "We went through a long initial phase where we were close to failure all the time." Schulten is professor of physics at the University of Illinois at Urbana-Champaign and director of the Theoretical and Computational Biophysics Group at the university's Beckman Institute.

> "There's a world of difference between developing code for yourself and developing code that you want to distribute," says Klaus Schulten.

For a tool to spread, it takes more than a good algorithm. From the start, someone has to build "disseminability" into the tool, with robust, flexible, and extensible code. Then, someone has to package the tool in a way that makes it accessible to a wide audience. Finally, someone has to publicize the tool, build a community of users, and support and maintain the tool.

In an ideal world, that "someone" would include a team of people with diverse skills—such as software engineers, technical writers, and marketers. But, in reality, it is often a scientist moonlighting as all of the above. Tool dissemination has traditionally been underappreciated and underfunded, making it hard for researchers to dedicate resources to tools beyond what's needed for their science. Fortunately, this situation is changing—with initiatives such as the NCBCs that recognize the importance of tool development and dissemination—but there is still a long way to go.

So how do scientists manage to do it right? *Biomedical Computation Review* spoke to a panel of individuals who have disseminated popular open source biomedical tools to find out what it takes to succeed and how they pulled it off.

Laying the Ground Work

The ingredients for successful tool dissemination have to be built into the tool's core from the start.

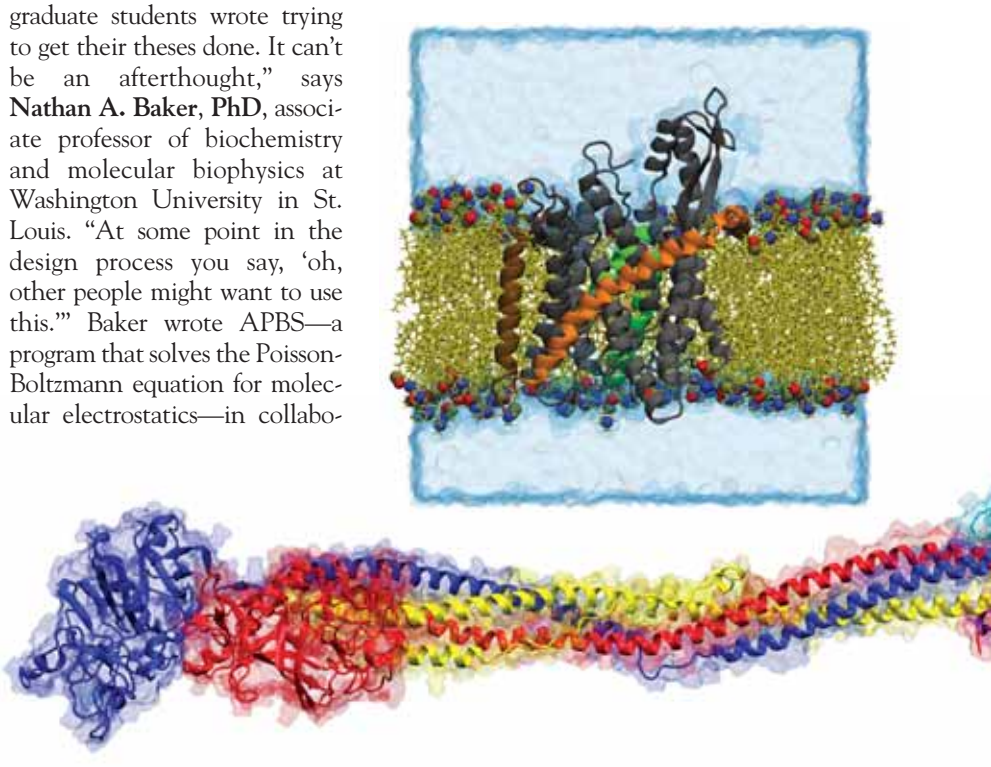
"You can't assemble a software package out of a bunch of code that your graduate students wrote trying to get their theses done. It can't be an afterthought," says **Nathan A. Baker**, **PhD**, associate professor of biochemistry and molecular biophysics at Washington University in St. Louis. "At some point in the design process you say, 'oh, other people might want to use this.'" Baker wrote APBS—a program that solves the Poisson-Boltzmann equation for molecular electrostatics—in collaboration with colleagues at the University of California, San Diego; the program is downloaded about 1000 times a month (http://apbs.sourceforge.net/).

When Baker realized that APBS offered something new that might be widely useful, he says, "I took most of what I'd written at that point and just deleted it and started over." A tool that is going out to others has to be built according to professional software design principles, he says. The code should be clean, bug-free, and robust; and it should be built in a flexible, modular fashion so that others can add to the tool and adapt it to their own problems.

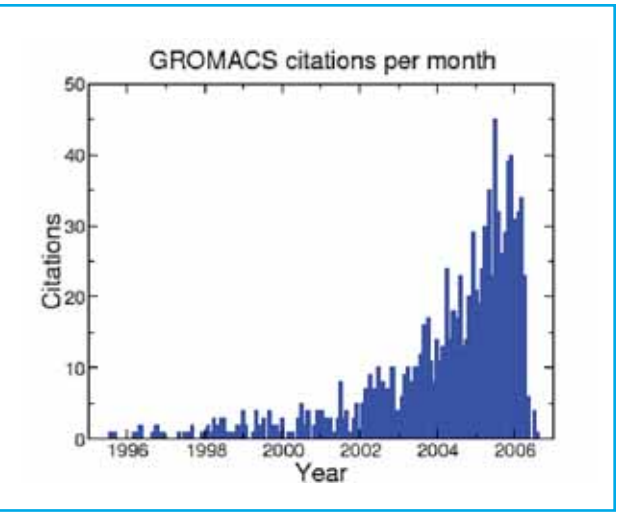
"There's a world of difference between developing code for yourself and developing code that you want to distribute," Schulten agrees. Establishing the proof of concept takes 10 percent of your time, whereas adhering to professional design principles takes 90 percent, he says. "And it is almost impossible to convince any normal scientist to spend that 90 percent." Professional programmers helped design VMD and NAMD, and they were a key factor in the tools' success, he says.

