DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

Structural Genomics

Exploring the 3D Protein Landscape

PLUS:

Clinical Decision Support:

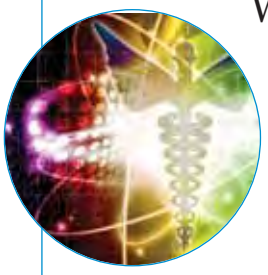
Providing Quality Healthcare With Help From A Computer

Winter 2009/2010

contents

ContentsWinter 2009/2010

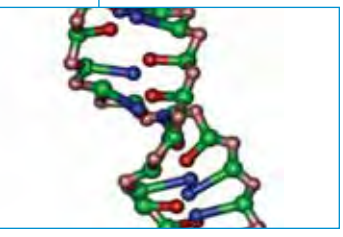
FEATURES

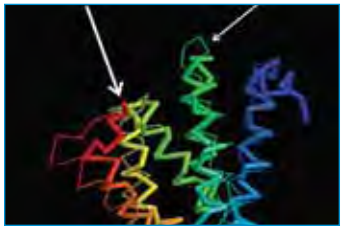


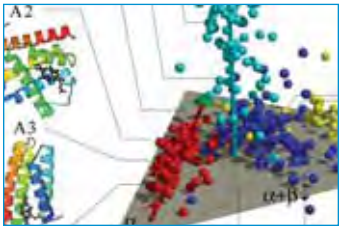
DEPARTMENTS

Cover Art: Created by Rachel Jones of Wink Design Studio. Background protein structures courtesy of the Protein Structure Initiative, NIH, NIGMS. Structure of anthrax protein (BA2930) courtesy of Wladek Minor, professor of molecular physiology and biological physics at University of Virginia, and The Center for Structural Genomics of Infectious Diseases, with funding from NIAID.

Page 19 to 28: Caduceus collage image is © Krishnacreations | Dreamstime.com.







Winter 2009/2010
Volume 6, Issue 1
ISSN 1557-3192

Executive Editor David Paik, PhD

Managing Editor Katharine Miller

Associate Editor Joy Ku, PhD

Science Writers
Denise Chen, Katharine Miller, Janelle Weaver, PhD, Jane Palmer, PhD, Gwyneth Dickey, Tia Ghose, Jennifer Welsh, Marissa Cevallos, Daniel Strain, Sandra M. Chung, Olga Kuchment, PhD, Adam Mann

Community Contributors
Ahmet Erdemir, PhD, Joy Ku, PhD, Gregory R. Bowman

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:

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)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Jim Gnadt, PhD (NINDS)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Yuan Liu, PhD (NINDS)
Richard Morris, PhD (NIAID)
Grace Peng, PhD (NIBIB)
Nancy Shinowara, PhD (NCMRR)
David Thomassen, PhD (DOE)
Jane Ye, PhD (NLM)

GuestEditorial

BY AHMET ERDEMIR, PhD

Recognizing and Encouraging Timely Dissemination

For scientific progress, dissemination is as valuable as publishing, if not more so. We need to recognize those who put in the effort and encourage all to do the same.

The availability of free and open access data, models, and software indisputably accelerates scientific progress. Unfortunately, dissemination necessitates organization, documentation, and quality assessment for ultimate impact. In addition, dissemination efforts require considerable resources and time. Constrained by the realities of funding and the requirements of academic publishing imposed on investigators, individual dissemination attempts are likely to be a low priority.

Those who have the motivation have many pathways to share their work. The material can be provided on personal websites, on a research laboratory's site, or through institutional repositories. Alternatively, or additionally, investigators can provide copies of datasets or models in publisher repositories following the work's peer-review and acceptance. But too often, researchers take the lengthy route of publishing before disseminating their work. Associating dissemination with peer-reviewed academic material does increase confidence in its utility. However, we seem to postpone the sharing of more usable and detailed information for months, if not years. Also, once the work is published and we get academic recognition, we may lose the motivation to clean up our raw data, document the details for other users, and provide the data and software on platforms for long-term access. For some of us, even when funders require dissemination, this doesn't rise to the same level of priority as publishing, which seems to be more important for our career advancement.

Not only does dissemination of scientific software and models receive low priority and attention, it's also poorly done: It can be nearly impossible to find a usable model on the Web. I recently used the keywords "knee" and "finite element" to look for knee models based on finite element analysis. This gave me 371 hits in PubMed and 86,800 in Google. And from those hits, I was unable to find a model that I could download and use. I might have been unlucky or maybe I didn't know where to look, but this experience certainly says something about the current practice of dissemination in my field. Are we scientists incapable of sharing? If open source software developers and the editors of Wikipedia figured out how to do it, why can't we? We seem to be emphasizing publishing and the search for funding over the need to openly share our work.

Are we scientists incapable of sharing? If open source software developers and the editors of Wikipedia figured out how to do it, why can't we? We seem to be emphasizing publishing and the search for funding over the need to openly share our work.

Maybe the solution is to establish recognition and rating systems for dissemination—similar to what exists for academic publishing. Let's assume that all data, models, and software, once documented and provided in a usable form, are worth sharing. We can provide a journal-like system with free and open access, where the submissions would have a digital object identifier; they would be indexed and made searchable in literature databases. We would also need to replace the traditional peer-review system with a more flexible rating system for timely dissemination. The information should be out in public immediately. The ratings could follow later, based on opinions from experts in the discipline—assigned by an editorial board for example—as well as on comments from the public. The dynamic rating can be associated with the indexing of the distribution and available in search results from literature databases, which will provide a quick assessment of quality for those who need the information. This may not be the only way, but it seems that we need new recognition and encouragement mechanisms for scientists to start sharing in a timely manner. □

DETAILS

Please send your feedback to Ahmet Erdemir, erdemira@ccf.org. Ahmet works in computational biomechanics at the Cleveland Clinic; he collects data, builds models, engages in software development, and tries to share them. He appreciates free and open access to other data, models, and software. An ongoing discussion on the proposed journal for dissemination can be found at http://www.imagwiki.org/mediawiki/index.php?title=Journal_for_Dissemination.

SimbiosNews

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

A Big Step Forward for OpenSim

With its initial release two years ago, OpenSim offered researchers a powerful open-source application for simulating movement. Simple enough to be used by high school students yet advanced enough to address complex biomechanical research questions, OpenSim has attracted thousands of users since then. Now, OpenSim 2.0 promises greater opportunities for customization, enabling users to extend existing algorithms and integrate their own new algorithms within the OpenSim framework.

"Until now, OpenSim was only configured for certain research questions. If your question didn't fit, you'd have to either use other software or recast your question as something OpenSim could answer," says **Matthew DeMers**, a mechanical engineering graduate student at Stanford University and member of the Simbios OpenSim development team led by **Scott Delp, PhD**, professor in bioengineering at Stanford University. "Now, people can extend OpenSim and use it to answer a wide variety of questions."

The new version of OpenSim provides an application programming interface (API) to allow researchers access to core OpenSim functionality. Outwardly, this is the most noticeable change; however, it has also been re-engineered for better performance and flexibility.

"The whole structure underneath has been redesigned," says **Samuel Hamner**, another OpenSim development team member and Stanford University mechanical engineering graduate student.

While the graphical user interface will look the same, the development team rewrote the underlying code so that it is built entirely on Simbios' biosimulation toolkit, SimTK, with its robust, high-performing computational components, such as integrators, optimizers, and contact models.

Attendees at the first OpenSim Developers Jamboree held in October had the opportunity to work with a pre-release version of OpenSim 2.0 and were excited about what it would enable them to do.

Ilse Jonkers, PhD, a professor at Katholieke Universiteit Leuven in Belgium, has several research projects that utilize OpenSim, including the development of a neural controller to generate simulations of walking that account for neural reflex activity, and not just the mechanics. One of the main improvements Jonkers noticed during the workshop was the ease in defining controllers. "That, to me, is a huge advantage that we will exploit in the coming months," she says.

"The API is richer and cleaner," says **Tom Erez**, a graduate student in computer science at Washington University in St. Louis who studies machine learning and motor control. He appreciates being able to access functionality like the SimTK integrators with their built-in error checking.

However, he says, "the most revolutionary thing I saw was the elastic foundation model." Through SimTK, the new version of OpenSim will provide contact models—such as the elastic foundation—so that a simulation will recognize and model the behavior of two arbitrarily shaped bodies, such as bones, when they come together. For Erez, it means more flexibility in the simulations he runs. "I can generate force simulations from scratch. I don't need recorded ground reaction forces, and I don't need to hack my own ground-foot interaction," explains Erez. "It's a big step forward."

The value of OpenSim and its continued enhancements is also clear to Jonkers. "Without OpenSim, I couldn't do my research," she says. □

DETAILS

OpenSim 2.0 was released in December 2009. To download the software and learn more about training opportunities, visit http://simtk.org/home/opensim.

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

NewsBytes

New Technology Reveals the Genome's 3D Shape

Try taking a human hair as long as Manhattan and cramming it—unsnarled—inside a marble. This is the challenge faced by a 2-meter-long strand of DNA as it folds into its compact array of 23 chromosomes within a cell's nucleus. Previously, scientists only theorized about how DNA squeezes inside a nucleus without becoming a hopelessly tangled mass. Now a new technique called Hi-C reveals that DNA packs knot-free into its chromosomal patterns by assuming a rare geometric shape observed in snowflakes, crystals and broccoli.

"We've developed a powerful new technique to look at chromosomes at an unprecedented resolution," says **Job Dekker, PhD**, cell biologist at the University of Massachusetts and coauthor of the study in the October 9, 2009 issue of *Science*. "What we found constitutes a breakthrough in our understanding of chromosome folding."

At the small scale, DNA wraps around proteins called histones and assumes its classical double-helix shape. At the large scale, chromosomes cluster in discrete sections within the nucleus called "territories." "Between the scale of chromosome territories and the scale of histones, effectively nothing has been known about the structure of the genome," says first author **Erez Lieberman-Aiden**, a graduate student in the lab of **Eric Lander, PhD**, professor of biology at the Broad Institute in Cambridge, Massachusetts.

Hi-C reconstructs an unbiased 3-D map of the entire genome. First, scientists soak a complete set of chromosomes in formaldehyde, which acts like glue to stick together parts of the genome that are close in 3-D space. Then they chop the DNA into a million pieces and perform massive parallel sequencing on the interacting fragments. Mapping software compares the sequences of attached fragments with a human genome reference sequence; based on the results, the scientists compute which parts of the folded DNA physically interact with each other.

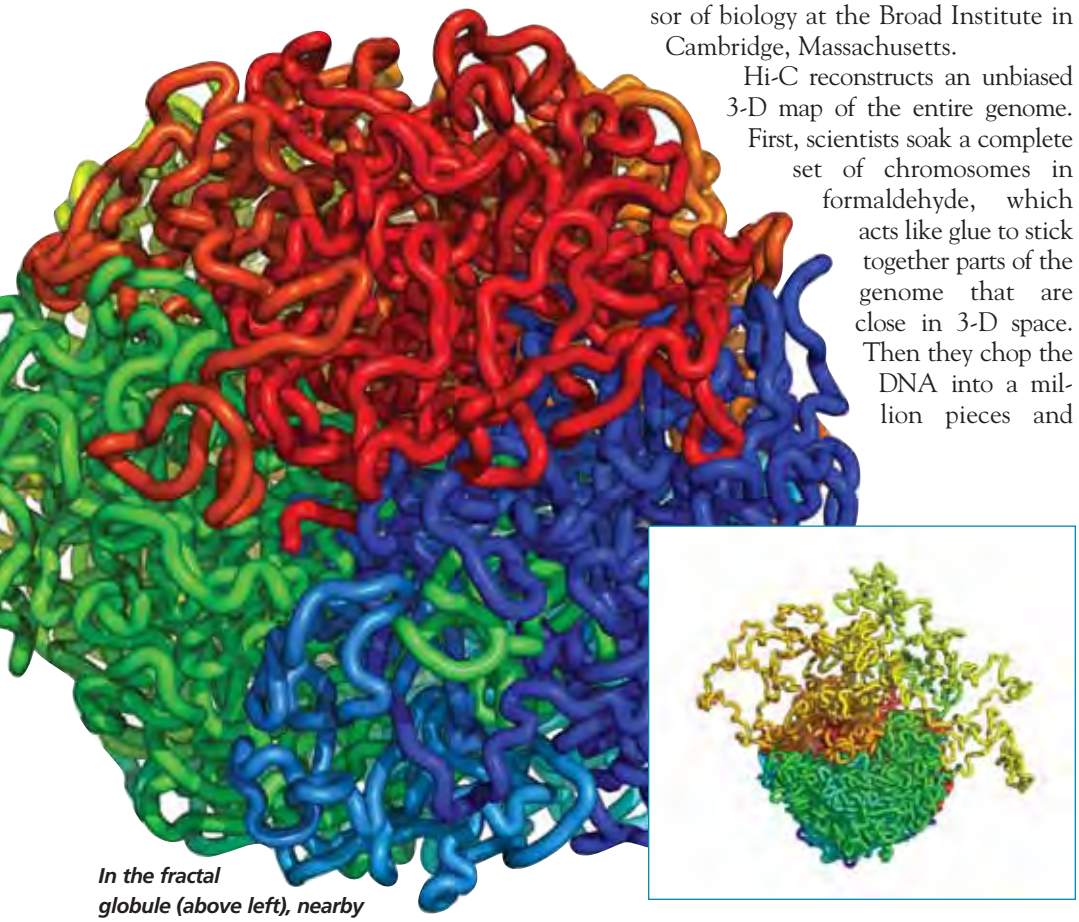
The team found that active, gene-rich and inactive, gene-poor sections cluster in separate parts of the nucleus. The active chromatin segments are like easily accessible papers spread out across a desk, whereas the inactive portions are densely packed, like folders in a file cabinet.

Simulations revealed that DNA assembles into dense fractal globules—structures that look alike at different levels of magnification, such as the intricate geometrical form of a crystal. Genes are easily accessible, but when they're not in use, the structure spontaneously collapses into a tight, knot-free bundle.

"This is the first spatial map of the genome," says **Tom Misteli, PhD**, cell biologist at the National Cancer Institute in Bethesda, Maryland. "It's a technical breakthrough that opens the doors to doing all sorts of interesting experiments."

Future experiments will investigate how the 3-D shape of DNA morphs depending on the activity of genes and disease states, like cancer. As genome sequencing becomes cheaper, Dekker says, it should be possible to obtain higher spatial resolution and even to reconstruct the shapes of individual genes.

—By Janelle Weaver, PhD



In the fractal globule (above left), nearby regions on a chain of DNA—indicated using similar colors—are packed into nearby regions in 3D space. The accessible DNA chain unravels easily (above right) because the globule lacks knots. Images courtesy of Leonid A. Mirny and Maxim Imakaev, reprinted from Lieberman-Aiden, E., et al., Comprehensive Mapping of Long-Range Interactions Reveal Folding Principles of the Human Genome, Science, *326(5950): 289-293 (2009), with permission from AAAS.*

How DNA Goes A'Courtin'

Until now, scientists have known little about how complementary single strands of DNA court one another before binding to form the classical double helix. But now, molecular dynamics simulations have identified that the binding—or hybridization—mechanism depends largely on the sequence of the DNA: Ordered sequences will meet and then slither lengthwise to find the correct match; but sequences that are random will connect at key sites then rapid-

ly assemble along the molecule's length.

"One would have thought that random sequences would have more difficultly hybridizing, and that is not necessarily the case," says **Juan J. de Pablo, PhD**, professor of chemical and biological engineering at University of Wisconsin, Madison. The work was published in the October 5 issue of the *Proceedings of the National Academy of Sciences*.

Scientists have previously tried to simulate the pathways by which DNA strands combine, but the models they used included too much detail to enable sufficiently long computations, de Pablo says. So De Pablo's group developed a highly simplified model, tested on experimental data, to capture essential details of the interactions between the base pairs of complementary strands of DNA. The researchers then simulated the process by which the single strands interact using molecular dynamics and Monte Carlo simulations, taking multiple "snapshots" of the double helix as it assembled. To the team's surprise, the path to a successful union depended crucially on the sequences of the molecules.

When the sequences of both single strands are ordered or repetitive, any two sites of base pairs can come together and the two strands slowly "slither" lengthwise until complementary base pairs match along the entire chain, says de Pablo. When the sequences are random, however, single sites located toward the center of the strands unite early. "The moment they come together, then the molecule just assembles perfectly and it does so very quickly," de Pablo says.

The results could influence the design of technologies that depend on the hybridization process, such as gene chips, de Pablo says. To engineer more efficient and reliable hybridization, researchers could use random sequences, which bind more efficiently and with fewer errors.

"This is an interesting step forward," says **Nadrian Seeman, PhD**, professor of chemistry at New York University. "No one had taken the time to track the pathway previously." Seeman has used the principle of random sequencing in his own hybridization studies, and he finds it reassuring to see it vindicated by the simulation data. "It does tell people who are designing sequences to avoid repetition in the sequences," he says.

—By Jane Palmer, PhD

Modeling Bacterial Comets

Rocketing within and between human gut cells, *Listeria monocytogenes*—a motile, foodborne bacterium—leaves a comet-like tail of actin protein behind it and makes us sick. Scientists have long wondered how actin allows the bacterium to puncture through multiple cells and evade the human immune system. A new computational model shows how rapidly accumulating actin at the back of the bacterium produces that force.

"Our simulation helps us understand the basic physical properties and mechanisms by which actin can produce force," says biophysicist **Mark Dayel, PhD**, a postdoctoral researcher at the University of California, Berkeley, and lead author of the paper published in the September 2009 issue of *PLoS Biology*. "We now have an explanation of why you get a switch from the initial pulse to smooth motion."

