

FEATURES

8 The Golden Age of Public Databases: Speeding Biomedical Discovery

BY REGINA NUZZO, PhD

17 Simulated Metabolism— A First Step Toward Simulated Cells

BY JULIE REHMEYER, PhD

DEPARTMENTS

1 GUEST EDITORIAL | FULFILLING THE PROMISE OF THE NIH ROADMAP THROUGH NATIONAL ENGAGEMENT BY THE NATIONAL CENTERS FOR BIOMEDICAL COMPUTING (NCBC) AND THE CTSA INFORMATICS BY BRIAN ATHEY, PhD

2 NEWSBYTES
BY KRISTIN SAINANI, PhD, CHANDRA SHEKHAR, PhD,
ROBERTA FRIEDMAN, PhD

- A Viral Closeup
- An In Silico Time Machine
- Bacteria Prepare Themselves
- Molecular Biology Wikis Launched
- Predicting Brain Response to Nouns
- A Finer Fat Model

24 SIMBIOS NEWS | ENHANCED FUNCTION RECOGNITION IN PROTEIN TRAJECTORIES OVER SPACE AND TIME
BY JOY KU, PhD

25 UNDER THE HOOD | HOW UPPER LEVEL ONTOLOGIES DEAL WITH FUNCTIONS AND OTHER REALIZABLE ENTITIES
BY MATHIAS BROCHHAUSEN, PhD

26 SEEING SCIENCE | SENSATIONAL SEQUENCES
BY KRISTIN SAINANI, PhD

COVER ART:

CREATED BY RACHEL JONES OF WINK DESIGN STUDIO USING A LIST OF DATABASES ADAPTED FROM THE MOLECULAR BIOLOGY DATABASE COLLECTION: 2008 UPDATE, BY MICHAEL Y. GALPERIN, NUCLEIC ACIDS RESEARCH 2008 36(DATABASE ISSUE):D2-D4.

PAGE 12 DATABASE IMAGE © ANDRES RODRIGUEZ | DREAMSTIME.COM • PAGE 15 IMAGE CREATED BY RACHEL JONES USING DATABASE IMAGE © ANDRES RODRIGUEZ | DREAMSTIME.COM • PAGE 17 IMAGE CREATED BY RACHEL JONES USING CIRCUITRY IMAGE (© SAILORMAN | DREAMSTIME.COM) AND E COLI IMAGE (© ERAXION | DREAMSTIME.COM)



Fall 2008

Volume 4, Issue 4
ISSN 1557-3192

Executive Editor

David Paik, PhD

Managing Editor

Katharine Miller

Associate Editor

Joy Ku, PhD

Science Writers

Kristin Sainani, PhD
Regina Nuzzo, PhD
Julie Rehmeier, PhD
Chandra Shekhar, PhD
Roberta Friedman, PhD

Community Contributors

Joy Ku, PhD
Brian Athey, PhD
Mathias Brochhausen, PhD

Layout and Design

Wink Design Studio

Printing

Advanced Printing

Editorial Advisory Board

Russ Altman, MD, PhD
Brian Athey, PhD
Dr. Andrea Califano
Valerie Daggett, PhD
Scott Delp, PhD
Eric Jakobsson, PhD
Ron Kikinis, MD
Isaac Kohane, MD, PhD
Paul Mitiguy, PhD
Mark Musen, MD, PhD
Tamar Schlick, PhD
Jeanette Schmidt, PhD
Arthur Toga, PhD
Shoshana Wodak, PhD
John C. Wooley, PhD

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

Office

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

Biomedical Computation Review is published quarterly by Simbios National Center for Biomedical Computing and supported by the National Institutes of Health through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from <http://nihroadmap.nih.gov/bioinformatics>. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Jennie Larkin, PhD (NHLBI)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Richard Morris, PhD (NIAID)
Joseph Pancrazio, PhD (NINDS)
Grace Peng, PhD (NIBIB)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)
Jane Ye, PhD (NLM)
Yuan Liu, PhD (NINDS)

BY BRIAN ATHEY, PhD

Fulfilling the Promise of the NIH Roadmap Through National Engagement by the National Centers for Biomedical Computing (NCBC) and the CTSA Informatics



For major team-based Roadmap initiatives, National Institutes of Health (NIH) officials expect grantees to look beyond the focus of their individual projects to build bridges not only among funded projects but also between themselves and the research community as a whole. These collaborations are an important part of creating a national biomedical computing infrastructure. And the National Centers for Biomedical Computing (NCBCs) and the Clinical and Translational Sciences Award Informatics programs (CTSA Informatics) are stepping up to the plate.

In August 2008, the leadership of the NCBCs held a third successful “all hands meeting” in Bethesda, MD. It is a tremendous credit to the vision of the NCBC Roadmap initiative (as specified by the Botstein-Smarr Report to the Biomedical Information Science Technology Initiative) that each of the Centers is now launched and productive. Direct feedback from NIH Director Elias Zerhouni and National Institute of General Medical Sciences Director Jeremy Berg points to the NCBC program as a “crown jewel of the NIH Roadmap.”

A major highlight of the all hands meeting was the talk by Simbios co-principal investigator Russ Altman, who described the many collaborative activities of the NCBC “Big P” (where P=“Program”). To the NIH, these additional activities constitute evidence that the NCBCs are helping to create a “national biomedical computing infrastructure.”

Meanwhile, another complementary component of

this national infrastructure is being created from the CTSA—an aggregation of Biomedical Informatics Programs of the National Center for Research Resources (NCRR) Roadmap initiative. This more than \$500 million per year program currently supports 38 sites and is projected to go to 60 sites within 2 years. This Roadmap program is intended to truly transform the way academic health centers do clinical and translational research. Interestingly, like the Big P of the NCBCs, the CTSA has its “Informatics Consortium.” Its first annual All Hands Meeting is scheduled for October 16, 2008, at NIH with more than 180 participants expected to attend.

To build a national biomedical computation infrastructure, it is also important to make good use of existing computational resources. To that end, three NCBCs (CCB, NCBO, and NCIBI) have recently been awarded administrative supplements by NIGMS to collaboratively support the creation of a “Biositemaps” protocol to address the issues of (i) locating, (ii) querying, (iii) traversing, (iv) composing or combining, and (v) mining biomedical computing and computational biology software tools and information resources on the Internet. This is a joint project of the “Yellow Pages/Resourceome” and Software Ontologies Working Groups, part of the NCBC Software and Data Interchange Working Group (SDWIG). This effort builds from the earlier, productive NCBC collaboration that created the “iTools” resource to organize and make web-accessible the NCBC software tools and data resources (published in *PLoS ONE* in 2008).

The CTSA Informatics Inventory and Resources Workgroup (IRWG) and the Biomedical Informatics Programs at the University of Pittsburgh Medical Center and the University of Michigan have also received an Administrative Supplement from NCRR to use the iTools and Biositemaps capabilities to organize the growing list of tools and data emanating from other CTSA working groups. This will lead to more effective and efficient communications for the CTSA Consortium overall, as well as produce useful tools for use at local CTSA sites.

These highly visible and potentially high-impact national collaborations bode well for the eventual fulfillment of one of the NIH Roadmap’s promise: to develop and sustain the nation’s capacity to perform biomedical research in the digital age. □

DETAILS

National Centers for Biomedical Computing:
www.ncbcs.org

Clinical and Translational Science Award:
(<http://www.ctsaweb.org/>)

Brian Athey, PhD is an associate professor of biomedical informatics at the University of Michigan and director of the Michigan Center for Biological Information. In addition, he is principal investigator of the National Center for Integrative Biomedical Informatics, which is one of the NCBCs, and co-chair for the CTSA Informatics Consortium.

NewsBytes

A Viral Closeup

The phi29 bacteriophage is an efficient infection machine—it fires its genome into a host bacterium, hijacks the host’s cellular equipment, and assembles an army of new viruses for its next mission. For the first time, scientists have produced sub-nanometer resolution pictures of the virus, revealing some striking new details—including an unexpectedly tight twist of DNA suggestive of how the virus springs into action. The results appear in the June issue of *Structure*.

“We use structure as a way to try and

understand how viruses function,” says **Timothy Baker, PhD**, professor of chemistry/biochemistry and molecular biology at the University of California, San Diego who led the collaboration between UCSD and the University of Minnesota. “The more we can learn from structure, the better we’ll understand the whole infection process and perhaps ways to circumvent it.”

Using computer reconstruction, Baker and his colleagues aligned roughly 12,000 electron microscope images of frozen viral particles at different angles and fused them into a 3-D picture of the assembled phage—including its head (either full of DNA or empty), its tail, and the head-tail connector. “You have to go through an iterative process of looking at all 12,000 images with respect to a model which is a cube of data that’s 900 pixels on a side. So the computational challenges are pretty severe,” Baker says. “This couldn’t have been done even a

few years ago, not without really dedicated supercomputer power.”

