DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

THE CELL IN 2010:
A MODELING ODYSSEY

PLUS:

MORE THAN FATE:
Computation Addresses Hot Topics in Stem Cell Research

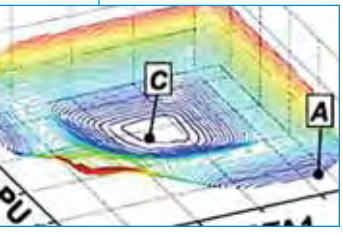
Spring 2010

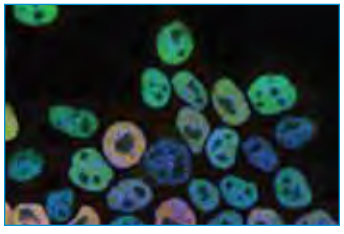
ContentsSpring 2010

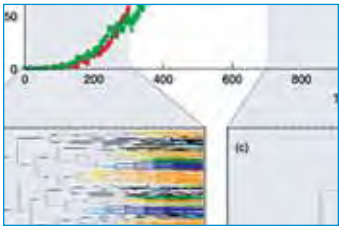
FEATURES

DEPARTMENTS

Cover Art: Created by Rachel Jones of Wink Design Studio using cell image © Promotive | Dreamstime.com.







Spring 2010
Volume 6, Issue 2
ISSN 1557-3192

Executive Editor David Paik, PhD

Managing Editor Katharine Miller

Associate Editor Joy Ku, PhD

Science Writers
Kristin Sainani, PhD, Katharine Miller, Louisa Dalton, Beth Skwarecki, Regina Nuzzo, PhD, Chandra Shekhar, PhD

Community Contributors
Peter Lyster, PhD, Joy Ku, PhD, Peter Eastman, 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
The 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)
Semahat Demir, PhD (NSF)
Jim Gnadt, PhD (NINDS)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Richard Morris, PhD (NIAID)
Grace Peng, PhD (NIBIB)
Nancy Shinowara , PhD (NCMRR)
David Thomassen, PhD (DOE)
Jane Ye, PhD (NLM)

GuestEditorial

BY PETER LYSTER, PhD, PROGRAM DIRECTOR IN THE CENTER FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, NIGMS

Update on Biomedical Computation at NIH

As a program manager in biomedical computing and computational biology at the National Institutes of Health, I field many questions, particularly from new investigators. They ask questions like: Where do I find out about research funding? How do I navigate all the information? Whom do I contact? I want to take this opportunity to share a few insights.

NIH does not have a top-down approach for biomedical computing and computational biology, but it does have a highly coordinated community. Individual institutes and centers develop initiatives or are assigned incoming applications for funding in biomedical research through the Center for Scientific Review Division of Receipt and Referral (CSR DRR).

If anything comes close to centralizing biomedical computing and computational biology at NIH it's the CSR DRR—a hub that literally sorts out applications and assigns them to the most appropriate institute or center as well as the study section. The people who do this are science administrators who use their expert knowledge and excellent judgment to identify the right home for each application. When applicants tell me that they're going to request assignment to a certain program director, institute or study section, I tell them: "If you're not sure what you're doing, don't get tangled up in all that—let the experts at CSR DRR handle it so you can concentrate on the science."

All study sections at CSR can potentially review applications for research funding that involve some computing. However, there are nine study sections that review applications with a significant amount of biomedical computing. These include mainline modeling and analysis (MABS), data and analysis (BDMA), health informatics (BCHI), neurotechnology (NT), genomics and computational biology (GCAT), macromolecular structure and function (MSFD), biostatistics (BMRD), biomedical imaging (BMIT), and microscopy (MI). These nine study sections really point to the importance of computing in biomedical research and that these research areas merit special focus.

The glue that holds a lot of this together is BISTI, the trans-NIH Biomedical Information Science and Technology Initiative (BISTI) consortium. BISTI, for example, coordinates a number of initiatives in biomedical computing and computational biology. It is also the administrative center for the National Centers for Biomedical Computing, which are part of the NIH Roadmap for Medical Research. This program and its affiliated collaborations have funded more than $150 million in research in the past five years, and the effort will continue through 2015.

There are plenty of other opportunities for research funding across a range of size and complexity, and you can find them all listed on the BISTI Web site. Last year, BISTI reissued four broad-based program announcements to support "innovations in biomedical computing." They cover a range of areas, from the development of enabling technologies and non-hypothesis-based research to specific research relating to the needs of a disease or research area of interest to a specific IC. Of course investigators can also use the regular investigator-initiated R01 mechanism for requests for funding that have substantial components of computing.

BISTI and other related programs across the institutes and centers play an important role in providing both contacts and coordinating initiatives—and this creates a lot of communication within the NIH community. When I receive an application that I think may be more appropriate for another institute, I will use the BISTI Web site to find the right contact and then discuss the best home for review.

continued on page 29

NIH does not have a top-down approach for biomedical computing and computational biology, but it does have a highly coordinated community. Individual institutes and centers develop initiatives or are assigned incoming applications for funding in biomedical research through the Center for Scientific Review Division of Receipt and Referral (CSR DRR).

CHANGES IN THE NIH GRANT APPLICATION AND REVIEW PROCESS:

Want to see what's going on lately in the effort to enhance peer review? Go to the NIH site http://enhancing-peer-review.nih.gov/.
Since January 25, 2010, all applications are submitted on new forms with shorter page limits. The new page limits (http://enhancing-peer-review.nih.gov/ page_limits.html) include a 12-page Research Strategy for most applications.

SimbiosNews

BY JOY P. KU, PhD, DIRECTOR OF DISSEMINATION FOR SIMBIOS

Protein Mechanica: Structural Modeling for the Experimentalist

Scientists sometimes find themselves up to their elbows in Styrofoam balls, pipe cleaners, and metal rods as they try to build models of the molecules they are studying. Now, they can exchange all that for the ease and precision of a computer. With the alpha release of the modeling software Protein Mechanica, researchers have a new option for constructing plausible models of molecules based on experimental data, such as x-ray crystallography and cryo-electron microscopy (cryo-EM), and then simulating and visualizing their conformations.

The model-building component is a unique aspect of the software, says **David Parker**, a recent Simbios student who developed the software as part of his PhD thesis. The software has been used to reproduce experimental observations of a myosin V, a protein that moves cellular cargo along actin filaments, but Parker points out that it is new and they are still validating it. "There's no software that does this kind of thing, so we are still trying to understand how to apply different coarse-graining modeling strategies. The more systems we work on, the better we understand how to parameterize the models."

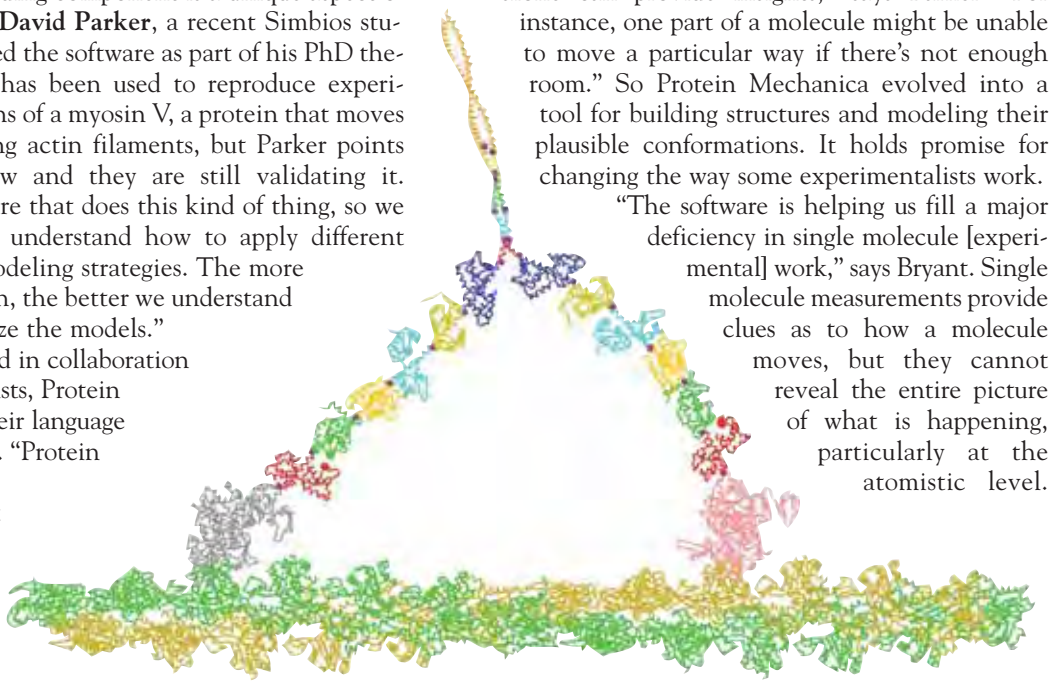
Designed for and in collaboration with experimentalists, Protein Mechanica uses their language to build the model. "Protein Mechanica will allow anyone to sit down and build these models," Parker says. "You use protein chain and residue identifiers and atom names, so in a script of only two lines you can build a complete mechanical model of a large molecular system."

With Protein Mechanica, researchers will be able to construct models using information from a variety of sources: crystallography, cryo-EM, secondary structure descriptions, as well as user-defined solid shapes, such as spheres and cylinders. This flexibility is useful since crystal structures typically contain only molecular fragments, requiring that different crystal structures be integrated or that missing pieces be represented by other means in order to produce a complete model.

By working closely with experimentalist **Zev Bryant, PhD**, an assistant professor of bioengineering at Stanford University, Parker identified what is really important, and it did not require full-blown dynamics. Experimentalists just need a tool that can quickly show them how different parts of a molecule interact mechanically. They do not need or want to spend weeks simulating the diffusion process of the molecule.

"Just the geometric constraints of the model alone can provide insights," says Parker. "For instance, one part of a molecule might be unable to move a particular way if there's not enough room." So Protein Mechanica evolved into a tool for building structures and modeling their plausible conformations. It holds promise for changing the way some experimentalists work.

"The software is helping us fill a major deficiency in single molecule [experimental] work," says Bryant. Single molecule measurements provide clues as to how a molecule moves, but they cannot reveal the entire picture of what is happening, particularly at the atomistic level.



A simulation of myosin V binding to actin, as modeled with Protein Mechanica. Courtesy of David Parker.

Modeling can help fill in the gap, as well as aid in the design of new molecules. Protein Mechanica is particularly useful because it enables comparisons between its predictions and experimental data at various levels of spatial resolution.

Bryant says his students can now quickly take any myosin they're working on, build hypothetical conformations, and compare measurements from the model with measurements taken from single molecule assays. "Protein Mechanica is very well-suited for that task and I think very expandable," he says. □

DETAILS

Protein Mechanica is freely available for download from https://simtk.org/home/protmech.

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

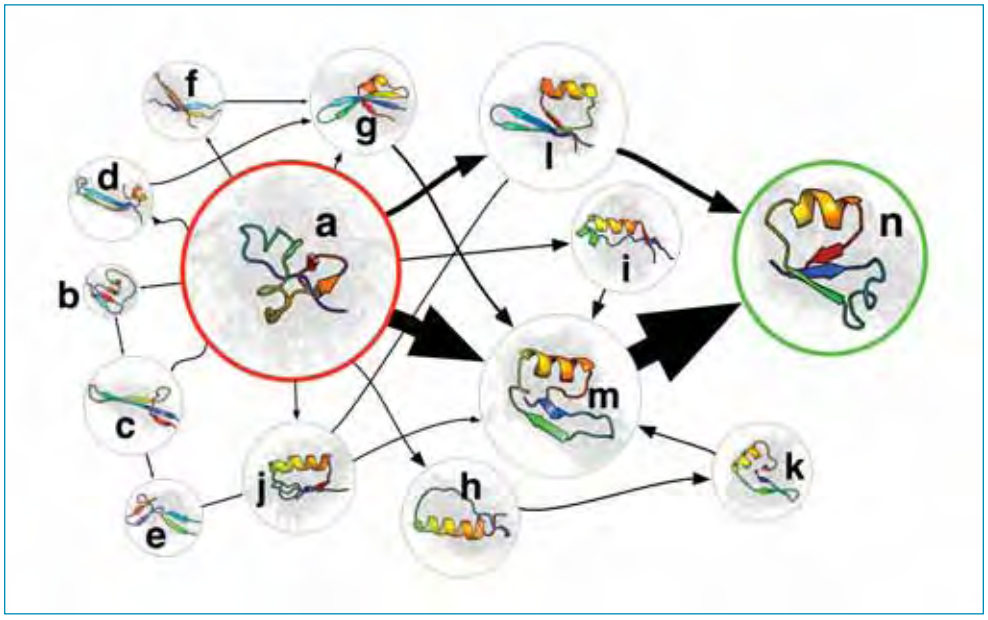
NewsBytes

Scientists Break Protein Folding Time Barrier

Scientists have now simulated protein folding at a timescale that begins to be relevant to biology: the millisecond. Indeed, the simulation busted through the millisecond time barrier to tackle the slowest folding protein yet studied—a 1.5 millisecond fold—using a combination of computational tools that provide both the requisite computational power and the necessary analytical methods for making sense of a slow, complicated folding event.

"It's kind of like a coming out party for a combination of technologies that have really started to mature," says **Vijay Pande, PhD**, associate professor of chemistry at Stanford University and co-author of the paper. Pande expects his team's technologies will be useful for simulating proteins important in misfolding diseases such as Alzheimer's and Huntington's disease. The work was published in the *Journal of the American Chemical Society* in January.

Protein-folding researchers have until now focused on a unique group of small, fast-folding proteins that fold in hundreds of nanoseconds or microseconds. This is great for simulating, but it is not characteristic of most protein-folding events. Pande's group chose to simulate a 39-amino acid chain called NTL9, which, like most proteins, dilly-dallies en route to its final structure. One side of the protein may partially fold, then unfold as another part misfolds. The process takes milliseconds or more.



A Markov State Model illustrates how NTL9 progresses from a fully unfolded state (a) to a fully folded state (n) through many different pathways. (This simple 14-state model illustrates only the most frequented folding pathways.) The larger the arrow, the more a path is traveled. The larger the circle, the greater a state's stability. Courtesy of Vijay Pande. Reprinted from Voelz, V., et al., Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9 (1?39),* Journal of the American Chemical Society, *132(5):1526-8 (2010) with permission from the American Chemical Society. A video is also available at http://www.youtube.com/watch?v=gFcp2Xpd29I

The computational power for the simulation came from Folding@Home, a distributed computing project that heaps together bits of donated computer time from individual systems located around the world. To fold NTL9, they relied particularly on the speedy graphical processing units (GPUs) within those computers, which sped up the simulations and made long folding trajectories possible.

To piece together the information from the different computers, Pande and his coworkers also devised a Markov State Models (MSMs) method. The approach merges myriad variations from thousands of successive protein-folding simulations and identifies a set of relatively stable conformations along the protein's many folding pathways. By choosing how many states to identify, whether fifteen or 100,000, researchers can dial in the degree of complexity they seek. It's like choosing the number of pixels in a photograph, Pande says. A small number of states gives a broad, coarse picture of the conformations and folding pathways of greatest frequency, while a larger number provides a more complex picture that can show specific protein movements in greater detail.

The MSM approach allowed Pande's group to see a real richness of range in the way NTL9 folds. NTL9 follows not just one or two pathways but many different paths to get to the final folded state. Pande expects to see similar heterogeneity in the way other proteins fold, and his group has created a tool called MSMBuilder to enable other groups to conduct a similar analysis of their own simulations.

Jed Pitera, PhD, a research staff member at IBM, says Pande's group found a way to build a statistically and physically accurate model of protein folding. "It shows off the state-of-the-art in studies of folding kinetics and reflects a maturation of the view of how protein folding happens," he says.

—By Louisa Dalton

Predicting Protein Complexes

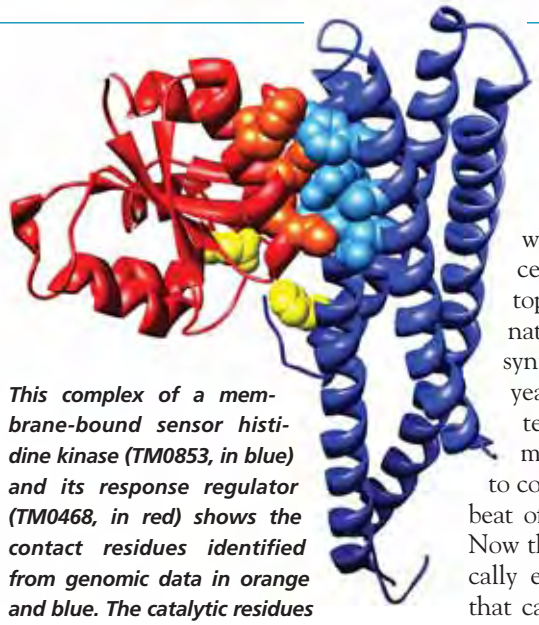
The zone where two proteins interact presents a possible target for drug design. But identifying possible drugs

requires a detailed understanding of the interface between the proteins. Computer simulation provides a useful tool for gaining such an understanding. But simulating protein complexes can be challenging, especially when the interactions are fleeting—such as when signaling molecules attach and detach in a flicker. Now, a new method can efficiently predict the structures of transient protein complexes from a combination of genomic and structural data.

"This is an entire approach to protein-complex structures based on several different computational methods," says **Hendrik Szurmant, PhD**, coauthor of the paper and an assistant professor of molecular and experimental medicine at the Scripps Research Institute. The work was published in *Proceedings of the National Academy of Science* in December 2009.

To determine the structures of proteins in complexes, researchers have used both homology modeling and purely physics-based molecular dynamics simulations. But both approaches have proven less successful than hybrid approaches. Szurmant and his colleagues developed their new hybrid approach using a two-component signaling system in bacteria as a test case. The system consists of a membrane-bound kinase that passes a phosphoryl group on to its response regulator within the cell. The team analyzed databases of genomic sequences for almost 9000 examples of the two co-evolving proteins looking for co-varying mutations. Their aim: to identify likely points of contact between the two players, under the theory that one protein's contact residue tends to match a mutation in its partner. Then they used those points of contact to combine the two proteins in a molecular dynamics simulation.