L. monocytogenes comes from contaminated produce or milk and infects epithelial cells in the gut. Using a membrane protein called ActA, the bacterium moves by continuously building a network of actin filaments from pieces of the host's cytoskeleton. To observe this system in action, scientists have reproduced the bacterial movement *in vitro* by coating tiny beads with ActA and putting them in a cell solution. Initially, actin fibers build from the surface of the bead, pushing old actin outward and forming a shell. But when the shell gets too big, it cracks and the bead bursts out, propelled forward by

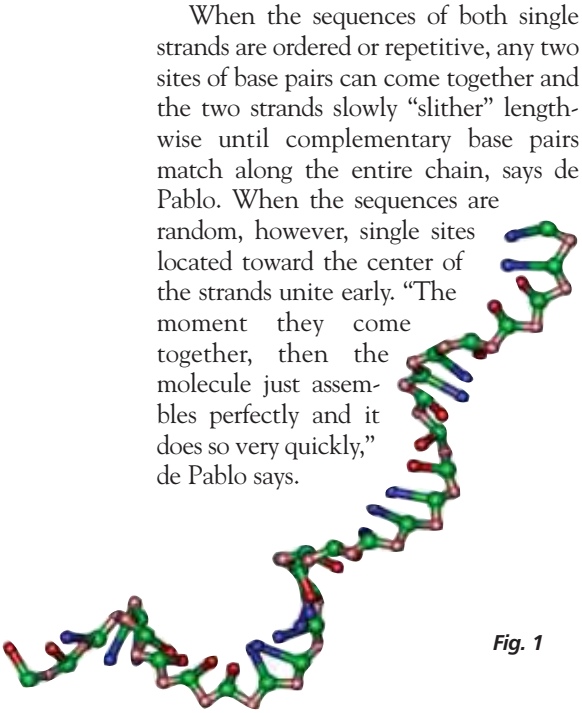


Fig. 1

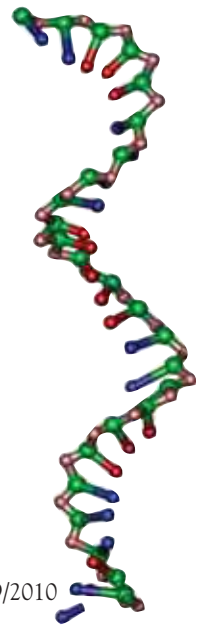


Fig. 2

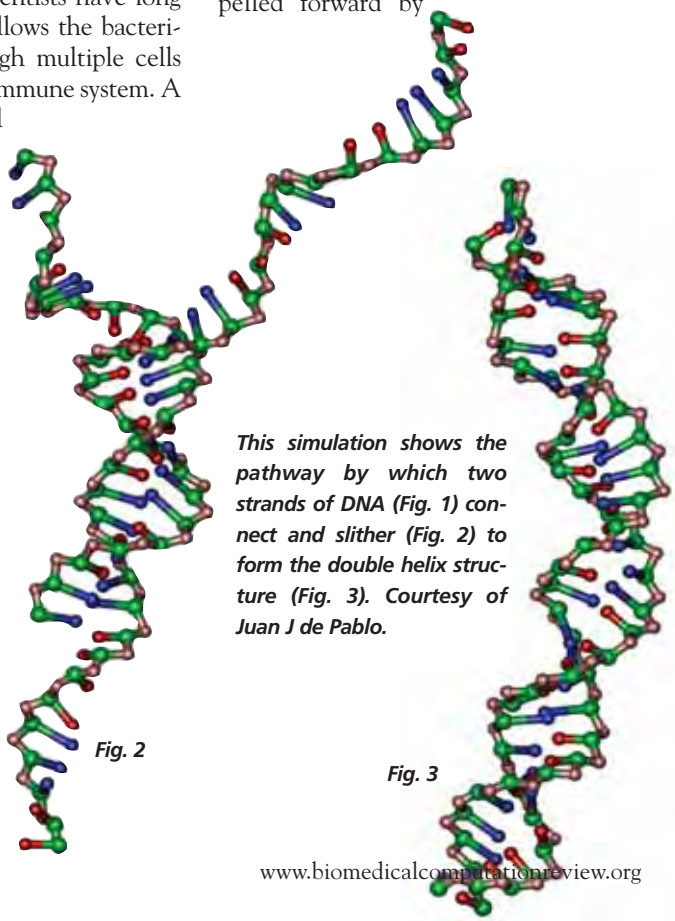
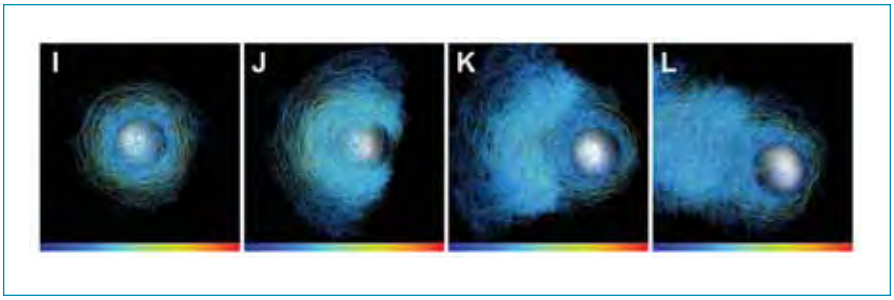


Fig. 3

This simulation shows the pathway by which two strands of DNA (Fig. 1) connect and slither (Fig. 2) to form the double helix structure (Fig. 3). Courtesy of Juan J de Pablo.



In this 3-D computer simulation time series, a bead representing the Listeria monocytogenes bacterium builds actin fibers at its surface before breaking out of its shell and moving forward, pushed by actin fibers accumulating at the back of the bead. Reprinted from Dayel et al., In Silico Reconstitution of Actin-Based Symmetry Breaking and Motility,* PLoS Biology, *7(9): e1000201 (2009), doi:10.1371/journal.pbio.1000201.

continual actin production. Until now, scientists thought that cracks in the outer shell spread inward and caused the shell to break. They also thought that the actin fibers stretched and then contracted behind the cell, squeezing it like a bar of soap.

To better understand these dynamics in detail, Dayel and his colleagues modeled the process, called "symmetry breaking." The simulation showed that the actin shell cracks from the inside, just above the surface of the bead, where tension of the actin is greatest. When the bead bursts out, surface actin accumulates against the shell left behind and pushes the bead forward, rather than squeezing as previously believed. The model then successfully predicted what would happen to the beads in novel situations, which Dayel verified *in vitro* by placing new bead shapes in different conditions. Dayel says the next step is to calibrate the model so scientists can measure forces that can't be measured *in vitro*. "We can extend its qualitative behavior to quantitative behavior, essentially allowing us to do virtual experiments," Dayel says.

"The combination of model and experiment has made a very compelling case that the mechanisms they're proposing are the real ones," says **Roger Kamm, PhD**, professor of mechanical engineering at the Massachusetts Institute of Technology. The model is "extremely simple, yet capable of capturing some fairly complex behavior," Kamm says.

—By Gwyneth Dickey

Cooking Cancer With Gold Nanoshells

Tiny gold particles that absorb laser light and convert it into heat are a promising therapy for destroying tumors. However, controlling the temperature of such gold nanoshells is crucial: The shells must get hot enough to kill tumor cells, but they must not scorch nearby healthy tissue. Now, researchers have developed a model that predicts how much these nanoshells raise the temperature of surrounding tissue.

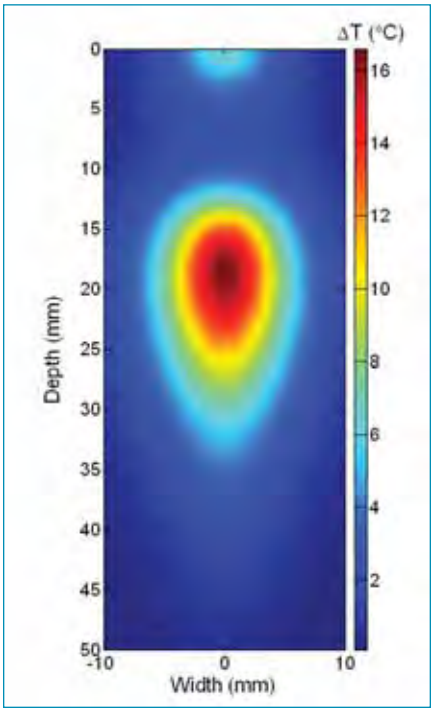
"When we tried to estimate how much heat is being generated from the process, we didn't have any good way to quantify it," says **Sang Hyun Cho, PhD**, a medical physicist at the Georgia Institute of Technology and senior author of the study in the October 2009 issue of *Medical Physics*. With the new model, researchers won't need to directly measure temperature with invasive temperature probes or magnetic resonance thermometry imaging, Cho says.

Other teams have modeled the temperatures of nanoshells in tissue. However, their models assumed that the nanoparticles spread out evenly, Cho says. "But we know that gold nanoshells are not uniformly distributed in tissue," he observes. Instead, the particles cluster tightly in some tumor regions and avoid others. That's because nanoshells travel to the growths within a tangle of misshapen blood vessels, but the vessels don't reach all parts of the tumor.

Using basic heat transfer principles, Cho's group created a computational model to calculate the heat generated by individual shells. At first, Cho assumed the nanoshells spread out evenly. But, unlike previous efforts, Cho's model is well-suited to capture the pattern of hot spots arising from a more realistic nanoshell distribution, he says.

The simulations captured the general heating profiles from past experiments but, Cho says, couldn't match the exact temperatures—probably because the team lacked good measurements of how much light is absorbed and scattered at the wavelength they used, thus affecting their calculations of the conversion to heat. His group plans experiments to pin down these values.

"They are doing very theoretically well-founded simulations," says **David Paik, PhD**, professor of radiology at Stanford University. The next impor-



Cross-sectional view of the temperature distribution in a tissue-like medium filled with gold nanoshells after three minutes of near-infrared laser treatment. Only the bottom layer of the medium (starting at a 12 mm depth) contains gold nanoshells. Reprinted with permission from* Medical Physics *36(10), 4665, 2009. doi:10.1118/1.3215536 (2009).

tant step is modeling heating in a more realistic nanoshell distribution, he says. "This is where their more computational approach would be a big win."
—By Tia Ghose

3D Angiogenesis Modeled

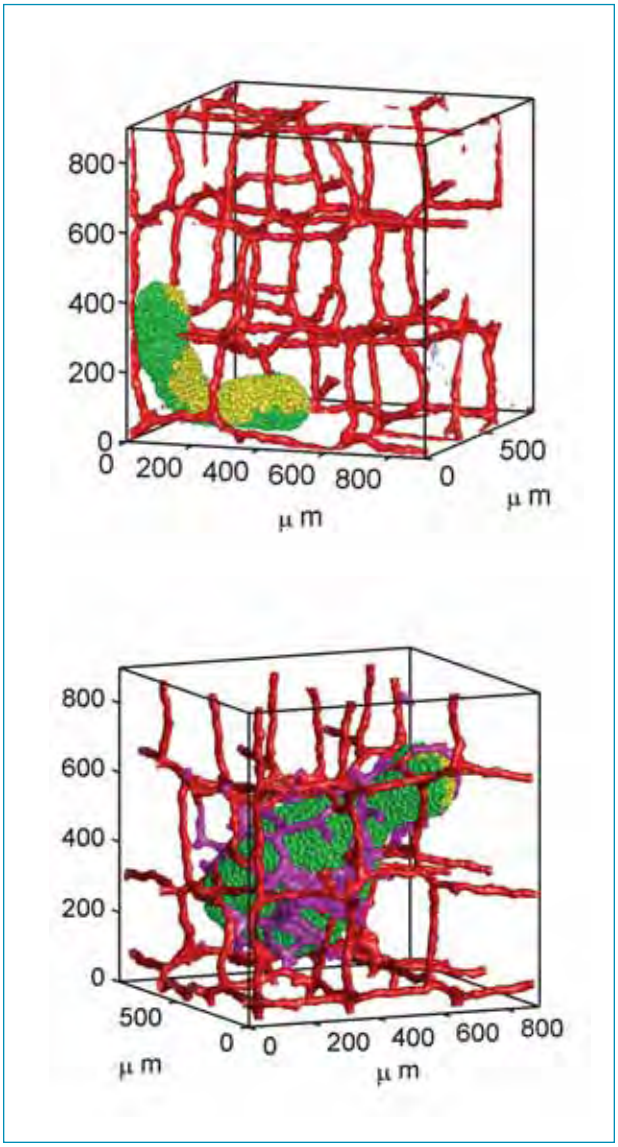
Researchers have successfully simulated how growing blood vessels affect the sizes and shapes of tumors using a 3-D model based solely on how cells behave—without reference to intracellular biochemistry. The simplified modeling system uses open-source cellular behavior "plug-ins" yet compares favorably with models laboriously coded from scratch. It also captures many essential details observed in real tumors.

"Building a computational model based on 10 to 15 behaviors is much easier than building one based on thousands of genes," says **Abbas Shirinifard**, graduate student at Indiana University's Biocomplexity Institute and lead author of the work published in the October 2009 issue of *PLoS One*.

The human body sprouts new blood vessels when they are needed. Cancer cells use this process—called angiogenesis—to their advantage. As a tumor grows, some of its interior cells become starved for oxygen and start emitting distress signals. In response, cells that create new blood vessels grow toward the distressed cells to provide them with oxygen and other nutrients. The result: a larger, actively growing tumor. Until now, researchers have modeled multi-cellular processes—such as angiogenesis—by painstakingly programming interactions among gene and protein cascades. Such models are not easily comparable between research groups, and take much longer to re-program with different conditions.

Shirinifad's team used an open-source platform called CompuCell3D (available at www.compucell3d.org) developed by **James Glazier, PhD**, and **Maciej Swat, PhD**. CompuCell3D models multi-cellular behaviors based on how each cell reacts to environmental conditions. The cells involved in tumor growth respond in defined ways, so Shirinifad and his colleagues modeled them using action-response rules, such as "If oxygen levels fall below X, send out Y signal," or "If protein X reaches concentration Y, divide." When they switched off the rule for cells to create new blood vessels in response to distress signals, the simulated tumors were small and irregular, with contours that followed the existing blood vessels. When they ran the simulation with angiogenesis "turned on," the resulting tumor grew large and rounded. These outcomes matched the appearance of such tumors in models programmed from scratch, as well as observations of real tumors treated with anti-angiogenesis compounds. In CompuCell3D, researchers can change and re-run such models in days—much quicker than if they were adapting a hand-coded algorithm, Shirinifad says.

"The exciting thing is the new technique," says **Mark Chaplain, PhD**, mathematics professor at the University of Dundee, Scotland. He notes that the current model lacks a proper simulation of blood flow; the simulated blood-vessel cells deliver oxygen itself rather than shuttling oxygen-rich blood. "If they develop this technique further by modeling blood flow, they will have a very powerful model," Chaplain says.
—By Jennifer Welsh



After 75 days of simulated growth, a tumor model looks quite different in the presence (A) and absence (B) of new blood vessel formation. Green cells in the tumor are actively dividing, while yellow cells are starved of oxygen. Red cells are blood vessel cells originally present in the model; purple cells are new blood vessel cells, present only in the model that supports angiogenesis. Axes are labeled in microns. Reprinted from Shirinifard, A, et al., 3D Multi-Cell Simulation of Tumor Growth and Angiogenesis, PLoS One, 4(10): e7190. doi:10.1371/journal.pone.0007190 (October 2009). Images provided by Abbas Shirinifard.

Improving the Sense of Touch for Surgical Robots

When a knife cuts into an organ, forces push back in ways that mechanical engineers can, to some extent, predict. But other factors are also at play: Ions shift in solution within cells, causing electromechanical changes that, researchers now say, can be predicted as

well. In a new model of soft tissue deformation, researchers for the first time include electromechanical changes as well as mechanical ones. The work could lead to better 3D surgical simulations and could ultimately provide surgeons at computer terminals with simulated feedback through surgical robot's controls.

"We want to bridge the gap between surgical simulation and surgical practice," says **Yongmin Zhong**, **PhD**, research fellow in mechanical and mechatronic engineering at the Curtin University of Technology in Perth, Australia. Zhong's novel way of modeling soft tissue deformation was outlined in the November 2009 issue of *Artificial Intelligence in Medicine*.

Robots lend a helping metal hand in surgery worldwide, cutting more precisely than trembling human fingers. But the surgeons behind the joysticks cannot feel how hard to push: slicing through fatty tissue feels the same as cutting through air. When cutting by hand, "you know how hard you're pushing, you know what damage you're doing, " says **Julian Smith**, **MD**, a heart surgeon at the Monash Medical Center in Melbourne, Australia, a co-author on the paper. "With robotic instruments, you get none of that."

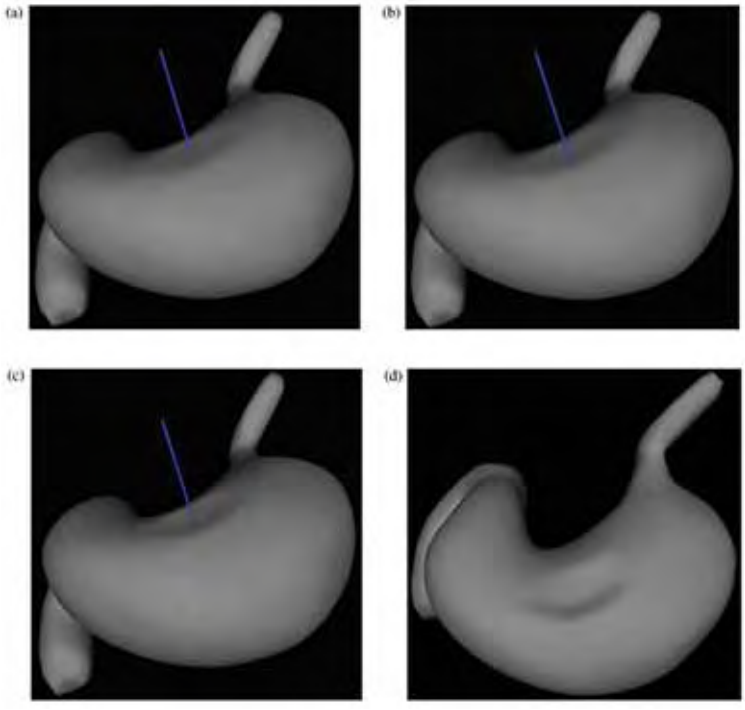
In previous attempts to provide a sense of touch in surgical simulations, researchers focused only on the mechanical force applied. While the mechanical force is important, Zhong explains, so are the electrical forces that come into play deeper within the tissue. For instance, charged particles like potassium swim in the plasma-like interstitial fluid between tissue cells, morphing the overall shape of the tissue.