The resolution achieved—8 Angstroms—was two-fold higher than ever before for an asymmetric virus (where researchers cannot exploit symmetry to reduce complexity). At this resolution, individual alpha helices (in the proteins that make up the head-tail connector piece of the virus) become distinguishable as tube-like structures. Baker’s team compared their picture of the viral head-tail connector with atomic-level models of this structure that were available from X-ray crystallography, and showed that the alpha helices matched up. “It helped us verify that what we were seeing in our map was in fact believable,” he says.

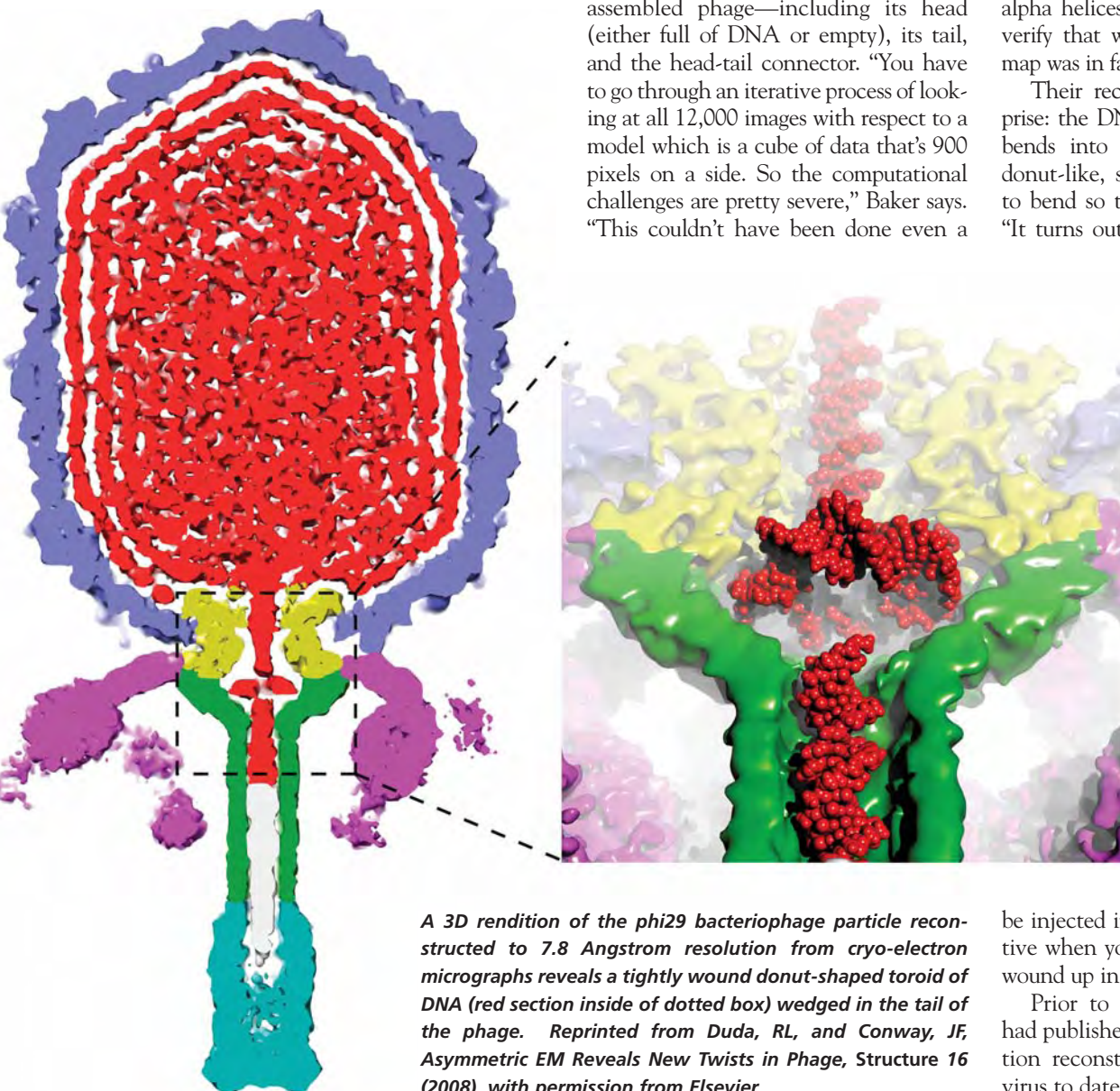
Their reconstruction revealed a surprise: the DNA in the tail of the phage bends into a tight coil—a toroid, or donut-like, shape. DNA isn’t expected to bend so tightly over short distances. “It turns out if you talk to people who

know something about DNA, they say it is possible,” Baker says. “They just haven’t seen it in a biological system like this before.”

“In terms of shock value, that was amazing,” comments **John E. Johnson, PhD**, professor of molecular biology at The Scripps Research Institute who occasionally collaborates with Baker. The bacteriophage must pack its DNA into a tiny space against tremendous forces, and Johnson speculates that the toroid may act as a plug to hold the DNA inside until it’s ready to

be injected into the host. “It’s so suggestive when you look at how this thing is wound up in this little cavity,” he says.

Prior to this work, Johnson’s team had published one of the highest resolution reconstructions of an asymmetric virus to date (17 Angstroms—as report-



A 3D rendition of the phi29 bacteriophage particle reconstructed to 7.8 Angstrom resolution from cryo-electron micrographs reveals a tightly wound donut-shaped toroid of DNA (red section inside of dotted box) wedged in the tail of the phage. Reprinted from Duda, RL, and Conway, JF, Asymmetric EM Reveals New Twists in Phage, Structure 16 (2008), with permission from Elsevier.

ed in *Science* in 2006). “We saw a lot of interesting things,” he says. “But this paper has pushed it to a higher level.”

—By *Kristin Sainani, PhD*

An *In Silico* Time Machine

In biology, many exciting events happen on the millisecond timescale—proteins fold, channels open and close, and enzymes act on their substrates. Atomic-level simulations of this duration are beyond the reach of current technology, but a new specialized computer called Anton—described in the July 2008 issue of *Communications of the ACM*—may change all this. Slated to be operational by the end of the year, the machine is projected to speed up molecular dynamics simulations 100-fold.

The basic goal is to be able to visualize, at the atomic level of detail, an entire biological trajectory, such as an anti-cancer drug (like Gleevec®) inactivating its target enzyme, says **David E. Shaw, PhD**, chief scientist of D.E. Shaw Research, the independent research laboratory that is creating Anton, and a senior research fellow at the Center for Computational Biology and Bioinformatics at Columbia University. Because it provides what might be thought of as a computational microscope, Anton is named after 17th century scientist Anton van Leeuwenhoek, known as the father of microscopy.

“Our machine only does molecular



Anton is designed for a specific task: molecular dynamics simulations. Here is one of the first Anton application-specific integrated circuits (ASIC), which arrived in January 2008. Reprinted from Anton, a special-purpose machine for molecular dynamics simulation, David E. Shaw, et al., Communications of the ACM 51:91-97 (2008) with permission from the ACM.

be executed simultaneously since each is dependent on the previous, but Anton uses 512 highly specialized chips working in parallel to speed up the massive calculations within each step.

“They’ve done a beautiful job, and there are a lot of intellectually interesting aspects to the approaches they’ve taken,”

just 10 days on 100,000 computers. “The approach not only gives access to long timescales, but having many trajectories allows you to do statistical testing, which you cannot do on a single trajectory,” Pande says. “Most of the questions that people in the field are interested in are inherently statistical questions,” he says.

Anton, slated to be operational by the end of the year, is projected to speed up molecular dynamics simulations 100-fold.

dynamics. It does it blindingly fast, but it’s pretty brittle and isn’t designed to do anything else,” Shaw explains. In molecular dynamics simulations, time is broken into discrete steps, each a few femtoseconds (10^{-15} of a second) of simulated time. At each step, the computer calculates the force exerted on each atom in the system (typically 25,000 to 100,000 atoms) and updates its position and velocity. The various time steps cannot

says **Vijay Pande, PhD**, associate professor of chemistry at Stanford University and director of the protein folding distributed-computing project Folding@home. Still, Pande advocates a different approach. Rather than simulating one long trajectory, which could take a million days on one general purpose computer, he simulates a large number of shorter trajectories and then merges them together with a clever algorithm. This may take

But according to Shaw, “The two approaches are very complementary and I think they may turn out to be useful for solving very different types of problems.” Combining many smaller trajectories is more efficient, he says. “But there are some cases in which you’d like to have confidence that what you’re seeing is one continuous, unbiased, physically realistic trajectory.”

Though other groups have previous-

ly attempted to develop specialized computers for molecular dynamics simulations, most efforts have failed to stay ahead of Moore's Law, which says that the speed of general purpose computers doubles every 18 months.

"The Shaw group's effort has been one of the most exciting examples of trying to do that to date," says Pande. "Since the machine isn't out yet, it's too early to say whether they have succeeded or not. But they've got a reasonable shot."

—By *Kristin Sainani, PhD*

Bacteria Prepare Themselves

When we see dark clouds, we might grab an umbrella before heading outside. We've long believed that showing such foresight requires a brain and complex information-processing capability. It turns out, though, that even microbes, which do not have brains or a nervous system, can learn to use cues from their surroundings to anticipate future events, according to a new research study based on both experimental and computational techniques.