"Our method brings the two proteins close together in a computationally very inexpensive way, then as a very last step the structure is refined in a molecular force field," says Szurmant. The combination of approaches minimizes computation time, he says, compared to methods that rely more heavily on molecular dynamics.



This complex of a membrane-bound sensor histidine kinase (TM0853, in blue) and its response regulator (TM0468, in red) shows the contact residues identified from genomic data in orange and blue. The catalytic residues that exchange a phosphoryl group are shown in yellow. Courtesy of Alexander Schug and Hendrik Szurmant.

When the researchers tested their methods on a complex whose structure had already been determined, the prediction was in excellent agreement with the known structure. They also tackled a then-unresolved complex from *Thermotoga maritima*, TM0853/TM0468. An x-ray diffraction structure of that complex has since been published, confirming many aspects of the prediction.

This technique could be used for other types of systems, says Szurmant, so long as enough sequence information is available for the genomic step to pick out statistically significant variations. "The approach relies on variability, so if the system is very conserved, one would need a lot more sequence," he says. The team's next step is to apply the method to other bacterial systems, and eventually to develop an online tool to make the approach available to other researchers.

This work shows that the combination of genomics data and molecular dynamics modeling seems to be sufficient to predict protein complex structures, says **Angel Garcia, PhD**, professor of physics at Rensselaer Polytechnic Institute. Garcia points out that the accuracy of the method is particularly impressive. He adds, "I think almost anyone that is working on a given complex is going to try this for their own pet system."

—By Beth Skwarecki

Synchronizing Cells

Without synchronized clocks—whether embedded in our body's cells or programmed into our desktop computers—any kind of coordinated activity is impossible. So after synthetic biologists succeeded last year in programming individual bacteria to keep time and blink rhythmically, they wanted to find a way to coax each bacterium away from the beat of its own idiosyncratic drummer. Now they've figured out how to genetically engineer a population of *E. coli* that can not only blink in unison, but also automatically synchronize itself.

"Often synchronization is achieved by enslaving multiple clocks, or oscillators, to one 'central command unit,'" says **Lev Tsimring, PhD**, associate director of the University of California, San Diego's BioCircuits Institute, who headed the research team with **Jeff Hasty, PhD**, associate professor of biology and bioengineering at UCSD. But that's putting a lot of eggs in one basket: If something goes wrong with the master clock, the whole system can collapse.

The team's solution was to make use of quorum sensing, in which cells communicate with each other by relaying small molecules between them. In their design, a genetic oscillator first drives engineered bacteria to turn fluorescent proteins on and off. Then the cells use quorum-sensing components to share information about the timing of their oscillations and adjust their cycles accordingly.

The work, which was published in *Nature* in January 2010. used computational modeling of the oscillators to quantitatively explain the experimental observations. For example, the researchers tweaked the computational model parameters to artificially prevent a certain molecule—which was thought to be involved in both the cells' timekeeping and communication—from penetrating the cell walls. Their results showed that without the molecule, the individual cells were indeed cut off from each other and their environment, and their clocks remained unsynchronized. And because there's

no way to confine this molecule within cell walls experimentally, "observing" this behavior was possible only through computational modeling, says **Tal Danino**, graduate student in the UCSD Department of Bioengineering and lead author of the study.

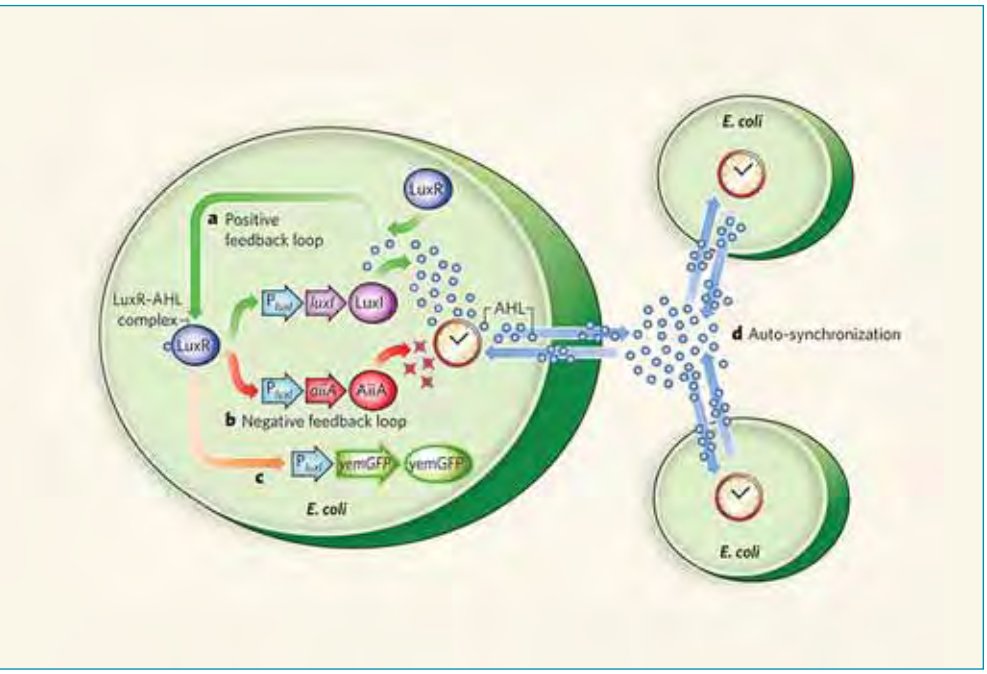
Computation is indeed a valuable tool for understanding gene networks, Hasty says. "We learned about time delay in gene regulatory networks, how signals propagate through colonies, and how interactions come together to synchronize behavior between cells." And with essentially only two genes at the heart of the synchronization mechanism, the system is a great demonstration of how small systems can generate very complex behavior. "It showed that you don't need a lot of genes in a network to get very interesting and rich dynamics, where all kinds of spectacular things can happen," he says.

"The complexity of the system is astonishing," says **Martin Fussenegger, PhD**, professor of biosystems science and engineering at the Swiss Federal Institute of Technology Zurich in Basel, Switzerland, who wrote an accompanying perspective on the study. Not only is the timing mechanism radically different from that of the central pacemaker in the brain, which uses one-way synchronization to control cellular clocks in remote tissue, but the cells manage to stay synchronized even while in constant motion and dividing every 20 minutes.

The bacteria can also be programmed to change their synchronized blinking rate in response to environmental triggers. This ability could lead to applications such as super-sensitive bacterial sensors that would flash more quickly in the presence of environmental contaminants, says **James Anderson, PhD**, program director for the Center for Bioinformatics and Computational Biology at the National Institute of General Medical Sciences within the National Institutes of Health.

But the immediate use of the work is more basic, Anderson points out. These researchers created computational models of the synchronization to drive both *in silico* and *in vitro* experiments of the synthetic biology, which in turn help refine the computational models even further. "What the synthetic biologists are doing now is helping us understand how the natural traits actually work at the same time that they're creating synthetic ones."

—By Regina Nuzzo, PhD

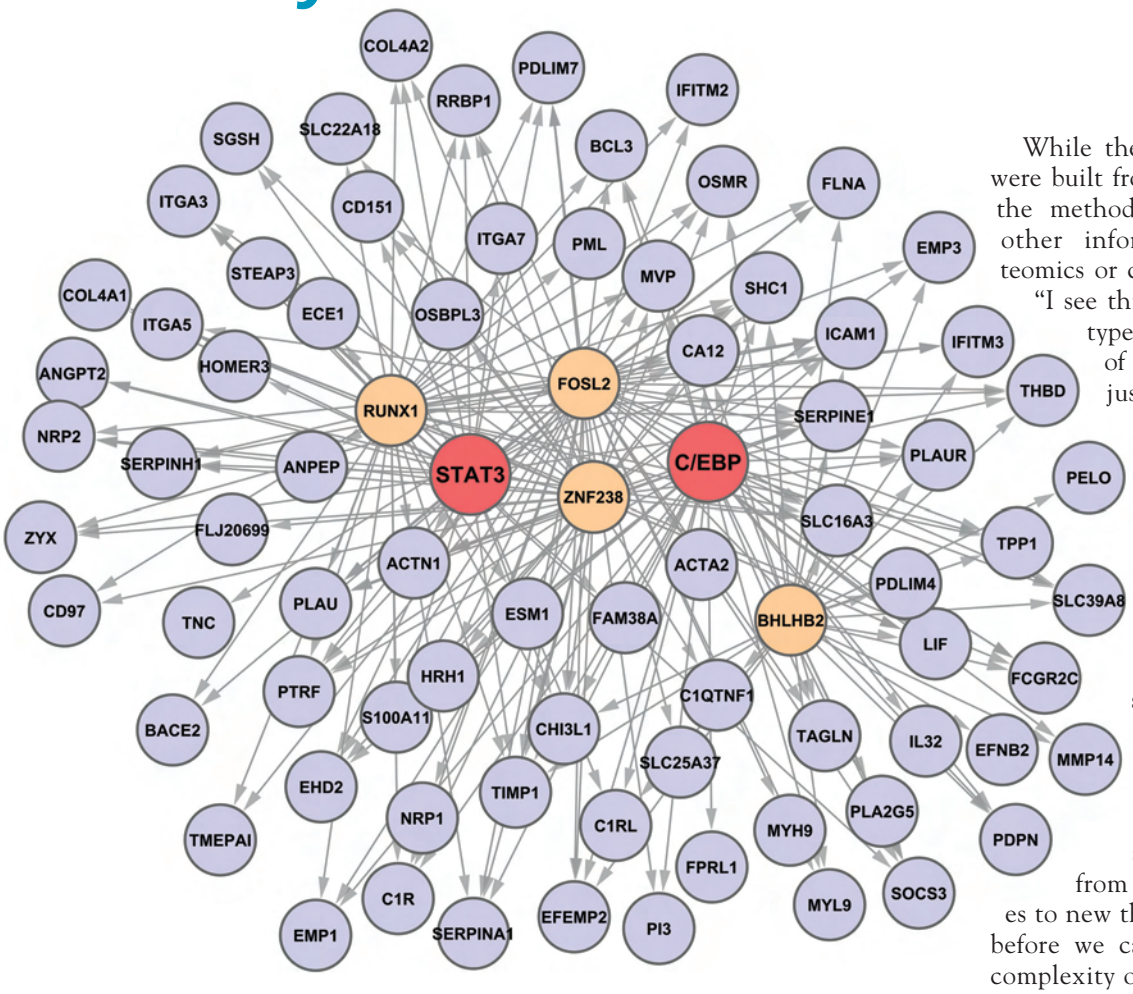


A team at UCSD built a network of genes and proteins in* E. coli *that acts as a molecular clock and can be synchronized across cells. A positive-feedback loop (a) triggers expression of a quorum-sensing gene that produces AHL, an intercellular communication molecule. At the same time, a negative-feedback loop (b) triggers a protein that degrades AHL and a green fluorescent protein (c) makes the waves of activity visible. The dynamic interactions of the positive- and negative-feedback loops produce regular pulses of AHL (d), which act as the metronome in the molecular clock. Since all the cells simultaneously send and receive AHL, they adjust and synchronize their clocks with each other. The result: coordinated fluorescent flashes. Reprinted by permission from MacMillan Publishers, Ltd: from Fussenegger, M, Synchronized Bacterial Clocks,* Nature *463, 301-302 (2010).

Reverse-Engineering Transcriptional Networks

A cell may change states several times in its lifetime—from a stem cell to a specialized cell, for example, or from a normal cell to a cancerous one. Each time this happens, a veritable army of genes must be raised to do the tasks needed by the new cell type. Now, researchers have successfully used computational approaches to identify the "master regulators" that, like generals, control the transformation of benign brain cells into the malignancies that cause high grade glioma, one of the most aggressive forms of brain cancer. The computational findings were then confirmed experimentally.

The work, which was published in *Nature* in February 2010, demonstrates the value that can come from reverse engineering molecular interaction networks for specific cell types. Coauthor **Andrea Califano, PhD**, professor of bioinformatics at Columbia University and director of the Center for the Multiscale Analysis of Genetic Networks (MAGNet), hopes to apply these methods to other questions of cellular transformation and development, particularly those relevant to dis-



This transcriptional network of high-grade glioma cells shows the two master regulators in red and other significant transcription factors in orange. Together, these transcription factors control about 80 percent of an HGG tumor's signature. Image courtesy of Columbia University, Califano Lab. Reprinted with permission from MacMillan Publishers, Ltd.: Carro, M.S., et al., The transcriptional network for mesenchymal transformation of brain tumours,* Nature *463, 318-325 (21 January 2010).

ease states such as cancer. "We can now ask what are the genes that control an arbitrary transformation," he says.

For a healthy cell to become the beginnings of a high-grade glioma (HGG) tumor, it needs to express a large number of genes that otherwise would never be activated. To find the key genes that produce that altered gene expression state, Califano's team first mapped out the regulatory logic of the most aggressive type of HGG cells using an information theory algorithm called ARACNE. The method can reconstruct regulatory networks from gene expression profiles of particular cell populations, even pruning out indirect interactions to determine which genes directly control others. Next, the researchers looked for genes in this network that were part of the tumor's signature – those that are highly expressed in HGG cells but not in normal brain cells. A handful of transcription factors emerged that together control about 80 percent of the characteristic genes. Two in particular, STAT3 and C/EBP, appeared to hierarchically control the others, even though they are expressed at levels so small they do not appear in the signature.

Further experiments, done with brain tumor experimentalist **Antonio Iavarone, MD**, verified the model, showing that activating the two genes simultaneously in neural cells causes the shift to a tumor-like cell. Likewise, silencing the genes together eliminated the malignant phenotype.

While the networks in this study were built from gene expression data, the method could also work with other information, such as proteomics or chromatin structure data. "I see this work as being a prototype of the power of this type of approach, but it's really just the beginning," says **Howard Fine, MD**, chief of the Neuro-Oncology branch at the National Cancer Institute.

Fine is also hopeful that the results of this work could lead to glioma treatments. "They've identified one small module within this very complex signaling network that is a cancer cell," he says. "This says to us, we might be able to translate findings from these kinds of approaches to new therapies for patients well before we can fully understand the complexity of the tumor cell."

— By Beth Skwarecki

Research Reproducibility from MSWord

A particular mashup of data and tools produces the unique results found in each computational biology publication. Now, researchers have developed a model system that gives readers—especially those lacking programming skills—the tools, data, and parameters they'd need to reproduce those results. Dubbed a "reproducible research system" (RRS), it lets the reader replicate original computational research directly from a Microsoft Word document.

"This effort was meant to show that the technology exists to make research reproducible by the non-programming user," says **Jill Mesirov, PhD**, director of computational biology and bioinformatics at the Broad Institute of the Massachusetts Institute of Technology. The work was described in a policy forum in *Science* in January 2010.

Often, to reproduce a computational biology research result, one must contact the original researcher to request the data and tools. Even then, the precise steps taken might be lost or unrecoverable. People have been struggling with this problem for more than twenty years, and several reproducible genomic analysis platform that provides access to more than 100 tools for gene expression analysis, proteomics, SNP analysis and common data processing tasks. In GenePattern, users' sessions can be captured and replayed. "The idea was to take the captured user session in GenePattern—with all the http://genepatternwordaddin.codeplex.com/) and to develop similar systems for other tools. "It's not one-size-fits-all," Mesirov says. "This is not about GenePattern or even this instantiation of reproducible research. It's about the need and the fact that you want reproducible research accessible to people

In this screenshot of an MS Word document, the user can open a GenePattern pipeline to reproduce the research. Courtesy of Jill Mesirov.

research systems already exist, but they are not widely used and require the user to do things that are "very much like coding," Mesirov says.

The RRS concept, as proposed by Mesirov and her colleagues, consists of two parts: an environment for doing the computational work that tracks the data, analyses and results and then packages them for redistribution; and a publisher, such as a standard word-processing software.

As an example system, Mesirov and her colleagues used GenePattern, a parameters and datasets—and embed that in Microsoft Word," Mesirov says. Luckily, from a technical point of view, the webservices architecture of GenePattern and the XML capabilities of Word "kind of meshed," she says.

With funding from Microsoft, the GenePattern RRS was developed. A user can link text, tables and figures to previously executed GenePattern pipelines. And readers can open up those pipelines from the document.

Mesirov invites people to try the GenePattern RRS (available online at who don't write code."

It's an exciting development, says **Kevin Coombes, PhD**, associate professor of biostatistics and applied mathematics at the University of Texas M. D. Anderson Cancer Center, where masters students are routinely trained to use a reproducible research system called Sweave. MSWord is already on peoples' computers, so the Genepattern RRS is potentially more useful than systems like Sweave that require some programming. At the same time, he says, there are sociological hurdles to

adopting reproducible research systems. "The software to unite these things is necessary to do reproducible research, but not sufficient. You have to get people to buy into it."

—By Katharine Miller

Decoding the Histone

To fit inside the cell nucleus, DNA molecules wrap around tiny protein spindles known as histones. These histones carry an intriguing biochemical code that helps decide a cell's destiny—whether it turns into a neuron or a lymphocyte, or turns cancerous, for instance. Decoding the so-called histone code is now faster and easier, thanks to a new system that combines innovative chromatographic techniques with advanced computer algorithms.

"What previously took a year, we can now do in a three-hour run, and get better results" says **Benjamin Garcia, PhD**, assistant professor of molecular biology at Princeton University, who co-led the effort with **Christodoulos Floudas, PhD**, of the chemical engineering department. The study appeared in the October 2009 issue of *Molecular & Cellular Proteomics*. The new system could advance our understanding of cell differentiation, stem cells, cancer, and other key problems in biology.

Each histone's tail region typically sports several chemical modifications such as methylations, acetylations, or phosphorylations. Individual modifications are known to activate or silence nearby genes, but their net effect—the histone code—remains unknown. This is partly because distinguishing the various histone "forms"—each carrying a distinct pattern of modifications—in a sample is very tricky; despite vastly different biological effects, they have very similar mass and structure.