Dressing Your Tool for Success: Accessible, Well Documented, with a GUI

To become widely used, tools also have to be accessible—which means open source, portable, well documented, and user-friendly.



VMD Visuals: (top) secY protein, (lower left) fibrinogen protein,(lower right) polio virus particle. Picture made by the molecular graphics software VMD. Despite initial challenges, VMD is now a clear dissemination success story. The software is even used in high school classrooms. Courtesy of: the Theoretical and Computational Biophysics Group, NIH Resource for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign.



Growing a Tool. The use of GROMACS software has spiked since 2000: There has been growth every month in the number of citations to one or more of the three GROMACS papers or the manual. Courtesy of Erik Lindahl.

"Science is about getting things out there," says **Erik Lindahl**, **PhD**, associate professor in the Center for Biomembrane Research and the department of biochemistry & biophysics at Stockholm University in Sweden. "Unless you have this great 10 million dollar idea that will make you a fortune, the last thing you want to do is to limit access to your work." Lindahl is a primary developer of GROMACS, a molecular dynamics simulation package developed at the University of Groningen, which has been cited more than 1000 times (http://www.gromacs.org/).

When GROMACS was released in the early 1990s, it was not open source—academic users had to sign a contract and industry users had to pay a fee. But the licenses were a hassle and Lindahl barely broke even paying for the secretary to handle them, he says. "So, we realized this wasn't really very smart."

When they moved GROMACS to open source, their user base quickly jumped from 1000 to 5000 and continued to climb from there. The community also started voluntarily fixing bugs and writing new modules and patches. "Everybody benefits from the openness. So I think overall it's been an incredibly positive experience for us," Lindahl says.

GROMACS follows the GPL-style open source license, which requires those who adapt the software to make their programs open source as well. Other tools in this article follow the less restrictive BSD-style license. "If I was starting from scratch, I'd seriously consider going with this completely open license," Lindahl says.

"BSD actually worked out quite well for us," says **Steve Pieper**, **PhD**, founder and CEO of Isomics, Inc., in Cambridge, MA, and the dissemination core PI for the NCBC NA-MIC (National Alliance for Medical Image Computing). The NA-MIC toolkit includes visualization software: VTK, ITK, and Slicer (http://www.na-mic.org/Wiki/index.php/NA-MIC-Kit). The BSD license has allowed medical imaging companies to incorporate bits and pieces of the software into their equipment—which gets the technology out where it can directly benefit patients, Pieper says.

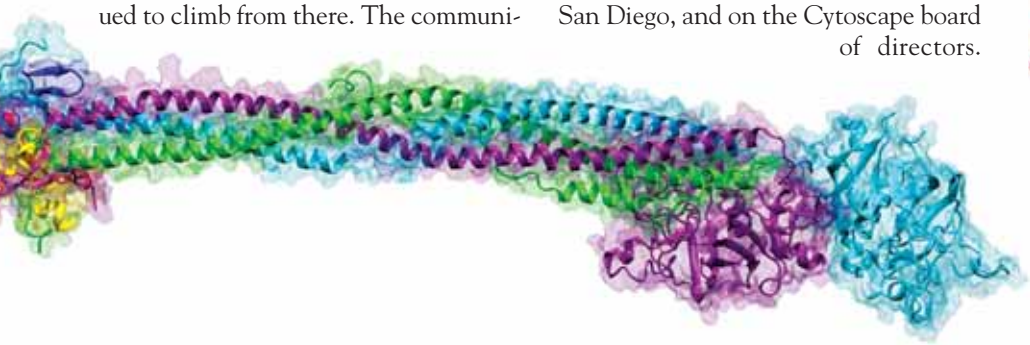
Cytoscape—which also follows the BSD license—has similarly been incorporated into several commercial software applications, says **Trey G. Ideker**, **PhD**, associate professor of bioengineering at the University of California, San Diego, and on the Cytoscape board of directors. Cytoscape is a software platform for modeling molecular interaction networks that gets about 3000 downloads per month (http://www.cytoscape.org/).

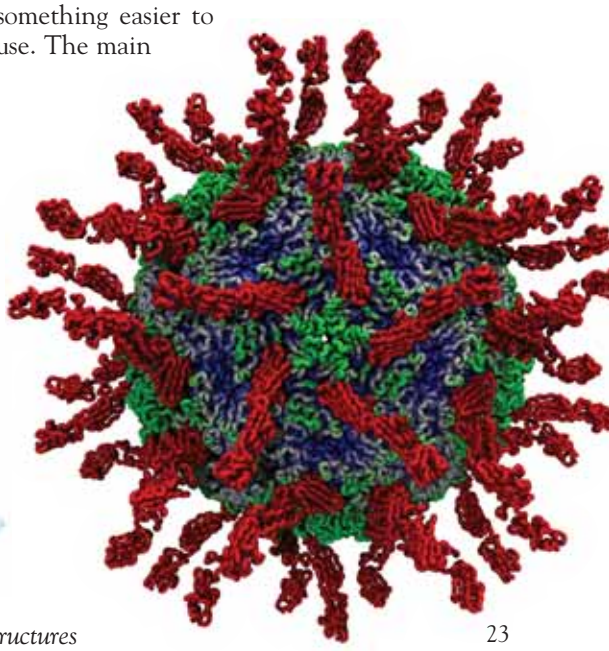
To be accessible, tools not only have to be free but also have to work on the computers that biologists are using, says **Thomas L. Madden**, **PhD**, a scientist at the National Center for Biotechnology Information at the U.S. National Library of Medicine. Madden helped transform UNIX-based BLAST into a tool that runs on multiple platforms, including Windows and Mac OS. BLAST is a sequence alignment tool and an undisputed tool success story—the original BLAST paper was the most highly cited biomedical paper in the 1990s (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

"A lot of bioinformatics tools are only made for Linux or Unix, but we've had just as many downloads of the PC version of BLAST as the Linux version," he says. "I think you can figure that just about every lab has a PC. So I don't think you can underestimate the importance of that."

Once users have a tool in-hand, if it is technically difficult or poorly documented, they are likely to seek out something easier to use. The main

"You can't assemble a software package out of a bunch of code that your graduate students wrote trying to get their theses done. It can't be an afterthought," says Nathan Baker.