"What we have shown is that microbes too have the intrinsic capacity for predictive behavior," says **Saeed Tavazoie, PhD**, an associate professor of molecular biology at Princeton University who published the study in

"What we have shown is that microbes too have the intrinsic capacity for predictive behavior," says Saeed Tavazoie.

the June 6 issue of *Science* with co-authors and Princeton colleagues **Ilias Tagkopoulos, PhD** and **Yir-Chung Liu, PhD**. "Indeed, this may be essential for their survival." The findings could have implications for infectious disease treatment and microbial applications in industry.

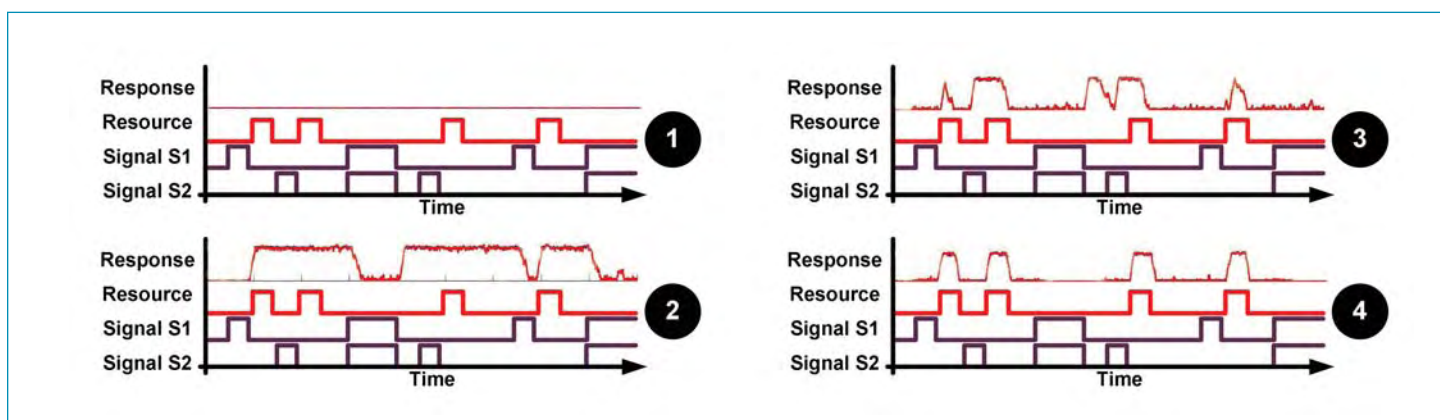
Escherichia coli (*E. coli*) normally adjusts its breathing to match the ambient oxygen level: In the open, the bacterium breathes oxygen; inside an animal's oxygen-poor gut, it doesn't. According to prevailing notions, this switch from aerobic to anaerobic respiration is a purely reflexive response to the drop in oxygen level.

But Tavazoie and his colleagues suspected the microbes wouldn't survive if they responded only when they were already oxygen-deprived. They proposed that, instead, *E. coli* senses warmth when it enters an animal's mouth, and uses this as an early cue to switch to anaerobic breathing. In laboratory experiments, the researchers found this to be the case: When the temperature rises, *E. coli* turns off many

genes needed for aerobic respiration. "By anticipating the subsequent lack of oxygen, it improves its chances of survival," says Tavazoie. "This is clearly predictive behavior." Moreover, when the researchers caused oxygen levels to rise shortly after an increase in temperature, *E. coli* evolved (over about 100 generations) to disregard warmth as a cue. "It rewires itself to forget the old association," says Tavazoie.

To explain how a microbe could evolve such complex behavior, the researchers devised a computational framework that mimics the essential aspects of microbe ecology. Modeled as a network of genes and proteins, a virtual bug in this virtual ecology lives and breeds when it has enough energy, or dies when it runs out of it. To gain energy, it has to be ready to eat, biochemically speaking, when "food" is available. But if it gets ready to eat and no food arrives, it wastes precious energy.

To help the virtual bugs, the researchers gave them different patterns of cues to indicate that food is coming. In one experiment, the bugs were fed



Predictive behavior of a simulated microbe species at different points along an evolutionary trajectory. The resource (food) is always given shortly after giving either, but not both, of the two signals (environmental cues). Initially (subplot 1) the response seems random relative to the food and cues. Eventually, however (subplot 4), guided by the

pattern of cues, the microbe evolves its feeding response to make it synchronize with the food availability. Courtesy of Ilias Tagkopoulos. Reprinted from the supporting online material for Predictive Behavior Within Microbial Genetic Networks, Ilias Tagkopoulos, et al., Science 320, 1313 (2008).

shortly after they got one of two different cues—but not if they got both cues at once. “To predict mealtimes accurately in this case, the microbes would have to solve a complex logic problem,” says Tagkopoulos, an electrical engineer associated with the Lewis Sigler Institute for Integrative Genomics. Sure enough, after a few thousand generations, a gastronomically savvy—and ecologically fit—strain of microbe emerged. The feeding response of such a fit bug (see figure) illustrates how interacting genes and proteins could evolve complex behavior.

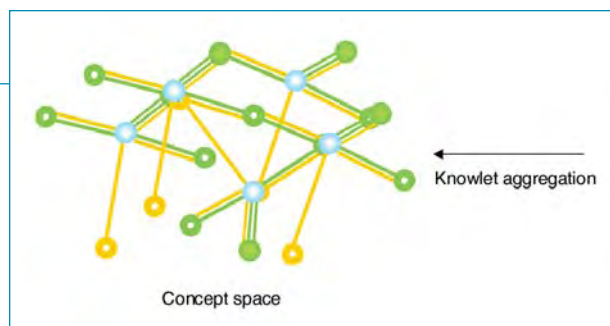
According to **David Reiss, PhD**, a computational biologist at the Institute for Systems Biology in Seattle, the researchers' computational framework is notable for incorporating more biological mechanisms than prior models did. He cautions, however, that even this model oversimplifies the behavior of real microbes. Nevertheless, Reiss says, the study is interesting and novel for showing that anticipatory behavior is not restricted to higher systems with decision-making capability.

—By **Chandra Shekhar**

Molecular Biology Wikis Launched

If you build it, will they come? That's the question on everyone's mind after the launch of two pioneering initiatives in community annotation: WikiProteins and Gene Wiki, announced, respectively, in the May 28 issue of *Genome Biology* and the July 8 issue of *PLoS Biology*. The efforts create a central repository of information on genes and proteins and call on the scientific community to keep it up-to-date and accurate.

“There's no way we can handle the current growth of knowledge with central annotation only,” says **Barend Mons, PhD**, who leads the WikiProteins effort. “I'm a big fan of the authoritative databases like UniProt, but we have to make them grow faster. So what we need is a shell around them of community annotation.” Mons is associate professor of human genetics at the Leiden University Medical Centre and



*Each unique biomedical concept in WikiProteins is attached to a “knowlet” or concept cloud, illustrated here. A concept (depicted as a solid blue ball) is associated with other concepts through facts (established relationships, depicted as solid green balls), co-occurrences (co-occurrences in sentences in PubMed, depicted as green rings), or implicit associations (overlapping concepts in their Knowlets, depicted as yellow rings). Reprinted from Barend Mons, et al., *Calling on a million minds for community annotation in WikiProteins*, in *Genome Biology* 2008, 9:R89.*

of medical informatics at Erasmus University, both in the Netherlands.

“WikiProteins is more than just a Wiki; it has the whole knowledge space hovering over it,” Mons says. Using text mining, WikiProteins imported structured content (adhering to computer-readable, controlled vocabularies) on 1.2 million unique biomedical concepts from existing databases, such as PubMed, Swiss-Prot, and Gene Ontology. The system also created profiles for about 1.6 million authors in PubMed, who are expected to serve as the knowledge guardians. “If you have 1.6 million people in PubMed publishing today and you have 1.2 million concepts in the Wiki, then roughly everyone could take one concept and make sure the page on that concept is correct. That's doable,” Mons says.

Gene Wiki operates within Wikipedia and, in contrast to WikiProteins, emphasizes unstructured content, such as free text and images, “more akin to a review article,” says **Andrew Su, PhD**, of the Genomics Institute of the Novartis Research Foundation, who leads the effort. Using data from Entrez Gene, the system added or amended about 9000 Wikipedia “stub” entries on human genes, which anyone can edit. “Being part of the larger Wikipedia community is certainly an advantage of this system. The people there are experts at welcoming newcomers, fighting vandalism, and formatting things correctly,” Su says.

Su and Mons have plans to collaborate. WikiProtein and Gene Wiki entries will be linked through a common “entry page” (likely hosted in WikiProteins),

making it easy to navigate between the systems. “This will allow users to take advantage of whichever system they feel comfortable with,” Su says.