Spectrometric heat map showing relative amounts of various modified forms measured from a human histone sample. The vertical axis indicates the chromatographic separation time, and the horizontal axis shows the mass/charge ratio of ionized histone forms. The relative amount of each form is color-coded in ascending value from blue to red. Forms separate out in 2D by degree of acetylations (0ac, 1ac, ...) and methylations (0me, 1me, ...), and then by the positions of these modifications. The new method identified more than 200 distinct histone forms from the sample, including ones never before associated with human cells. The method is sensitive enough to distinguish between an acetylation and a trimethylation—two modifications that differ in mass by only a few parts per million. Reprinted from Young, N.L. et al., High throughput characterization of combinatorial histone codes, **Molecular & Cellular Proteomics***, 8(10):2266 - 2284 (2009).*

The new system tackles this with an advanced chromatography process that induces different histone forms in a sample to separate out over a remarkably short 2 to 3 hour period. The emerging histone molecules are then analyzed by tandem mass spectrometry. A typical histone sample might yield thousands of spectra, each carrying contributions from one to three histone forms. To unscramble this, a computer algorithm finds the optimal mix of forms that best matches each spectrum. Combining these results from all the spectra yields an accurate tally of the identities and relative amounts of the forms present in the sample, says Floudas.

The approach successfully identified nearly 200 distinct forms in a histone sample, including some never before seen in human cells. It is sensitive enough to distinguish between modifications with nearly equal masses; indeed, it even teases apart forms that differ merely by the position of a single modification.

"If you want to characterize histone modifications on a large scale, and do it very quickly, this is the way to do it," says University of Wisconsin chemist **Joshua Coon, PhD**. The method is an important technical and methodological advance, agrees **Michael Washburn, PhD**, of the Stowers Institute for Medical Research in Kansas City, Missouri. Washburn cautions, however, that the method will have a real impact only if other researchers succeed in implementing it. Due to their complexity, proteomics techniques are hard to replicate, he notes.

Next, Garcia says, his team will use the approach to unravel the histone code governing cellular phenomena such as stem cell differentiation and cancer. "We've shown we can measure modified histone forms, but there's so much to do now," says Garcia. "This is really the beginning of some true biological breakthroughs."

—By Chandra Shekhar, PhD □

MORE THAN FATE:

Computation Addresses Hot Topics in Stem Cell Research

By Katharine Miller

To the casual observer, stem cells offer the almost magical promise of—Voila!—turning into exactly the kind of cell needed to repair an injured spinal cord or replace a damaged organ. And despite the political issues that swirl around the topic, new research findings fuel the public's hope that the stem cell miracle is right around the corner. Since 2005, scientists have gotten better and better at converting adult skin cells into pluripotent stem cells capable of becoming any cell-type in the body.

But beneath these exciting results lies a far more subtle truth: Although researchers can produce desired cell lines in the lab, they don't always understand the underlying mechanisms. "It works, but we don't necessarily know why it works," says **Ingo Roeder, PhD**, group leader at the Institute for Medical Informatics, Statistics and Epidemiology at the University of Leipzig in Germany.

Stem cells are complex creatures, responding to external and internal cues with an array of cellular changes, including alterations in gene expression, DNA methylation, alternative splicing, microRNA (miRNA) expression, and post-translational modification of proteins. Researchers need to understand these intricacies before they can control stem cells for clinical purposes. But as researchers start to track all of these changes in cells over time, the vast quantity of accumulating data overwhelms the human mind. Researchers can even lose track of what hypotheses they are testing. The only way to make sense of it all is with computation.

"Computation allows you to distinguish between hypotheses in systems where we don't always have all the information we need," says **Peter Zandstra, PhD**, professor of biomaterials and bioengineering at the University of Toronto in Canada.

Above: Differentiated cells (stained red) have developed to surround the tightly packed colony of smaller pluripotent cells (stained green). Photo courtesy of John Butler, Dr. Jeanne Lawrence, Dept of Cell Biology, UMASS Medical School.

The scenario runs something like this: As experimental results accumulate, stem cell researchers start developing theories about why stem cells do what they do. Computational biologists then develop computer or mathematical models to examine the theories in a rigorous way, to guide further experiments. "That's in my mind where a lot

of the power is," Zandstra says.

Using computational models, researchers are gaining traction toward understanding what makes a stem cell a stem cell; how gene expression drives stem cell differentiation; why studying stem cell heterogeneity is important; and, ultimately, how stem cells control their fate.

WHO AM I:
The Stem Cell Identity Crisis

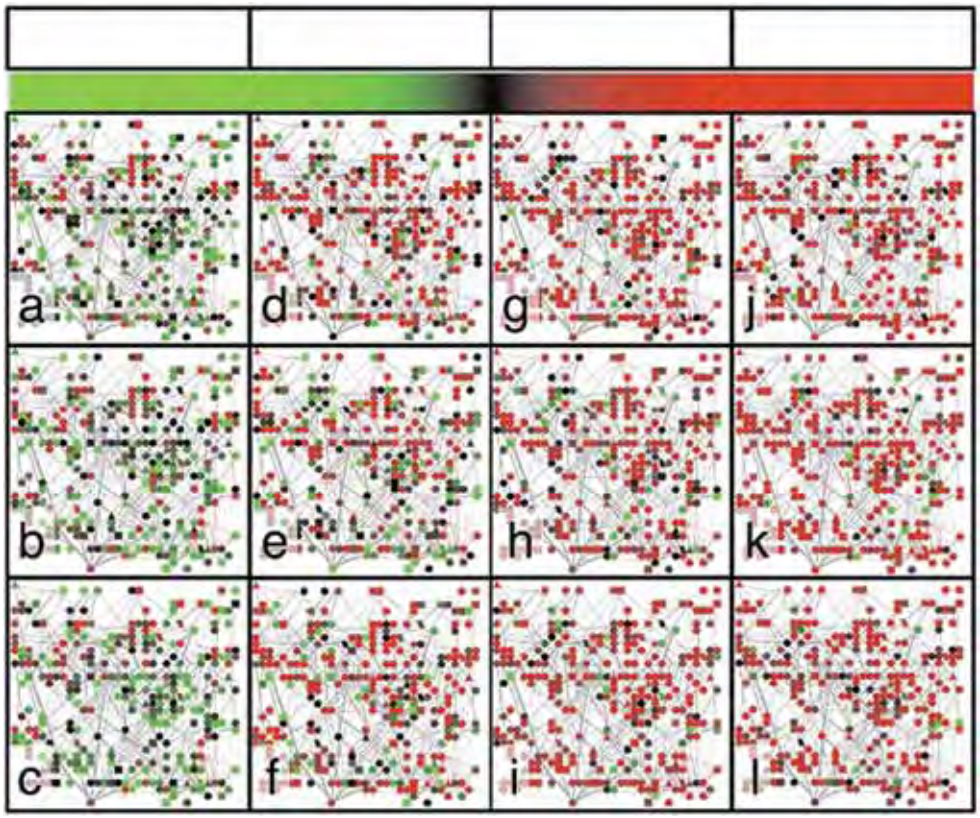
Stem cells seem to know who they are, but it can be hard for humans to tell them apart. Stem cells come in two general types: pluripotent stem cells that can give rise to all cell types in the body, and multipotent stem cells (such as those in bone marrow or the brain) that have a smaller repertoire of options. There are also adult cells that have been induced to resemble stem cells; and cancerous cells that exhibit stem-like traits. And there are many shades of gray in between—stem cells in the process of differentiating; or adult cells in the process of reverting into stem cells. So when stem cell researchers work with cell cultures, it's not always an easy matter to determine what kind of cells lie before them. They need a way to determine whether they've successfully pushed stem cells to differentiate or induced differentiated cells to become pluri- or multipotent. In essence, they need a test for stem cell status.

Jeanne Loring, PhD, professor of developmental neurobiology and director of the Center for Regenerative Medicine at The Scripps Research Institute in La Jolla, California, is taking a bioinformatics approach to address this problem. She and her colleagues built an enormous database (which they call "the stem cell matrix") that contains data on gene expression, microRNA expression, DNA sequencing, and epigenetics (DNA methylation), among other things. Using the data from 22 samples and a machine-learning algorithm, they taught a computer how to identify stem cells. When applied to 66 test samples, the algorithm clearly separated pluripotent stem cells into a class by themselves. "They are a different category from all other cells—as distinguishable as white rocks from black rocks," Loring says.

Since the work was published in *Nature* in 2008, the clusters have held up using additional data. Loring's team has now applied their analysis to more than 500 samples—with data on expression of 40,000 genes per sample and 37,000 DNA methylation sites per sample. "The more information we get, the simpler the answer is," Loring says. "We get rid of the noise."

Loring's lab plans to offer a simple method to help scientists determine what cell type they are looking at right now. "Researchers will send us a sam-

> "Computation allows you to distinguish between hypotheses in systems where we don't always have all the information we need," says Peter Zandstra.



Using a program called Matisse, developed by researchers in Israel, Loring and her colleagues superimposed gene expression data from her lab's stem cell matrix on known protein-protein interactions. This filtered the data down to the key set of interacting proteins that were commonly all up-regulated in pluripotent stem cells—a network they dubbed the PluriNet. This figure displays the activity of the PluriNet in three different samples of (in columns from left to right) four different cell types: nonpluripotent stem cells, tumour-derived (cancerous) pluripotent cells, induced pluripotent stem cells and pluripotent embryonic stem cells. The pluripotent cell lines (right two columns) resemble one another and are quite different from the other two). Reprinted with permission from McMillan Publishers, Ltd, Muller, FJ, et al., Regulatory networks define phenotypic classes of human stem cell lines,* Nature *455, 401-405 (18 September 2008).

ple or gene expression data, and we can tell them which category their cells are most like," she says.

THE ATTRACTOR LANDSCAPE:
The Lay of the Land for Stem Cells

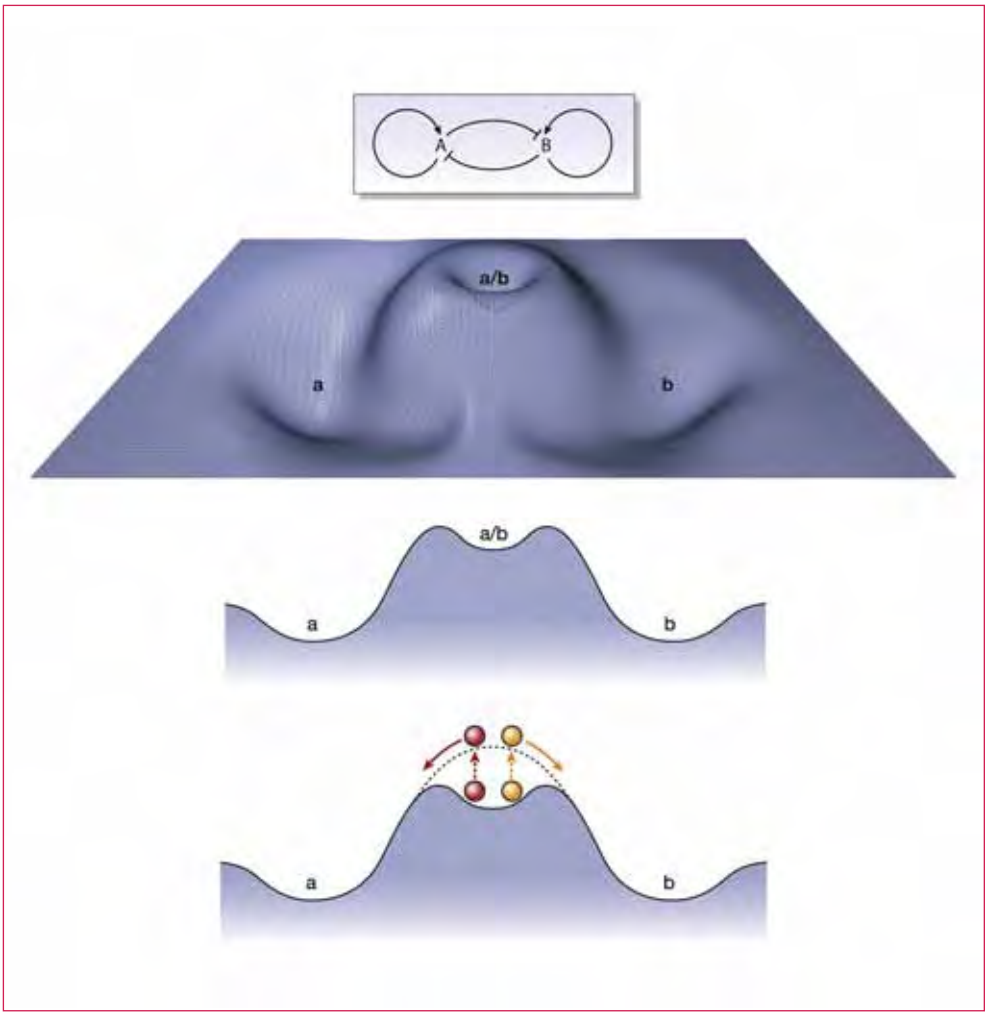
For more than 40 years, cell biologists have described stem cell differentiation in terms of a metaphorical energy landscape. The cell's gene expression state—essentially the transcriptome—can be stable in multiple different combinations. When it finds a stable state, it stays there, like a marble stuck in a valley on the landscape. The starting "well" is the pluripotent stem cell. Certain driving forces then push the cell up and out of those wells and across ridges into new low areas. Called attractors, these low areas represent specific differentiated states, such as neurons or blood cells.

Many computational researchers still consider this depiction useful only as a metaphor. But a few are taking it further. "It's actually a very mathematical thing," says **Sui Huang, PhD**, associate professor of biology at the University of Calgary in Alberta, Canada. Within every cell, there's a network of genes interacting with other genes. These interactions generate gene expression patterns that define the cell's state—i.e., whether it's a stem cell or has differentiated into some other kind of cell.

Some possible states are more likely to be stable than others. For example, if gene A inhibits gene B, then a pattern where A and B are both highly expressed is very unlikely to be stable. So, in theory, if you understood the wiring diagram for all gene interactions, including which genes inhibit or activate which other genes, you could predict the likelihood that the network permits a particular gene expression pattern. And, says Huang, "that probability would give you the derivation of the landscape." There is a caveat, Huang says. "From a physics point of view, it's not really an energy landscape because living systems are not equilibrium systems, but the intuition of landscapes is wonderful."

Of course, to compute the landscape, Huang says, you would have to know the details of the wiring diagram—which genes activate each other

The cell's gene expression state—essentially the transcriptome—can be stable in multiple different combinations. When it finds a stable state, it stays there, like a marble stuck in a valley on the landscape.



In this gene regulatory network, two transcription factors (A and B) activate their own expression and mutually inhibit each other. A mathematical model of this network produces a landscape with three "attractors"—states of gene expression that are stable. One (a/b) in which there is low expression of both A and B, would be a stem cell-like state, whereas the valleys a or b constitute stable states in which A or B are exclusively expressed. If the stem cell's state is somehow destablilized, the upper valley becomes a hilltop (dotted black line) rather than a valley, inducing the cell's state to move toward one of the differentiated (committed) states. If A and B expression fluctuate within the stem cell valley, then those near the lip of the valley will be the first to commit because they are already prone to flow in a particular direction. Reprinted from* Cell Stem Cell, *Vol 4 issue 5, Enver, T., et al., Stem Cell States, Fates, and the Rules of Attraction, pages 387-397 (2009), with permission from Elsevier.

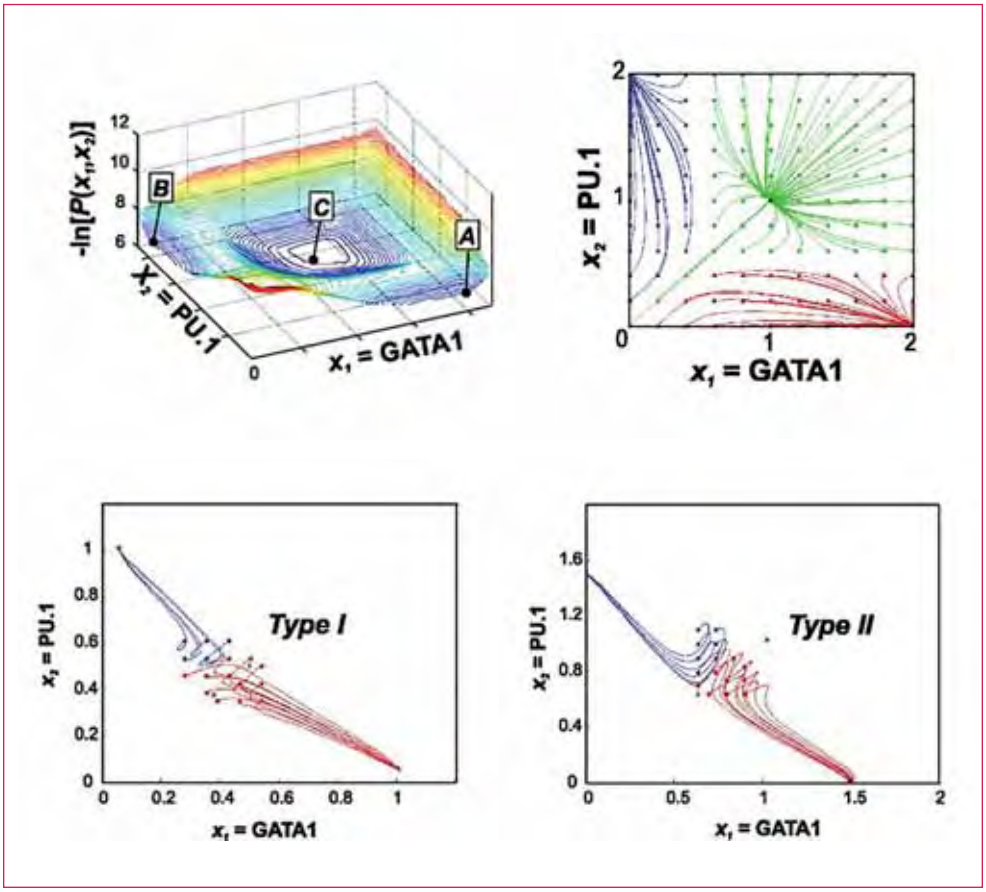
and how. And, he says, "getting that wiring diagram is a big, big step."

So far, Huang and his colleagues have modeled the landscape with a two-gene interaction network. The genes function as a stem cell switch for hematopoietic (blood-forming) stem cells: transcription factor PU.1 drives them to become white blood cells, while GATA1 drives them to become red blood cells. Each transcription factor activates itself and inhibits the other.