By including the diffusion of charged particles in a set of sophisticated mathematical expressions, Zhong showed how prodding tissue shoves like-charged ions together, creating electrostatic repulsion. The model shows that this repulsion makes it harder to cut the soft tissue because it pushes back on the knife. Zhong tackled the equations with an artificial cell neural network, a much zippier problem-solver than numerical algorithms because the "cells" number-crunch as a team, instead of iteratively. It's the computational equivalent of six people jointly solving a jigsaw puzzle instead of taking turns. Such quick computational solutions are critical in a surgery, Zhong notes, because doctors cannot work with a time lag.

"They did a very good job and it's closer to what we can get in the real world, but it doesn't mean the problem is solved," says **Xiaobu Yuan**, **PhD**, associate professor in computer science at the University of Windsor. For example, poking the stomach causes it to shrink because it's connected to the nervous system, but the new model doesn't take that into account.

Smith plans to test if the model matches reality by putting animals under the knife. "The model is yet to be applied," says Smith, but it has "outstanding potential."

—By Marissa Cevallos



Researchers at the Curtin University of Technology used a cellular artificial neural network to simulate how soft tissue deforms under pressure, say, from a surgical knife—as shown here in blue deforming a virtual human stomach. Reprinted from Zhong, Y, et al., An electro-mechanical based deformable model for soft tissue simulation,* Artificial Intelligence in Medicine*, 47(3):275-288 (2009), with permission from Elsevier and Dr. Yongmin Zhong.

Conducting Medical Research from Electronic Health Records

To discover links between genes and disease, researchers typically recruit individual patients with and without the disease of interest; have them sign consent forms; take their medical histories; and analyze their blood samples. As well as being time-consuming and expensive, it can be hard to get a large enough sample of patients. But now researchers have shown there might be another way—using electronic medical records to identify patients with the

desired phenotypes and then obtaining their anonymized leftover blood samples to test for genetic information.

"We showed that we can actually conduct full-blown association studies to find the right patients with the right phenotypes and connect them to the right samples," says **Isaac Kohane, MD, PhD**, professor at Harvard Medical School and director of i2b2 (Informatics for Integrating Biology and the Bedside), the National Center for Biomedical Computing that conducted the study published in the September 2009 issue of *Genome Research*. "It's soup to nuts work."

With the help of natural language processing (NLP), the i2b2 researchers set out to use a large, available, cheap data pool: the electronic medical record archives for 2.6 million patients at Partners Healthcare System in Massachusetts. Although doctor's notes are notoriously unstandardized, NLP tools can break them into their smallest components, analyzing parts of speech and how words are joined. The i2b2 team sought to identify pools of patients with rheumatoid arthritis, asthma, secondary illnesses and risk factors for asthma (for example, smoking history). Along the way, clinical experts gauged the accuracy of the process and helped refine search terms. "It takes three to four months of iteration with expert clinicians until we get it just right," Kohane says. In addition, the researchers developed a system to access anonymously saved leftover blood samples from the identified populations to use for future studies requiring genetic data.

And the NLP tools did a pretty good job: Of about 98,000 patients identified as having asthma, 82 percent of the time the experts reviewing the files concurred in that diagnosis; 90 percent of the patients identified with a history of smoking had such a history; and of the 4,618 NLP-identified rheumatoid arthritis sufferers, 92 percent had definite arthritis (according to expert review) while 98 percent probably did. By studying these electronic patients, the researchers successfully reproduced several results from past clinical research. And while the clinical studies had paid an average of $650 to characterize and obtain blood samples from each patient, i2b2 spent $20 to $100.

Medical Record Snippet	*Smoking History*
SOCIAL HISTORY: The patient is married with four grown daughters, uses tobacco, has wine with dinner.	Positive
SOCIAL HISTORY: The patient is a nonsmoker. No alcohol.	Negative
SOCIAL HISTORY: Negative for tobacco, alcohol, and IV drug abuse.	Negative
BRIEF RESUME OF HOSPITAL COURSE: 63 yo woman with COPD, 50 pack-yr tobacco (quit 3 wks ago), spinal stenosis, ...	Positive
SOCIAL HISTORY: The patient lives in rehab, married. Unclear smoking history from the admission note...	Insufficient data
HOSPITAL COURSE: ... It was recommended that she receive ...We also added Lactinax, oral form of Lactobacillus acidophilus to attempt a repopulation of her gut.	Insufficient data
SH: widow,lives alone,2 children,no tob/alcohol.	Insufficient data

After lengthy training, i2b2's natural language processing software scans clinical histories, tagging words and phrases that describe smoking history and making a diagnosis (right-hand column). With training, the NLP tools were able to equate "smoking history" with "smokes often," distinguishing both from "non-smoker." Clinical experts also reviewed random results and computer scientists refined the search terms to clarify ambiguities like "tob." Reprinted from S. Murphy, et. al, Instrumenting the health care enterprise for discovery research in the genomic era,* Genome Research *, 19(9): 1675–1681 (2009).

"This paper represents very encouraging results using free open-source software," says **Chunhua Weng, PhD**, assistant professor of biomedical informatics at Columbia University. She says the next step is to include information such as how long an individual smoked or when symptoms began in patient descriptions. Kohane agrees, noting that researchers are working to include time-varying data in i2b2's model.

—By Daniel Strain

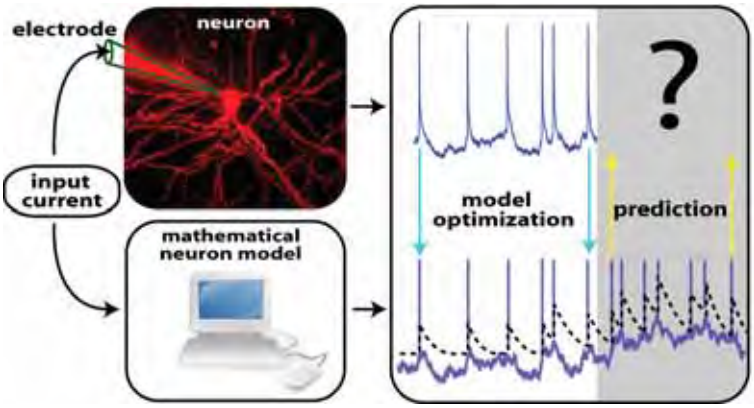
Neuron Models: Simpler Is Better

During the summer of 2009, the International Neuroinformatics Coordinating Facility in Stockholm dangled a nearly $10,000 cash prize in front of neuron modelers and challenged them to do better. And they did. The winners of the competition, which was described in the October 16, 2009 issue of *Science*, produced a neuron model that became more accurate as they stripped away pieces of a much more complex starting model.

"It was amazing for us physicists to see the description become simpler as we tried to make the performance better," says **Shigeru Shinomoto, PhD**, a physicist at Kyoto University in Japan who, along with two of his former students, snagged the grand prize.

Modeling the electrical behavior of individual neurons is crucial to understanding how thought and other cognitive functions arise in complex neuronal networks. Current neuron models can predict some neuron behavior, but with limited accuracy and at high computational cost.

The international competition has grown from eight entrants in 2007 to 33 this year and included teams around the world. "We had different people from different backgrounds using methods we would never have thought of," says **Wulfram Gerstner, PhD**, a computational neuroscientist at the Ecole Polytechnique Federale in Lausanne, Switzerland who co-authored the *Science* paper.



To set up one of the challenges for the neuron modeling competition, an artificial current was injected into a live neuron (upper left) and the resulting electrical activity was recorded for 60 seconds (blue trace, top right). Competitors used data from the first 38 seconds of the recording to fine-tune the parameters of a mathematical neuron model receiving an identical current injection (purple trace, lower right). Model performance was measured by the percentage of spikes correctly predicted in the final 22 seconds of the recording. Graphic courtesy of Richard Naud.

Contestants had to predict the precise timing of electrical spikes in individual neurons from different parts of the brain. Since different neurons can respond differently to the same signal, competitors used the first 38 seconds of data from a neuron to adjust their model parameters to better fit that neuron. They used the freshly tuned model to predict spikes in the subsequent 22 seconds of data. Shinomoto's winning model predicted 59.6 percent and 81.6 percent, respectively, of the spikes from two different neurons.

Electrical activity in a real neuron spikes when its membrane potential passes a set threshold value. Shinomoto's model neuron has an adapting threshold that increases immediately after a spike and decays exponentially to its initial value. The decay is modulated by two time constants of 10 ms and 200 ms, chosen to reflect the timing of ion currents in the neuron membrane.

The competition will evolve with the field, Gerstner says. Computational neuroscientists will soon draw on an emerging body of molecular knowledge to improve their models, says **Erik De Schutter, PhD**, a professor of computational neuroscience at the Okinawa Institute of Science and Technology. Advanced molecular techniques should reveal the physical structures and electrical properties of neurons in much greater detail than is currently known. These data may help modelers account for the effects of variations in temperature and chemical conditions and in the physical structures of the neurons.

"Neuron modeling is still a work in progress," De Schutter says. "It's much more difficult than we thought."

—By Sandra M. Chung

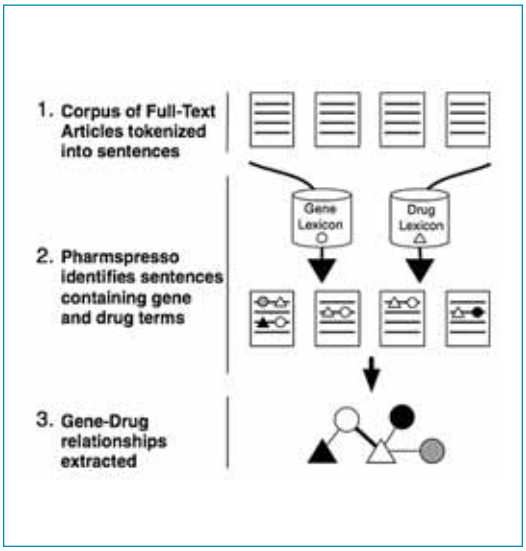
Trawling for Drug-Gene Relationships

When a drug saves one person but makes another ill, a bitter lesson in genetic differences often follows. With many such lessons already under our collective belts, researchers are using existing knowledge to predict additional drug-gene relationships as a way to forestall future calamities. A new software program can trawl published papers for gene-drug relationships, plug those relationships into known genetic networks, and predict which genes are likely to affect a patient's response to a drug.

"Our contribution is using text mining and taking decades of research and folding that in to inform the prediction," says **Yael Garten**, biomedical informatics PhD candidate in the lab of **Russ B. Altman, MD, PhD**, at Stanford University and a lead author of the work. "We showed that this is as good as and sometimes even better than manual curation," in which scientists painstakingly enter published drug-gene interactions into a database. Garten will present the team's research January 2010 at the Pacific Symposium on Biocomputing in Hawaii.

The previous version of the algorithm, designed by Altman and others, relied more heavily on manual labor. Called PGxPipeline, it employed a database of gene-drug relationships manually compiled from scientific articles by a team of scientists at Stanford Medical School. PGxPipeline wove these relationships into an orderly web, along with a database of gene-gene interactions and other data, to predict how strongly each of 12,460 genes affects response to a specific drug.

The team has now cut PGxPipeline loose from the manually created drug-gene database, automatically mining the information from published papers. This faster, cheaper method will inform the drug-gene rankings with constant updates from new literature. The manual-curation- and text-mining-based



The text-mining-based version of PGxPipeline automatically dissects journal articles into component sentences and marks where a drug or a gene is mentioned. Reading the sentence syntax and vocabulary, it tracks the interactions between drugs and genes. A network/web of interactions is established (bottom), in which the thickness of each edge corresponds to the number of articles that support the interaction. The web of relationships is later enhanced using a database of gene-gene interactions and other information. Image reprinted from Garten, Y., Tatonetti, N., & Altman, R., Improving the prediction of pharmacogenes using text-derived drug-gene relationships, Pacific Symposium on Biocomputing, Hawaii, January 2010.

versions of PGxPipeline predicted with similar accuracy a test set of 682 drug-gene interactions. And the text-mining-based version was slightly better at identifying genes that play the largest roles in response to a specific drug.

Garten hopes to use the revised PGxPipeline to parse all relevant scientific literature for drug-gene relationships. Better predictions will save researchers time in deciding which of the possible interactions to test in the lab and eventually influence how doctors prescribe drugs, she maintains.

"There is an emerging trend in bioinformatics to combine information from curated databases with information extracted from text," says **Tom Rindflesch, PhD**, principal investigator for the semantic knowledge representation project at the National Institutes of Health in Bethesda, Maryland. "This is an excellent example."

—By Olga Kuchment, PhD

Scientific Discovery Through Video Games

When it comes to folding proteins, even modern supercomputers don't always get things exactly right. Enter FoldIt, an online video game that harnesses the human brain's natural pattern-recognition abilities to tweak computer oversights. Since its release in May 2008, FoldIt has captivated a core group of several thousand dedicated players. Contestants manipulate three-dimensional protein chains into the best configuration they can find, exposing effective and previously unknown algorithms. In recent months, the puzzles have focused on medical applications. For example, a puzzle released in October called "Finding Home" asks players to bind a potential gene therapy tool—a homing endonuclease—to DNA. In another, called "Pack the Holes and Fight Cancer," gamers will help design a protein that could activate a new kind of cancer drug.

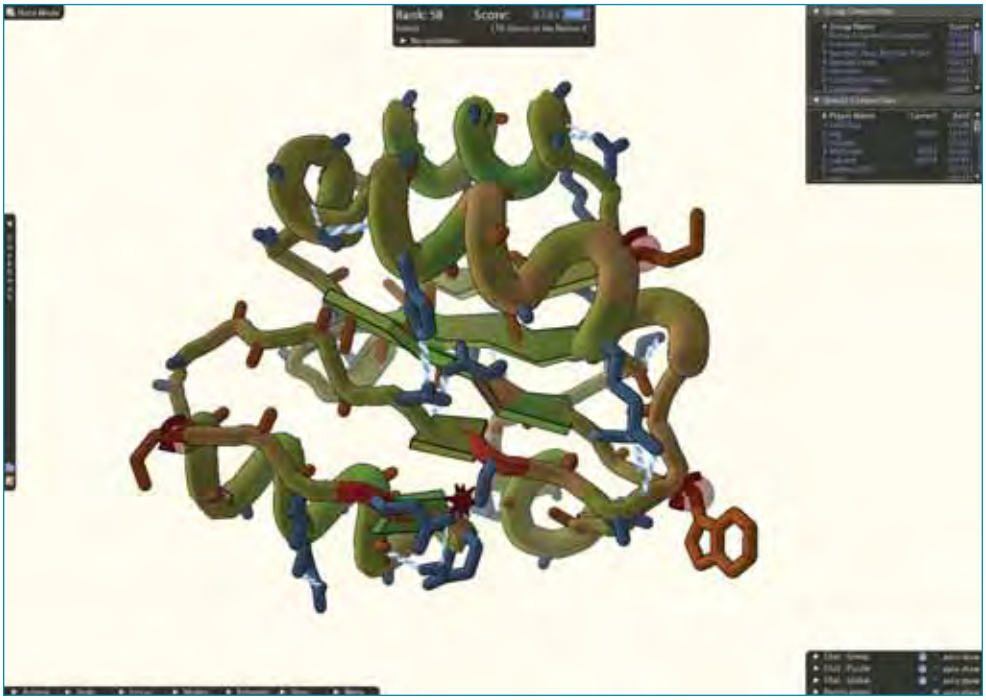
"The players, most of whom are non-experts, have sort of become protein scientists," says **Adrien Treuille, PhD**, assistant professor of computer science at Carnegie Mellon University. Treuille helped create FoldIt with a team at the University of Washington led by graduate student **Seth Cooper**, computer scientist **Zoran Popović, PhD**, and biochemist **David Baker, PhD**.

Researchers often must correct obvious errors in computer-folded proteins. FoldIt was developed to allow amateurs to spot and fix these computer inaccuracies. Players rack up points by pulling, wiggling, and tweaking a polypeptide sequence into the most chemically and physically accurate orientation. Most gameplay has concentrated on uncovering new folding algorithms, but FoldIt's current focus is producing player-designed proteins that can interact with particular biological targets, such as a small DNA strand. The game's creators recently released a puzzle asking players to generate a better design for human fibronectin, a protein used to mimic antibodies. One player modified fibronectin's peptide chain in a way that may turn out to be more stable than the original. Chemists at the University of Southern California are currently fabricating the novel structure for testing.

"FoldIt is a seminal and important project," says **David P. Anderson, PhD**, research scientist at the University of California, Berkeley Space Sciences Laboratory who created an online astronomy volunteer project called Stardust@home. But he encourages the team to focus more on hard scientific data in the future. "I hope they are able to quantify what they've actually done," he says.

Despite such concerns, Treuille thinks other researchers might imitate FoldIt's approach to computational analysis. "Everywhere you look in science there's labor that could use many people," he says. Treuille believes that similar projects could draw on the power of crowds while entertaining and educating the public.