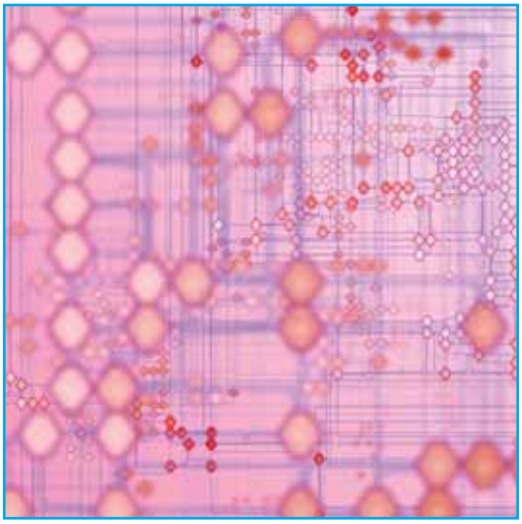
"Unless you have this great 10 million dollar idea that will make you a fortune, the last thing you want to do is to limit access to your work," says Erik Lindahl.

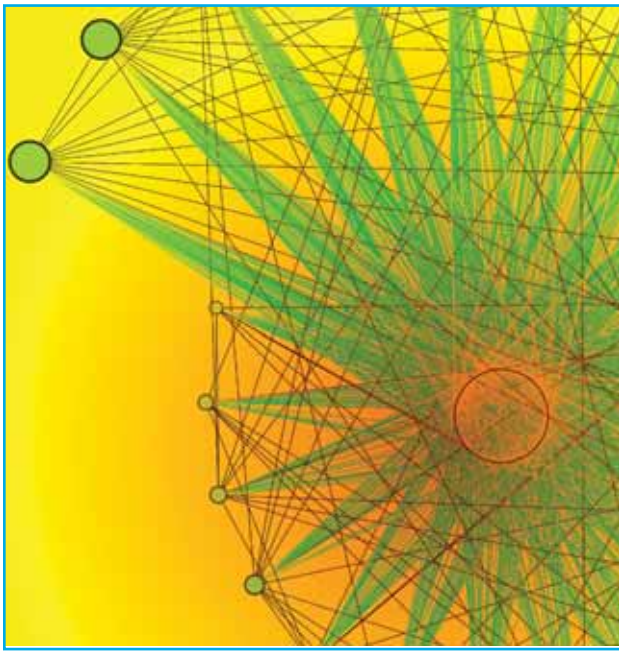
reason scientists flock to commercial alternatives for open source software is not because of superior performance (often the opposite is true), but because of a great user interface and great documentation, Lindahl says. Open source tools often fall short on these aspects.

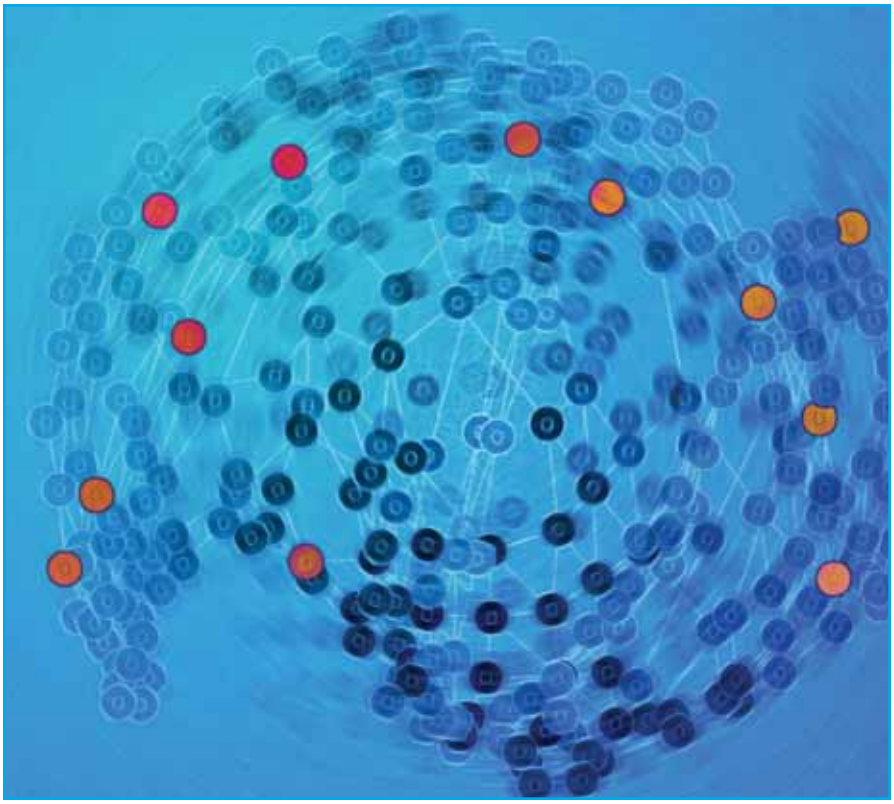
"I'm a sucker for good documentation. If there are not clear PDFs with graphics, I'm extremely unlikely to use it," says **Raymond R. Balise, PhD**, a biostatistical programmer at Stanford University, who uses the open source statistical package R, which has hundreds of thousands of users (http://www.r-project.org/). But the best programmers are usually not the best writers, he says. "So you have brilliantly designed elegant packages—and then good luck reading the documentation."

To help make the documentation more user-friendly, several of our interviewees advocate "learn by example" tutorials, which lead users step-by-step through common research problems.

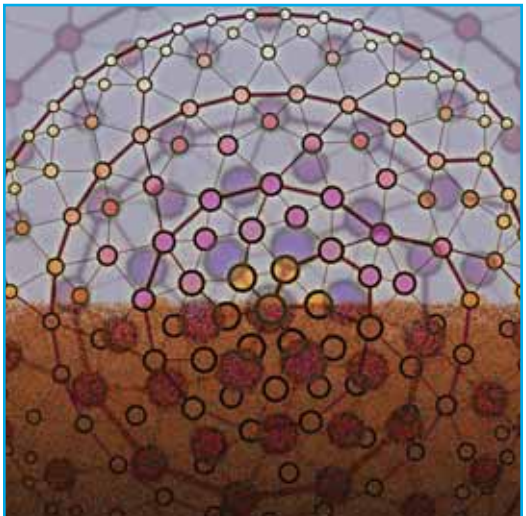
Many potential users are also deterred by the lack of a graphical user interface (GUI). For example, Baker says of APBS: "It's no worse than the other command-line computational biology tools. But I would say that maybe 80 percent of our audience would prefer to interact with it in some other way."