For a network with two genes, the third dimension is the elevation, which equals the probability of each possible expression pattern of PU.1 and GATA1. This elevation is very hard to compute even with just two genes, Huang says. "You cannot derive a mathematical equation to give the shape of the landscape," he says. "You have to use brute force."

Looking at more than two genes would require an even greater number of dimensions. "If you have 100 genes then your state space has 100 dimensions and the elevation is the 101st dimension," Huang says. "That's hard to picture, and computing that landscape would be very intense."

But even Huang's two-component model provides useful insights. For example, his team modeled the trajectories that the cell takes from a particular undifferentiated state to the differentiated state. Surprisingly, it's not necessarily a straight line; the levels of PU.1 and GATA1 fluctuate and even loop around before the cell moves toward the attractor. By modeling this, and comparing the *in silico* trajectories to those determined experimentally, the researchers better understand how the two genes are interacting in the cell to decide its fate.



In this illustration of the attractor landscape for a two-gene network, (top left) the blue wells representing hematopoietic stem cell state C as well as the attractor states A and B for erythroid and myeloid (red and white blood cell) cells respectively. At the top right, the state space for the PU.1/GATA1 system shows how the cell state is driven by the network dynamics from various possible starting states (points on the grid) toward the attractors: Blue lines move toward the myeloid state, red toward the erythroid state, and green toward the metastable hematopoietic stem cell state. Adjustments to the model parameters will shift the size of the area represented by each color. According to one theory, differentiation occurs when the multipotent state (green) becomes unstable—essentially disappearing so that the cell must choose between the two possible fates. As shown at bottom, Huang and his colleagues also modeled the trajectories that the cell might take from an initial stem cell state (in the center of the state space) to each corner, given different sets of initial conditions. These trajectories were then compared to actual trajectories observed in the lab. Reprinted from* Developmental Biology*, 305:695-713 (2007), Huang, S., et al., Bifurcation dynamics in lineage-commitment in bipotent progenitor cells, with permission from Elsevier.

INDIVIDUALS MATTER: Understanding Stem Cell Heterogeneity

Much of stem cell biology relies on population statistics—for example, the average gene expression levels of a vast numbers of cells grown in a culture. But stem cells exhibit a surprising degree of individuality even within such cultures. Gene expression varies greatly; cells divide asymmetrically; some cells die out while others reproduce themselves endlessly. Moreover, adding certain chemical cues to stem cells will induce some, but not others, to differentiate.

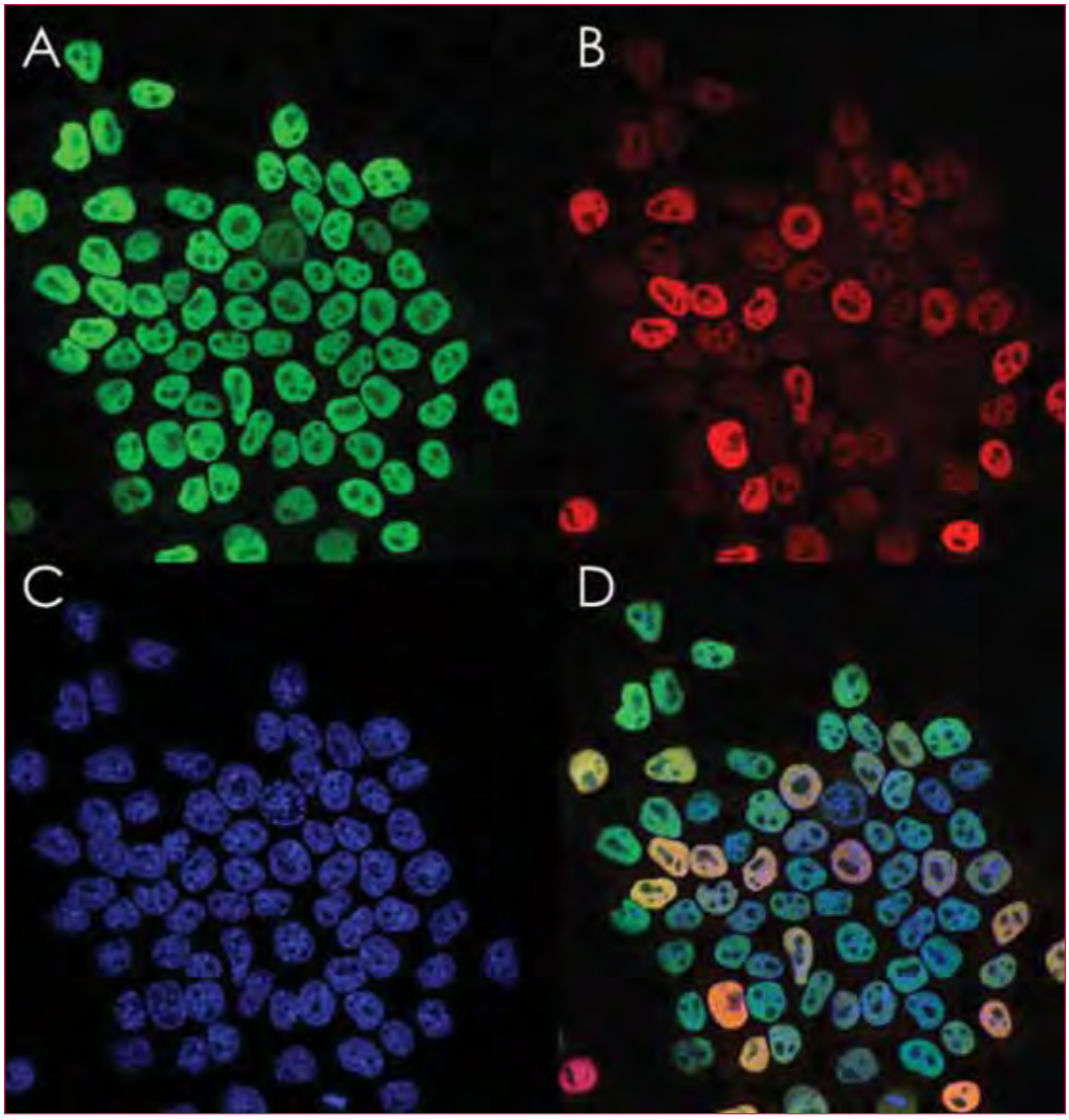
"Outliers matter in biology," Huang says. "Science has a tendency to operate with averages—average populations, average females, average males, but individuals are important. All you need is one cell behaving differently and it has consequences for the organism."

"Outliers matter in biology," says Sui Huang. "All you need is one cell behaving differently and it has consequences for the organism."

Heterogeneity also means that any given attractor is actually a cloud, rather than a point, on the landscape, Huang says. "So we need to have the statistics for thousands of individual cells to get the landscape," he says. "We need to follow individual cells."

To do that, researchers turn to microfluidic devices that can rigorously control the environment surrounding individual cells without washing them away or moving them around. In time-lapse experiments, a camera can capture gene expression levels in individual cells while also tracking cells as they divide, die, differentiate, or remain pluripotent. The data retrieved can help confirm or deny predictions from computational models.

For example, Ingo Roeder used a system of differential equations to model the wide fluctuations observed in Nanog expression in individual mouse embryonic stem (ES) cells. The work (unpublished) suggests two possible explanations: either cells themselves fluctuate between two possible stable states (induced by random perturbations—essentially noise); or the state itself is oscillating (such that the state is never really stable). There is no theoretical way to distinguish between these two scenarios based on average population statistics, Roeder says. Rather, experimentalists will need to monitor temporal changes in Nanog in individual cells over time.



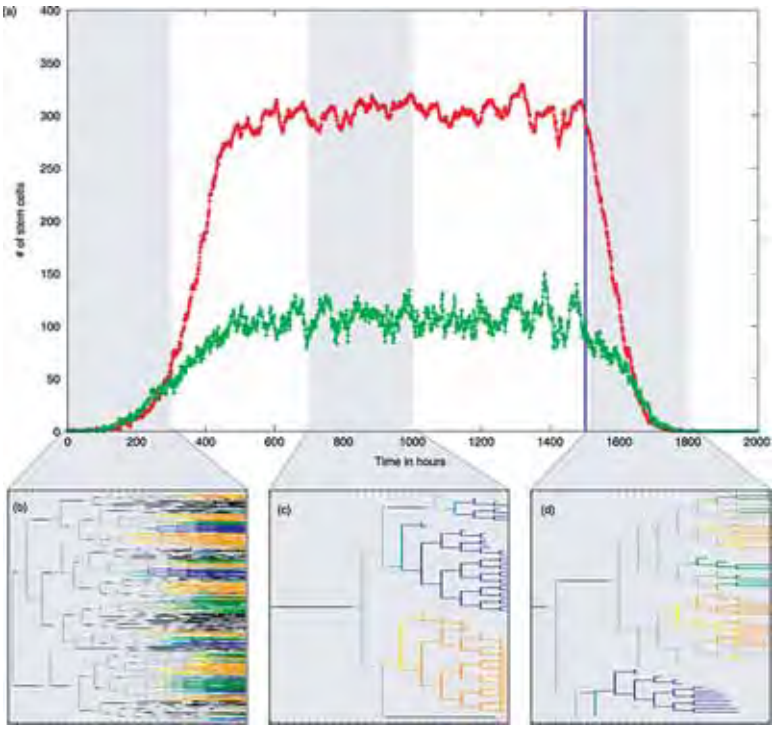
Monitoring individual stem cells shows that they are heterogeneous. Here, the same set of cells is stained for expression of the genes Oct4 (A) and Nanog (B), both known to be important transcription factors for the maintenance of the stem cell state. The DNA was also stained with a blue stain (C) and these three images overlain on one another (D), dramatizing the heterogeneity of the cells. In particular, Nanog expression levels (red) differ substantially within the cells. Image courtesy of Austin Smith. Reprinted with permission from Roeder, I. , and Radtke, F., Stem cell biology meets systems biology, **Development** ***136, 3525-3530 (2009).***

Monitoring individual cells over time can also generate the cells' genealogical trees. But, says Roeder, "There is currently no set way to computationally analyze those genealogical trees." So Roeder decided to get a jump on that problem before collecting data. Using a computer model of hematopoietic stem cell organization, Roeder and his colleagues simulated an array of possible cell genealogical trees *in silico*. The cells can self-renew, die, or differentiate. To be realistic, the simulations include random noise.

The simulations showed that changes in the growth conditions (*in silico*) altered the shape of the genealogical tree. The computer can distinguish proliferating cells from cells in a steady state or in decline, and can recognize asymmetry. Going forward, the goal is to determine whether different cell scenarios generate unique tree "signatures" that a computer can spot. These will have to be validated against the real

"One of the major advantages of computer modeling is that you can try lots of different scenarios and then narrow down the possibilities for explaining certain behavior," Ingo Roeder says.

genealogies of cells. "With the experimental data, we will see the trees and use the simulations to estimate back to learn the underlying mechanism," Roeder says.



This chart shows the population dynamics of proliferating (red) and quiescent (green) cells as they move through three different scenarios: expansion, homeostasis, and terminal differentiation (i.e., dying off over time). Characteristic genealogies for each scenario are shown below. To the human eye, these genealogies are clearly different. Roeder and his colleagues sought to train a computer to spot distinct tree signatures. Reprinted with permission from John Wiley & Co., Glauche, I., et al., A novel view on stem cell development: analysing the shape of cellular genealogies,* Cell Proliferation *42, 248-263 (2009).

For stem cell research, the benefits of this work may still be a ways off. But ultimately, Roeder says, "One of the major advantages of computer modeling is that you can try lots of different scenarios and then narrow down the possibilities for explaining certain behavior."

IT'S FATE:
Switches and Beyond

Many computer models of stem cells focus on the question of fate: How does the stem cell decide whether to remain pluripotent or differentiate into another kind of cell? "The most useful models recognize that decisions inside the stem cell are collective decisions of networks of interacting biological molecules," says **David Schaffer, PhD**, professor of chemical engineering, bioengineering, and neuroscience at the University of California, Berkeley. So modelers build networks from what they know about interactions in the stem cell and then set the models in motion to see what happens.

"A model is really a statement of hypothesis that aggregates our knowledge of how the system behaves," Schaffer says. "You're either right or not right. And when you compare the model predictions to experimental data, if you're not right, then you know you're missing something. That then motivates experiments to determine what you don't understand."

It's often an iterative process, Schaffer says. "The model summarizes what we know but also guides experimentation so we can best learn more about the system experimentally."

Until fairly recently, it was difficult to construct models of stem cells because there weren't enough data available. But that is changing, Schaffer says. Now, to build a model of a network inside a stem cell, one can start with information about networks in other systems (such as in yeast, which is well understood), hypothesize the interactions that might be occurring, and then mesh that with all kinds of data being collected from stem cell systems, including protein expression data, miRNA expression levels, protein post-translational modifications, protein phosphorylation, signal transduction, and, increasingly, any or all of these types of data as a function of time.

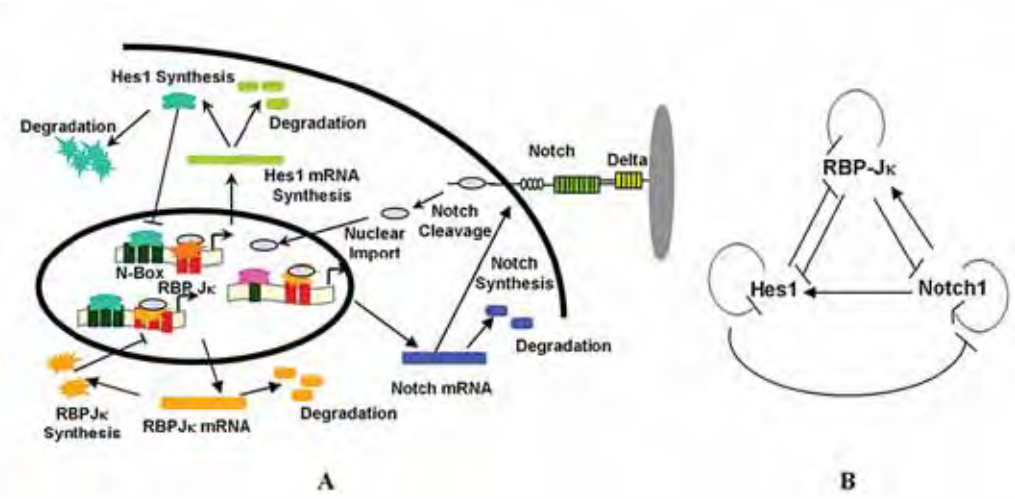
In a 2004 paper, Schaffer and his colleagues created a model to explore the dynamic behavior of the sonic hedgehog (Shh) gene regulatory network—a network known to function as a cell fate switch in certain contexts. The model used differential equations to track the rates of change in concentrations of network participants as well as the rates of protein synthesis and degradation. The model showed that the system functioned as a digital all-or-nothing switch that is not easily reversed.

In 2009, Schaffer's team also modeled the Notch signaling pathway, known for its involvement in cell fate decisions during development and adulthood. They found that the Notch system also acts as a bistable switch, but they identified a factor that could change the system into an oscillator. Thus, the Notch system can be adjusted to exhibit different behaviors depending on the context. The work is currently being validated experimentally.

But these models do not necessarily represent the absolute truth. As so often happens in science, new information can come to light, requiring changes in the model. In a 2006 paper, **Carsten Peterson, PhD**, professor of biological physics at Lund University in Sweden, and his colleagues modeled three key transcription factors involved in embryonic stem (ES) cell self-renewal—Nanog, Sox2, and Oct4. The model showed that the three could—on their own—function as a bistable switch to maintain stem cell pluripotency.

Now, it appears that Oct4 activation is just an early step in the process, triggering the opening up of the chromatin region around Nanog and several newly identified transcription factors that play a key role in the switch. "On the very top you have these epigenetic things happening," Peterson says. "That adds another dimension to the whole modeling perspective."

Besides extending the model to



In this schematic of the Notch1-RBP-Jk-Hes1 signaling network (left), each arrow represents a term or event in the differential equation model including transcription, translation, mRNA and protein degradation, nuclear import, TF binding, receptor-ligand binding and receptor processing. At right, a schematic of the positive and negative feedback loops of the Notch1-RBP-Jk-Hes1 network. (-\|) represents repression and (-.) represents activation of target genes. Reprinted from Agrawal, S., et al., Computational Models of the Notch Network Elucidate Mechanisms of Context-dependent Signaling,* PLoS Computational Biology, *May 2009.

include epigenetics, Peterson says, they are also trying to include mechanical interactions that play a role in ES cell differentiation and migration. "It turns out that mechanics are not negligible," Peterson says.

When ES cells are inside the egg, some start to change into endoderm—the cells that form an outer shell around the inner ES cells. But initially, the cells that are changing are "like salt and pepper," Peterson says. "They are all over the place." To understand how the endoderm develops, he added different adhesion properties to the two types of cells in his model. And the computational result matched the experimentally observations: the endoderm cells move out, leaving the stem cells inside. Friction alone, without any influence from chemical cues, was enough to properly separate the two different types of cells. The next step will be to determine whether ES and endoderm cells actually exhibit different adhesion qualities and what genes cause those traits to develop.

Peterson and his colleagues are also modeling aspects of the hematopoietic stem cell system and finding interesting features that resemble locks. After the straightforward PU.1/GATA1 switch has been activated (described above), the hematopoietic system gets more complex. "Here's where a mathematical model can help," says Peterson. His model, published in 2009 in *PloS Computational Biology*, suggests that downstream genetic players interact with one another and also send feedback to the PU.1/GATA1 switch, preventing changes in previously made decisions. "There have to be locks on the way down to make sure it's irreversible," he says. And it's crucial to understand that irreversibility if researchers want to induce hematopoietic stem cells from differentiated blood cells.

NEIGHBORS MATTER: Modeling Stem Cell Interactions

Many models of stem cell switches look at the circuitry inside the switch without considering what threw the switch in the first place. But stem cells' fate decisions depend, at least in part, on changes in the environment. Zandstra and his colleagues are modeling one key environmental component—cell-cell interactions—within the hematopoietic system.