—By Adam Mann □

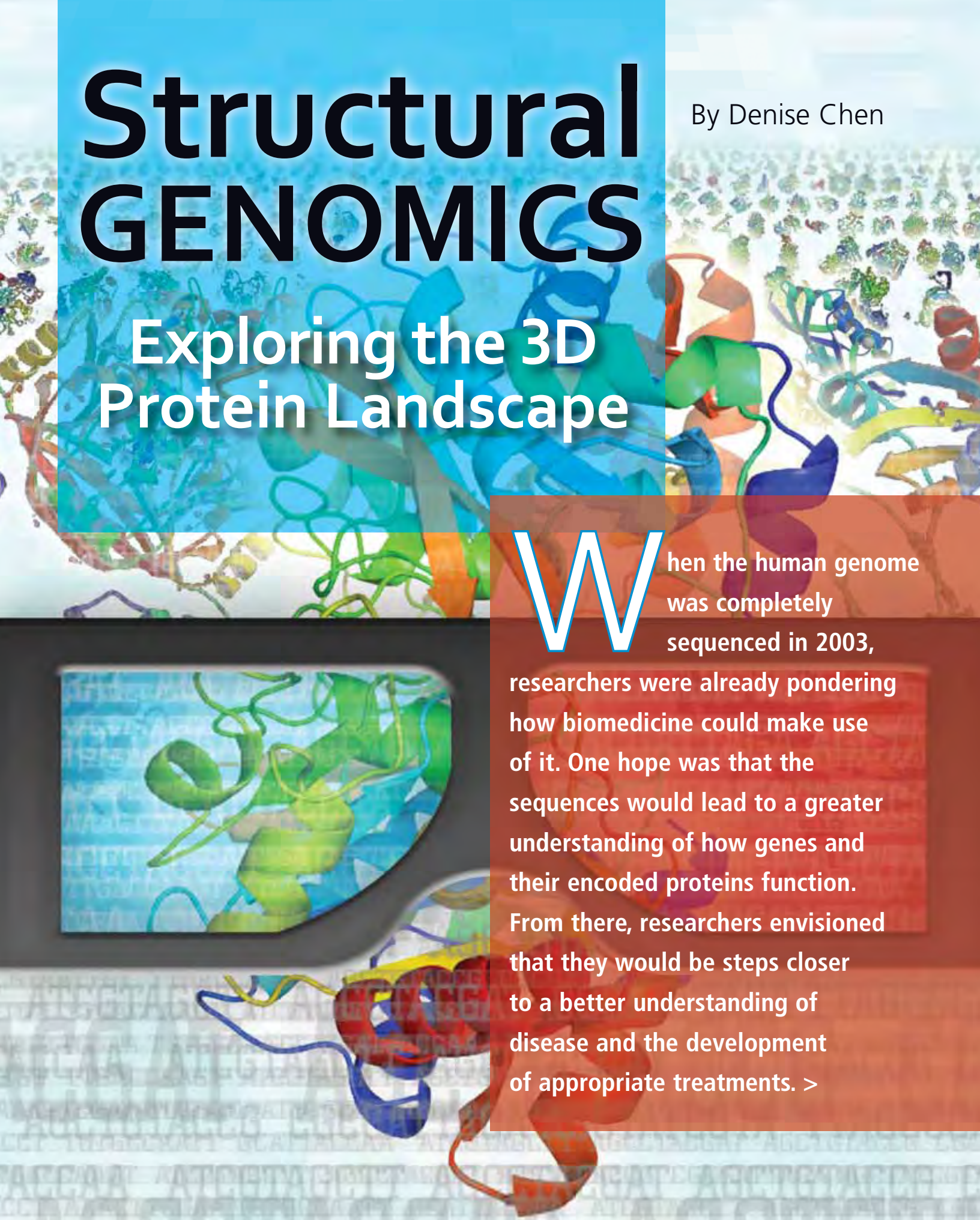


A team at the University of Washington designed the online game FoldIt to improve protein-folding algorithms. Players maneuver polypeptide chains, such as this 2HSH sequence, into their lowest energy configuration to get the highest score. Image courtesy of Seth Cooper at the University of Washington.

Structural GENOMICS

Exploring the 3D Protein Landscape

By Denise Chen



When the human genome was completely sequenced in 2003, researchers were already pondering how biomedicine could make use of it. One hope was that the sequences would lead to a greater understanding of how genes and their encoded proteins function. From there, researchers envisioned that they would be steps closer to a better understanding of disease and the development of appropriate treatments. >

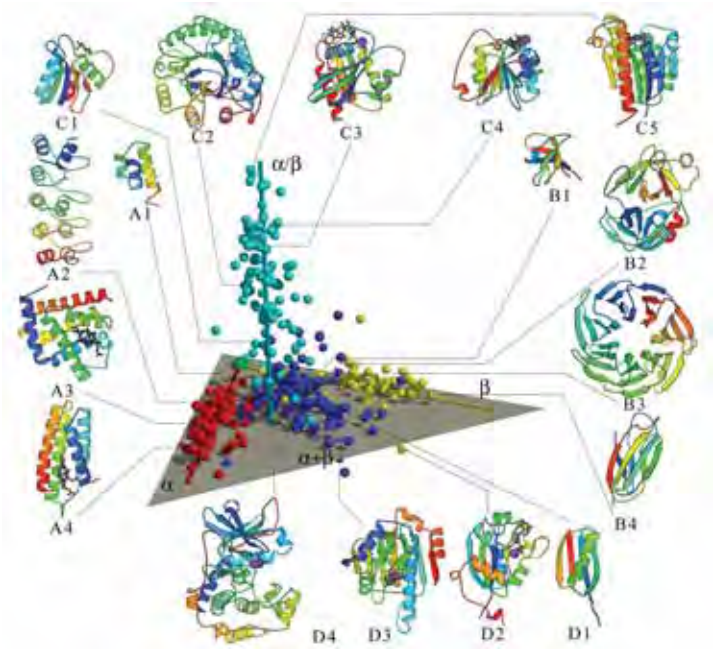
But a fuller understanding of proteins' functions within the human body depends on determining those proteins' structures. And as the number of known gene sequences grew, many scientists realized they could not catch up simply by determining protein structures one by one. So a group of scientists embarked on a strategic plan to uncover the three-dimensional structures of all the proteins that these genes encode.

This endeavor is called structural genomics. "The original question for which structural genomics came into being was: 'Can we translate the sequence of everything into the structure of everything?'" says **Peter Preusch, PhD**, acting director of the Protein Structure Initiative at the National Institute for General Medical Sciences (NIGMS).

The primary motivator of structural genomics is the sheer speed with which genomic sequence data is accumulating. Structural determination in traditional structural biology laboratories can't possibly keep up, researchers say. Fortunately, unlike sequences, which are nearly infinite in number, "there may be a finite number of different shapes that proteins actually adopt to perform their functions in the cell," says **Ian Wilson, DPhil**, professor of structural biology at The Scripps Research Institute and director of the Joint Center for Structural Genomics (JCSG).

In fact, a 1992 *Nature* paper estimated that the majority of proteins belong to no more than 1,000 families. Thus, researchers reasoned that it might be possible to unveil the universe of protein structures through a combination of experimental structure determination and computational structure prediction. And although upwards of 10,000 protein families have now been identified, uncovering the protein structure universe remains feasible.

In pursuit of this goal, ten years ago, NIGMS made a major investment to fund and spearhead a coordinated public structural genomics effort in the United States. Known as the Protein Structure Initiative (PSI), the program established four research centers and several specialized centers. The plan: to determine structures faster and cheaper; improve computational methods for predicting protein models; and ultimately develop innovative strategies for delivering useful structural information to the greater biological community.



Dimensions of the Protein Universe. Protein structures are displayed here along axes signifying secondary protein structure elements: strictly α helices or β sheets, both α and β, or combinations of α and β. The more complex and highly structured proteins reside at the extreme ends of the axes. In 1992, researchers estimated the number of protein families at around 1000, but the size of the protein universe has turned out to be much larger than predicted--as exemplified by the 23rd release of the Pfam database listing over 10,000 protein families. Source: NIGMS image gallery: http://images.nigms.nih.gov/index.cfm?event=viewDetail&imageID=2367. Courtesy of Berkeley Structural Genomics Center, PSI.

"The original question for which structural genomics came into being was: 'Can we translate the sequence of everything into the structure of everything?'" Preusch says.

In each of these areas, the PSI has made great strides. Before the PSI launched, determining the structure of a relatively complex protein was a major task, requiring the efforts of a graduate student for several months or even years, says **Keith Hodgson, PhD**, professor of chemistry at Stanford and head of the JCSG structure determination unit. Today, at each of the four main PSI centers, "a structure is turned out every few days," he says. At the same time, protein structure prediction helps fill in the gaps between known and unknown structures, bringing us closer to knowing the "structure of everything." This increased coverage of the structure space is transforming the field of biology, making it possible to assemble all of the structures in a particular pathway and visualize the interplay between them; or screen multiple structures to determine what they will bind; or carefully study the structures of proteins involved in disease.

Progressing through the Pipeline: From Sequence to Structure

Structure determination consists of multiple steps including cloning, expressing, and purifying a protein, finding appropriate conditions for crystallizing the protein, performing structural analy-

sis by techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, and analyzing the resulting data. This can be a long and arduous process requiring an enormous investment of time, labor, and money with no guarantee of success. Thus, a key initial goal of the PSI was to make this process more efficient.

In the PSI's first five years (2000-2005, often referred to as PSI I: Pilot Phase) four large-scale pilot centers were created: the Joint Center for Structural Genomics (JCSG), the Midwest Center for Structural Genomics (MCSG), the New York SGX Research Center for Structural Genomics (NYSGXRC), and the Northeast Center for Structural Genomics (NESG). These centers developed high-throughput X-ray crystallography and NMR spectroscopy pipelines. Using automation and robotics, they consolidated and refined all of the individual protein production and structure determination steps. "It really is like a pipeline where you start at one end with a sequence, and out of the end of that pipeline comes a three dimensional structure," Hodgson says.

During the PSI's second five years (2005-2010)—known as PSI 2: Production Phase—the centers' pipelines churned out large quantities of previously unknown protein structures. Six specialized centers were also established to focus on the structural determination of proteins that are technically trickier to crystallize, such as membrane proteins, protein complexes, and eukaryotic proteins. The goal: to cover the "structure space."

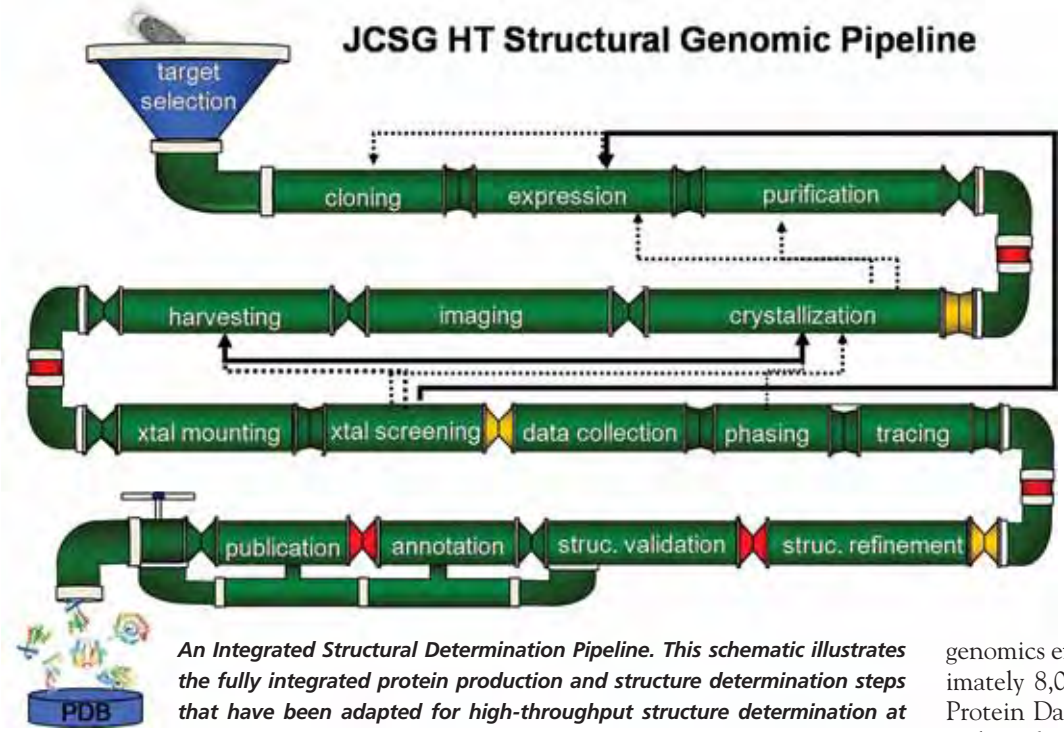
In picking protein targets, the PSI sought to complement what others were doing. "We've been trying to look into areas where nobody's really looking yet," Wilson says.

In picking protein targets, the PSI sought to complement what others were doing. "We've been trying to look into areas where nobody's really looking yet," Wilson says. Thus, novel protein targets that share less than 30 percent sequence identity with proteins of known structure comprise 70 percent of the focus at each center; the remaining targets are proteins deemed important by the biological research community.

To generate a target list of novel proteins, the PSI bioinformatics group used the publicly available Pfam (protein families) database and other automated protocols. "You've got to use informatics and computational biology to do structural genomics—first of all, to pick the right targets," says **John Norvell, PhD**, former director of the PSI at NIGMS. Pfam groups protein families by functional domains found within protein sequences. It uses protein sequence alignment entries and hidden Markov models to probabilistically determine how well a particular protein sequence matches with known families. In this way, PSI researchers identified protein sequences belonging to families with little structural information and targeted those for structural determination.

The approach has paid off. Working together to conquer the list of target proteins, the PSI centers reached their goal of solving more than 3,000 novel structures during PSI 2. In fact, over the last ten years, worldwide structural genomics efforts have deposited approximately 8,000 protein structures in the Protein Data Bank (PDB), the primary archive for structural biology.

Structure determination using PSI pipelines can now even be done by offsite researchers from the general science community. "You can run the synchrotron beamline, collect the data,



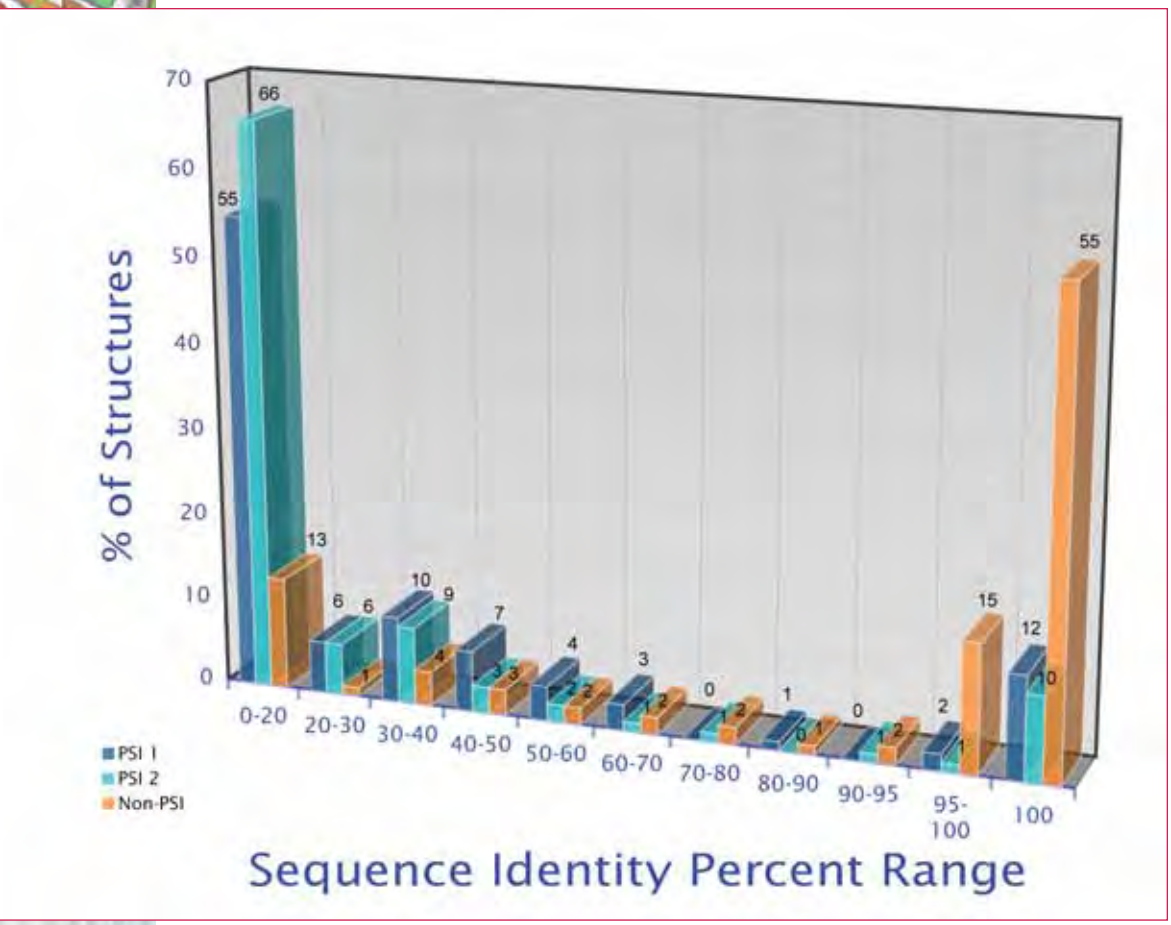
An Integrated Structural Determination Pipeline. This schematic illustrates the fully integrated protein production and structure determination steps that have been adapted for high-throughput structure determination at the PSI large-scale centers. All the steps of the pipeline are tied together by a common bioinformatics framework that enables feedback. For example, if the crystal screening step cannot identify usable crystals for structure determination, this information will be communicated to an earlier stage such as the crystallization step, where appropriate modifications will be made that help increase the procedure's likelihood of success. Courtesy of Marc Elsliger and Ian Wilson, JCSG.