Getting bench scientists to participate will be a challenge, Mons says, but he believes the incentives are high. The WikiProteins system

mines PubMed for new information daily, finds new explicit and implicit associations—such as predicting protein-protein interactions—and alerts scientists of all edits and updates to concepts in their purview. “I hope it becomes a daily part of their knowledge

“I'm a big fan of the authoritative databases like UniProt, but we have to make them grow faster. So what we need is a shell around them of community annotation,” says Barend Mons.

discovery process,” Mons says. Since its launch, WikiProteins has also received requests to enable users to enter data as unstructured, free text, which should lower the barrier to participation.

Another factor that may boost participation is the development of ways to trace authorship for each entry, so that authors can get credit for their work and readers can assess the reliability of content. A recent proof of this possibility was demonstrated in

“WikiGenes,” (not to be confused with the GeneWiki!) a project described in *Nature Genetics* in September 2008. WikiGenes was developed by **Robert Hoffmann, PhD**, at the Massachusetts Institute of Technology. It’s part of his Memoir project, which has, he says, “the ambitious goal to create a free collaborative knowledge base for all of science—where authorship matters.”

Though the creators of the various Wikis have not yet formally quantified participation, Su says that there’s been an uptick in Gene Wiki activity since the *PLoS Biology* paper came out. “It gives me hope that the system is right and that the framework is there, so if we are tapping into a desire in the community to share knowledge and harness community intelligence, then we have the structure to do it now.”

More information is available at: www.wikiprofessional.org (WikiProteins) and http://en.wikipedia.org/wiki/Portal:Gene_Wiki (Gene Wiki).

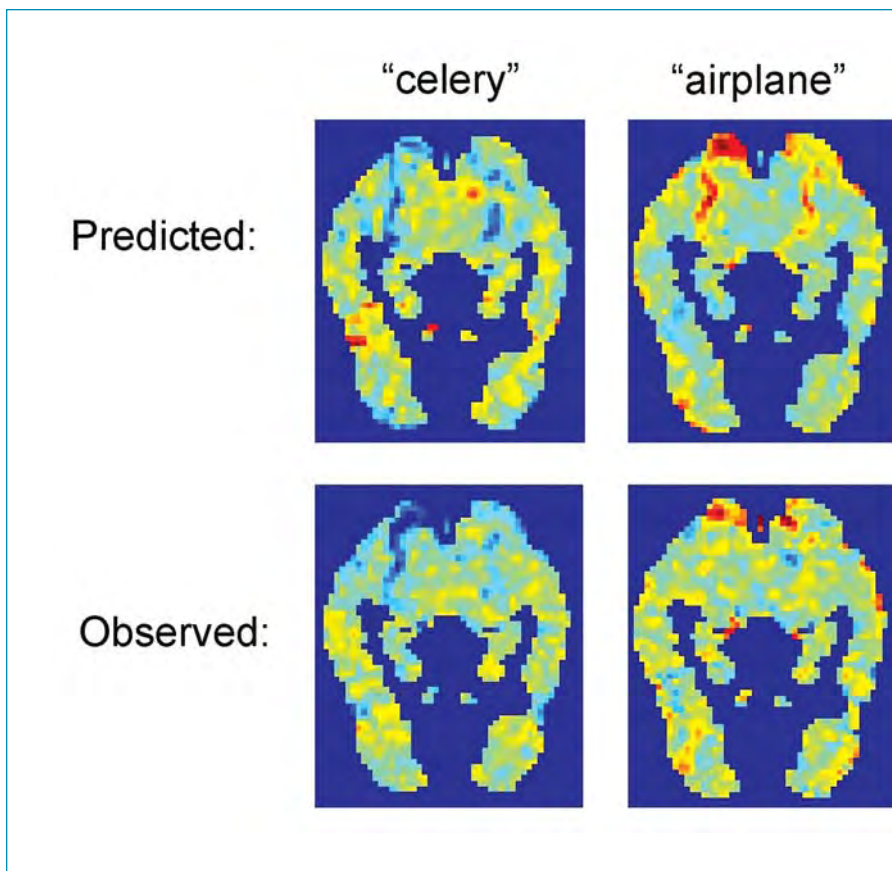
—By *Kristin Sainani, PhD*

Predicting Brain Response To Nouns

Thinking of a noun—a peach, train, or bird, for example—activates specific parts of the brain. Now, scientists have trained a computer to predict such activation patterns. The achievement represents a step toward understanding language processing and could one day contribute to treatments for cognitive decline.

“If we had a better model of how the

brain represents language, we’d be better able to make sense of disorders like dementia,” says **Tom Mitchell, PhD**, a professor of computer science at Carnegie Mellon in Pittsburgh and lead author of the research published in the



Brain activation patterns in response to nouns: The computer algorithm predicted the response to newly encountered words with 77% accuracy. Courtesy of Tom Mitchell. From Mitchell, TM, et al., Predicting Human Brain Activity Associated with the Meanings of Nouns, Science, 320 (5880): 1191 (2008) DOI: 10.1126/science.1152876. Reprinted with permission from AAAS.

May 30 issue of *Science*.

Functional magnetic resonance imaging, or fMRI, registers changes in blood flow within peoples’ brains as they are asked to do a specific task—

computer to produce fMRI images like those generated by humans. The training process uses two sources of data: fMRI images collected from nine people viewing 60 nouns; and a database

The computer model was able to produce a pattern of brain activity in response to words it had never before encountered with greater than 70 percent accuracy.

such as thinking of a specific word. Since 2000, Mitchell and **Marcel Just, PhD**, professor of psychology at Carnegie Mellon and co-director of the Pittsburgh Brain Imaging Research Center have collaborated to train a

(derived from a trillion words of text from the Internet) describing pairings of nouns and the verbs that accompany them most frequently in written text. Noun-verb pairings are the basis of language, as anyone knows who has

raised a toddler, Mitchell notes.

Once trained, the computer model was able to produce a pattern of brain activity in response to words it had never before encountered with greater than 70 percent accuracy. “We now have a model that is capable of extrapolating beyond the data on which it was trained,” Mitchell says. For example, after training, the model could predict that a food noun would provoke activity in the area of the brain mediating eating sensations, the so called gustatory cortex: “peach,” for example, frequently occurs in English paired with the verb “eat.” Similarly, a noun will activate motor areas of the brain to the degree that it co-occurs with the verb, “push,” or cortical regions related to body motion to the degree that it co-occurs with “run.”

Harvard cognitive psychologist **Alfonso Caramazza, PhD**, cautions that the model may be imperfect. He says it fails to capture an area of the brain that is damaged in semantic dementia, one form of brain damage in which people cannot understand the meaning of words. “Our understanding of concepts, and representation of this information in the brain, is not only sensory-motor,” Caramazza says. Evolution likely has sculpted our brains to react appropriately to inanimate things that may be either potentially dangerous or pleasurable. Emotional areas of the brain respond differently to a hammer than to a dog, he points out.

“These are deep questions to which no one has the answers, so one should be cautious,” Caramazza says, adding, “I think (the Pittsburgh team) would agree, these tools are in their infancy and we are only beginning to know how to use them.”

—By **Roberta Friedman, PhD**

A Finer Fat Model

When it comes to heart disease risk, “bad” and “good” cholesterol—also known as low density lipoproteins [LDL] and high density lipoproteins

[HDL]—do not tell the whole story. These particles that carry fat through the blood can be broadly classified based on their density, but they actually vary widely in their composition and clinical risk. A new computational model, described in the May issue of *PLoS Computational Biology*, allows scientists to see this diversity for the first time, providing additional information to aid in diagnoses and treatment planning.

“We look at lipoprotein profiles in greater detail in order to find possibly

Unlike previous models of blood lipid metabolism, Hübner and colleagues modeled the whole spectrum of individual lipoproteins.

relevant abnormalities in the lipid values that would remain undetected by looking only at LDL or HDL,” says lead author **Katrin Hübner, PhD**, a post-doctoral research fellow at the University of Heidelberg who completed much of the work while a PhD student at the Charité University hospital in Berlin. The model has several potential clinical applications.

Unlike previous models of blood lipid metabolism, which considered just four lipoprotein density classes (very low, low, intermediate, and

high), Hübner and colleagues modeled the whole spectrum of individual lipoproteins—by combining any of three proteins (apoB, apoA, and other) and three fat molecules (cholesterol, triglycerides, and phospholipids) in varying amounts. The particles undergo 20 reactions, including particle birth from the liver, particle death from cell uptake, and transfer of fats between particles.

In initial simulations, Hübner and colleagues generated virtual blood lipoprotein profiles that closely matched experimental values from healthy individuals. Then they tweaked the parameters in their model to mimic three known lipid disorders. For example, to simulate familial hypercholesterolemia, which involves a malfunctioning LDL receptor, they decreased the rate of cellular uptake of apoB-containing particles (which are recognized by the receptor) by 75 percent. The simulations accurately reproduced the characteristic lipid profiles of the three diseases.

The model could help pinpoint the underlying molecular defect in patients with abnormal lipid profiles of unknown origin, Hübner says. It could also be used to predict the impact of specific treatments, such as drugs or lifestyle changes, on a patient’s lipid profile.