The hematopoietic system is quite remarkable. Every day, hematopoietic stem cells in the bone marrow produce tens of millions of red blood cells as well as an appropriate number of white blood cells and platelets—all the different cellular components of blood. To be so reliable day in and day out, researchers believe the system must—at least in part—be tightly controlled by soluble factors secreted by blood cells. Because the process is poorly understood, researchers have a hard time growing hematopoietic cells in culture—a prerequisite to further research.

In a 2009 paper, Zandstra's team made an initial foray into modeling how cell-to-cell interactions control hematopoietic self-renewal and differentiation. "We've built theoretical models of feedback systems where stem cells give rise to progeny through a series of fate decisions," he says. At this point, he says, "We're starting to understand the structure and connectivity of the cell-to-cell networks and what determines whether the stem cell population proliferates or differentiates."

In addition, Zandstra's team developed a way to test the model in a cell culture system by removing different cell types along the way—cutting out various feedback loops. "This has been very fruitful," he says. "By understanding the intercellular networks and controlling them, we can

"We're starting to understand the structure and connectivity of the cell-to-cell networks and what determines whether the stem cell population proliferates or differentiates," Zandstra says.

grow these cells far better than you could before."

The underlying principles in Zandstra's model should be applicable to stem cell systems beyond blood.

PREPARING FOR THE CLINIC

Although computational modelling of stem cells might not directly lead to therapies that treat Parkinson's disease or Alzheimer's, Zandstra says, the models help weed through potential solutions to find those with the greatest impact. By understanding how stem cells make decisions, we gain the ability to control those decisions. "You can't get the clinical outcomes without the increased control," he says.

> "Once you have good models of how cells are maintained as well as transition or differentiate, you can start to think about how the various parts of the network are druggable," Schaffer says.

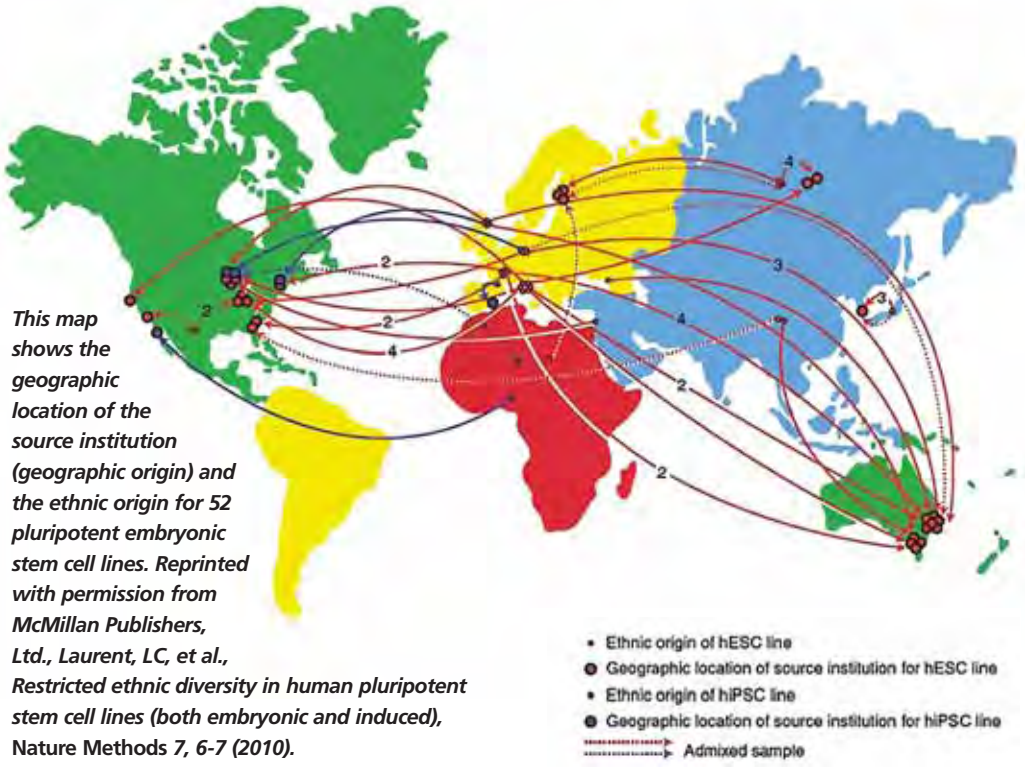
Schaffer agrees. "I really view my job as measure, model, manipulate," he says. "Once you have good models of how cells are maintained as well as transition or differentiate, you can start to think about how the various parts of the network are druggable." □

STEM CELL DIVERSITY AND DRUG TESTING

Someday soon, pharmaceutical companies will be converting stem cells into liver cells they can use for testing drug toxicity, says Loring. "It's the wave of the future. There's no better way to test a drug on liver than to grow liver cells in a dish and dump the drug on them." This approach could help drug companies better understand variations in the way people react to drugs. But there's a problem looming, Loring says: Almost all of the preclinical work with embryonic stem (ES) cells uses Caucasian cell lines.

In a 2009 paper published in *Nature Methods*, Loring and her colleagues used a Bayesian analysis of ES cell genotypes to determine the ethnic background of existing ES cell lines. What they found—the dominance of Caucasian cell lines—springs from the cells' source in in vitro fertilization (IVF) clinics. "Embryonic stem cells made from embryos discarded at IVF clinics are almost all Caucasian and East Asian," Loring says. There are almost no African stem cells, she says. "So our soapbox is that if the pharmaceutical industry is going to start using pluripotent stem cells, it needs to incorporate diversity."

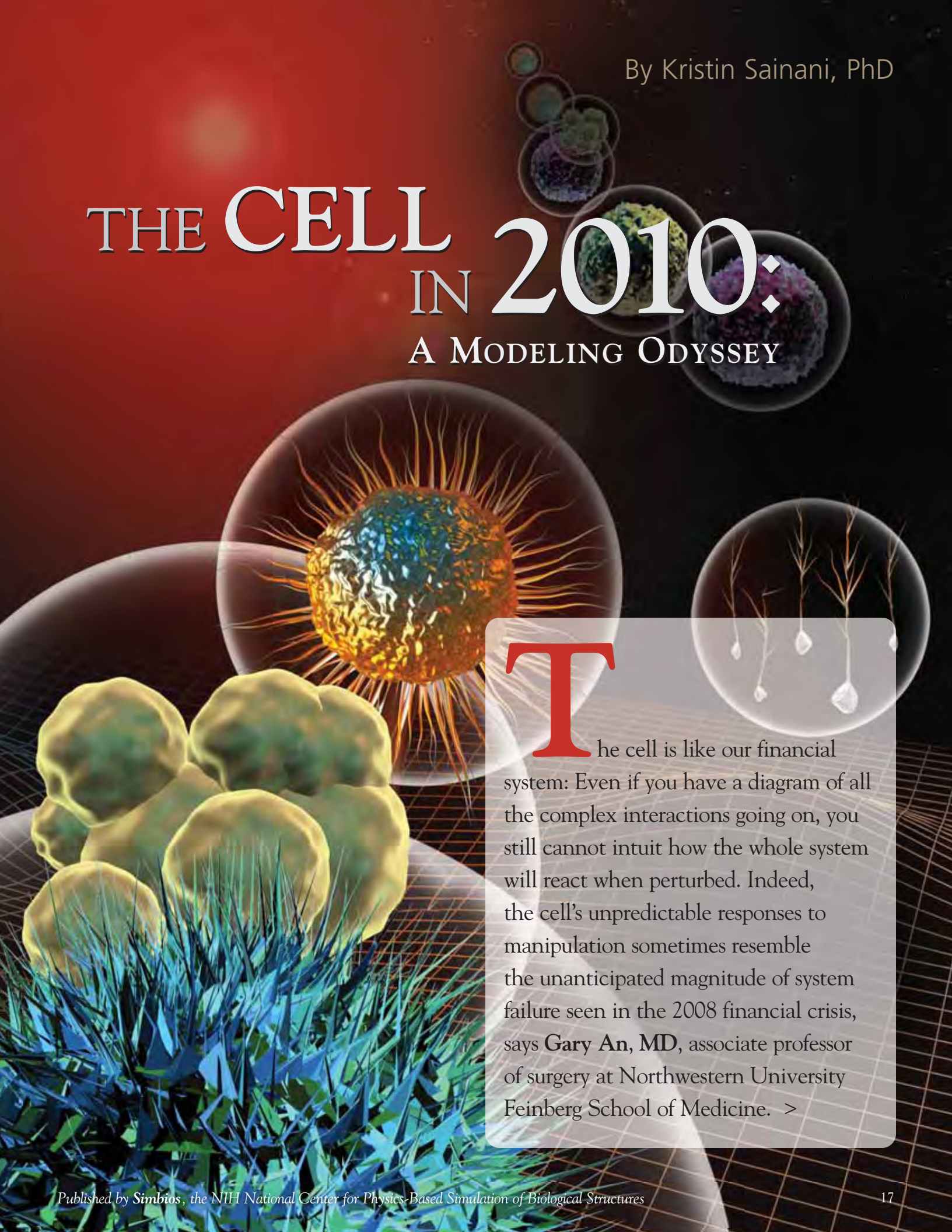
"This is more important than anything I've ever done," Loring says. If all the preclinical work is done on Caucasian cell lines, then the pharmaceutical companies might release drugs that are toxic to some people and don't work on others. Loring hopes her paper puts some pressure on pharmaceutical companies. "They need early assays for toxicity that capture the diversity of people." The *Nature Methods* paper took a first step in that direction, publishing the creation of an induced stem cell line from skin cells of a Yoruba (Nigerian) individual.



This map shows the geographic location of the source institution (geographic origin) and the ethnic origin for 52 pluripotent embryonic stem cell lines. Reprinted with permission from McMillan Publishers, Ltd., Laurent, LC, et al., Restricted ethnic diversity in human pluripotent stem cell lines (both embryonic and induced),* Nature Methods *7, 6-7 (2010).

By Kristin Sainani, PhD

THE CELL IN 2010:
A MODELING ODYSSEY



The cell is like our financial system: Even if you have a diagram of all the complex interactions going on, you still cannot intuit how the whole system will react when perturbed. Indeed, the cell's unpredictable responses to manipulation sometimes resemble the unanticipated magnitude of system failure seen in the 2008 financial crisis, says **Gary An, MD**, associate professor of surgery at Northwestern University Feinberg School of Medicine. >

With hundreds of trillions of atoms, thousands of proteins, and a host of tiny organs, motors, and highways that often interact in non-linear ways, the cell is a rich target for computational modeling. But modelers and cell biologists haven't traditionally worked together. "In the past I think a lot of really interesting mathematical modeling was going on, but I'm not sure how closely tied it was to the biologists' consciousness," says **Steven Altschuler, PhD**, associate professor of pharmacology at Southwestern Medical School.

This is slowly changing. "Now is a time when both sides are realizing it's a good thing to get together. And I think a lot of progress is happening," Altschuler says.

Greater integration stands to benefit both cell biology and biomedical modeling alike.

Cell biologists need modeling to understand how genes, proteins, and pathways work together to make the cell go. "To me, it's no longer possible to even imagine thinking about these problems properly without using models as a crutch," says **Ed Munro, PhD**, assistant professor of molecular genetics and cell biology at the University of Washington. "There are simply too many moving parts and too many interactions for your brain to synthesize."

Even with relatively simple models, Munro says his intuition about what will come out of a simulation is wrong much of the time. "I'm often completely surprised," he says. "That tells me that if we're limited to assembling verbal explanations for the things we study, then we're in trouble."

At the same time, modelers need cell biologists. Traditionally, modelers have focused on either the molecular level (genes and proteins) or the macro level (tissues and organisms). But some are arguing that when it comes to multi-scale modeling, it makes the most sense to start in the middle—at the cell level. After all, molecular interactions coalesce at the level of the cell, and tissues are just a bunch of cells acting together.

"When we're thinking about multi-scale systems in biology, many people either start at the very smallish level or they start at the tissue level; I think very few people have thought of the cell as the main point. But the cell is the basic unit of life," says **Jenny Southgate, PhD**, professor of molecular carcinogenesis at the University of York in the United Kingdom.

"When we're thinking about multi-scale systems in biology, many people either start at the very smallish level or they start at the tissue level; I think very few people have thought of the cell as the main point. But the cell is the basic unit of life," says Jenny Southgate.

What follows are examples of how cell-centered models are adding fundamental insights into our understanding of cell behaviors—including how cells divide, eat, sense, move, cooperate, travel, and battle injury—as well as helping modelers bridge from the molecular to the tissue and organism levels. These models range in scale from single-cell to multi-cell, but all have implications for the basic life sciences as well as for diseases, such as cancer, heart disease, and sepsis.

MODELING THE CELL:

Beyond Biochemistry

Modelers have traditionally treated the cell as a bag of chemicals, focusing on signaling networks, such as positive and negative feedback loops. These models have led to important insights. But the biochemistry isn't happening in a vacuum; reactions unfold within, and are influenced by, the cell's heterogeneous physical environment. To truly understand cell behavior, you have to account for the physics and geometry.

"People normally think about biochemical networks and pathways. That's what systems biology is about. But, in addition to that, there's polymer physics, membrane transport, electrophysiology, electrical events, cell mechanics, and the forces in adhesion," says **Leslie M. Loew, PhD**, professor of cell biology and of computer science and engineering at the University of Connecticut Health Center, and one of the creators of Virtual Cell, a well-known cell modeling program (www.vcell.org).

"When people say that they want to model the cell, they're mostly talking about what's happening in time; very few modelers try to think about what's happening in space. And not only space, but also mechanical processes, like forces and movements," says **Alex Mogilner, PhD**, professor of neurobiology, physiology and behavior and of mathematics at the University of California, Davis.

But incorporating space and mechanics is challenging, Mogilner says. Several software programs can model simple diffusion in a relatively nice geometry, but that doesn't capture the reality of the cell. "The inside of the cell

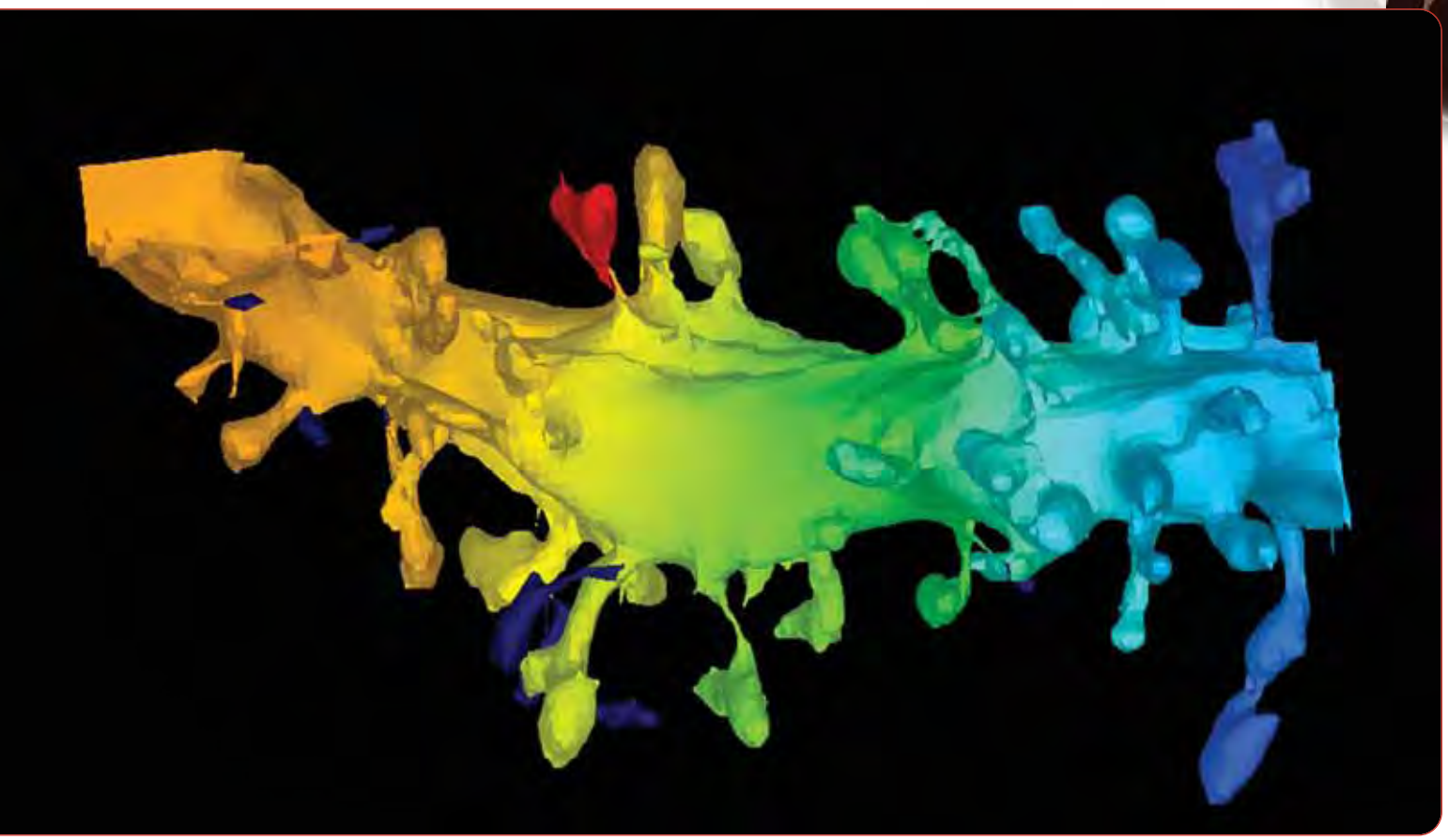
is cluttered with all sorts of debris—cytoskeleton, organelles, and other stuff. In addition to diffusion, there's also directed transport by molecular motors. Plus, diffusion may happen in the bulk of the cytoplasm or in the plane of the membrane. It's very difficult," Mogilner says. Virtual Cell has developed the ability to model diffusion along a membrane and in complex geometries. These capabilities are state of the art.

Spatial modelers make other simplifications as well, such as modeling in two dimensions or treating cells as perfect circles. But some are trying to bridge to 3-D or account for versatile and changing cell shapes. Virtual Cell allows continuum models in 3-D; and another cell modeling program, MCell (www.mcell.psc.edu), can do discrete stochastic simulations in 3-D.