Cytoscape Pathways (Including background image on page 21). Pictures generated from Cytoscape, software for visualizing complex molecular interaction networks. Cytoscape follows a "non-viral" open source license, which allows companies to incorporate the software into their own commercial tools. Many companies now rely on Cytoscape as a critical part of their tools. Courtesy of: Vuk Pavlovic and Benjamin Elliott, the University of Toronto.



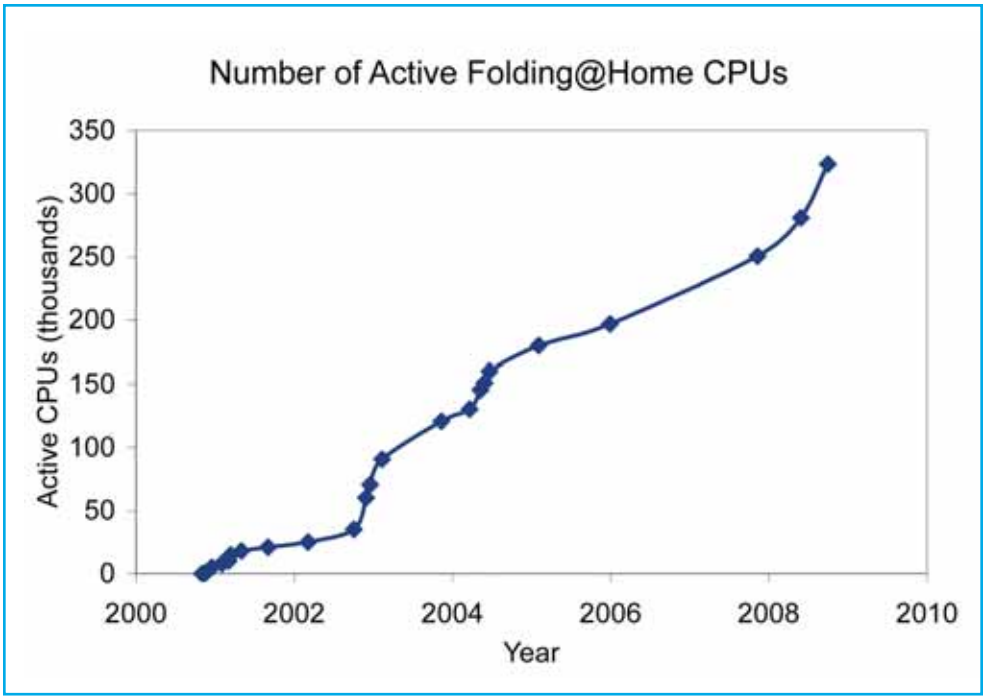
Similarly, R is a great tool for mathematicians and statisticians who are used to difficult programming languages, but telling physicians or biologists to "learn to program" just doesn't fly, Balise says. To make tools accessible to a wider audience, you need to wrap a nice GUI around the package and build in checks and balances to alert users if they're doing something wrong, he says.

Tool developers often resist these steps for fear that they will have to sacrifice power and flexibility for usability. An easy-to-use GUI-based interface is too constraining for research-driven tools, such as R and Bioconductor, that need to keep up with the cutting edge of science, says **Martin Morgan, PhD**, a core developer for Bioconductor, an R-based tool for analyzing high-throughput genomic data that has tens of thousands of users (http://www.bioconductor.org/). These tools may never be a satisfactory solution for a general audience, says Morgan, who is also a staff scientist and director of the Bioinformatics Shared Resource at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

But usability can evolve, even if the tool was designed for expert users. For example, community developers have spontaneously added GUIs onto several programs—including R Commander for R, and PyMOL and VMD plugins for APBS. Core developers may also revisit usability as a tool matures. For example, BLAST's core developers have become more focused on ease of use in recent years, particularly for the BLAST webpage interface, Madden says.

In rarer instances, developers consider usability from the start. This was the case with GenePattern, says **Jill Mesirov, PhD**, director of computational biology and bioinformatics and chief informatics officer at the Broad Institute of MIT and Harvard. GenePattern is an analysis program for genomic and proteomic data, which also captures users' steps in a reproducible pipeline; the package, released in 2004, already has thousands of users (http://www.broad.mit.edu/cancer/software/genepattern/).

From the beginning, GenePattern's developers recognized that they were targeting two audiences: "We have a number of computational scientists who do a lot of their own coding. We also have a lot of bench biologists who want to do analyses but don't want to write code. And why should they?" Mesirov says.



Power Surge. The number of active computers running Folding@home has surged since 2000. Courtesy of: Vijay Pande, Stanford University.

So, they developed the program to be modular and flexible for expert users, including allowing it to interface with standard programming languages such as MATLAB, Java, and R; but they also provided a point-and-click GUI.

"I think it really is the non-programming community that has made the package so popular," she says. "We get emails from both types of users, and we get really effusive ones from the non-programming users, because they say 'Wow, this really lets me use all these sophisticated tools and I can do it on my own,'" Mesirov says.

Connecting to Your Audience

The next step in tool dissemination is the actual dissemination—connecting the tool to users. This means not only getting the word out about the tool but also "selling" it.

"There is a mentality that if the tool is good enough it will speak for itself," says Stanford University's **Joy Ku, PhD**, director of dissemination for Simbios and its tools, including SimTK Core, a toolkit for physics-based biological simulations (http://simtk.org/home/simtkcore), and OpenSim, a package for modeling musculoskeletal movement (http://simtk.org/home/opensim). But, in many cases, particularly for complex tools, you really need active outreach to show people how the tool applies to them and how to use it, she says.

Outreach often starts with a publication that announces the tool. In the early days, people discovered BLAST primarily through the publication and word of mouth, Madden says. BLAST solved a key problem, so it was obvious how it was useful. Nowadays, "lightweight" outreach on the web can also go a long way, he says. You can reach many potential users with little cost through newsgroups, email lists, bloggers, and even random web searches.