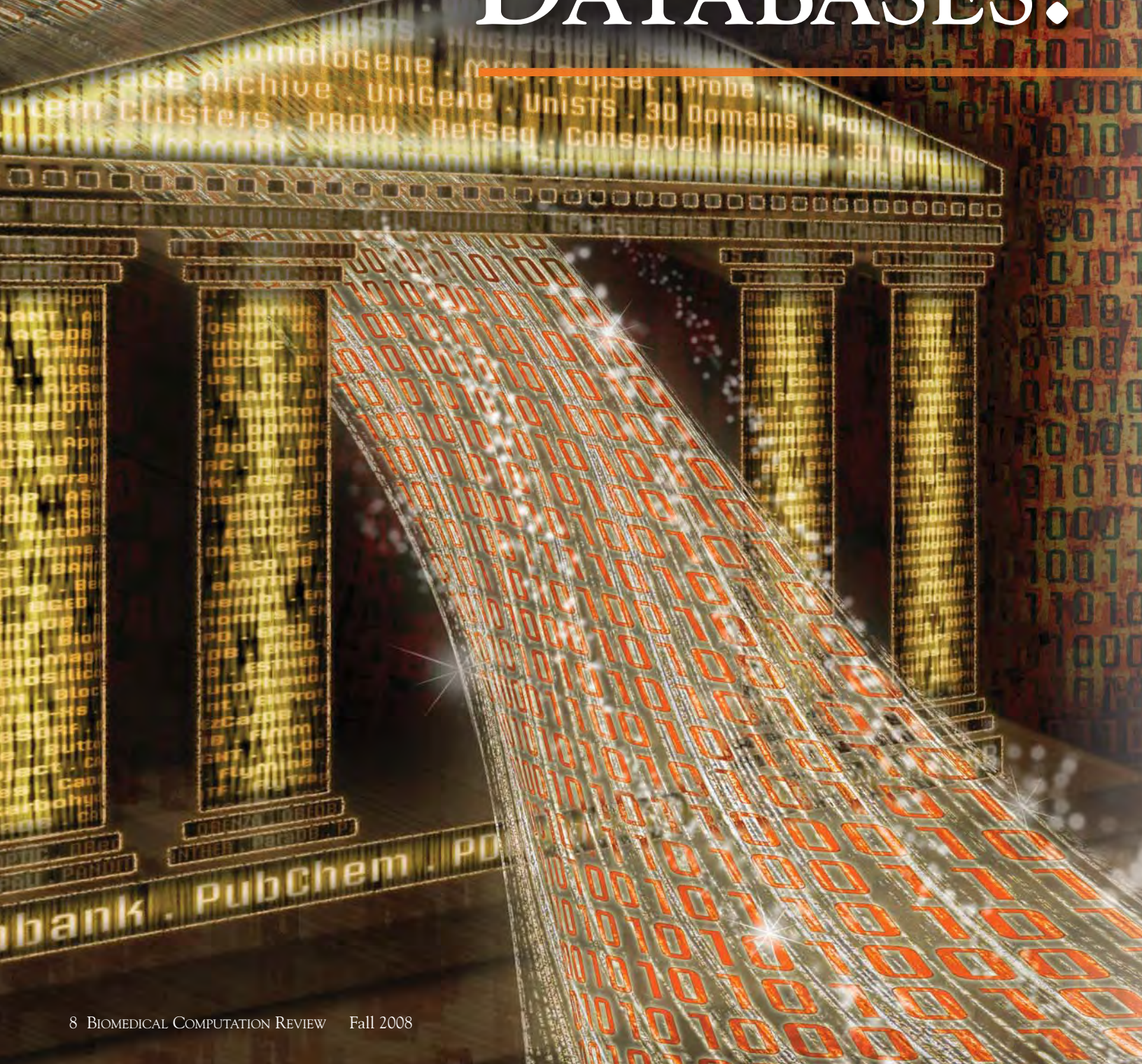
“This work addresses an important issue in modeling lipoprotein metabolism, which is the heterogeneity of lipoproteins,” says **Brendan O’Malley, PhD**, Project Leader of Systems Biology of Lipid Metabolism at Unilever Corporate Research in the United Kingdom, who also works on lipoprotein modeling (using a different approach).

“This is one of the first works in this area, so there’s still quite a lot of work to be done,” he says. For example, the model needs to be further validated with high quality patient data. But, in the future, it could lead to improved diagnostics and personalized treatments for cardiovascular disease, he adds.

“It’s not ready for the clinic yet,” Hübner agrees. “But we’ve made a promising first step.”

—By **Kristin Sainani, PhD** □

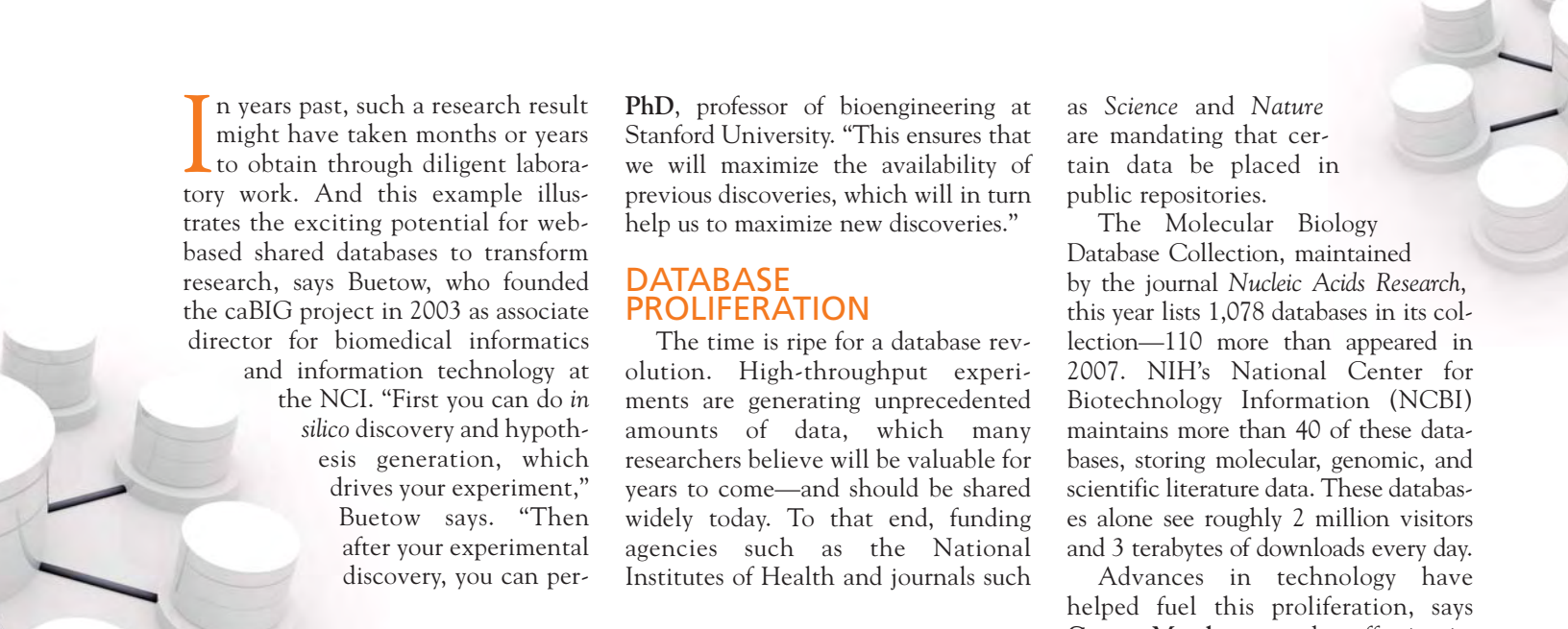
THE *Golden Age* OF PUBLIC DATABASES:



Speeding Biomedical Discovery

The setting: a scientific conference in January 2008. The speaker, **Bruce Ponder, MD, PhD**, an oncology professor at Cambridge University, is describing a previously unknown link between a particular gene (FGFR2) and breast cancer. A prominent researcher in the audience, the late **Judah Folkman, MD**, raises his hand to propose a hunch: could another gene (for endostatin) in the same network also be related to breast cancer? The speaker doesn't know.

After the session, another audience member, **Kenneth Buetow, PhD**, pops the question into a public database, the National Cancer Institute's Cancer Biomedical Informatics Grid (caBIG; <http://cabig.nci.nih.gov/>), a web accessible collection of interoperable software tools and data sources. Voilà! The information highway kicks out a preliminary research result: variants of the endostatin gene are associated with breast cancer and can be protective against the disease. >



In years past, such a research result might have taken months or years to obtain through diligent laboratory work. And this example illustrates the exciting potential for web-based shared databases to transform research, says Buetow, who founded the caBIG project in 2003 as associate director for biomedical informatics and information technology at the NCI. “First you can do *in silico* discovery and hypothesis generation, which drives your experiment,” Buetow says. “Then after your experimental discovery, you can per-

PhD, professor of bioengineering at Stanford University. “This ensures that we will maximize the availability of previous discoveries, which will in turn help us to maximize new discoveries.”