"Ten years ago, there were people who would almost remember all the structures in the PDB. … But we're exactly in this stage where this type of old style analysis is no longer sustainable," Godzik says.

carry out the whole experiment from your own laboratory sitting in front of your own desktop or laptop," Hodgson says. "All you have to do is get your samples here and you can get FedEx to do that for you."

Where will scientists go from here? "Ten years ago, there were people who would almost remember all the structures in the PDB," says **Adam Godzik, PhD**, associate professor of bioinformatics at the Burnham Institute for Medical Research in La Jolla, California. "At some point this breaks down—you can memorize 300 structures, 500 structures, but you can't memorize 50,000. We're exactly in this stage where this type of old style analysis is no longer sustainable."

"What is still lacking are tools and, in some sense, even concepts of how to analyze large numbers of structures," Godzik says. Out of the small handful of structural alignment programs for doing just that, the Godzik group has written two of them, including the Flexible structure AlignmenT by Chaining Aligned Fragment Pairs with Twists (FATCAT) method. FATCAT improves upon its predecessor by accounting for structure flexibility and rearrangements. However, Godzik says, the structure alignment field is still very young and many concepts remain to be refined.



The Focus on Novel Families. This graph shows the percentage of structures contributed to the PDB by the PSI and other sources from 2000 to 2008. Each group of three bars represents how similar the sequences of the new contributions are to known structures. As shown here, the PSI determined the structures for novel protein families at a far greater rate than did other researchers during this time period. For example, of the total PSI deposits during this time (divided into the PSI1 and PSI2 phases, represented by the blue and aqua bars, respectively), most share less than 30 percent sequence identity with known structures (leftmost bars). On the other hand, over half of non-PSI deposits (represented by the orange bars) during this same timeframe had 100 percent sequence identity with known structures (rightmost bars). From "Investigators' White Paper" from the Future Structural Genomics Initiatives meeting held by NIGMS in October 2008. Courtesy of Peter Preusch.

Expanding the Protein Universe: Improved Tools for Structure Prediction

While structure determination has been evolving, so too has a complementary field: structure prediction. "To make a real impact, you've got to pick the right targets and then use modeling to expand the structural information to many more sequences," says Norvell. Thus, structural genomics can leverage structure prediction to help fill in the gaps.

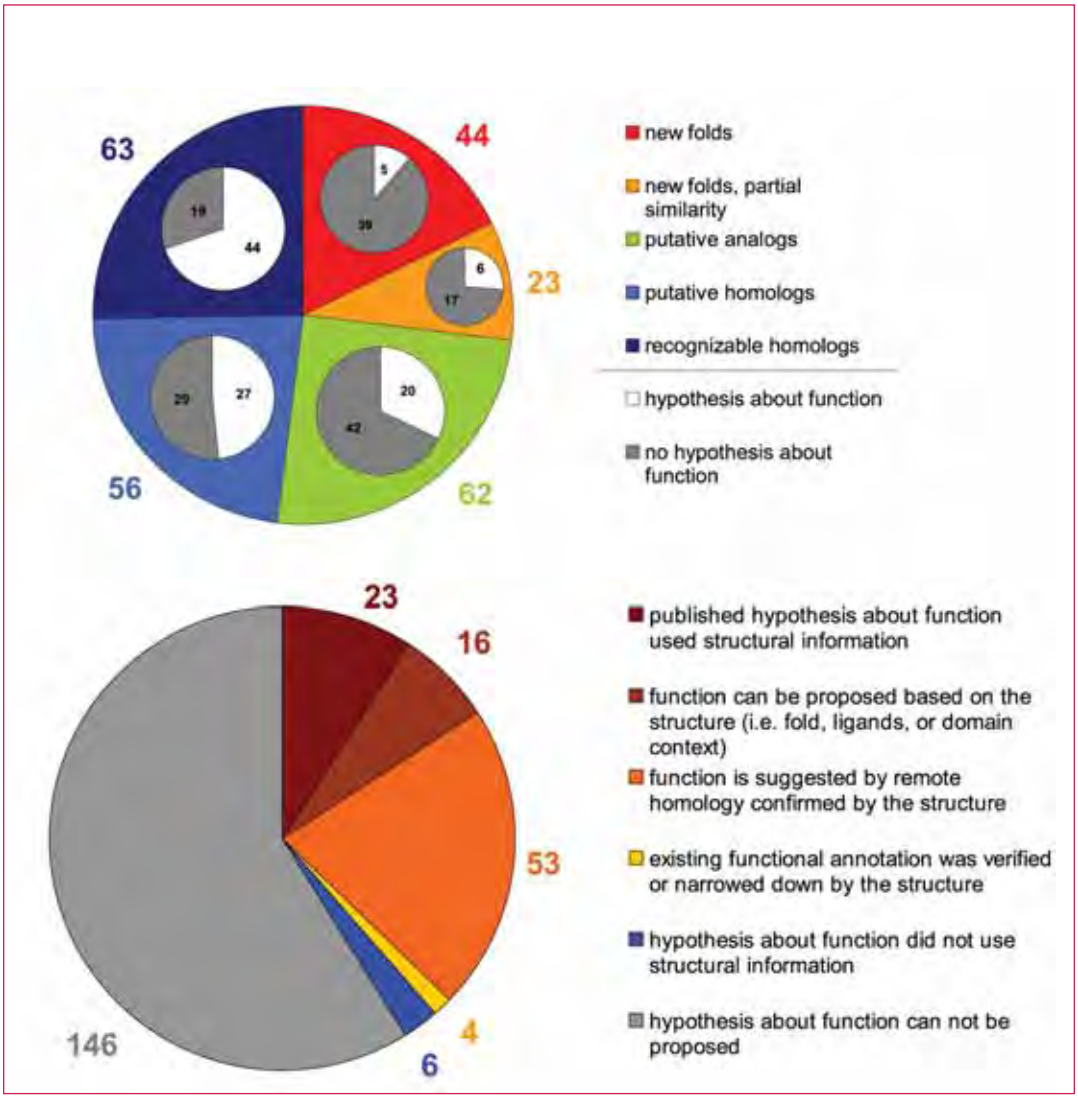
Structure prediction aims to accurately predict protein structures directly from their primary sequences, without wet lab experimentation. This prediction can be done by taking a

"knowledge-based" approach which gathers hints from known structures used as templates or a "physics-based" approach which starts from scratch using first principles to explore the possibilities of protein folds. The knowledge-based approach, also called homology modeling, is essentially "designing new buildings as better old buildings," says **Michael Levitt, PhD**, professor and chair of computational structural biology at Stanford. "The idea is that it has worked, so you can reuse it in a different combination."

High-throughput structure determination efforts have increased the number of known protein folds in sequence alignment databases, making it more likely that a protein with unknown structure will produce matches with sequences of known proteins that can serve as a template to then predict higher quality structures. Thus, structural genomics efforts contribute to homology modeling. "You're running on the same computers, same codes, but the database on which it runs is much larger now," says **Nir Kalisman, PhD**, a postdoctoral researcher of structural biology and computer science at Stanford.

In turn, structural genomics has benefited from structure prediction efforts, which leverage known structural information to fill in gaps in the structure space. "The main idea is that we really can get large scale coverage of all the structure space by sampling strategically, getting experimental structures of particular representatives, and then modeling around that using homology modeling techniques," says **John Moult, DPhil**, professor at the University of Maryland Biotechnology Institute.

Currently, scientists are able to learn from structures generated from the best of both the structure prediction and the structure determination worlds. "For any structure that's determined using X-ray crystallography or NMR, the model that you get is very highly reliable, the gold standard," says **Helen Berman, PhD**, professor of chemistry at Rutgers University and director of the PDB. Homology modeling, on the other hand, might be less certain, but still provides useful information, she says. Moult agrees: "A rough structural



Illuminating Protein Function via Structure. The Pfam database currently contains 2,247 families of "hypothetical proteins"—proteins with unknown functions or that are uncharacterized. In a 2009* PLOS Biology *paper, researchers looked at 248 of these families that were solved by the PSI to better understand regions of the yet unexplored protein universe that these families represent. The top pie chart breaks down the hypothetical proteins into subgroups based on their structural similarity and homology to known structures, ranging from proteins composed of new folds (red slice) to proteins with recognizable homology to known structures (dark blue slice). Within each of the five slices are mini pie charts showing the percentage of structures within each category for which hypotheses about their functions exist (white). What emerges is a relationship between structural similarity and homology and hypotheses about function: the greater the degree of structural similarity and homology to known structures, the more likely a functional hypothesis can be formed for that protein family. The lower pie chart further demonstrates that known structural information can facilitate inferences of function. From Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, et al. 2009 Exploration of Uncharted Regions of the Protein Universe.* PLoS Biology *7(9): e1000205. doi:10.1371/journal.pbio.1000205.

"To make a real impact, you've got to pick the right targets and then use modeling to expand the structural information to many more sequences," Norvell says.

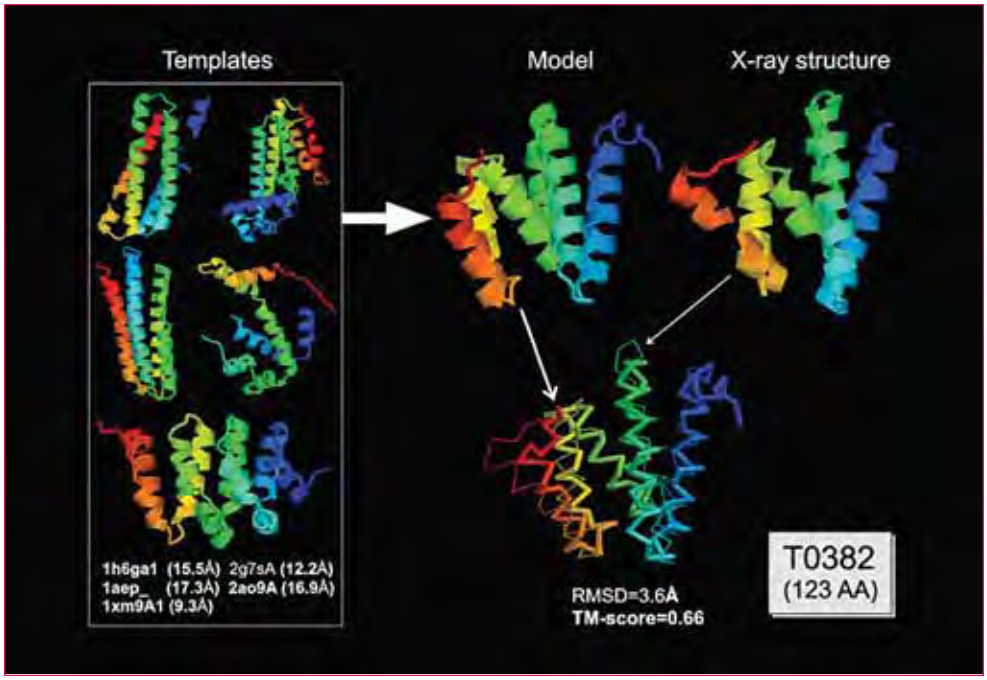
model based on the distant relationship by homology will be enough to give you some idea—albeit a crude low resolution idea—about function." And of all the publicly available gene sequences in GenBank, Moult estimates that more than 50 percent could be modeled at some rough level.

Yet Moult believes that for something like structure-based drug design, a very atomically detailed protein model is still required. Improving structure prediction to that level of precision will require further advances in computational methods and a better understanding of physical chemistry, he says.

On the other hand, a recent study by **Andrej Sali, PhD**, professor of bioengineering and therapeutic sciences at the University of California, San Francisco, and his colleagues illustrated that homology modeled proteins do nearly as well as X-ray crystal structures at deducing proteins' functions.

In order to directly evaluate the strength of current methods, the structure prediction community holds a competition every two years—the Critical Assessment of Techniques for Protein Structure Prediction (CASP). "CASP gives an objective way for many groups and methods to be compared on a level field," Kalisman says. Since 1998, the top performing modeling tool in CASP is ROSETTA, developed by **David Baker, PhD**, professor of biochemistry at the University of Washington. ROSETTA uses a fragment assembly method, taking short fragments from existing protein structures as guidance for modeling an unknown structure. The structures produced by ROSETTA get closer and closer to matching crystal structures all the time.

But another barrier to structure prediction remains: a cultural one. Biologists who work with structures want to know how reliable a predicted structure is—and that's often unknown, or at least unstated. Additionally, many structure prediction programs are large software packages that require a lot of computing power and are not accessible by the non-structural biologist. It's a problem, Kalisman says, because there's little outreach from the structure prediction folks to the biology community. "Biologists could definitely benefit much more if there was a better interface between most structure prediction algorithms and how biologists can approach them."



Fragment Assembly Algorithms for Structural Prediction. At the two most recent structure prediction competitions (CASP7 and CASP8), an algorithm called I-Tasser ranked as overall winner and outperformed human expert groups. Developed by Yang Zhang, PhD, associate professor of computational medicine and bioinformatics at the University of Michigan, I-Tasser uses fragment assembly as one step in a three-step procedure to model an unknown protein structure. In this example, I-TASSER used multiple algorithms to generate five templates (left panel) with secondary protein structure elements that best matched the query protein sequence T0382. These templates were then reassembled and refined to produce a structural model that only deviated from the experimental X-ray structure by 3.6 Angstroms. Reprinted with permission from Wiley Publishers, from Zhang, Y., Template-based modeling and free modeling by I-TASSER in CASP7 (2007). **Proteins** ***69(Suppl 8):108-17 (2007).***

MOVING BEYOND THE PIPELINE: THE PSI'S SPEED AND PRODUCTIVITY MAKE A DIFFERENCE FOR BIOMEDICINE

By increasing the number of available protein structures at a rate much faster than previously possible, PSI leaders believe structural genomics will hasten research advances in many areas of biomedicine. Indeed, there are signs that this is already happening.

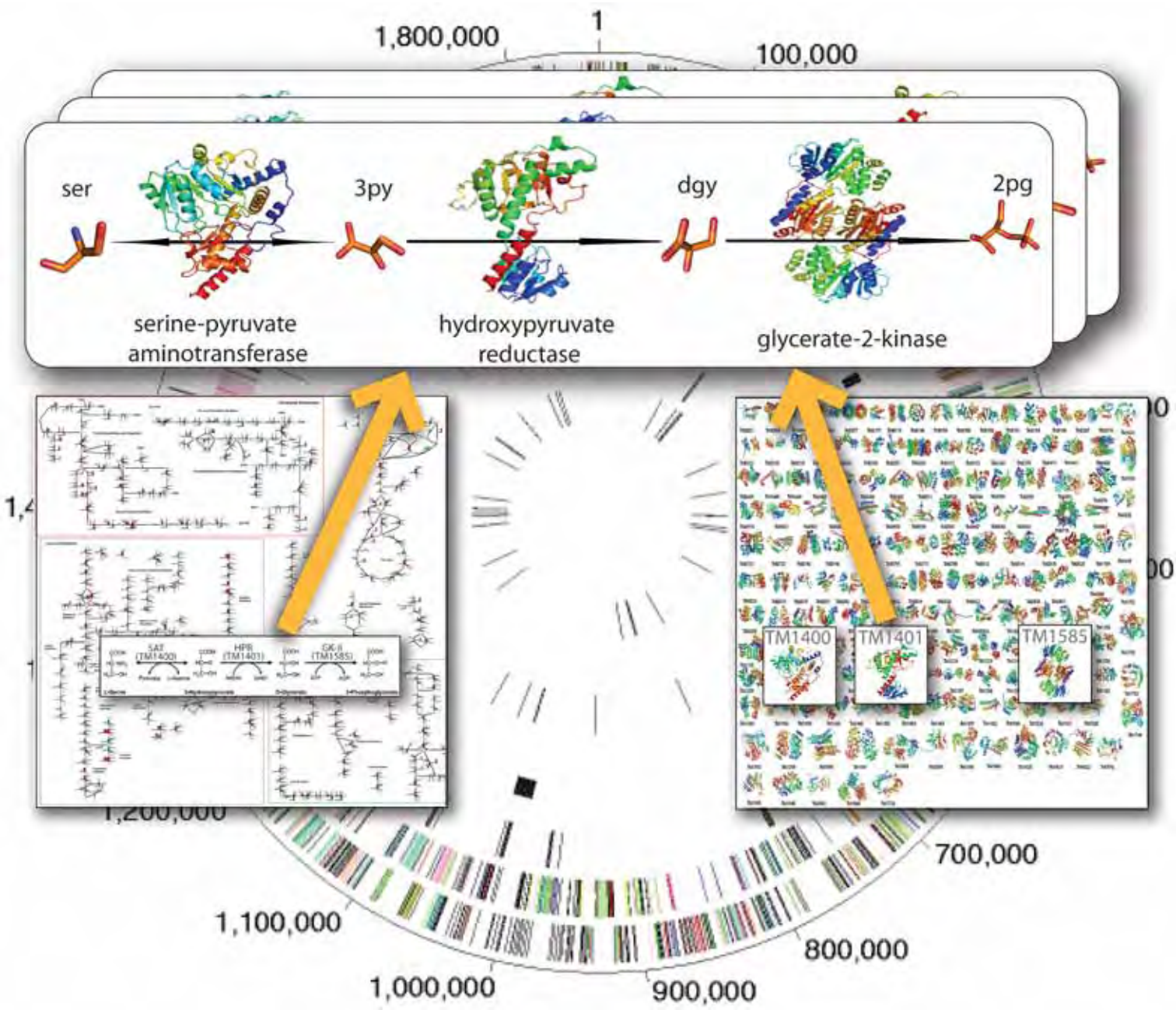
In a large collaborative project between one PSI center (NYSGXRC) and the Enzyme Specificity Consortium (ENSPEC), researchers proactively selected 535 proteins for structure determination. The proteins came from two structurally similar protein families (the amidohydrolase and enolase protein families) that catalyze a broad set of chemical reactions. To date, the NYSGXRC has completed X-ray crystallography for 75 of these proteins and modeled many more. To demonstrate the potential utility of these structures, the researchers performed *in silico* docking on one of them—the *Thermatoga maritima* amidohydrolase enzyme, Tm0936—to determine the enzyme's function, which was previously unknown. In the work, published in a 2007 *Nature* paper, thousands of configurations and conformations of molecules were docked into Tm0936 and ranked for fit. The top ranking compounds indicated that Tm0936 bound to and modified the structure of adenosine. The researchers then determined the crystal structure of Tm0936 in complex with one of the top ranking compounds and found only minor differences from the prediction, confirming its function. This is one example of how new approaches, in combination with the wealth of information from structural genomics, can lead to new insights.