"I'm a sucker for good documentation. If there are not clear PDFs with graphics, I'm extremely unlikely to use it," says Raymond Balise.

"One thing that worked very well for us is the web," Schulten agrees, speaking about VMD and NAMD.

"That really was a godsend because it's basically like we have a shop and our shopping window is the web." he says. "It's so easy to do and you reach so many people."

Non-programming users of GenePattern send effusive emails, says Jill Mesirov, "because they say 'Wow, this really lets me use all these sophisticated tools and I can do it on my own.'"

To promote FFTW ("the Fastest Fourier Transform in the West")—a general-purpose tool that performs Fourier transforms, which are often used in molecular dynamics simulations—creators **Steven G. Johnson, PhD**, assistant professor of applied mathematics at MIT, and **Matteo Frigo, PhD**, chief scientist and founder of Cilk Arts, focused heavily on web outreach. They posted benchmarks comparing their software with other FFT implementations; added FFTW links on websites that list FFT programs, as well as on sites that catalog free-software projects (such as freshmeat.net and directory.fsf.org); advertised on mailing lists; created their own mailing list; and answered questions on online discussions about FFTs, including providing links to FFTW and other free FFT software.

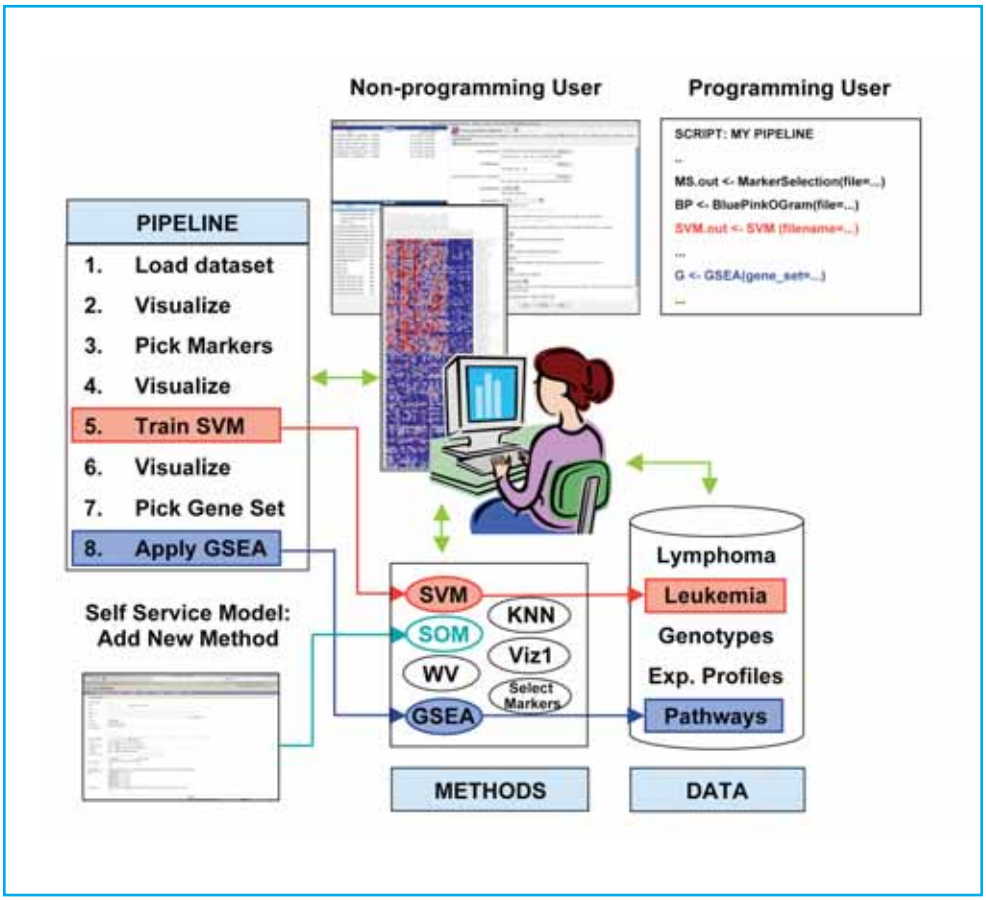
"Eventually, people began posting questions specifically about FFTW, and it was especially important to respond to these—having a support presence on public forums reassures people that the software works and is actively maintained," Johnson says. FFTW is now downloaded about 10,000 times a month (http://www.fftw.org/).

Active mailing lists and online forums help draw in new users, support existing users, and build a sense of community. "I frequently get much better support from open source mailing lists than you get from vendors," Lindahl says. Answering emails about the tool also goes a long way: "We've received over 10,000 email messages about FFTW over the past 10 years, and responded to a large fraction of them," Johnson says.

Beyond the web, more "heavyweight" outreach includes training sessions, workshops, and conferences. For example, Simbios and NA-MIC as well as other NCBCs hold training events at conferences and stand-alone workshops for developers and general users. Cytoscape developers run tutorials at the major bioinformatics conferences and some major disease conferences. It's hard to convince scientists to spend time running training sessions rather than improving the tool, Pieper says. So, it's important to involve people who are specifically interested in and passionate about teaching, he advises. R, Bioconductor, and Cytoscape hold their own annual conferences (funded primarily by corporate sponsors and paying participants), which help advertise the tools as well as bring developers together. "There's definitely a community, and the whole mentality of working as an international team is huge for R," Balise says.

High school teachers and college professors also promote tools in their classrooms. With VMD, "it became so user friendly that it could actually trickle down to college and high school education," Schulten says. "We were very fortunate that these outreach efforts were essentially ripped out of our hands. So now there are many efforts, and we just happily receive the news."