DATABASE PROLIFERATION

The time is ripe for a database revolution. High-throughput experiments are generating unprecedented amounts of data, which many researchers believe will be valuable for years to come—and should be shared widely today. To that end, funding agencies such as the National Institutes of Health and journals such

as *Science* and *Nature* are mandating that certain data be placed in public repositories.

The Molecular Biology Database Collection, maintained by the journal *Nucleic Acids Research*, this year lists 1,078 databases in its collection—110 more than appeared in 2007. NIH’s National Center for Biotechnology Information (NCBI) maintains more than 40 of these databases, storing molecular, genomic, and scientific literature data. These databases alone see roughly 2 million visitors and 3 terabytes of downloads every day.

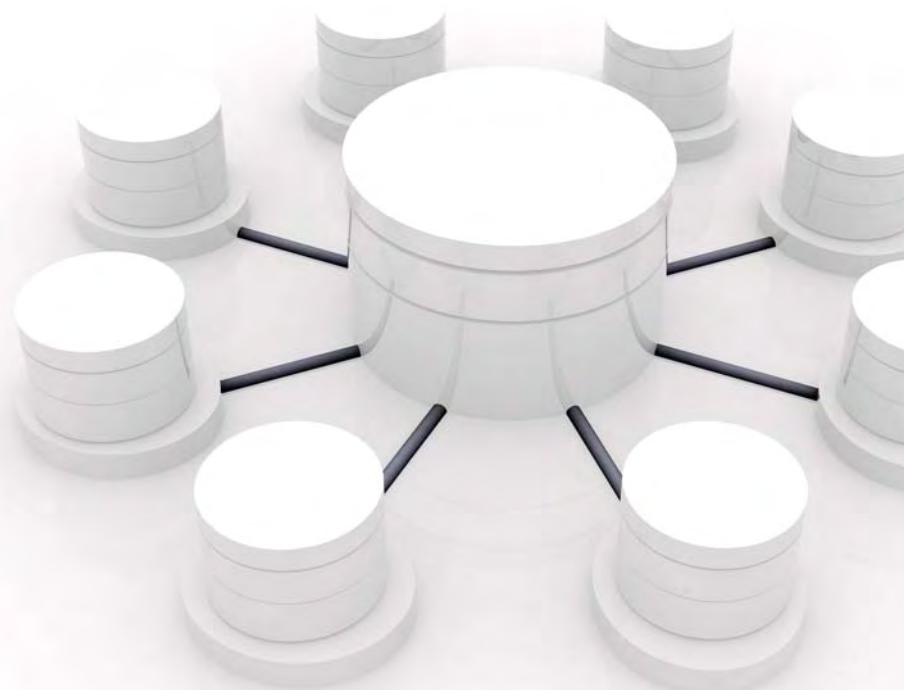
Advances in technology have helped fuel this proliferation, says **George Moody**, research staff scientist in the Harvard/MIT Division of Health Sciences and Technology. He is the architect and caretaker of PhysioNet (<http://www.physionet.org>), a growing archive of freely accessible collections of digitized physiologic signals and time series measurements and related open source software. “For the types of databases that PhysioNet is mostly concerned with, the instruments that gather the data are almost invariably digital now. Our databases are large, but storage is cheap, and adequate network bandwidth is also cheap. So we can afford to collect them and make them available, and users can download them for little or nothing.”

There are also cultural changes behind these trends, says **Atul Butte**,

“It seems like a no-brainer that a portion of our investment in biomedical research should be in the archiving, annotation and maintenance of the resulting data and knowledge,” says Russ Altman.

form *in silico* validation and extension. Essentially, we can more meaningfully join the beginning and end of an experiment through information technology.”

The impact of public databases on the research process is slowly becoming known—with effects on not only *how* the work is done but also on *what* kind of research is done in the first place. Before the golden age of public databases will be able to fully translate into innovative medical advances, however, certain challenges will need to be overcome. But those who work extensively with databases say the benefits will outweigh the costs. “It seems like a no-brainer that a portion of our investment in biomedical research should be in the archiving, annotation and maintenance of the resulting data and knowledge,” says **Russ Altman, MD**,



“The database pioneers have proven their value,” says Teri Klein. “With better understanding and acceptance of databases comes greater usage.”

MD, PhD, assistant professor of medicine at Stanford University, who uses public web-based databases extensively in his research. Increasingly, science is influenced by new movements in “openness,” he says—open-source software, open-access publishing, and so on. This coincides with an increased culture of sharing what were previously proprietary tools of the biomedical trade, such as reagents and protocols. Sharing data is a natural extension of that movement, he notes.

And successful current-generation databases can thank previous database projects for blazing the path, says **Teri Klein, PhD**, senior scientist in the department of genetics at Stanford University, and director of The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB; <http://www.pharmgkb.org/>), which integrates, aggregates and annotates genotype and phenotype data, pathway information and pharmacogenetics. “The database pioneers have proven their value,” she says. “With better understanding and acceptance of databases comes greater usage.”

One of the oldest and most successful of these pioneering databases is NCBI’s GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>), which recently celebrated its 25th anniversary. An annotated collection of all publicly available DNA sequences, GenBank contains nearly 83 billion gene sequences from more than 260,000 different species.

One of the reasons for GenBank’s success is its partnership with journals, according to research by **Nathan Bos, PhD**, senior staff research

scientist at the Johns Hopkins Applied Physics Laboratory, whose chapter, “Motivation to contribute to collaboratories: A public goods approach,” will soon appear in a book called “Scientific Collaboration on the Internet.” Most genetics journals now require authors to deposit their sequence data into GenBank as a prerequisite for publication. This partnership began in the late 1980s as a way to encourage researchers to deposit data directly, rather than rely on GenBank’s staff to input published sequences by hand.

In fact, this system appears to be the most effective in solving the “public goods” problems, Bos concludes. The social dilemma around public databases is that it is difficult to motivate researchers to freely give away their hard-earned data—even though such sharing is ultimately for the greater good of the entire community and, therefore, beneficial for the researchers themselves. The partnership between journals and GenBank works because it ties rewards and sanctions together for the researchers, Bos says.

Of course, not all databases have such a clear mandate. In some ways, biomedicine has been slow to adopt information technology, Buetow says—even though the same tools have already transformed other sectors.

Yet without having access to integrated data resources, Buetow says, “we in biomedicine are going to hit a wall.”

Many biomedical phenomena are complex and need a systems-level approach, for which large shared databases are a natural tool. “We already are increasingly aware, for example, that cancer emerges through complex networks of alterations and we’re going to need combinatorial therapies,” he says. “But it’s beyond the capacity of a single human neural network to be able to integrate all that information. We need this complex network of information sources.”

SPEED AND SYNERGY

Breakneck speed is one of web-based databases’ biggest attractions. Data tasks that would have previously required researchers’ valuable time—to track down, request, transport, and enter—can now be accomplished with a few clicks, even for a casual visitor.

For genetics researchers, quick and easy research verification through databases like GenBank is more than just a luxury, says Nobel laureate **Richard Roberts, PhD**, director of New England Biolabs, Inc. and director of the restriction enzymes database

The social dilemma around public databases is that it is difficult to motivate researchers to freely give away their hard-earned data—even though such sharing is ultimately for the greater good of the entire community.

“Without having access to integrated data resources, we in biomedicine are going to hit a wall,” says Ken Buetow.

REBASE (<http://rebase.neb.com>). “It is possible to check a new sequence against all known sequences within a very short time frame and know you haven’t missed anything. This is very important for avoiding duplication and knowing when your data and inferences truly are new,” he says.

Big research projects can also be accelerated through integrative databases. For example, in 2006, the team of **Howard Fine, MD**, chief of the Neuro-Oncology Branch at NCI’s Center for Cancer Research, published a paper in *Cancer Cell* showing that stem cell factor (SCF) is critical in the genesis of malignant gliomas, the most common form of brain tumors. They had reached the conclusion through exhaustive *in vitro* and *in vivo* studies, Buetow says. But today, he points out, the same conclusion could easily be reached through synthesis of data in the Repository of Molecular Brain Tumor Data (REMBRANDT; <http://rembrandt.nci.nih.gov>), which Fine launched in 2005 to archive information on gene expression, copy num-

ing to get a grant before beginning to look at the data can mean the difference between doing a project or not.”

The best databases offer benefits beyond simple speed, however. Exploring multiple connections in the data can lead to a unique synthesis of knowledge. For example, a large, multi-national team of scientists recently used the data available in the Cancer Genome Atlas (TCGA; <http://cancergenome.nih.gov/>), which houses multidimensional molecular cancer data. The team found that the molecular etiology of glioblastoma, the most aggressive kind of brain tumor, was characterized by a combination of factors: gene mutations, copy number changes, epigenetic silencing, and expression alterations. The work was published in *Nature* in September 2008. “If you could only look at one dimension of that data, and didn’t have the other data accessible in electronic resources, you’d never be able to see this,”

now we can start with public data. Then we figure out a new, useful, and valuable question we can ask and answer. And a completely different question can be asked of the data when it’s put together, beyond the initial question asked. It’s a shift in the scientific method.”