Another example of the PSI's impact is the JCSG's human gut microbiome project, which focuses on impor-

tant pathways relevant to human health. Scientists hope that a better structural understanding of the proteins found in bacteria that populate the human gut will lead to the development of targeted drugs and therapies for human diseases. Even the study of proteins from deep ocean vents—far afield from the human gut—has the potential to aid in treatment of disease. For example, classical thymidylate synthase (TS) plays an important role in DNA synthesis and repair—and has been targeted in chemotherapy treatments for cancer. But the version of TS in *T. maritima* bacteria that live in thermal vents in the ocean turns out to have a completely different fold. Indeed, this enzyme has a different functional mechanism and has now been found in some pathogenic bacteria. The researchers suggest that a drug targeting the bacterial protein could prove to be a safe antibiotic because the human version is not homologous.

MAKING USE OF THE FOREST OF STRUCTURES TO ADVANCE BIOMEDICINE

Even as scientists have begun to capitalize on the large numbers of available structures, structural genomics researchers hope to take



Shaping a Metabolic Network. Godzik and his colleagues used experimentally determined and computationally modeled protein structures to reconstruct the central metabolic network of the bacterium **Thermatoga maritima.** ***By including the three-dimensional structures (lower right panel) of the proteins involved in the central biochemical reactions (lower left panel), they discovered a strong degree of conservation in protein folds that compose enzymes involved in similar reactions (top panel). From Ying Zhang and Adam Godzik, Burnham Institute. From Zhang, Y, et al., Three-Dimensional Structural View of the Central Metabolic Network of*** **Thermotoga maritima,** **Science** ***325:1544-1549 (2009), reprinted with permission from AAAS.***

things to a new level altogether.

"Classical structural biology focuses on individual proteins, so it's sort of looking at each tree separately," Godzik says. "Through changes in scale, what this becomes is looking at a forest—you suddenly see all the structures together and you start analyzing and comparing large groups of structures."

In recent work published in the September 18, 2009, issue of *Science*, Godzik and colleagues took the first leap in this direction. They constructed a comprehensive model of the metabolic network of thermophilic bacterium *T. maritima* that includes all the three-dimensional protein structures. For the first time, Godzik says, "we have a huge biological network which can be simulated and viewed as a mini cell *in silico*."

To build the model, Godzik's team had to first identify all the proteins in the metabolic network by extracting relevant information from more than 150 publications, and then subjecting that list to *in silico* analyses to identify gaps or redundancies that had to be resolved manually. Of the complete set of 478 proteins in the *T. maritima* metabolic network, 120 structures had been experimentally determined in part by the PSI JCSG. Using homology modeling, among other computational techniques, the researchers predicted the structures of the remaining 358 proteins. Standard methods produced the structures of 95 percent of these proteins, but the last few percent took a lot of effort, Godzik says. "Getting to 100 percent coverage was a huge challenge." And the quality of the predicted structures varied. For example, about 190 were comparable to low-resolution, experimental structures, while about 52 were merely approximate and others were somewhere in between.

But that effort proved worthwhile, Godzik says, because it led to a number of insights, perhaps most significantly, into the evolution of protein structures and organisms. The model they had constructed demonstrated that a small number of folds are represented in a majority of the proteins involved in the metabolic reactions of *T. maritima*. In fact, of the 478 proteins, including a total of 714 domains, there were only 182 distinct folds. And proteins involved in similar biochemical reactions have a higher probability of adopting similar folds. All of this supports the idea of structural conservation in nature, and to a much larger degree than researchers expected.

With this project, researchers also challenged the conventional thinking that accompanies structure determination. "When we first submitted our paper, the first question that came from the editor was 'If this is a structural biology paper, what is the main structure you're talking about?' And we said, 'Well, there's no main structure; there are 478 main structures,'" Godzik says. "Both technological and conceptual changes are what structural genomics has brought to the table."

THE NEXT CHAPTER OF STRUCTURAL GENOMICS: STEPPING OUT INTO THE PUBLIC

In 2008, the Structural Genomics Knowledgebase (PSI SGKB) (http://kb.psi-structuralgenomics.org) was launched to integrate all the results from the PSI and make them available to the public along with an array of technology, protocols, and software. "The PDB has the structures. The SGKB has the structures and the sequences and the functional annotation and different technologies that allow you to get the structures," Berman says. "You have everything where you can find it in order to begin making new hypotheses and gaining new understanding."

The launching of SGKB signifies an important shift in the evolution of the PSI, says **Emily Carlson** of the NIGMS Office of Communications and Public Liaison. "It's gone from being a group of grants to being an actual research network where the researchers are sharing information and they're collaborating in ways that hadn't been done before. Not just within the PSI, but within the field and community in general."

By encouraging public access to solved protein structures and providing over 150 different resources at the PSI SGKB, the structural genomics community is showing its commitment to transforming structural data into meaningful information of use to the greater biological community, says **Michael Sykes, PhD**, postdoctoral researcher at The Scripps Research Institute. "It is not sufficient to determine structure for structure's sake. The scientific community needs to use these structures to make inroads into understanding the fundamental principles of biology."

The coverage of "structure space" will continue to be an aim of structural genomics, but the next phase—called PSI Biology instead of PSI 3—is shifting directions. The aim: To bring structure and function studies back together again and to connect biologists with the PSI effort, Preusch says. "The new thing is partnerships. We want to bring in people who have a biological problem of significant scope for which solving a large number of protein structures is necessary to really move the problem forward." □

"Classical structural biology focuses on individual proteins, so it's sort of looking at each tree separately," Godzik says. "Through changes in scale, what this becomes is looking at a forest—you suddenly see all the structures together and you start analyzing and comparing large groups of structures."

By Katharine Miller

Clinical Decision Support

Providing Quality Healthcare With Help From a Computer

In a classic cartoon, a physician offers a second opinion from his computer. The patient looks horrified: How absurd to think that a computer could have better judgment than a human doctor! But computer tools can already provide valuable information to help human doctors make better decisions. And there is good reason to wish such tools were broadly available. >

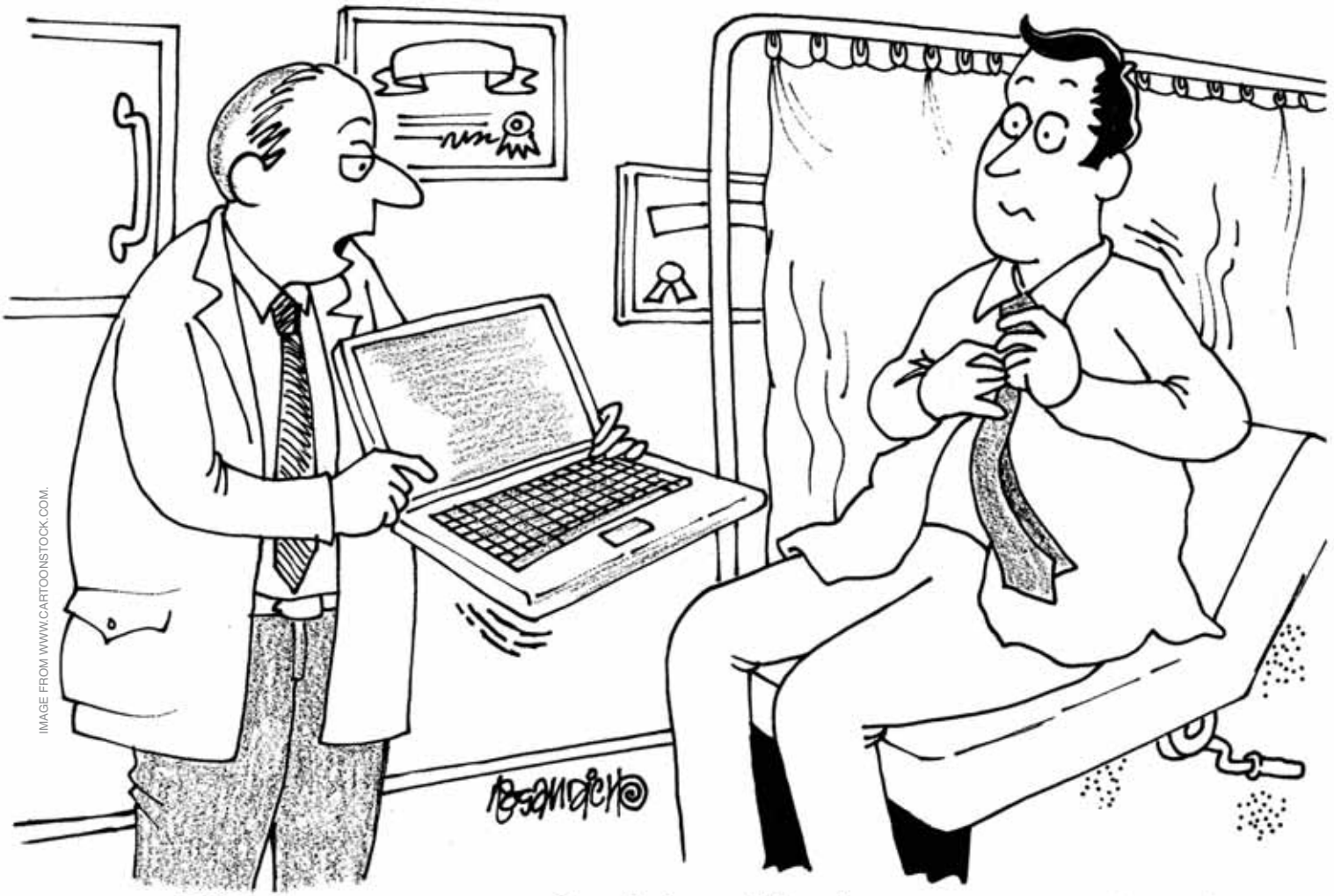


IMAGE FROM WWW.CARTOONSTOCK.COM.

"If you want a second opinion, I'll ask my computer."

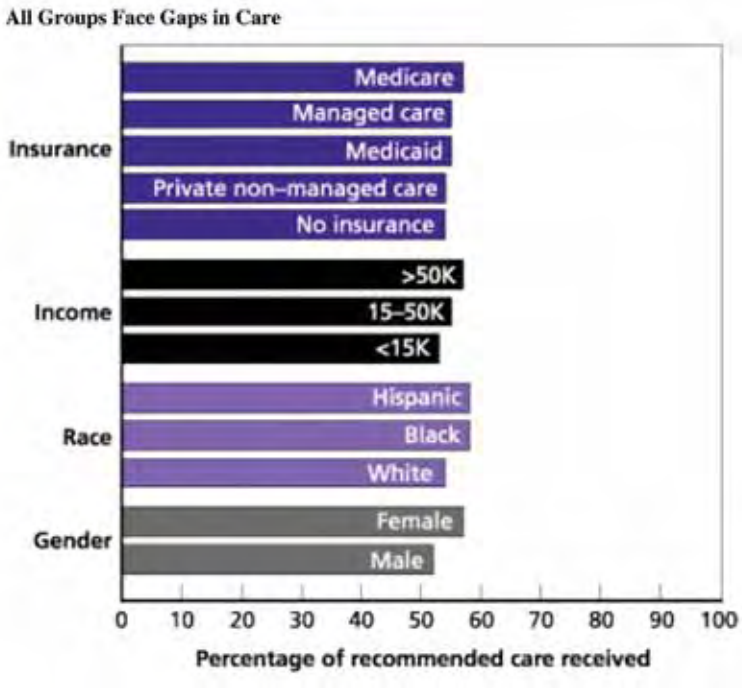
About half of the time, doctors fall short of providing quality medical care as defined by national guidelines, according to a 2003 paper in the *New England Journal of Medicine*. In addition, patients leave their doctors' visits with an average of 1.6 unanswered questions. "That's too many," says **Blackford Middleton, MD**, assistant professor of medicine at the Harvard Medical School and corporate director of clinical and informatics research and development at Partners Healthcare System in Boston. And because medical professionals have incomplete knowledge or incomplete information about a patient, "we order too many tests, patients are called back, and sometimes bad things happen," Middleton says. "It's embarrassing. That's why I get up every day and run to work."

Why the hurry? Middleton and his colleagues are trying to build a safer, higher quality, and lower cost health care system right now. And one way to do that, he says, is through well-designed clinical decision support (CDS) systems connected to a nationwide knowledge base of best medical evidence.

Plenty of doctors are dubious about the value of CDS systems. They say they don't need it; that they are experts in their fields; that they know their patients well, or that their practice is of high quality. "Regrettably, that's not supported by the evidence," Middleton says.

Some say it's also becoming humanly impossible to provide best-evidence medicine without computer support. "There's too much to know, too much to do; people are overwhelmed," says **John Fox, PhD**, a professor in the department of engineering science at the University of Oxford in the United Kingdom.

And the problem is only going to get worse. In addition to staying on top of the 6 million pages of research published in books and journals each year, physicians may soon have to keep track of several hundred thousand genetic variants that could become relevant in medical practice, Middleton says. "It will be impossible for the unaided mind to compute what to do for the patient sitting before you," he says. Health care providers will need decision support tools.



According to a 2006 Rand report, overall, adults receive about half of the recommended care they should get. These findings were based on a quality score assigned to each patient based on the number of times in a two-year period that the patient received the care recommended across all of the conditions the patient had, divided by the number of times the patient was determined to need specific health care interventions. As shown here, the findings were equally true regardless of gender, race, or income level. Reprinted with permission from* The First National Report Card on Quality of Health Care in America, *RAND Research Brief RB-9053-2 (2006).

Today, such tools range from simple alerts, to computerized guidelines that provide recommendations based on electronically stored patient data, to systems that visualize patient data over time or over entire patient populations. Some of these are well-established in various medical institutions around the country; and some are being developed in academia. Others are making the transition to the commercial arena.

So how do we get from where we are now to an efficient national decision support system? It will require incentives for physicians and hospitals to install electronic medical record systems—with both carrots and sticks, Middleton says: carrots to get them to purchase the systems and sticks to make sure care actually improves. It will require a shift in our understanding of what a doctor does—"instead of the final authority, the doctor should be seen as a knowledge manager who helps the patient make the right decision using modern computing tools and decision aids," Middleton says. It will require a better understanding of what decision support can and can't do—from the simplest rules to the most complex algorithms—and a way to determine the safety of the system itself. It will require taking the best and most effective academic efforts and bringing them into the commercial arena. And, Middleton says, it will require a national knowledge repository accessible through patients' electronic records, perhaps as a public web service.

It won't happen overnight. But it's enough within reach that Middleton keeps running to work in the morning.

Building the Plumbing and Standardizing the Data

Only 15 to 20 percent of physician offices and hospitals in the United States use electronic medical records (EMRs), but that will soon change: the stimulus package passed by Congress in early 2009 is investing nearly $20 billion to incentivize physicians to install EMR systems. "This is (hopefully) a one time frame-shift payment large enough to allow us to wire the country," Middleton says.

With hundreds of EMRs to choose from, one question is how to ensure that physicians purchase systems that are interoperable. So, as part of wiring the country, the Office of the National Coordinator for Health Information Technology (ONCHIT) is also pushing for data standardization.

"The missing piece is rigorous standardization of the data points that might be used from an electronic medical record for decision support," says **Clem McDonald, MD**, a CDS pioneer now at the National Library of Medicine within the National Institutes of Health.

Data can be recorded in ways that confound a computer. For example, to remind a patient to get a mammogram, the computer needs to know when she last had one. But perhaps she had it at a different site, so the system doesn't know about it. Or perhaps the physician entered the response to a yes/no question ("Have you had a mammogram in the last year?"), in which case the computer still doesn't know when the reminder should be sent. Or perhaps the EMR contains the patient's mammograms but can't easily determine which of many X-ray records actually represents the most recent mammogram.

Or suppose a health care institution wants to contact all of its diabetic patients to inform them of a new treatment option. How does the computer know if a patient is diabetic? Maybe the EMR says so, but perhaps not. So maybe the system looks at whether a patient is on insulin—but there are multiple codes for different types of insulin. Or maybe it looks at lab tests to see if the person is hypoglycemic—but that could be a temporary occurrence.

"Humans will usually make sense of things even if the data is not standardized," McDonald says. "Whereas the computer has to be able to get at the data." The problem lies at the interface between the core logic and the facts that feed it—which varies a lot depending on the context and the hospital, McDonald says.

McDonald is hopeful that ONCHIT's push toward standardization will help. "Everyone invents things in their own way," McDonald says. "Once that's solved, all the rest of it gets easier."

"Humans will usually make sense of things even if the data is not standardized," McDonald says. "Whereas the computer has to be able to get at the data."

Smart Alerts—Getting Beyond Simple Rules

Decision support is not a new idea. Indeed, says **Jonathan Teich, MD, PhD**, chief medical informatics officer for the health sciences division at the publisher Elsevier, "We've already gone through the hype cycle." Several big studies in the 1990s showed that CDS can prevent medical errors in various circumstances. And, after a 2000 Institute of Medicine report documented the wide extent of medical errors ("To Err is Human"), electronic health record companies created computerized ordering systems and alerts—pop-up windows or alarms that signal the physician should think twice before taking a specified action. "So part of CDS has become mainstream," Teich says.