Distributed computing efforts are all about outreach, since researchers must convince the general public to download and run their tool. Coverage in

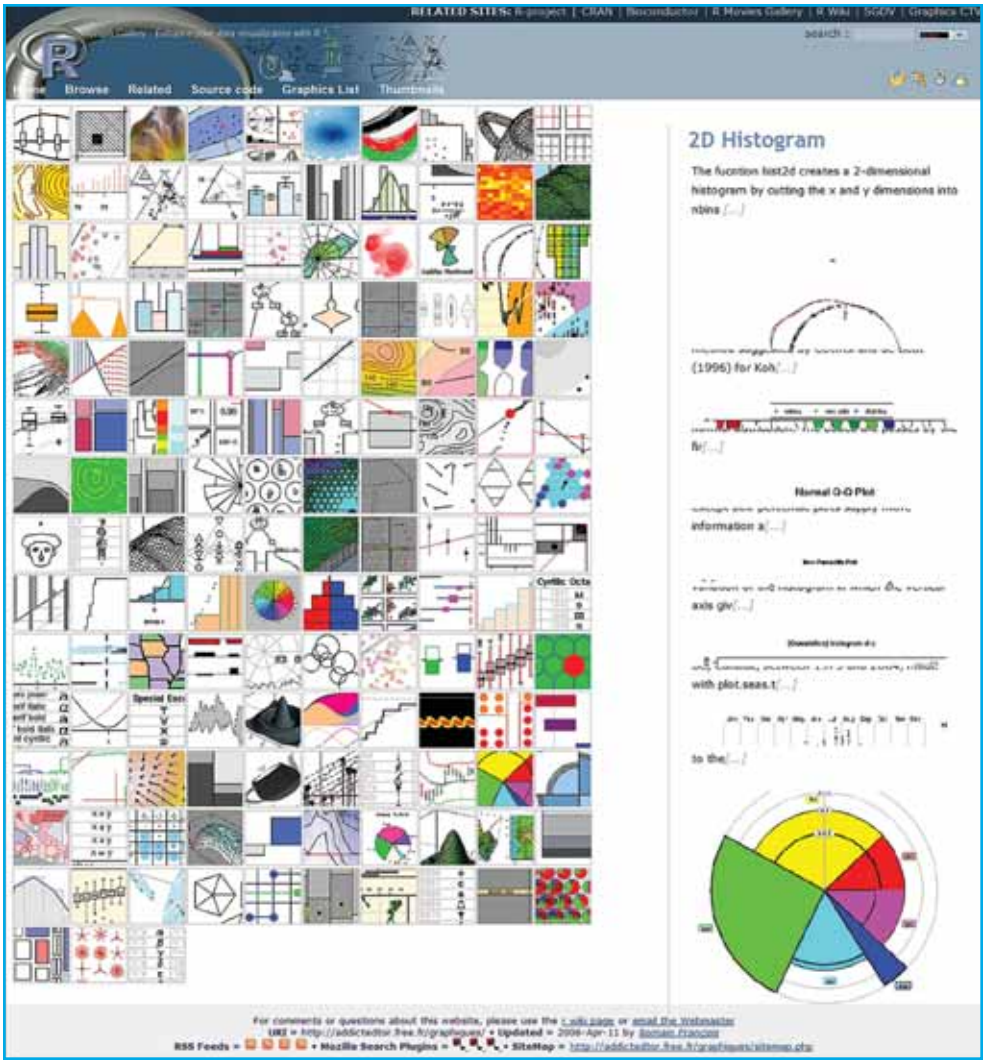


Building a Pipeline. The GenePattern tool helps expert and non-expert users analyze genomic and proteomic data, while capturing the steps in a reproducible pipeline. The tool was built with non-expert users in mind, which has been a major factor in the popularity of the tool. Reproduced from Reich M, GenePattern 2.0,* Nature Genetics *(2006) 38:500-501, supp. fig. 1.

the popular press (Time, CNN, and the New York Times, for example) helped generate buzz for Folding@home (http://folding.stanford.edu/), a distributed computing project at Stanford University led by **Vijay Pande, PhD**, associate professor of chemistry. Distributed computing also uses competition to stir up interest—participants collect points based on the amount of computing power they contribute. Capturing the high score is reminiscent of holding the high score on Asteroids at your local video arcade back in the eighties, but this is on a much grander scale, Pande says. "It's something on a very high profile site, where you can be number one out of hundreds of thousands."

Competitions are something we'd like to explore, Ku says. Already, Simbios runs a traditional grant competition for seed projects, which generates interest in and awareness of their center. "Ultimately you're only going to fund a small percentage of applicants, but all the applicants have to become familiar enough with what you're doing," she says. A similar approach could be used for software.

So, which of these outreach efforts is most effective? Until this year, we've just been going by an intuitive feel for what works, Ku says. But, in an effort to improve dissemination, they collected eight months of data on how people find their software project repository Web sites, simtk.org. The breakdown is: 29% word of mouth; 25% publications and conferences; 24% web search; 13% mailing lists and newsgroups; 9% other mechanisms (including use in the classroom, *Biomedical Computation Review*, and links on other Web sites). Word of mouth leads the way, but it accounts for less than one-third of hits—so more active outreach is vital.



R Gallery. Community developers have written so many graphical programs for data visualization in R that it's hard to keep track of them; here the programs are cataloged visually for easier access. Contributions from the community have been critical to R's growth and success. Screenshot from the R Graph Gallery, http://addictedtor.free.fr/graphiques.

Making it Happen

Successful tool dissemination can be lengthy and costly, and it requires diverse skills, such as programming, writing, marketing, and teaching. So how do scientists support these efforts?

"Up to now it's frequently been the case that you're kind of moonlighting," Lindahl says. "One problem both in Europe and in the States is that it's hard to get funded only for software development." Many tools are supported using bits and pieces of resources scrounged from science-driven grants as well as many hours of volunteerism—from professors, graduate students, postdocs, and community members. Under this piecemeal model, there's no money to hire professional programmers let alone technical writers or outreach coordinators. Lindahl says he'd "nudge" postdocs to turn code they wrote for their research into formal GROMACS modules. Pande says he and his graduate students have to work 60 to 70-hour weeks to keep Folding@home going. "It's just a lot of work to be running something like this," Pande says. Johnson says he and Frigo did most of the legwork for FFTW themselves over the years, despite many other time commitments.

"One problem both in Europe and in the States is that it's hard to get funded only for software development," says Lindahl.