For example, in research currently under publication review, Butte and his team integrated and analyzed publicly available data in the NCBI Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) and the Unified Medical Language System (<http://www.nlm.nih.gov/research/umls/>) to discover a genetic similarity between Duchenne muscular dystrophy and heart attacks. This finding might have clinical value, he says, because while there are no drugs currently developed to treat muscular

“Public data collections free us from the need to recreate them many times over,” Moody says.

ber alterations and clinical information from several thousand patients with malignant gliomas.

In a broader sense, the right databases can speed a researcher’s entire career along, Moody says. “Simply being able to begin a study with suitable data already in hand can mean eliminating the first two years of what would have been a three-year project,” he says. “For a young researcher, graduate student, or researcher seeking to broaden his or her experience, not hav-

Buetow says. “The conclusions are an emergent property of being able to see all the pieces together.”

In some cases, databases can upend the typical research cycle. “A lot of times, a scientist starts with a question, then collects data and answers the question,” Butte says. “But

“Simply being able to begin a study with suitable data already in hand can mean eliminating the first two years of what would have been a three-year project,” says George Moody.

dystrophy, there are several available to treat heart attacks. If the pathways are similar, the heart attack drugs might be helpful in treating muscular dystrophy. And the conclusions were reached by mining publicly available data.

Databases can also divert where researchers devote their efforts. “Public data collections free us from the need to recreate them many times over,” Moody says, “and that means that priorities shift to favor collecting novel types of data, making better use of scarce resources rather than replicating existing databases.” Collections can be expanded in ways that balance depth and breadth, he says. “For example, researchers can collect data from populations complementary to those already well-represented, gather multidimensional data sets that can lead to insights about relationships among variables, or else make use of existing data in novel ways.”

A ripple effect in other fields is becoming clear, as the biomedical community taps into quantitative disciplines for help in dealing with the vast amount of data generated. “This has created unprecedented demand for advanced computational tools and interdisciplinary expertise to capture, store, integrate, distribute and analyze data,” Klein says.

The reach of databases extends beyond the laboratory. For example, clinical cardiac arrhythmia analysis has benefited enormously from databases, including the MIT-BIH (Massachusetts Institute of Technology/Beth Israel Hospital) Arrhythmia Database and the American Heart Association Database for Evaluation of Ventricular Arrhythmia Detectors. “Nowadays it is taken for granted that computers can do a reasonably decent job of detecting important cardiac arrhythmias,” Moody says, which

is a task with utility in both the clinic and the laboratory. “Without shared annotated databases, we wouldn’t have reliable arrhythmia detectors.”

Databases are changing how people work with each other, too. Perhaps most importantly, shared data let researchers in different centers and countries collaborate in novel ways, Klein says. For example, the international warfarin pharmacogenomic consortium (IWPC), which PharmGKB helped broker in 2007, merged datasets totaling more than 5,000 patients from 11 countries and four continents. Their goal was to develop an algorithm for dosing warfarin, an anticoagulant. This merging of data had many benefits, she says: an increased impetus for data sharing, better quality control, and greater statistical analysis power. To address concerns about ownership, the consor-

tium first made the dataset available only to consortium members, Klein says; the entire dataset will be released upon publication of the manuscript.

Being reviewed by many eyes can also increase the value of data, in much the same way that software is improved through an open-source approach. “When many motivated observers examine the same data, their analyses can be compared,” Moody says. “Not only do we learn more about the data as a result of peer review of the data, but we learn more about the analytic methods themselves, about their strengths and weaknesses.”

Sometimes a new database can ignite a new research focus, Moody says. For example, in 2000, Moody and his team created a public, annotated database of polysomnography data and issued an open challenge to the scien-

“A lot of times, a scientist starts with a question, then collects data and answers the question,” says Atul Butte. “But now we can start with public data. Then we figure out a new, useful, and valuable question we can ask and answer.”

“Not only do we learn more about the data as a result of peer review of the data,” says Moody, “but we learn more about the analytic methods themselves, about their strengths and weaknesses.”

“Some people feel current data are too noisy,” says Butte. “I argue they are good enough. As Voltaire said, ‘Perfection is the enemy of the good.’”

tific community: find ways to diagnose sleep apnea using a single ECG signal—a cheaper and less intrusive technique than standard polysomnography methods. “What surprised us was that at least a dozen research teams from around the world took up the challenge,” Moody says. Now clinicians can diagnose sleep apnea with commercially available, clinically certified software that uses their methods, he says, and researchers can also use open-source software based on these methods in their own studies. Better yet, Moody says, “because the researchers had worked independently on a common problem with common data, new collaborations among them formed easily.” The challenge is now an annual event with a different topic and data collection each year.

Even beyond collaborations, freely available data now means that a broader universe of people—not just well-funded labs—can join the research process. Researchers in developing countries have easy access to the types of data they could not afford to generate themselves, Buetow says. For example, he points out, only a little more than half of the usage of GenBank stems from the United States. “We are now able to tap into a global biomedical community of innovative thinkers, such as the billions of imaginations that are present

in India, China, Latin America and in other places,” he says. “Our capacity to solve problems should grow exponentially.”

HURDLES, BOTH TECHNICAL AND CULTURAL

Still, many technical challenges remain in the widespread adoption of databases. The most prominent plague is probably that of noise: inaccuracies in entries and annotations can greatly reduce the value of a dataset. “Some biologists think that there has been a proliferation of databases with low-quality information,” says Altman, whose lab developed PharmGKB. “The quality of annotations and curation is absolutely key for the reliability of the databases.”

In some cases, big (and potentially noisy) repositories are essential—and useful. But it helps if small annotated collections containing the same sort of information also exist. For example, the Broad Institute’s ChemBank and NCBI’s PubChem both house small-molecule structures and screening data. PubChem relies on submission of data and structures from outside sources; ChemBank data are generated and annotated internally.

distinct advantages, says **Paul Clemons, PhD**, director of computational chemical biology research at the Broad Institute of Harvard and MIT.

For some researchers, however, the issue of noisy data isn’t a crucial one. Butte says his approach regarding data accuracy is the same as President Reagan’s during the Cold War: “Trust, but verify.” In data mining studies, it’s not hard to throw out data suspected of having errors. Plus, having several labs contributing similar data into a public database will end up increasing their reliability, he says. “Some people feel current data are too noisy. I argue they are good enough. As Voltaire said, ‘Perfection is the enemy of the good.’”

As mountains of data continue to grow, helping researchers reach them in practical ways will become increasingly difficult. Data need to be both accessible and integrated into other data sets, says **Mark Ellisman, PhD**, professor of neurosciences and bioengineering at the University of California at San Diego and director of Biomedical Informatics Research Network (BIRN). “We need more effective ways to bring data together on the fly in ways that can be visualized and understood by a researcher,” he wrote in Fall 2005 in *The National Academies’ Issues in Science and Technology*. One approach is to prescribe the specific meta-data entities to be used by all sources, such as is done by caBIG. The

“It’s easy to get funded to build a database, but it is hard to get funding for maintenance,” Altman says.

ChemBank also adds value to its data through the storage of other information, such as plate locations, raw screening data, field-based metadata and standard experiment definition. Although PubChem is tremendously useful as a large repository, ChemBank’s annotation and curation offers some

other is developing flexible methods based on dissimilar standards, such as is done in BIRN. Both approaches have benefits and can provide fertile areas of research, he says.

Yet many challenges facing database use are cultural, not technical. Adapting to the needs of clinical sci-

“There is an argument that the NCBI should have all major databases since they are likely to last forever,” Altman says.

entists is one such hurdle. Although results of clinical tests are increasingly being captured in electronic health records, the incorporation of clinical data into large public web-based databases still lags, largely due to privacy concerns and clinical researchers’ unwillingness to share. “We don’t have an equivalent of GenBank for de-identified patient records,” Butte says, but he believes that could change, as he wrote in a perspective in *Science* in April 2008. For example, he says, although clinicians and hospitals might view clinical data as a trade secret, health care networks can pool de-identified data—thus de-identifying the source of the data as well as the individual records themselves. There are projects currently underway to achieve an integrated clinical database, Butte says, particularly at Informatics for Integrating Biology and the Bedside (i2b2), a National Center for Biomedical Computing based at Partners HealthCare System in Boston, Massachusetts. But, he notes, more effort will be needed to make this database available to the entire research community.