"We have really great technology for simple rules," says **Mark Musen, MD, PhD**, professor of medicine at Stanford University. "Making a situation-action rule that says 'if the patient has a penicillin allergy, don't give penicillin' is easy." The problem is getting beyond these simple rules, he says.

Simplistic alerts are among the most annoying types of CDS to implement, Teich says. They can lead to "alert fatigue," where doctors start ignoring alerts because they receive too many of them.

But some institutions around the country are taking alerts to the next level. For example, Intermountain Health in Utah developed a system that helps doctors determine the right dose of the right antibiotic. "It's a complicated space," Musen says. "You have the patient; the bugs in the environment (and their drug sensitivities); renal function; liver function; severity of illness; and all of these play into what antibiotic you use in a given context." It's a lot of data to be looking at,

"Making a situation-action rule that says 'if the patient has a penicillin allergy, don't give penicillin' is easy." The problem is getting beyond these simple rules," Musen says.

yet the system summarizes that data and proposes candidate drugs. "That's a more advanced system that's out and looks very promising," Musen says.

The Intermountain Health system also functions in the background, monitoring patients continuously, says **R. Scott Evans, PhD**, senior medical informaticist in the department of medical informatics at Intermountain Healthcare and professor of biomedical informatics at the University of Utah. It evaluates every new piece of data that comes into a patient's EMR to determine whether a staff member should receive an email message or page notifying them about it. The system also compiles a report of all "reportable" infections and sends it off to Utah's public health department. And it monitors for adverse drug events, so that if a nurse records a rash or hives, or a lab result indicates a doubling of creatinine, an indicator of kidney function, an alert goes to the appropriate pharmacist suggesting that perhaps the patient should be checked for a drug allergy.

And alert fatigue has not been a real problem, Evans says. "Intermountain is very careful to only create alerts for high priority problems or situations in which the potential harm is large. In those situations, healthcare providers tolerate false positives more than they would otherwise." Intermountain also makes sure that alert emails include pertinent data in them, so healthcare providers can determine if something is happening or not. For doctors, Evans says, "the worst thing that can happen is 'I wish I'd known earlier.'"

Initially, some doctors resist the decision support system, Evans says, "until it provides them with some information that prevents harm. Then they become advocates for it." It helps that the system has been evolving over 30 years. "The alerts that stay in place are those that the staff and docs really want," he says.

Initially, some doctors resist the decision support system, Evans says, "until it provides them with some information that prevents harm. Then they become advocates for it."

And the system has been good for patient care. Adverse drug events are down, and pre-operative antibiotics are delivered at the right time.

Intermountain also added alerts to ventilators and IV pumps. For example, if there's a problem and no one responds within 10 seconds, the alert takes control of every computer monitor in that division, showing the room number and sounding an unmistakable audible alarm. It has been a huge success, with few patients disconnected longer than a minute. Previously, situations such as a patient's hospital room door being closed during a late night shift with few staff around could result in long-term harm to the patient because of the delayed response to an alert. "That no longer happens now," Evans says.

Bed	Patient name	Age	LOS	Orders			Vent	SBT Screen	SBT Trial	DVT	SUP	RASS Ord.	RASS Pt.	HOB	swab	feeth	hyS'x
3002B	T, V W	72y	6 d		flowsheet	MAR	v	F		v	v	-4	-4	30			
3003X	N, D	60y	17 d		flowsheet	MAR	v	F		v	v	0	-2	45			
3004B	T, P L	64y	34 d		flowsheet	MAR	v			v	v	-1	-1	30			
3005A	C, D E	61y	7 d		flowsheet	MAR				v	v	0	-1	30	v	v	
3005B	B, J	66y	7 d		flowsheet	MAR	v	F		v	v	-1	-3	30			
3006X	W, A A	20y	66 d		flowsheet	MAR	v			v	v	-1	-2	30			
3007X	W, L E	49y	9:14		flowsheet	MAR				v		0	-1	30			
3008X	P, J L	69y	50 d		flowsheet	MAR	v	F		v	v	0	0	30			
3009X	R, C	72y	15 d		flowsheet	MAR	v	F		v	v	-1	-2	30			
3011A	P, J E	88y	9 d		flowsheet	MAR				v	v	0	0	45	v	v	
3011C	J, W D	69y	2 d		flowsheet	MAR				v	v	0	-1	30			
3011D	P, P J	55y	10 d		flowsheet	MAR	v	P	P	v	v	0	-3	30			
3011E	R, R E	74y	9 d		flowsheet	MAR				v	v	0	0		v	v	
3011F	N, E Y	55y	3 d		flowsheet	MAR				v	v	-1	0	30	v	v	
3012A	S, J D	56y	14 d		flowsheet	MAR	v	F		v	v	0	0	30			
3012B	R, M	63y	10 d		flowsheet	MAR	v	F		v	v	-2	-2	30			
3013A	N, B D	60y	8 d		flowsheet	MAR	v	F		v	v	-3	-2	30			
3013B	H, S M	66y	16 d		flowsheet	MAR				v	v	0	-1	30	v	v	

Vanderbilt's Process-control dashboard shows real-time feedback for ventilator management in the ICU. For each task the staff must perform for each patient, the dashboard shows a green, yellow or red light indicating whether the task was performed on time according to the guidelines. Source: Stead and Starmer, Beyond Expert-based Practice. Pp. 94–105 in* Evidence-based Medicine and the Changing Nature of Health Care. *2007 IOM Annual Meeting Summary, Washington, D.C.: National Academies Press (2008). Reproduced with permission from Bill Stead.

Another example of a CDS system that's gone beyond simple alerts and has helped save lives is one developed at Vanderbilt University. As nurses, therapists and doctors follow a standardized guideline for ventilator management in the intensive care unit, the status of each patient against the plan shows up on a "dashboard" screen. For each action that has to be taken, the dashboard displays a red, yellow, or green light indicating whether the plan is on track, is in need of attention, or is off track. No actions are lumped together. This provides everyone with a real-time updated measure of how well the team is performing on the guideline. "The thing that makes the difference, is visualizing any gap while the team still has time to take corrective action," says **Bill Stead**, **MD**, associate vice chancellor for strategy/transformation and chief information officer at Vanderbilt University Medical Center.

Identifying the Right Thing to Do: Making Clinical Guidelines Computable

Handling Complexity

Many medical guidelines are too complex for simple rules to handle. That's particularly obvious in the area of chronic disease, where a condition evolves over a period of time and might eventually involve multiple diseases co-occurring. Such complex guidelines necessitate a more complex clinical decision support system.

To that end, Musen's group creates abstract computerized representations of clinical care plans that unfold over time. The idea is to create a decision support system that not only suggests to doctors how to follow a standard plan for chronic disease care but also how to deviate from it in appropriate ways.

The backbone of Musen's work is a task-specific architecture called EON. "This kind of architecture allows you to get above situation-action rules and talk about problem solving in terms of bigger building blocks," Musen says.

For example, EON can handle tasks such as decomposing an abstract plan into its constituent parts to make it actionable. Take the guideline: "if there's been a period of uncontrolled hypertension and the patient has been treated with a given therapy, then consider adding a second line agent." To translate this abstract rule into a concrete action, such as prescribing a particular drug, EON would determine if the precondition holds (there has been a period of uncontrolled hypertension); whether there's been a primary treatment but no second agent; and if so, what's the right second agent to add given the patient's current drugs, allergies, drug sensitivities, and so on.

Using the EON architecture, Musen's lab worked with **Mary Goldstein**, **MD**, at the at the Palo Alto Veterans Administration Medical Center to create a program called ATHENA-CDS. It helps doctors treat hypertension pursuant to guidelines developed by the Joint National Commission on Hypertension. ATHENA-CDS looks at a patient's data and previous therapies (in the EMR) and recommends treatment according to the guidelines—while retaining the flexibility to deviate as needed.

Dealing with Ambiguity and Going Commercial

Medical guidelines are rife with ambiguities and qualifications. Phrases such as "consider this," or "keep in mind that you might want to do this instead of that," are difficult to put into code, says **Milton Corn**, **MD**, deputy director for research and education at the National Library of Medicine. "The ambiguity that humans handle is very difficult for computers to handle," he says.

In addition, there is a great deal of inconsistency in the way recommendations are written. A recent study led by **Richard Shiffman**, **PhD**, of Yale University examined about 1275 guideline recommendations derived from the National Guideline Clearinghouse and found that 32 percent of them did not include a reliably identifiable recommended course of action, and more than half did not indicate the strength of the recommendation. As part of the GLIDES project (GuideLines Into DEcision Support) at Yale University, a project funded by the Agency for Healthcare Research and Quality (AHRQ), Shiffman and his colleagues are trying to demonstrate that practice guidelines that include such ambiguities can actually be transformed into computer-based CDS through a systematic and replicable process. Their demonstration projects involve pediatric asthma and obesity.

"If we could apply statistical methods, many would regard that as the ideal way to do [decision support]," Fox says. "But we can't. We don't know what the numbers are."

Fox works on decision support tools that deal with ambiguity in a different way—through the logic of argument. There's a lot of uncertainty in medicine, Fox says—uncertainty about what is wrong with someone or how to treat them or what tests to do. And old-fashioned logic won't do the job because it doesn't have any uncertainty in it. "Things are either true or false," he says, "whereas decision-making in medicine is rarely about truth. It's about what is likely or preferable."

"If we could apply statistical methods, many would regard that as the ideal way to do it," Fox says. "But we can't. We don't know what the num-

bers are." So Fox developed a language called PROforma that provides a way of reasoning about what may be the case or what ought to be done. "It's reminiscent of the way people think," Fox says. It involves evaluating the arguments for or against a course of action. "And it lets you model any kind of medical decision or clinical process," Fox says. "It's very powerful and versatile for such a simple language."

PROforma lies behind Arezzo, a decision support tool sold by the British company Infermed. Arezzo provides computer interpretable protocols for managing a patient over time. Rather than develop its own knowledge base, Infermed uses guidelines provided by third-party sources such as medical publishers; professional bodies such as the academies of the various specialty services; or the National Institute of Clinical Excellence (Britain's equivalent of the AHRQ). Typically, a government payer or other significant healthcare organization sponsors the development of an Arezzo guideline. As ambiguities arise, Infermed works closely with experts at the sponsoring organization. "Any questions around these ambiguous or vague recommendations are really exposed quickly when you're trying to execute them," says **Robert Dunlop, MD**, the clinical director at Infermed. "This engages the people who will be using the content so they are more likely to use the system subsequently."

PROforma, provides a way of reasoning about what may be the case or what ought to be done. "It's reminiscent of the way people think," Fox says.

At the Royal Free Hospital in London, a multidisciplinary team of physicians routinely uses a PROforma decision support system called MATE. MATE supports more than a dozen decisions that the team has to make, including decisions about surgery, chemotherapy, adjuvant and radiotherapy. The system also determines patient prognosis for each of the therapy options being considered and automatically identifies patients for recruitment into clinical trials. Here, Dr. Vivek Patkar (a breast surgeon who developed the PROforma knowledge base) drives the application at the front left of the room. Photo courtesy of CREDO project (http://www.cossac.org/projects/credo) and Mo Keshtgar (Principal Investigator of MATE trial in breast cancer).

New Zealand is one of the largest users of the Arezzo system. Twenty-five percent of the country's family physicians use Arezzo, and last year they accessed the guidelines more than one million times. The guidelines are hosted on a central server and linked through Web services to all five of the EMR systems used by New Zealand's family physicians. "When the family physician opens a patient record, the system will contact Arezzo through the patient data and recommend the next steps in the patient's treatment," Dunlop says. The system is focused mostly on chronic diseases, particularly those that tend to co-occur such as diabetes, hypertension, kidney disease, and ischemic heart disease. The New Zealand system also includes referral triage and management of accidents and injuries.

In the commercial arena, Arezzo is somewhat unusual in providing a system that goes beyond alerts and situation-action rules. "When we talk to customers, we have to explain that the value proposition we bring is very different from these other systems," Dunlop says. "Once we get past that knowledge barrier, people realize there is nothing else like it in commercial use."

Dunlop also distinguishes Arezzo from Musen's work at Stanford. "It's not like an algorithm where you have a specific pathway you follow," Dunlop says. Because illness doesn't follow a specific pathway (from A to B and B to C), the engine has to be able to navigate to whatever part of the guideline content

is relevant to the patient at a particular time. "What if the patient doesn't start at A or is suddenly at Q?" he says. "Arezzo fits the guideline to the patient rather than the other way around."

Like Intermountain Health's "background" decision support system, Arezzo can also trigger guidelines without the doctor having to realize that he needs them. "We call that the guardian angel approach," Dunlop says, "where you ensure that if the patient record is updated, with a path to Arezzo behind the scenes, then Arezzo will send an alert that the patient might need to be reviewed according to these guidelines."

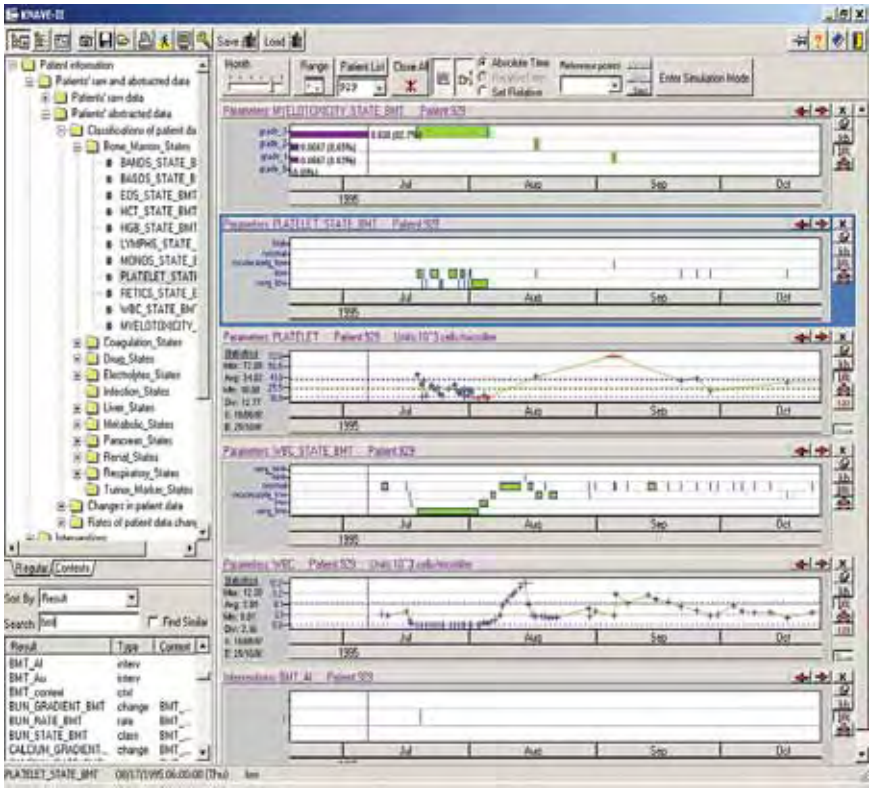
According to Dunlop, New Zealand doctors are finding that Arezzo delivers what they need. "It's our experience that the physicians are the hardest to convince but they become the greatest advocates once it's in production," he says.

Guessing What the Doctor's Doing: Using Computerized Guidelines Unobtrusively

Another option is to let the computer figure out what a doctor is trying to do and then help him or her with that task. **Yuval Shahar, MD, PhD**, professor of medical informatics at Ben Gurion University in Israel is developing tools that can determine which—if any—guideline a physician is trying to follow. "We can compare the temporal pattern of care over time to see if the physician is actually using any guideline." Thus the computer might observe that the doctor is trying to apply a particular anti-hypertensive guideline (JNC7) because the physician's actions fit that guideline the best. The computer might then intervene to say (in a constructive fashion), "if that's what you're trying to do, then let me point out that you should now really consider switching medications." It's a way of non-obtrusively using artificial intelligence to watch the physician's actions and try to help them. Shahar has created such a guideline library. "This is still under development," he says. "It will lead to, I hope, a new kind of medicine in the 21st century."

Data Analysis Support: Helping Docs Understand the Patient

CDS systems are not just about computerized guidelines, though; there is a large amount of data available in patient records that could be harnessed to improve medical decisions. This is particularly true for chronic disease. In the U.S., 80 percent of healthcare system expenses are due to chronic illness—even though chronic illness affects only 25 percent of the patients. Because patients with hypertension, cancer, AIDS, or kidney or heart disease are being treated and monitored for a long time, they generate a lot of data. "We need to help doctors grasp the significance of these data, says Shahar. "Without that, we risk treating



Shahar and colleagues at the VA Hospital in Palo Alto, California, found that doctors could more quickly and accurately answer key questions about cancer patients' status when data was visualized over time. Here, the KNAVE II knowledge browser shows the bone-marrow transplantation ontology on the left with panels containing raw clinical data and their abstractions on the right. Panels are computed on the fly and displayed when a raw or abstract concept is selected within the left hand browser. Here we see visualizations (top to bottom) of a transplant patient's bone-marrow toxicity (myelotoxicity) states, platelet-count states and white blood cell (WBC) states over a four-month period. Users can zoom in on specific time periods or select icons to the right. The "KB" icon, for example, defines the concept in that panel. Reprinted from Martins, S.B., et al., Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data, **Artificial Intelligence in Medicine** ***43, 17-34 (2008) with permission from Elsevier.***

these patients in a non-optimal fashion or spending too many funds (unnecessarily), or both."

Interpreting and Visualizing Clinical Data Over Time

Shahar is therefore working to visualize patient data over time in ways that doctors will find helpful. "The key is to apply medical knowledge to these data, thus displaying meaningful concepts emerging in the patient's data over time," he says.