As modelers account for more and more of the cell's physical realities, it seems that, by necessity, models will get more complex and detailed. This isn't always the case, however. Models can range from all-inclusive models that attempt to perfectly mimic the cell to conceptual models that describe the cell in caricature, Mogilner says. Though it may seem that more detail would always be better, in fact there is a tradeoff between complexity and insight. All-inclusive models have a direct correspondence with experiment and tend to be more accessible to biologists and physicians, but they may add little to overall understanding.

"You can take biology, which is a big black box, and turn it into an accurate

"When people say that they want to model the cell, they're mostly talking about what's happening in time; very few modelers try to think about what's happening in space. And not only space, but also mechanical processes, like forces and movements," says Alex Mogilner.



Modeling in Space. Programs like Virtual Cell allow researchers to model the spatial realities of the cell, such as diffusion on a membrane. This Virtual Cell simulation shows lipid signaling and diffusion on a protrusion of membrane on a neural cell (called a "spiny dendrite"). Courtesy of Sherry-Ann Brown, University of Connecticut Health Center; published in: Brown, S., F. Morgan, J. Watras, and L. M. Loew. 2008. Analysis of phosphatidylinositol-4,5-bisphosphate signaling in cerebellar Purkinje spines.* Biophysical Journal *95:1795-1812.

simulation, which in itself has become a big black box," Altschuler says. In contrast, he says, conceptual models "give you a glimpse into something really fundamental."

Conceptual models "give you a glimpse into something really fundamental," Steven Altschuler says.

How a Cell Divides: Harnessing the Wildness of Microtubules

When a cell divides, it assembles an intricate piece of machinery called a "mitotic spindle" that physically separates the chromosomes. Chromosomes are pulled apart by filamentous rods, called microtubules, anchored on either side of the nucleus, at the centrosomes. One of the fundamental questions of mitosis is how this spindle assembles. Mathematical modeling has been instrumental in answering this question because it is difficult to experimentally follow and perturb individual microtubules, Mogilner says.

Microtubules are dynamic polymers that can rapidly shed or add proteins to their unanchored end. It's known that microtubules find the chromosomes through a "search-and-capture" process: they randomly grow and shrink from the centrosomes until, by chance, they encounter a chromosome and hook it.

In an influential paper four years ago, Mogilner and his colleagues showed that the process cannot be completely random. They built a comprehensive model of spindle assembly, including hundreds of microtubules (represented as rods that grow and shrink in different directions) and tens of chromosomes (represented as randomly oriented cylinders dispersed throughout a spherical nucleus). Their simulations showed that a purely random search-and-capture would not be fast enough to assemble the spindle in the 15 to 20 minutes it takes the cell. Instead, a "biased" search-and-capture was required—molecular motors direct microtubules to grow in areas where they are more likely to bump into chromosomes.

In a follow-up paper in *PNAS* in 2009, Mogilner's team ran simulations that probed not only the speed of biased search-and-capture, but also its accuracy. The result: there were errors in a whopping 70 percent of microtubule-chromosome attachments (for example, when a chromosome is captured by only one microtubule or by two microtubules from the same pole). In real life, cell division is highly accurate. So this revealed that the cell must use some kind of error-correction mechanism.

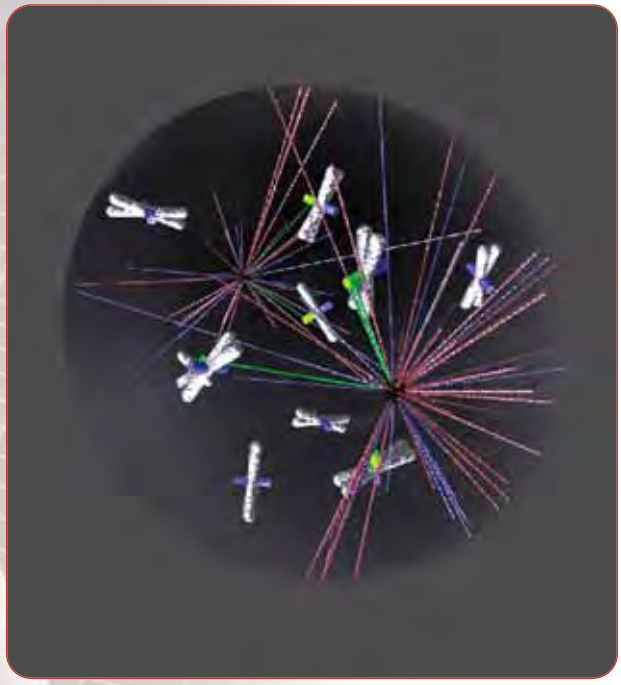
They simulated a number of plausible mechanisms but "so far, what we are finding is almost nothing can explain totally fast and accurate assembly," Mogilner says. Their model provides constraints for researchers exploring alternative error-correction mechanisms, he says.

Once microtubules have accurately captured the chromosomes, they line them up evenly at the equator of the nucleus. What's unclear is how the microtubules, which start at highly varied lengths, manage to even themselves out. "The question is: how do you harness the wildness of the microtubules, which would otherwise be inclined to grow and shorten very randomly and willy-nilly?" says **David Odde, PhD**, professor of biomedical engineering at the University of Minnesota.

In a 2008 paper in *Cell*, Odde and his colleagues used a Monte Carlo simulation to predict that an unidentified molecular motor must regulate microtubule length. Simulations showed that deleting this protein would cause microtubules to grow too long and uneven, and overexpressing it would cause microtubules to grow too short and to cluster near the poles of the nucleus. His graduate student, **Melissa Gardner**, then identified the protein experimentally: kinesin-5, a motor protein not previously recognized as a player in microtubule assembly.

The model shows that kinesin's mode of action is really simple, Odde says. The longer a microtubule becomes, the more places kinesin—which promotes disassembly—can attach to. "It evens the game out. It just keeps penalizing the ones that keep getting out ahead of the others," Odde says.

The finding has implica-



***Search and Capture.** Visualization of a computer simulation of microtubules (growing in blue, shortening in red, captured in green) searching for chromosomes during mitotic spindle assembly. Courtesy of: Raja Paul and Alex Mogilner, University of California, Davis. Reprinted from Paul, R., et al., Computer simulations predict that chromosome movements and rotations accelerate mitotic spindle assembly without compromising accuracy,* **PNAS** *106(37) 15708-15713 (2009).*

tions in cancer, as it means that anti-kinesin drugs—which are already in clinical trials—could help control tumor growth by disrupting a critical step in mitosis.

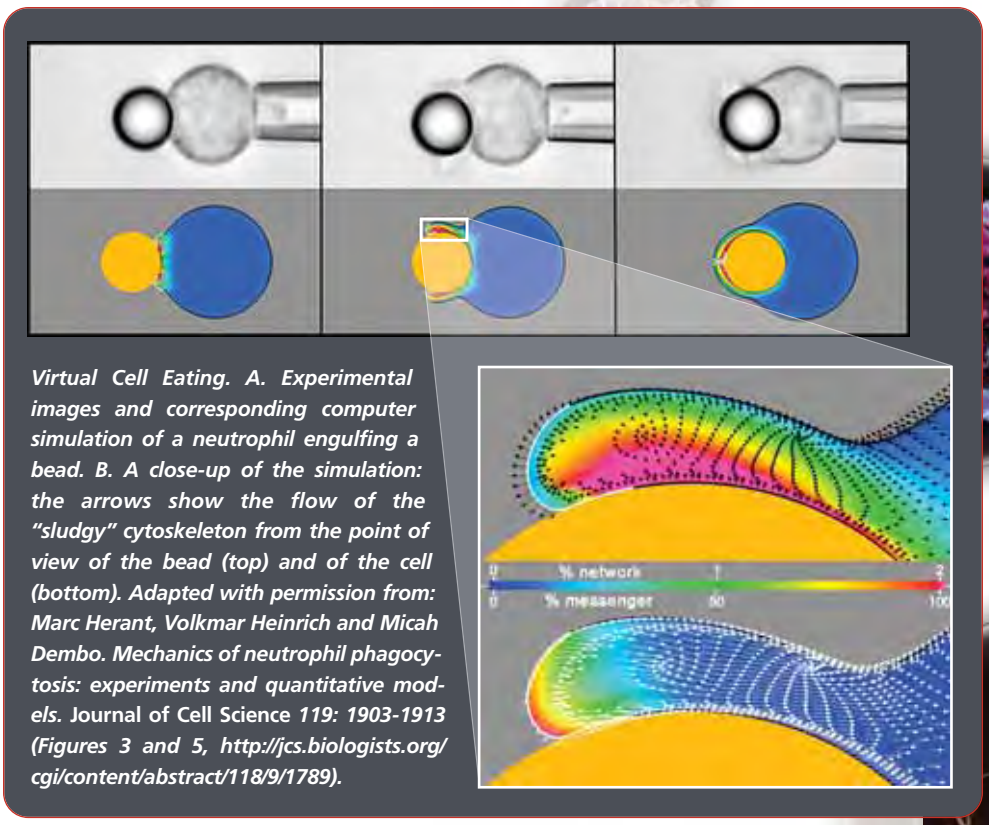
How a Cell Eats: Protruding Hands and Fingers

Single-cell organisms obtain nutrients via a process called cell eating, or phagocytosis. Using its cytoskeleton—dynamic filaments including actin and microtubules—the cell wraps itself around a particle until it's fully engulfed. Cells of the immune system use the same process to destroy bacteria and yeast and to clean up debris. "Without the phagocytosis of yeast, you would be fermented within a day or so," says **Micah Dembo, PhD**, professor of biomedical engineering at Boston University.

"Though the components of cell eating have been well worked out, mechanistic explanations are lacking," Dembo says. "We want to know: what are the forces that the cell is producing? How is the cell pushing? How hard is it pushing? Where is it pushing? Is it pulling? How does it orchestrate its little hands and fingers to do something like phagocytosis?"

Dembo has built a model of phagocytosis for neutrophils (a type of white blood cell) in collaboration with **Volkmar Heinrich, PhD**, an associate professor of biomedical engineering at the University of California, Davis, and **Marc Herant, PhD**, a research assistant professor of biomedical engineering at Boston University. Rather than model the cytoskeleton components as individual proteins or rods, "we believe at its basis, the cytoskeleton is just kind of a gooey glop," Dembo says. "It's got intermediate filaments in there; it's got actin in there; it's got microtubules in there; it's got water; it's got endoplasmic reticulum; it's got big chunks like granules and lysosomes; and the nucleus is a big rock in there. We think of it as a sludge, which, to a good approximation, can be regarded as a creeping fluid." They use a system of partial differential equations to keep track of the forces exerted by and on this viscous fluid as it moves within the cell.

In a paper in the *Journal of Cell Science* in 2006, Dembo's team reported that neutrophils use two key interfacial forces to eat a bead: a protrusive force and an intrusive force. The cytoskeleton and the cell membrane repulse each other (the protrusive force), causing a gap to open between them; as cytoskeleton polymerizes in the gap, this causes fingers of cytoplasm to jet out around the bead. At the same time the cytoskeleton and cell membrane attract each other (the intrusive force), causing cytoskeleton to build up near the membrane; as this excess cytoskeleton depolymerizes, this sucks the bead into the cell.



Virtual Cell Eating. A. Experimental images and corresponding computer simulation of a neutrophil engulfing a bead. B. A close-up of the simulation: the arrows show the flow of the "sludgy" cytoskeleton from the point of view of the bead (top) and of the cell (bottom). Adapted with permission from: Marc Herant, Volkmar Heinrich and Micah Dembo. Mechanics of neutrophil phagocytosis: experiments and quantitative models.* Journal of Cell Science *119: 1903-1913 (Figures 3 and 5, http://jcs.biologists.org/cgi/content/abstract/118/9/1789).

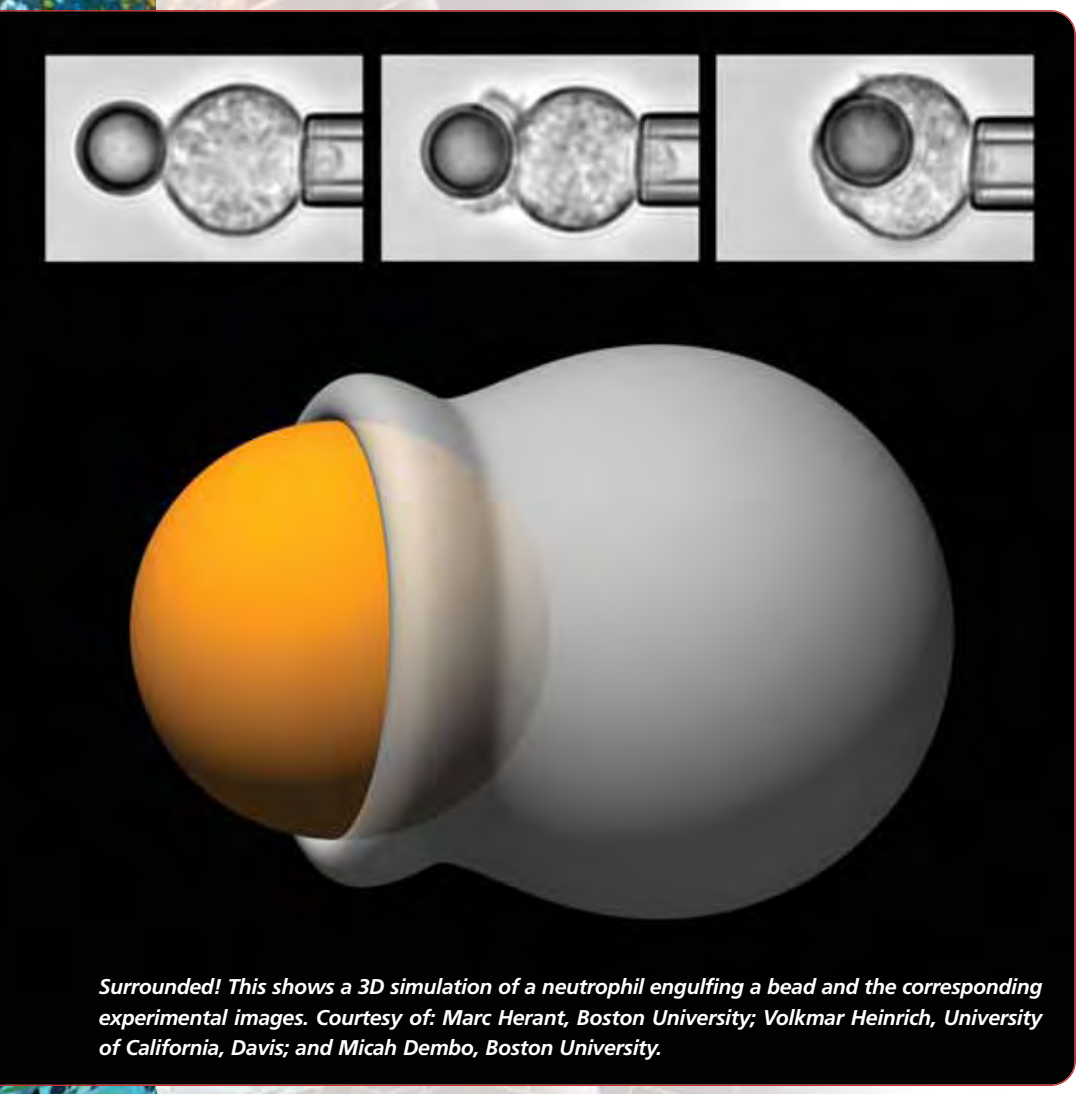
"We want to know:
What are the forces that
the cell is producing?
How is the cell pushing?
How hard is it pushing?
Where is it pushing? Is it pulling?
How does it orchestrate
its little hands and fingers to do
something like phagocytosis?"
Micah Dembo says.

Surprisingly, when the same neutrophil eats a yeast particle, it loses its ability to generate the intrusive force. "It has to slowly wrap its fingers around the yeast without any sucking in the cell is trying to grab it." The researchers don't really know why this happens, but perhaps the yeast particle has a defense mechanism that blocks the intrusive force.

"Until you model it and think about it, you never realize how clever the cell is and all the problems that the poor cell is facing to do these things," Dembo says.

motion," Dembo says. "So the cell is trying to make a big enough hand, and it will eventually manage to do that. But in the meantime the yeast is getting pushed away [by the protrusive force] as

"I love this kind of thing because until you model it and think about it, you never realize how clever the cell is and all the problems that the poor cell is facing to do these things," Dembo says. "Without the modeling, you would just be looking at pictures of cells eating things."



Surrounded! This shows a 3D simulation of a neutrophil engulfing a bead and the corresponding experimental images. Courtesy of: Marc Herant, Boston University; Volkmar Heinrich, University of California, Davis; and Micah Dembo, Boston University.

How a Cell Senses: Feeling the Environment

The cell's environment plays a critical role in directing cell behavior. In a landmark 2006 paper in *Cell*, researchers showed that the mechanical properties of the environment alone—just its elasticity, nothing biochemical—can influence cell fate: for example, a stem cell grown on a very stiff substrate becomes a bone cell whereas the same stem cell grown on a soft tissue becomes a brain cell. Follow-up experiments showed that substrate stiffness also directly affects cell shape, motility, growth, and malignancy. "The fundamental question is: how do they sense the stiffness?" Odde says.

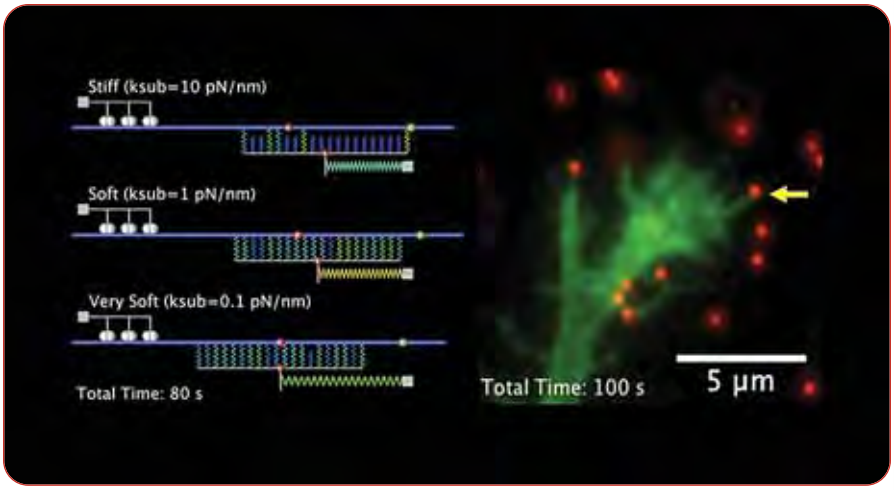
Cells bind to and interact with their environments (typically, the extracellular matrix) through proteins called integrin receptors. These receptors cluster in the cell membrane to form "adhesion complexes" that link the cell's actin cytoskeleton to the matrix and play a key role in cell movement and cell-to-matrix communication.