Tool upkeep and dissemination are also undervalued when it comes to academic promotion—making it even harder to justify dedicating scarce time

and resources to these endeavors. "Academic credit for maintaining software is not the same as producing publications," says BioPerl developer **Jason E. Stajich, PhD**, Miller Research Fellow in the department of plant and microbial biology at the University of California, Berkeley. BioPerl is a programming toolkit for processing sequence data. It has been cited more than 500 times (http://www.bioperl.org/wiki/Main_Page). Stajich worked heavily on BioPerl before and during his graduate studies but, as he transitions to a faculty position, he needs to focus more on his science; and many other developers are in the same situation. "We'd like to do more outreach, but it requires a critical mass of people who actually have time to do that," he says.

To augment the piecemeal model of tool dissemination, some groups have formed non-profits. For example, Stajich and his colleagues formed the Open Bioinformatics Foundation, which provides infrastructure for BioPerl and related projects, such as BioJava and BioPython. Similarly, the Cytoscape Consortium provides an umbrella for the institutions involved in Cytoscape core development. The non-profit model can help with logistics, including accepting donations and running conferences.

Other tools in this article have managed to obtain tool-specific funding, which was likely instrumental in their success. For example, APBS, GenePattern, and some members of the Cytoscape Consortium have been funded through NIH's R01 program for "software development and maintenance" (which has been available since 2002). GROMACS has also obtained recent funding through the European Union. The funding gives us the ability to reply to user requests within 24 to 48 hours and to develop tutorials, Baker (of APBS) says. "Without that funding, that just wouldn't happen. It would be like, as with most previous funding, an afterthought in some grant: 'Oh, and by the way, I guess we'll keep this tool limping along.'"

As part of the NCBCs, Simbios and NA-MIC have specific funding for tool maintenance and dissemination. "One of the things that's great about the NCBC program is that there's funding to do actual training events," Pieper says. Finally, Schulten has had long-standing (two decades of) tool-specific funding through an NIH P41 grant—which specifically funds technology development. These funds allow him to hire professional programmers and run training events.

Measuring Success and Reflecting on Failure

The final step in tool dissemination is evaluation—measuring how well the efforts are going.

"It is extremely difficult to measure the popularity of a free software project like FFTW," Johnson says. Citations provide a rigorous measure of success, but these take time to accumulate. So, our interviewees also track softer measures including: registered users, downloads, mailing list subscribers, mailing list activity, Web site visits, conference attendees, and the number of plugins added to a tool.

This article focuses on tools that succeeded. But, for every success story, many more tools have failed. In a recent editorial in *PLoS Computational Biology*, founding editor-in-chief **Philip E. Bourne, PhD**, a professor of pharmacology at the University of California, San Diego, and his colleagues describe their efforts to track down 14 software programs (for partitioning proteins into domains) described in published papers. Eight programs were not even accessible in a usable form, let alone widely used and popular. Given the difficulty of the task and the lack of rewards, it's not surprising that so many tools languish into obscurity—but it is wasteful and reflects poorly on the biomedical computing community. "There's a huge amount of resource that goes into making these things, and so much of it is just lost." Bourne says.

Fortunately, funding agencies and journals are beginning to acknowledge the importance of tool upkeep and dissemination. In the past few years, the National Science Foundation (NSF) and NIH have "come around to the idea that software is not something to be dabbled with," Pande says. Lindahl has also noticed an increase in tool-specific funding. Journals could also help alter the reward system, Bourne says. *PLoS* is contemplating a software section where papers will only be published if the software is deposited in an open source archive such as sourceforge.net or bioinformatics.org. Online journal editors or readers could simply add a comment to papers when the software is no longer available, Bourne says. "That would sort of be a black mark against the author, so I think that might encourage the author to make the software available longer."

Even with more incentives and resources, tool dissemination will still be a challenge. Despite sufficient resources and a proven track record in tool dissemination, Schulten says his latest tool, BioCore (http://www.ks.uiuc.edu/Research/biocore/), is teetering on the edge of failure. BioCore is a collaborative work environment for biomedical research, supporting tasks such as co-authoring papers and sharing molecular visualization results. The program hasn't taken off yet, in part because scientists are reluctant to try new technology, he says. But Schulten is determined to showcase the tool more and run more training events. "We have to put more energy into these efforts," he says.

Success requires persistence, Lindahl agrees. "Don't give up in the beginning. It takes a while to build these communities." □

In the past few years, the National Science Foundation (NSF) and NIH have "come around to the idea that software is not something to be dabbled with," Vijay Pande says.

Under TheHood

BY KARTIK MANI, PhD

Network-based Approaches to Prediction of Disease Genes

The recent surge of high-throughput experimental data, such as gene expression microarrays, offers a profound opportunity to gain a more detailed understanding of the genes involved in the progression of disease. While initial analyses of these data used statistical techniques to identify genes capable of distinguishing disease tissue from normal (biomarkers), researchers are now turning to the analysis of gene interaction networks to address this problem.

Gene interaction networks may be developed from several sources including manual curation, high-throughput experiments (such as yeast 2-hybrid), literature mining and reverse engineering algorithms. They can include many different types of interactions as well (complexes, regulatory, signaling, etc). Integrating and analyzing all of this information to discover genes relevant to disease requires network-based algorithms. Thus far, such algorithms fall into three general (though not necessarily mutually exclusive) categories. The first predicts protein complexes, rather than individual genes, associated with the disease phenotype. The second identifies key regulators (transcriptional, signaling, or other), which control a large set of genes differentially expressed in the disease state. The third, the focus here, relies on the fact that interaction networks are themselves dynamic and may change from a normal to disease state. Thus, if one identifies interactions that have actually changed between phenotypes, one might then work backwards to identify genes that could prove promising for further investigation.

We will detail two examples of the third category, both of which incidentally use an information-theoretic approach. The first defines a concept called synergy, which measures the cooperative effect of two variables on the state of a third. The two variables in this case are genes (*G1* and *G2*), and the third is a binary state variable representing disease or normal (*D*). Formulaically, this can be represented as the difference between *I(G1,G2;D)* (the cooperative effect) and the sum *I(G1;D) + I(G2;D)* (the individual effects), where I is mutual information. Biologically, synergistic interactions imply that the combined state of the two genes affects disease, while individually the genes have a far lesser or no effect. This algorithm computes this quantity across all gene pairs represented on the input microarray data, and a "synergy network" is generated from the highest scoring interactions. When applied to publicly available prostate cancer data, this approach showed the *RBP1I* gene participating in a large number of synergistic interactions. This finding along with others indicated that the progression of prostate cancer is linked with oxidative stress and inhibition of the apoptosis pathway, consistent with previous hypotheses.