In a study published in 2008 in Artificial Intelligence in Medicine, Shahar collaborated with Mary Goldstein, MD, and **Suzanne Martins, MD**, at the Veteran's Administration Hospital in Palo Alto, California, to test a tool aimed at helping physicians determine whether a cancer patient's chemotherapy treatment needs adjusting. They first asked oncologists to identify the questions they'd need answered in order to make such a determination. The questions ranged from simple to complex: Is there anemia? Is there low white blood cell count? Is there a continuous period of liver dysfunction? Is there any pattern of organ toxicity (defined as having 2 out of 3 organs involved)? "It's not simple at all, at least for humans doing it on their own," Shahar says. "It requires some really difficult conceptual and cognitive work in putting these values together and drawing a conclusion." Shahar developed a tool that could visualize the answers to these questions by applying clinical knowledge to raw data, such as hemoglobin values, or raw liver enzyme measurements.

After training for only 10 to 20 minutes, physicians were timed as they answered the questions using three different sources of information (in randomized order): a traditional paper patient record; an Excel spreadsheet containing the patient data; or the data as visualized by Shahar's tool (KNAVE II), which showed not just the data but also patterns and abstractions of the data. "They were looking at an interface that displayed the patient's problems through a filter of knowledge," Shahar says.

Using paper or Excel records, the physicians often took 15 minutes or more to answer all of the more difficult questions, and they did so accurately only 57 percent of the time. By contrast, the KNAVE group answered each of the difficult questions in about 10 seconds—the same time as for the easy questions—and answered with 92 percent accuracy. Since physicians typically see a patient for at most seven to eight minutes, this difference really matters, Shahar says.

"Essentially, humans are probably not very good at noticing temporal trends from clinical or other types of time-stamped data in a spreadsheet," Shahar says.

> "Essentially, humans are probably not very good at noticing temporal trends from clinical or other types of time-step data in a spreadsheet," Shahar says.

Shahar's overall methodology—called Knowledge-Based Temporal Abstraction (KBTA)—can be applied in many areas. He has used it for AIDS therapy, for monitoring children's growth, and for diabetes care. "In these cases, one picture is really worth a thousand words," he says.

Analyzing Patient Population Data

In addition to providing a picture of individual patient data, CDS can be used to analyze data on patient populations to provide better care. This often evolves from an institution's quality assurance and compensation program. For example, Partners Healthcare looks at outcomes for groups of patients as part of the physician compensation scheme, Middleton says. Doctors get a bonus if they are up to snuff for quality measures for certain groups—say diabetes patients or heart disease patients. "The docs love it because they get to see how they are doing compared to peers or national benchmarks," he says. Once such a quality system is in place, decision support can be applied to patient populations. Doctors can also drill down from the population to the individual level. For example, they could identify patients in the population who are outliers in terms of how they are responding to treatment. "It helps target attention on those people who need it most," Middleton says.

As McDonald points out: "It's valuable to see everything at once and act on that collective rather than one off. There might be more leverage in doing that rather than focusing on the person in the office or the ones who happen to call a lot."

Nationwide Knowledge Representation

Many of the cutting edge CDS systems described here are available at only a handful of key institutions such as Intermountain Health, Partners Healthcare in Boston, and the Veterans Administration system. But with the stimulus package funding EMRs all around the nation, the potential for widespread CDS will soon exist. The question is: How to make it happen in the most efficient and effective way?

Middleton says it would cost $25 billion per year for physicians to do the knowledge engineering themselves to put the knowledge needed for CDS into their EMRs. He and his colleagues at Partners Healthcare therefore propose a more cost-effective option: Creation of a national knowledge

repository—a federal facility that delivers knowledge artifacts in a form that every EMR can access and use for decision support.

Using a two-year grant from the AHRQ, Middleton launched the Clinical Decision Support Consortium to build a prototype system. "We'll aim to stuff it full of knowledge from Partners, from the Riegenstrieff Institute, from the VA, from Kaiser hopefully, and other members of the CDS consortium," he says. "And we'll build web services off of that knowledge repository, so that a remote EMR in Iowa could subscribe to a publicly available Web service and benefit from the knowledge repository."

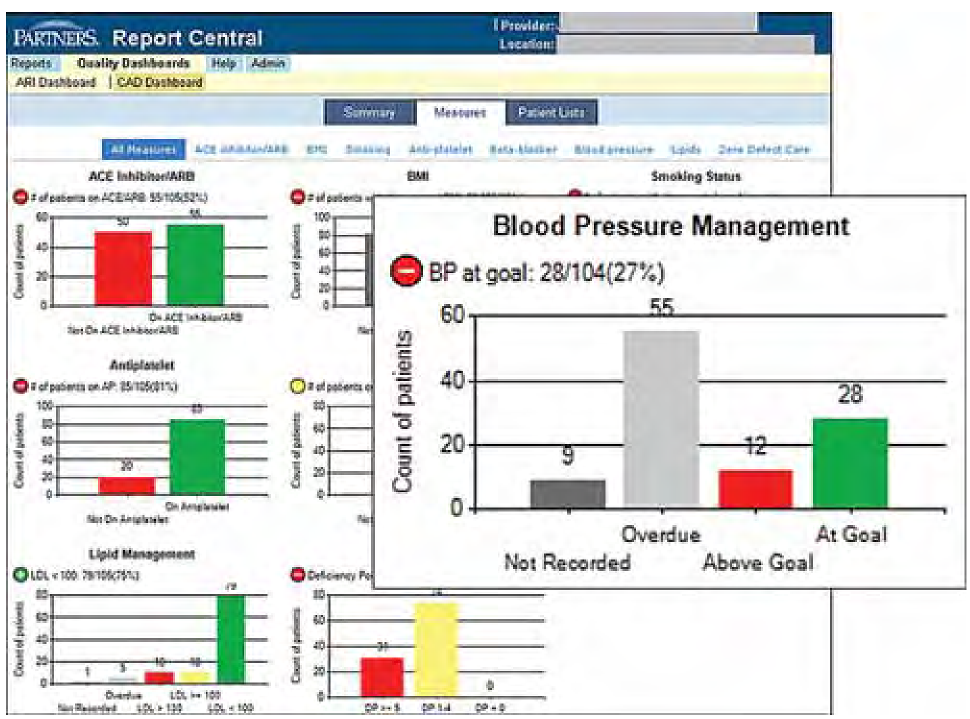
The tricky part is in the knowledge representation, which is still an active area of research, Middleton says. The most widely used method is the so-called Arden Syntax, which describes a way to procedurally represent knowledge so it can be used in rule-based systems in EMRs. "It has a host of problems and issues, but is still the best-known example of how to describe knowledge in a way that many people can use and uptake," he says. "I think as the world moves more toward a service-oriented architecture, it will be easier to represent knowledge in Web services that can be subscribed to in a service catalog."

Interestingly, Partners' pilot system is still largely built on simple rules. Middleton says he's dealing with a different problem than Musen and Fox. "It's more about accessing knowledge and using it remotely than it is about expressing knowledge in a knowledge formalism," he says. "I hope in the end that we converge on a general theory of knowledge representation that is both practical and addresses the theoretical limitations of approaches tried to date."

Teich, at Elsevier, shares Middleton's urge to make decision support practical to a large segment of the population. In academia, Teich says, you are developing a system in a controlled environment where you can be there every day to make adjustments and tweaks. "When you do it on a larger scale, you have to create things that are more flexible—that can be adjusted and operated by people who may not have full-scale informatics training," he says.

So although Elsevier is trying to do relatively advanced decision support,

"I hope in the end that we converge on a general theory of knowledge representation that is both practical and addresses the theoretical limitations of approaches tried to date," Middleton says.



At Partners Healthcare in Boston, the Quality Dashboard can display blood pressure measurement information for an entire population of patients. Physicians can drill down into these data to pay appropriate attention to the patients who are not at their blood pressure goal or schedule appointments for people whose blood pressure has not been recently measured. Courtesy of Blackford Middleton.

Teich says, "we're also looking at what drives care needs and quality needs at thousands of hospitals. It's a different focus than at an academic institution."

To Teich, the goal is to have a repository of CDS tools that hospitals can download to the local EMR where they'll start running. You could even have knowledge bases scattered all over the world and tools that integrate those knowledge bases with patient records at any medical center.

"That's the way life should be," he says. "We need that kind of drop-in CDS service if we're going to make this work at 6000 hospitals and many thousands of ambulatory practices."

The Grand Challenge of Scaling Up

But even if the country is wired with EMRs and has downloadable CDS knowledge representation, there remain some grand challenges for CDS research, according to a January 2009 report by the National Research Council of the National Academies. One of those is to get beyond dealing with patients primarily as a series of transactions, and to provide an overarching view of the individual patient, says Stead, who co-authored the report. "Think what it would be like to examine the world with 1000 ground robot views but no satellite view," Stead says. "That's in essence the problem with today's health care IT." The report identifies several "Grand Challenges" for the field, one of which is to provide "patient-centered cognitive support."

At Vanderbilt, Stead and his colleagues are constructing a number of role-specific views of the patient, trying to focus people on the things they need to see, he says. "These are definite steps in the right direction, but they don't scale up to solving the problem." The challenge is to make models work at the different scales of biology so that the computer can figure out how to construct views of patients with varying combinations of conditions—without anyone having to sit down and write any specific programs. "I've not come across that," he says.

Indeed, says Musen, scaling up seems extremely difficult in decision support. The knowledge bases behind these tools typically deal with one medical problem at a time, Musen says. "If you've seen one, well then you've seen one," he says.

"Think what it would be like to examine the world with 1000 ground robot views but no satellite view," Stead says. "That's in essence the problem with today's health care IT."

Fox agrees: "These systems have to be lovingly—and in some cases, painfully—crafted by hand," he says. "And while these systems can be very useful, we're a long way from automation of any systems that have the versatility of a human clinician." On the other hand, he says, once you have a powerful tool like PROforma or Protégé (a tool developed by Musen), you can apply experience and knowledge to build new applications faster and faster.

Funding for the Future of Decision Support

Meeting the grand challenges will require a significant investment. "Unfortunately," Stead says, "we haven't had a focused national research agenda in this space."

Some of the problem is institutional. The NIH doesn't view health care delivery and efficiency as a major part of its mission, Corn says. The NLM has invested in the field because of its interest in how computer science can help in health care delivery and efficiency. "But once we've demonstrated that something can be done, is efficient, and might be safer, we're in no position to make it happen," he says. "That has to be taken care of outside the Institutes."

According to Middleton, what's missing is a single place within the NIH where clinical informatics and its related specialties—including computational science, cognitive science, and information science—are being studied in a coherent and aggressive way. The NLM and AHRQ fund some important work, and the new Office of the National Coordinator for Healthcare IT is a great thing, he says, but there's no clear focal point for coordinated research activities in information technology for healthcare, and there should be.

"There are research problems in this space," Stead says, "that are as important as the Human Genome Project a decade ago." □

Under TheHood

BY GREGORY R. BOWMAN

Understanding Molecular Kinetics with Markov State Models

Atomistic simulations have the potential to elucidate the molecular basis of biological processes like protein misfolding in Alzheimer's disease or the conformational changes that drive transcription or translation. However, most simulations can only capture the nanosecond to microsecond timescale, whereas most biological processes of interest occur on millisecond and longer timescales. Also, even with an infinitely fast computer, extracting meaningful insight from simulations is difficult because of the complexity of the underlying free energy landscapes. Fortunately, Markov State Models (MSMs) can help overcome these limitations.

MSMs may be used to model any random process where the next state depends solely on the current state. For example, imagine exploring New York City by rolling a die to randomly select which direction to go in each time you came to an intersection. Such a process could be described by an MSM with a state for each intersection. Each state might have a probability of 1/8 of going to each of the four neighboring intersections, a probability of 1/2 of getting stuck at a red light (e.g., staying at the current state), and a probability of 0 of going directly to any other intersection. Drawing such a model would result in something resembling a road map with speed limits replaced by probabilities. Of course, the probabilities of going North, West, East, or South at a given intersection don't have to be the same; they just have to sum to one—because you have to go somewhere.

MSMs for molecular kinetics are conceptually similar to our road-map example but instead of intersections, the states now correspond to basins in the free energy landscape governing the dynamics of the molecule. These states are referred to as metastable states because a molecule is more likely to stay in a particular state than to transition to a new one. Each state may also have many more connections than a typical intersection because of the enormous number of degrees of freedom in most biomolecules.

In our road-map example, one can easily imagine defining states and their connectivity by referring to satellite images and road signs. MSMs for molecular kinetics, however, must be inferred from simulation trajectories (like molecular dynamics trajectories). It's like being asked to draw a map of New York City from GPS coordinates taken at regular intervals by a few drivers. Fortunately, we can make great headway by recognizing that it should be possible to quickly transition between conformations in the same free energy basin (or metastable state) while transitions between different basins will be slow because they are separated by significant free energy barriers. Thus, we can build MSMs by grouping conformations that can reach one another quickly. These groups of conformations become the states of our model and we can simply count transitions between states in our simulations to determine the probabilities of going from one to another.

MSMs for molecular kinetics have many advantages over other approaches. Simply inspecting an MSM can provide an intuition for the dynamics of the system and calculations performed with the matrix representation of MSMs, plus a few representative conformations from each state, make it possible to quantitatively compare with experimental measurements, like fluorescence relaxation curves. MSMs also provide a means of aggregating the data from many simulations into a single model. Moreover, just as it is possible to build up a road map by assembling information representing different locations, a larger, more comprehensive molecular kinetics map can be assembled from many shorter simulations. While MSMs have mostly been used to understand conformational changes on the microsecond to millisecond timescale with atomic resolution, an exciting future direction will be to use them to address ever larger, slower, and more biologically relevant systems. □

MSMs for molecular kinetics must be inferred from simulation trajectories.... It's like being asked to draw a map of New York City from GPS coordinates taken at regular intervals by a few drivers.

DETAILS

Gregory R. Bowman is a PhD student in Vijay Pande's lab at Stanford University. He is the primary developer of MSMBuilder, a freely available tool for the automated construction and analysis of MSMs (https://simtk.org/home/msmbuilder), and is currently using it to understand protein and RNA folding.

DIVERSE DISCIPLINES, ONE COMMUNITY
Biomedical Computation 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

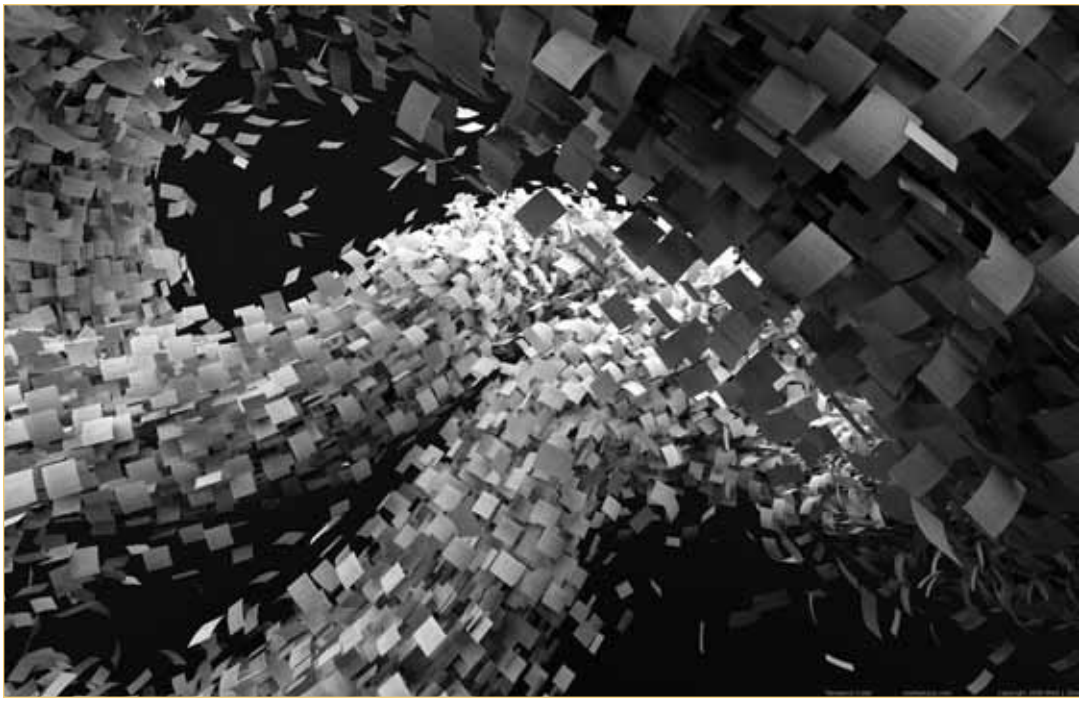
seeing science

SeeingScience

BY KATHARINE MILLER

Fluid Code

In 1999, **Mark J. Stock, PhD**, took his first accidental step toward becoming an artist. He used a 3-D rendering program to help him debug some code. "The first image that came out was beautiful," he says. "There were images within the data that shouldn't really be there." Just as people find shapes in clouds, Stock says, the computer serendipitously creates things we recognize, but with more possibilities. After eight years of creating art with his computational tools, Stock says he has come to understand the medium the way a painter understands paint. "Code is the farthest thing from a picture," he says. "But I now have a much better idea of what the image will look like when I write the code." □



For his PhD dissertation in aerospace engineering at the University of Michigan, Stock used vortex particle methods to simulate flow. "The particles move around and carry properties with them," he says. But most particles in such simulations are like billiard balls bouncing off of one another. Stock's particles, on the other hand, interact in a more complicated way, giving them directionality and curvature. This accomplishment required a lot of work and produced 600 pages of code. Here, Stock displays each revision of those 600 pages of code, using the code itself to simulate the movement of those pages in space. "I wanted to show that there's a tremendous amount of energy that goes into making sure the simulation solves the physical equations of motion—and that this energy can turn one type of data into a picture." Courtesy of Mark J. Stock, www.markjstock.com.