In a December 2009 paper in *PLoS Computational Biology*, **Daniel A. Hammer, PhD**, professor of bioengineering and of chemical engineering, and his colleagues, revealed a "simple calculation that shows why substrate elasticity affects the biology so strongly." They modeled the cell membrane and the substrate as lattices of springs and the integrins as individual springs that can diffuse along the cell membrane, cluster with each other, bind to the substrate, and pull on the membrane and substrate.

In simulations, they found that as you make the substrate stiffer and stiffer, it drives receptor clustering. "If the receptors remain distributed, then they have to pull up the substrate at many locations, and that's energetically very unfavorable on stiff surfaces," Hammer says. "What they'd rather do is get together in a cluster and then pull up the surface just in small regions."

The extent of clustering is directly correlated with cell activation. "I think the effect of substrate mechanics on cell biology is nothing more than this physical chemistry of driving clustering in these receptor patches," Hammer says.

The work has important implica-



Sensing Stiffness. LEFT: This computer simulation provides one possible explanation for how cells sense the mechanical stiffness of their environment. As myosin motors pull on actin bundles, molecular clutches (modeled as springs) engage and disengage with the substrate (also modeled as a spring). Stiff substrates have little give, and thus the clutches frequently slip and disengage; soft substrates can stretch and move with actin, so the clutches remain engaged longer. RIGHT: The motor-clutch model was tested against a series of experiments; for example, cell traction can be measured by labeling neurons (green) and soft substrates with fluorescent beads (red). Chan CE and Odde DJ, Traction Dynamics of Filopodia on Compliant Substrates,* Science*; 322: 1687-1691 (2008). Reprinted with permission from AAAS.

tions for cancer, because tumors are stiffer than normal tissues; and this stiffness promotes malignancy and growth. For example, breast tumors get stiffer and stiffer as they progress. "It used to be thought that this was an effect of breast cancer, but now people are starting to think that it might be one of the causative determinants of breast cancer," Hammer says.

In a 2008 paper in *Science*, Odde and his colleagues similarly used modeling to explore how the cell senses stiffness as it moves across a substrate. They modeled actin filaments as individual rods, and integrins and substrate molecules as individual springs. They found that more springy substrates can stretch and move with actin as the cell moves, so the clusters of integrin—which act like motor clutches—remain engaged longer. But less springy substrates have little give, and thus the clutches slip and disengage more frequently.

"So, cells, through that motor clutch system, actually have the innate ability to sense stiffness. How they actually read it out for these decisions that they make is now the next problem. And we're moving on to that and trying to apply it to brain cancer cells and how they migrate," Odde says.

How a Cell Moves: Crawling on Substrates

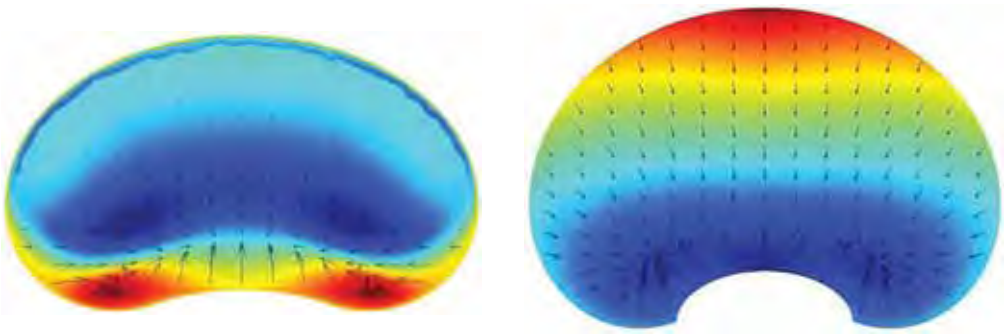
Cells move by crawling along substrates, propelled by actin filaments—which add proteins to one end and shed them from the other (called "treadmilling"). Actin polymerizes at the leading edge of the cell, pushing forward a protrusion of cytoplasm, which grabs hold of the substrate via clusters of integrins. Then the back of the cell detaches from the substrate and is pulled forward by the contraction of the actin cytoskeleton. Though the general principles are well understood, specific details are lacking; for example, it's unclear what determines a moving cell's shape and speed.

Mogilner's team devised a simple model to explain movement in fish keratocytes, fan-like cells that are among the fastest moving animal cells. "It turned out that a very simple mechanistic model, with very few equations, describes everything," Mogilner says. As actin polymerizes at the leading edge, it pushes on the cell membrane, causing tension all along the membrane (which does not stretch). This force, in turn, pushes back on the growing actin filaments. Actin density is highest in the middle of the leading edge, so the force per filament is lowest here, and actin grows rapidly. Actin density is lowest at the sides, so the force per filament is high here, which restricts polymerization. The work was published in *Nature* in 2008.

The model predicted that the higher the ratio of actin in the center to actin in the sides, the more canoe-shaped the cell would be and the faster the cell would move. These predictions were borne out by experiment.

"The equations are very enlightening because they connect the biochemistry (the kinetics of actin cytoskeleton) with the geometry (the shape) and with the physics (the forces and movements)," Mogilner says. "So I think this is a very cool thing."

Like Mogilner's model, most models of cell movement are two dimensional. This is a problem, because 3-D is not simply an extension of 2-D, says **Muhammad Zaman, PhD**, assistant professor of biomedical engineering at Boston University. In 2-D models, the cell interacts with the substrate only on one side. But when a cell moves in the body, it interacts with the extracellular matrix on all sides. "In reality a cell does not have a top or a bottom or a ventral or a dorsal surface; reactions happen all



Cells on the Go. Computer simulations of motile fish keratocyte cells. The color represents actin density (red/hot=high; blue/cold=low) and the arrows represent the flow field. The cell on the left is more canoe-shaped and moves faster due to the pattern of actin flow, whereas the cell on the right is rounder and moves slower. Courtesy of Raja Paul and Alex Mogilner, University of California, Davis.

over the surface," Zaman says. Thus the relevance of 2-D models for biological processes *in vivo* "is very limited if not completely inaccurate," he says. "More often than not, we find that the 2-D paradigms break down completely."

Unfortunately, most experiments are conducted in 2-D—on glass or plastic plates—which creates a severe bottleneck for would-be 3-D modelers.

3D Obstacle Course.
Computer rendition of a 3D extracellular matrix. The red fibers are collagen fibers that surround the cell; the cell must navigate through these during migration and invasion. Courtesy of Muhammad Zaman, Boston University.

"Modeling and experiments go hand in hand. It's very hard to publish or think about 3-D if you don't have any real data to compare it to," Zaman says. To counter this problem, Zaman's team measures cells moving through 3-D gels derived from *in vivo* sources.

Using these data, they built the first 3-D model of cell migration, a comprehensive, multi-scale model. At the lowest level, they zoom in on individual snippets of proteins in the cell and matrix, solving Newton's force equations for these snippets. "So you're looking for the right conformations that will bind, that will attach, that will stretch, things like that," Zaman says. Then they zoom out, feeding relevant information from the lower level into higher level models that solve similar force equations for proteins, protein complexes, or whole cells (with continuum rather than stochastic equations). Grid computing provides the computational power to run such large simulations.

In a 2005 paper in *Biophysical Journal*, Zaman's team explored how altering the 3-D environment affects cell velocity. Others had predicted that if you increase ligand density in the matrix—that is, give integrins more points where they can attach—this will give the cell a better grip and allow swifter motion. But, surprisingly, they showed that there is an optimal ligand density, after which speed decreases. At this point, the back of the cell experiences difficulty detaching, creating drag. "That was counterintuitive, but we showed that experimentally indeed it was the case. And the match was not only qualitatively accurate, but also quantitatively accurate," Zaman says. The validation of their computational prediction appeared in *PNAS* in 2006.

Their work may have practical implications for cancer. For example, there is a relationship between the collagen density in a woman's breasts and her chance of developing invasive breast cancer. It may be that, at optimal collagen densities, rapid cell movement increases the potential for invasion and metastasis.

MODELING MANY CELLS:

Bridging to Tissues and Organisms

The aforementioned models focus on the behaviors of single cells. But cells rarely act alone. To truly understand cell biology and to bridge to tissue and organism biology, multi-cell models are needed.

Though several approaches for multi-cell modeling are available, agent-based modeling is gaining momentum. Unlike traditional continuum models, which treat groups of cells as homogenous masses, agent-based models treat cells as individual autonomous entities. Besides capturing the heterogeneity of cells and their interactions, agent-based models facilitate collaboration between biologists and modelers.

"The cell really is an autonomous unit. It lends itself very well to agent-based modeling, where you have the one-to-one relationships between the computational model and the actual cell," says Southgate, a biologist who works closely with modelers. "For cell biologists, that's important, because you

"The cell really is an autonomous unit. It lends itself very well to agent-based modeling, where you have the one-to-one relationships between the computational model and the actual cell," says Southgate.

can immediately see the relationship between the modeling and the cell."

Rod Smallwood, PhD, professor of computational systems biology at the University of Sheffield in the United Kingdom, agrees. "Because you can talk about a computational object as if it was a physical object, this seems to make the discussions with cell biologists a lot easier. It seems much more intuitive to be able to talk about cells as if you have physical objects interacting with each other rather than to talk about sets of differential equations," he says.

Agent-based cell models also fill an important and largely untapped niche in multi-scale modeling: the middle-out model. The models can easily embed molecular-level modules, such as signaling networks—allowing them to scale down; at the same time, the collective behavior of cells falls right out of the simulations—allowing the models to scale up.

How Cells Cooperate: Growing Into Tissues

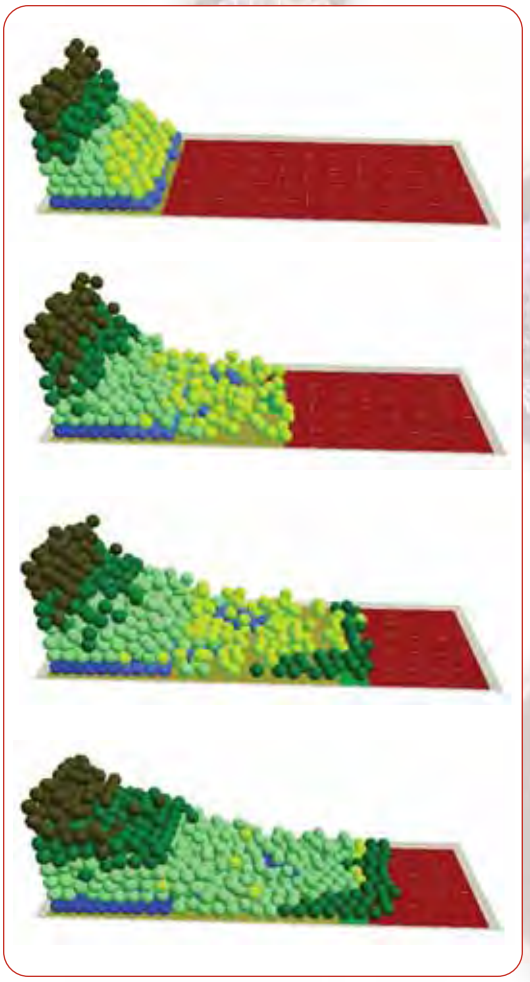
Cell cooperation plays a key role in promoting tissue growth during development and inhibiting it later in life. Cells bind to and interact with each other through surface receptors called cadherins. Mutations in the cadherins have been linked to cancer.

Southgate's team studies cell-to-cell interactions in human bladder epithelial tissue aided by agent-based modeling. In their model, rules govern whether each cell bonds to other cells, grows, divides, migrates in two dimensions, or dies. For example, each cell's probability of binding to its neighbor is proportional to the local calcium concentration. The local signaling milieu is determined by a series of mathematical models linked to the agent-based model. "We often adopt other people's pathway models, deriving rules that we then incorporate into the agent-based models," Southgate explains.

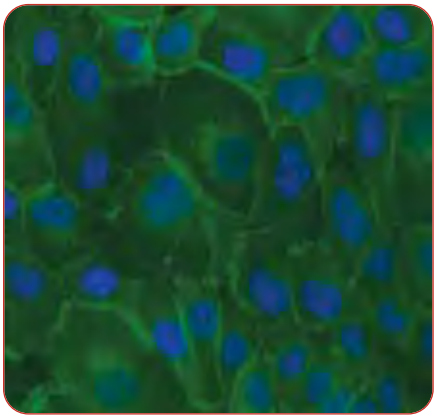
In a 2010 paper in the *Journal of Theoretical Biology*, Southgate's team introduced anti-social cells—cells lacking functional cadherin—into their models to see how they would influence normal cells and affect population behavior. In some situations, just a few anti-social cells could influence the growth of the entire population. The model illustrates one way that cancerous cells can disrupt the growth behavior of normal tissue.

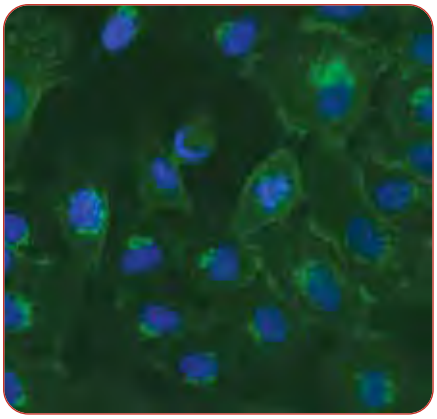
Cell cooperation is also important in wound healing. To heal a wound, cells migrate into the rift and multiply to fill the gap. The process is governed by both cell-to-cell and environment-to-cell signaling.

Smallwood and his colleagues are working out the details using 3-D, multi-scale, agent-based models. The agents are cells that can bond, migrate, divide, or differentiate. External modules determine cell signaling and resolve the forces between cells. "So there are models of particular cell signaling pathways that others have created that you can download. The functions that control cell transitions can be culled from these external models," Smallwood says. "Things move in time steps and at the end of each time step, the forces are resolved and the position and size of the cell is updated." To make the calculation computationally tractable, they model the behavior of 10,000 cells—just a fraction of the million cells involved in wound healing, but enough to capture the fundamental



***Incomplete Repair.** An agent-based simulation that shows why wounds greater than 2 centimeters across cannot heal spontaneously. Different colors represent different cell types: blue cells are keratinocyte stem cells; they change to light green as they migrate and proliferate and then to dark green as they differentiate. When the wound (red) is too big, the cells differentiate and stop moving before they can fill the gap. From: Tao Sun, Salem Adra, Rod Smallwood, Mike Holcombe, Sheila MacNeil. Exploring hypotheses of the actions of TGF-β1 in epidermal wound healing using a 3D computational multiscale model of the human epidermis.* **PLoS ONE 4(12): e8515. doi:10.1371/journal.pone.0008515.**





***Anti-Social Cells.** These bladder epithelial cells are labeled with a fluorescent antibody to E-cadherin (green), with nuclei stained blue. The top panel shows the normal pattern of E-cadherin concentrated to junctions between cells, whereas cells in the bottom panel have been genetically modified to disrupt E-cadherin and create anti-social cells. Courtesy of Jenny Southgate, University of York.*

biology, Smallwood says.

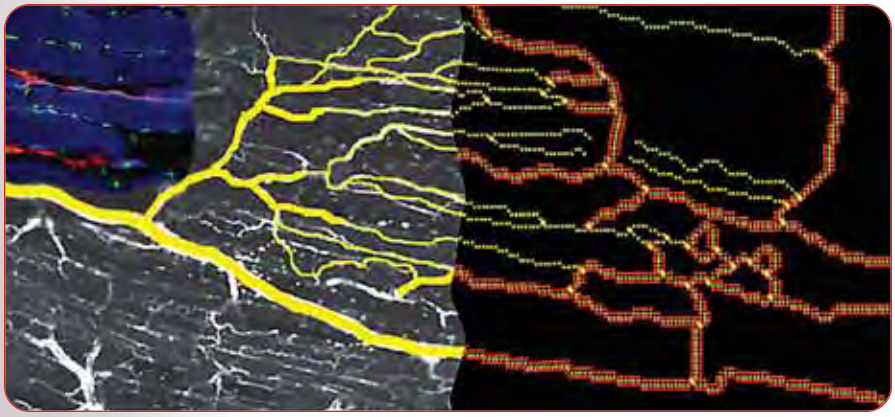
In a paper in press with *PLoS Computational Biology*, Smallwood's team used simulations to explain, for the first time, why wounds wider than two centimeters cannot heal spontaneously. The reason: cell-to-cell signaling drives the cells to first start migrating and then to differentiate; once they differentiate, they can no longer move. If the distance the cells have to migrate is too great, they differentiate before they have filled the gap. "If you can't move on any more, you're not going to heal. So that's quite interesting. You can actually see the critical reason why the wound doesn't heal," Smallwood says.

This suggests that it might be possible to get large wounds to heal if you could override the cells' differentiation rules, he says.

How Cells Travel: Trafficking in the Bloodstream

When the body is injured or invaded, immune cells travel through the bloodstream to the site of injury. They exit the bloodstream through a precise set of steps: first, they roll along blood vessel cells, then they halt to a stop, and, finally, they slide through the blood vessel wall. The process is orchestrated through adhesion molecules on both the vessel cells and immune cells (selectins and integrins), as well as signaling molecules called cytokines. A fundamental question is how cells decide where to stop in circulation.