The second algorithm, Interactome Dysregulation Enrichment Analysis (IDEA), computes the mutual information between two genes across a large, diverse dataset, including or excluding one particular disease phenotype (P). Formulaically, this test is represented as the difference (ΔI) between $I_{all}(G1;G2)$ and $I_{all\text{-}P}(G1;G2)$, where I_{all} includes all sample points, and $I_{all\text{-}P}$ excludes the phenotype P. Biologically, a positive or negative ΔI implies that these two genes have gained or lost an interaction in the phenotype P respectively (e.g., an oncogene "loses" its ability to be regulated in cancer). The genes participating in a statistically significant number of these interactions are then selected. When applied to data from three primary B cell lymphomas, IDEA correctly predicted the known oncogenes reported in the literature (e.g., MYC in Burkitt's Lymphoma), as well as effector genes not identified by differential expression analysis.

These network-based approaches, along with others, have shown promise in more accurately delineating the mechanisms of disease progression. Like any new class of methods, however, there are drawbacks. First and foremost, there is no "gold standard" of gene interactions that can be used, although the knowledge base is growing rapidly. They often require large training sets or sample diversity to be effective, which may not always be available. Lastly, computational complexity may limit their applicability.

Nevertheless, the application of networks and these algorithms to the identification of disease-causing genes remains an exciting new area of computational biology. Expect to see several new network-based approaches emerge as the body of high-throughput and interaction-based data continues to grow.

If one identifies interactions that have actually changed between phenotypes, one might then work backwards to identify genes that could prove promising for further investigation.

DETAILS

Kartik Mani received his PhD in Biomedical informatics at Columbia University, working in the Multi-Scale Analysis of Genomic and Cellular Networks (MAGNet) Center under the direction of Dr. Andrea Califano. His research focused on the application of interaction networks to gene-disease association, and culminated in the development of the IDEA algorithm described above. He is currently pursuing his MD at the Albert Einstein College of Medicine in Bronx, NY.

REFERENCES

1. Watkinson, J., X. Wang, et al. (2008). BMC Syst Biol 2: 10.

2. Mani, K. M., C. Lefebvre, et al. (2008). Mol Syst Biol 4: 169. □

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by Simbios, an NIH National Center for Biomedical Computing

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

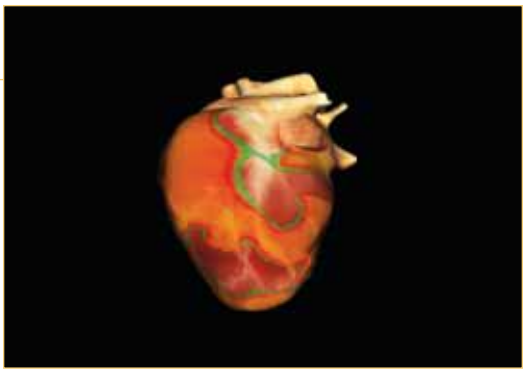
Biomedical Computation Review
Simbios An NIH National Center for Biomedical Computing
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

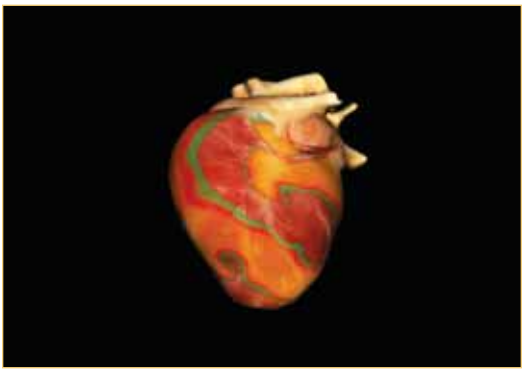
seeing science

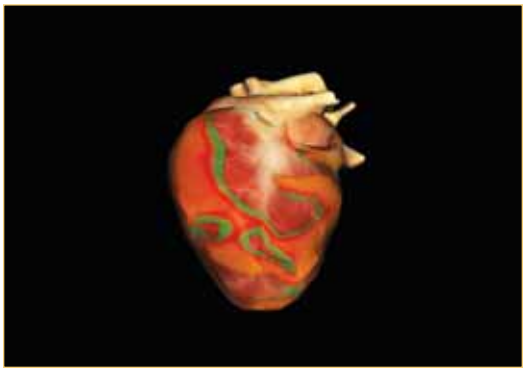
SeeingScience

BY KATHARINE MILLER

Visualizing Ventricular Fibrillation







Unsynchronized twitching of the heart's ventricles—known as ventricular fibrillation—kills about 300,000 Americans yearly. Its underlying cause: electrical spiral and scroll waves that propagate through the heart. Simulation and visualization are playing an important role in understanding that process.

In a novel approach to a review of the research, **Flavio Fenton, PhD**, and **Elizabeth Cherry, PhD**, research associates in biomedical sciences at Cornell University, simulated and visualized what's currently known about how electrical spiral waves propagate through the heart to cause tachycardia (rapid heart rate) and fibrillation. The work was published in the December 2008 Visualization in Physics focus issue of the *New Journal of Physics*. □

Cherry and Fenton simulated electrical spiral waves through the three-dimensional heart. In the Java Applet of this 3-D simulation, we see a so-called "mother rotor" spiral wave on the front of the heart. Although this might suggest a single spiral wave that would cause only tachycardia (rapid heart beat), the 3-D heart can be rotated in the Java applet to show the breakup of the wave on the back of the ventricles—a sign that this heart would begin to quiver or twitch uncontrollably in fibrillation. When this happens, no blood gets pumped to the body or lungs.

Images reprinted with permission from EM Cherry and FH Fenton, Visualization of spiral and scroll waves in simulated and experimental cardiac tissue,* New Journal of Physics *10 (2008) 125016, Figure 33d Java applet. Also visit http://thevirtualheart.org.