Shayn Peirce-Cottler, PhD, assistant professor of biomedical engineering at the University of Virginia, studies immune cell trafficking with agent-based computational models. Cells drift, adhere, roll, stop, or enter tissues based on concentrations of simulated cytokines and adhesion receptors. The cells are embedded within a simulated microvascular network—complete with pressure, flow velocities, and wall shear stresses—that shuttles cells around the body. It's a complex system. The researchers have to keep track of the cells in time and space, monitoring the state of hundreds of chemokines and cell surface receptors as well as the cells' behaviors, Pierce-Cottler says. The models are two dimensional, since moving to 3-D would make them computationally intractable at this point, she says.



Traffic in the Bloodstream. Agent-based models in conjunction with* in vivo *experimental models are used to study the recruitment of circulating cells in the microvasculature of ischemic muscle. The left panel shows a confocal micrograph image of the macrovessels (yellow) and microvessels (blue and red) in mouse muscle; immune cells (monocytes) are stained in green. The right side is a screenshot from an agent-based model of this same system. Courtesy of Shayn Peirce-Cottler, University of Virginia.

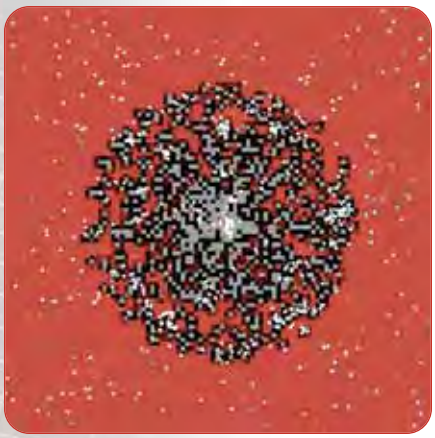
Peirce-Cottler's team is exploring the build up of plaques in the arteries (arteriosclerosis). Because inflammation is a major contributor to arteriosclerosis, it turns out that the trafficking of immune cells (particularly monocytes) to plaques plays a critical role in their initiation, progression, and eventual rupture. Peirce-Cottler and others believe that microvessels—the small blood vessels that feed into large vessels—may be an important conduit of monocytes to plaques. They are using simulations to tease out the relative contribution of monocytes from the microcirculation versus the macrocirculation.

"That's hard to quantify experimentally, because you need to have a system where you're tracking individual cells *in vivo* and watching to see, when a monocyte shows up in a plaque, where does it come from. And technically speaking, we just don't have the tools to be able to do that," Peirce-Cottler says. "That's the great thing about computational models. You can actually follow an individual monocyte and say 'hey, where did you come from?'"

How Cells Battle Injury: Testing Drugs *In Silico*

A major insult to the body, such as an overwhelming infection or injury, can cause a condition called sepsis: The immune system goes into overdrive, leading to collateral damage of otherwise normal tissue, subsequent organ failure, and death. In the 1990s, researchers reasoned that since certain cytokines incite immune cells, administering anti-cytokine drugs would cure sepsis. But they were wrong. "It turns out that none of the drugs worked, and some of them actually hurt people," says Gary An, who is a trauma surgeon and ICU doctor at Northwestern University Feinberg School of Medicine.

Frustrated by these failures and the lack of effective treatments for his sep-



sis patients, An turned to computational modeling "as a means of addressing the bottleneck in translational research." It was clear that sepsis exhibited complex behaviors that could not be predicted through reductionism and linear thinking alone, he says. However, his path to computational research had a significant hurdle.

"I was not a computer science or a math guy at all; I hadn't taken anything in those areas since high school. So the computational bar was kind of high," he says. Fortunately, he discovered an agent-based modeling toolkit called StarLogo that was designed for teaching kids, and thus was very intuitive.

"The results of a cell biology paper are: I take this cell; I stimulate it with this particular compound that performs this particular function; I then see how the cell responds. Those sorts of behaviors can be converted to rules and computer code for agent-based modeling relatively straightforwardly."

He built agent-based models of sepsis and used them to run *in silico* drug trials based on actual clinical studies. The agents are the immune and blood vessel cells at the blood-to-vessel interface. The cells change states based on cell-to-cell interactions, the presence of mediators such as cytokines, and the influence of drugs. When enough of the blood vessel cells are injured, then the simulated person dies.

In a paper in *Critical Care Medicine* in 2004, he simulated what would happen if you treated populations of *in silico* patients with various anti-cytokine drugs. He showed that mortality rates were 30 to 40 percent, no better than standard treatment. He also tested different combinations of the drugs (which some had hypothesized were needed to override redundancies in the immune system), as well as various doses and durations of treatment, but nothing worked.

"By running the computational models, you identify that the disease state itself is very, very stable and resistant to change," he says. "When you simulate the intervention, you get this sort of pebble in the stream effect where you might see a little bit of a result initially, but the flow of the system is such that it basically swallows up your intervention and it doesn't have any effect."

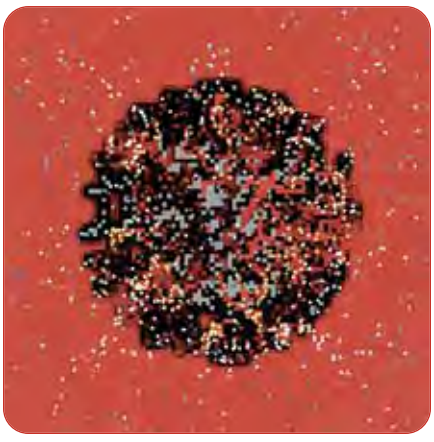
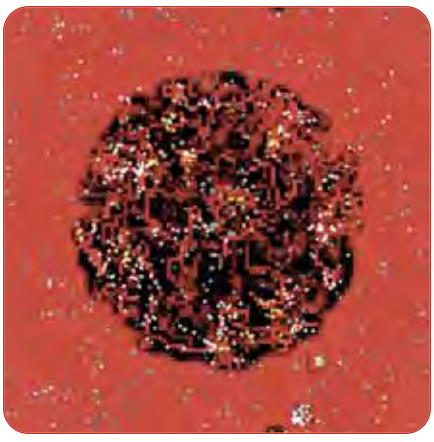
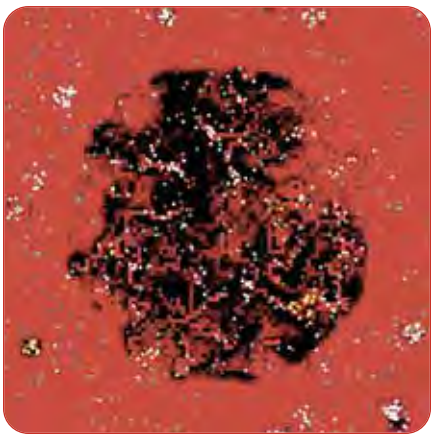
"System-level computational models are invaluable in identifying these types of unexpected behaviors, and will play a critical role in addressing the challenges of developing effective therapeutic interventions," An says.

"System-level computational models are invaluable in identifying these types of unexpected behaviors, and will play a critical role in addressing the challenges of developing effective therapeutic interventions," Gary An says

Bringing Modeling and Cell Biology Together

Despite these recent successes in pairing cell biology and computational modeling, the two fields remain only loosely integrated. Breaking down these barriers will take long-term collaborations, Zaman says. For example, his lab comprises half experimentalists and half modelers. Yet, he says, "I still see it in many of my students that it takes a long time before they can speak a common language."

"We need a more integrated environment, not only for the computations to be more powerful, but also for the experiments to be more probing and much more quantitative," Zaman says. "I think the burden of responsibility is on both sides." □

Sepsis Explosion. (Lower opposite page and below) These serial screenshots from a 2-D agent-based simulation of inflammation and sepsis follow the progression from infection, to initial immune response, to cell death and the start of healing. Upon infection with bacteria (gray areas), the healthy blood vessel cells (red) become damaged (dark red) or die (black). Gradually, inflammatory cells (white neutrophils) gather near the bacteria and become activated (yellow or other colors). The inflammatory cells gradually clear the bacteria, allowing healing to occur. Courtesy of Gary An.

Under TheHood

BY PETER EASTMAN, PhD

Efficiently Evaluating Mathematical Expressions with OpenCL Code

OpenCL is a cross-platform language for doing general purpose computation on graphics processing units (GPUs) and other massively parallel architectures. One of its most interesting features is the fact that the compiler is built into the runtime. This means that while a program is running, it can programmatically generate the source code for new computational kernels, compile them, and execute them at full speed on the GPU.

Because of this feature, OpenCL provides a unique opportunity to build both flexibility and high performance into a piece of software. We use this ability in OpenMM, a library for running molecular simulations on high performance architectures, including those that support OpenCL. Users can specify arbitrary mathematical expressions for calculating the interaction energy between particles in their simulation and have them transformed on the fly into OpenCL kernels that can be run on a GPU. To get optimal performance, this transformation must be done carefully.

Consider the following expression which describes a Lennard-Jones nonbonded interaction between two particles:

$$E(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right)$$

where r is the distance between the two particles and ε and σ are parameters of the force field describing the interaction. The simplest approach to transforming this expression into source code is a one-to-one translation of mathematical operations to OpenCL instructions:

```
energy = 4*epsilon*(pow(sigma/r, 12)-
pow(sigma/r, 6));
```

Users can specify arbitrary mathematical expressions for calculating the interaction energy between particles in their simulation and have them transformed on the fly into OpenCL kernels that can be run on a GPU.

This is clearly an inefficient way to perform the computation. The first thing we notice is that the ratio `sigma/r` is being calculated twice. We should identify common subexpressions, compute them only once, and assign them to temporary variables so they can be reused. In fact, the easiest way of doing this is to create a new temporary variable for *every* subexpression. For each piece of the expression we translate, we first check whether an identical one has already been processed, and if so, simply use the existing temporary variable. This produces the following OpenCL source code:

```
float temp1 = sigma/r;
float temp2 = pow(temp1, 12);
float temp3 = pow(temp1, 6);
float temp4 = temp2-temp3;
float temp5 = epsilon*temp4;
energy = 4*temp5;
```

This may look much wordier and harder to understand, but that isn't important. We're generating it to be read by a compiler, not a human!

The next problem we notice is the use of the `pow()` function, which is a slow way to calculate small integer powers. Building up the power through repeated multiplication is much faster. The trick is to decompose the power into a sum of powers of 2, such as 12=8+4, then compute the powers of 2 by repeatedly squaring a multiplier:

DETAILS

Peter Eastman, PhD, is a software engineer at Simbios (http://simbios.stanford.edu) and a key developer of OpenMM (http://simtk.org/home/openmm). He recently wrote a small C++ library for parsing and analyzing mathematical expressions called Lepton (http://simtk.org/ home/lepton) that has been incorporated into OpenMM.

```
float temp2;
{
  float multiplier = temp1;
  multiplier *= multiplier;
  multiplier *= multiplier;
  temp2 = multiplier;
  multiplier *= multipier;
  temp2 *= multiplier;
}
```

We are using only four multiplications to calculate a 12th power, which is much faster than the `pow()` function. Similarly, we can calculate the 6th power with three multiplications. But we can do even better by combining both of them into a single evaluation:

```
float temp2;
float temp3;
{
  float multiplier = temp1;
  multiplier *= multiplier;
  temp3 = multiplier;
  multiplier *= multiplier;
  temp2 = multiplier;
  temp3 *= multiplier;
  multiplier *= multipier;
  temp2 *= multiplier;
}
```

We are now calculating both powers at once with only five multiplications!

The final important optimization is to translate all expressions at once as a single unit. The above example shows only the expression for the energy, but in OpenMM we need to calculate the derivative of the energy as well. The two expressions share many subexpressions. For example, the derivative includes $(\sigma/\varepsilon)^{11}$ and $(\sigma/\varepsilon)^5$, so by translating both expressions together, we can compute four different powers at the same time.

In practice, we find these techniques work extraordinarily well for generating optimized OpenCL code to evaluate mathematical expressions. Our preliminary benchmarks with OpenMM show that the automatically generated GPU kernels are only a few percent slower than hand-tuned versions. At the same time, the user gains enormous flexibility to select the precise interactions they want in their simulations. □

Guest Editorial

cont'd from page 1

Every application that gets exchanged like this goes through CSR DRR.

While there is no central institute or office for biomedical computing and computational biology at NIH, there is a very vibrant and organic entity.

Now you're probably wondering about the outcomes of all these activities. In the last six years, the four broad-based BISTI announcements funded a total of 297 research grants in the amount of $355 million. In addition, the Continued Development and Maintenance of Software announcement funded 106 research grants in the amount of $160 million. In that same period, 5560 unique grant applications were reviewed in the informatics study sections (MABS, BDMA, BCHI, NT, GCAT, MSFD, BMRD, BMIT, MI and Continued Development and Maintenance special study section), and of these 1330 (24 percent) were funded.

For early stage investigators who want to add to these numbers by submitting successful grant applications, I offer the following advice:

- Team up with experienced mentors who can help you through the science and logistics of the NIH process.
- Talk to NIH program staff about your ideas. You can identify the appropriate contacts from the BISTI funding page/funding contacts link, http://www.bisti.nih.gov/funding/index.asp.
- Visit the BISTI Web site, which offers many useful resources, including a list of ongoing government programs, initiatives and public-private partnerships dealing with multiscale modeling, ontologies and data management, mathematical biology, systems biology, and numerous other biomedical informatics or computational biology efforts.
- Whether a new or seasoned NIH investigator, always focus your applications on the science because, after all, biomedical and health-related research is the NIH mission. □

Seeing Science

cont'd from page 30

rational approach to predicting gene function in *Arabidopsis thaliana*, a plant widely studied by plant geneticists. Dubbed AraNet, the work was published in the February 2009 issue of *Nature Biotechnology*. Marcotte and Lee are currently using the same approach to study gene function in humans.

"The idea is that we're making functional links between genes based on their behavior in a lot of different assays," Rhee says, including microarray analyses, protein-protein interactions and inferences from animal orthologs culminating in 24 different data sets.

The researchers started by analyzing pairs of genes with known function in order to set a baseline score for inferring related function. They then looked at about 27000 *Arabidopsis* genes—most of which are uncharacterized—to identify possible gene-gene associations among them. "By then asking 'what are the functions of the neighboring genes?' we can try to infer the functions of the uncharacterized genes," Rhee says. When her team experimentally tested the predictions for three uncharacterized genes, two out of the three had functions that were predicted by the network.

Rhee is interested in using inferences from AraNet to narrow down the candidate genes involved in complex traits. Although she'll be doing this work in plants, Rhee says the approach will be applicable to all organisms. She's also curious about uncharacterized genes that are connected only to other uncharacterized genes. "Perhaps we can use the network to characterize some undiscovered processes."

Ideally, Rhee says, researchers will combine AraNet's predicted functions with their own knowhow to try to design the best sorts of experiments to conduct. It's like rational drug design, she says: "You're using all the available information to be as systematic as possible in designing your experiments. This is a good application of systems biology." □

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

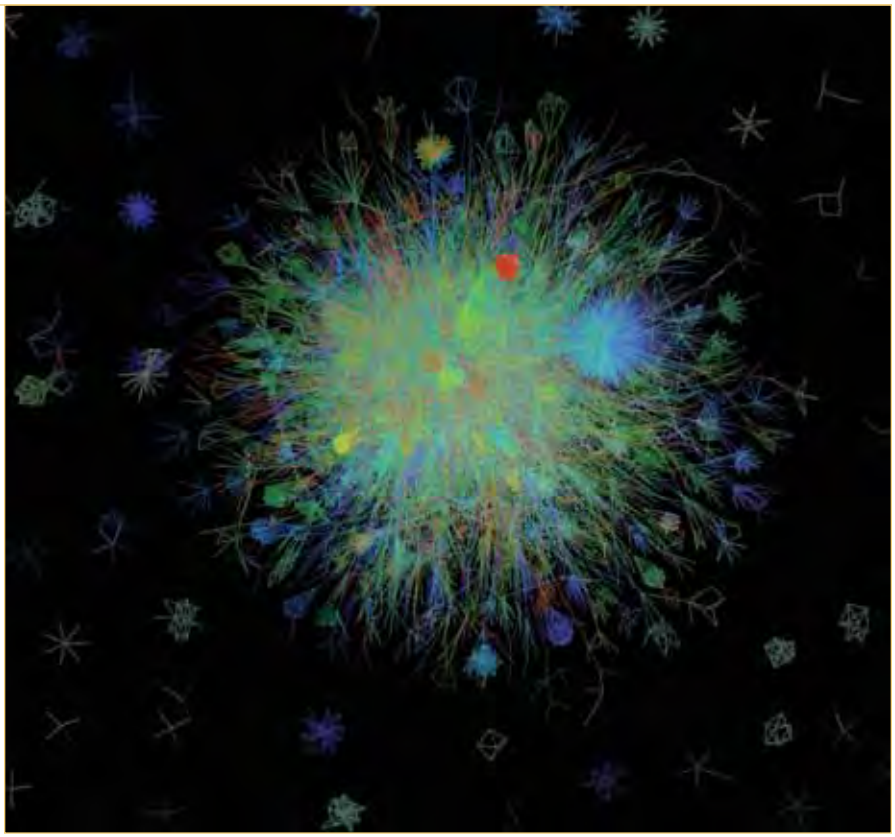
A Tipping Point for Function Prediction

There comes a tipping point in systems-biology studies of gene function where knowing some genes' functions can, using a computational approach, help hone in on the functions of other genes. That point has already been reached for yeast and *C. elegans* but is just now being reached for systems where functional information is more sparse—such as in plants and humans.

"There are still a lot of plant genes with unknown functions," says **Sue Rhee, PhD**, in the plant biology department at the Carnegie Institution for Science. "We need more sophisticated ways to characterize what these genes are doing."

So she and her colleagues, including **Edward Marcotte, PhD**, at the University of Texas, Austin, and **Insuk Lee, PhD**, at Yonsei University, South Korea, modified the *C. elegans* and yeast algorithm for use in systems with less complete data. This produced a

continued on page 29



This functional network of* Arabidopsis *genes shows the top 10% of the functional links identified by AraNet. Each line represents the connection between two genes and is colored to reflect the likelihood score for a relationship between the paired genes' functions: Red means a high score, blue is low. For example, the red area in the middle top of the figure represents the ribosomal complex, while the large blue cluster to the right represents the phosphatases, which have a weak relationship to one another although they share enough biological behavior to be linked. Image courtesy of Sue Rhee, Edward Marcotte and Insuk Lee.