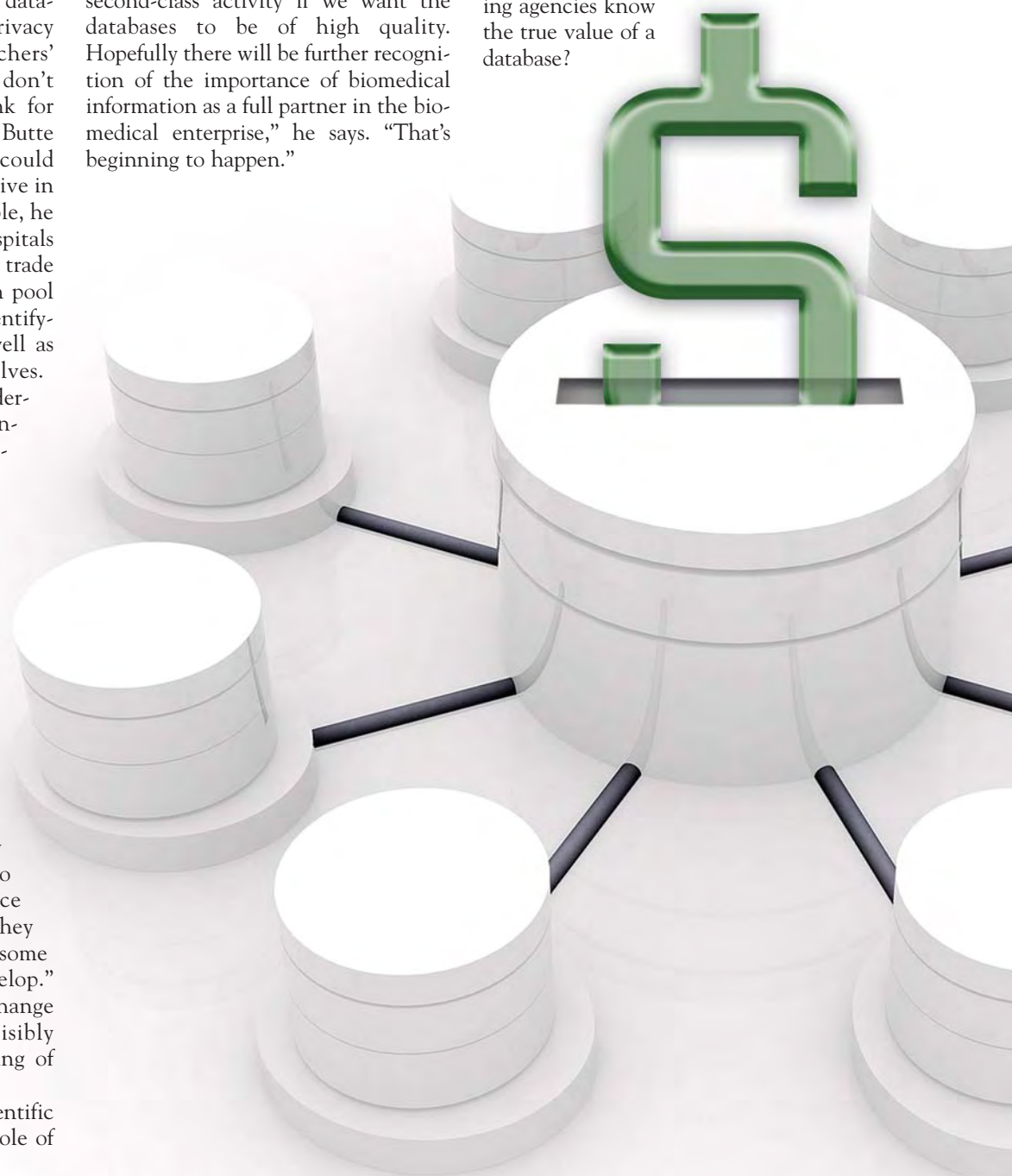
Scientific competition is another cultural obstacle. “Given the way research is funded, many researchers are justifiably hesitant to share their data,” Moody says. “They worry about giving those who compete with them for scarce research funds a look at what they themselves have had to spend some of those scarce funds to develop.” Funding agencies can help change this culture, he says, by visibly rewarding the responsible sharing of data among researchers.

Funding agencies and the scientific community can help boost the role of

the informatics field, Buetow says. “We will need to recognize the true scientific benefit of creating, maintaining and using these databases. That can’t be a second-class activity if we want the databases to be of high quality. Hopefully there will be further recognition of the importance of biomedical information as a full partner in the biomedical enterprise,” he says. “That’s beginning to happen.”

QUIRKS OF FUNDING

In a time of increasing competition for biomedical resources, the question of money looms large. How can funding agencies know the true value of a database?



The simplest method, perhaps, is to examine the citations a database garners in the scientific literature, which provides an indication of its level of use. For example, the annual number of publications based on the MIT-BIH Arrhythmia Database, which has been available since 1980, continues to increase over time. But in general, that's not enough, Altman says. "The reliability of citations to various databases is very low, and often the citation is to the paper whose results are in the database, and not to the database itself."

Ironically, relying on citations would punish the most popular databases. "Widely used and well known databases often don't get cited anyway," Roberts points out. "It becomes assumed that people know what they are and where to find them." And in general, statistics can be misleading. "It can be difficult for funding agencies to assess the worth of a database using traditional peer review mechanisms," Roberts says. "Often the study sections or panels that review database grants lack the expertise to provide a critical assessment. I think there should be a special mechanism set up to review all databases and one that mainly uses expert assessments."

Rather than relying on traditional academic metrics, funding agencies might also do better to turn to commercial metrics when evaluating the worth of collaborative databases, Buetow says. For example, caBIG started when NCI realized that each of its 63 designated cancer centers was independently generating its own information infrastructure. Now with a common infrastructure, a direct return on investment can be calculated, he says, by measuring the difference between the cost of caBIG and the costs of each group collecting data and developing tools individually.

Still, measuring the hypothetical cost of each research team creating its own database isn't perfect, Moody says.

In some cases, it can even substantially undervalue the database. That's because peer review of shared data leads to better quality data, he says, plus the use of the same data in multiple studies generates objective comparisons and insights—which is added value that wouldn't otherwise be measured.

Indeed, many researchers cite funding for maintenance as a top challenge facing the future of public web-based databases. "It's easy to get funded to build a database, but it is hard to get funding for maintenance," Altman says. This is because federal research agencies' desire for novelty is built into their infrastructure. As a result, he says, it's hard to compete with exciting new research ideas.

In the early stages of a database, it's easy to show a connection with a particular research project that moves science forward, Clemons says. "Ironically, once something becomes more useful to more people, it's really harder to pin down a particular beneficiary and show how their grant really benefited from this

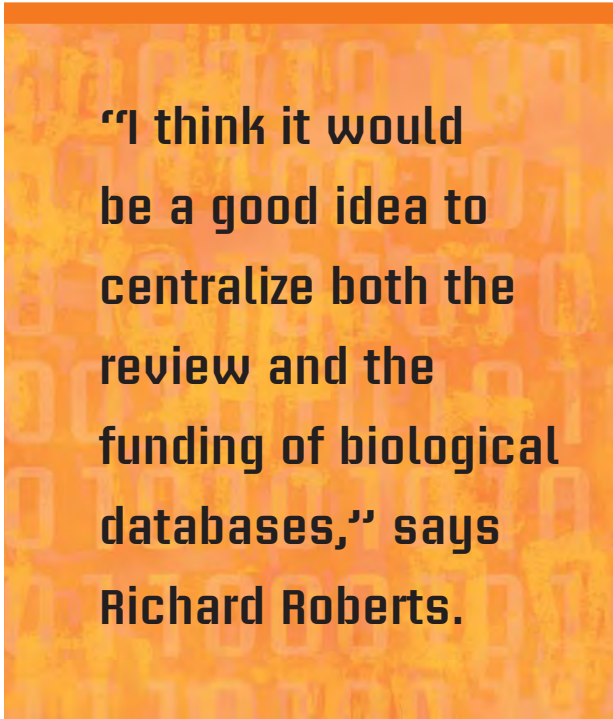
now being distributed throughout the research community and not specifically in the sponsor's area of interest—cancer.

This instability might hurt some databases more than others. "There is an argument that the NCBI should have all major databases since they are likely to last forever," Altman says. Databases created by individual research teams are more vulnerable, since they must re-compete for grant funding every five years. "Why would someone put data in there if the existence is not guaranteed?" he says. On the other hand, he acknowledges, a competitive-funding system might be the very situation that fosters innovation in database technology.

One solution would be to have long-term funding competitively but readily available after a research team has already established a useful database, Roberts says. "It is unreasonable to expect a database team to undergo the vagaries of peer review every three years or so. One poor review and absence of funding can wreck the database," he says, because a suspension of even a year or two in data collection and team continuity severely harms the enterprise. "Since factual databases like GenBank are now critical to modern biology, the government needs to make sure they continue without interruption," he says. "I think it would be a good idea to centralize both the review and the funding of biological databases."

New business models might also help, Buetow says. For example, many modern libraries are starting to consider raw databases to be primary information resources in addition to their collections of books. Also, "groups like Google are actively courting the immortalization of key reference datasets" in non-biologic fields, he says, and they're interested in hosting some raw biomedical datasets.

As the technology and culture of biomedicine continue to change, so too will its practice of storing, sharing, and synthesizing data. Teasing apart the factors driving the evolution may not be simple. "I think large public databases are a symptom of changes in science and they themselves are also changing the face of science," Klein says. □



"I think it would be a good idea to centralize both the review and the funding of biological databases," says Richard Roberts.

activity and how continued support should continue to include a focus on software development." For example, in its early days, ChemBank was funded by NCI, but as the database became more useful to more types of researchers, showing its sponsors that it was an appropriate investment became more difficult; the database's benefits were

By Julie J. Rehmeyer, PhD

SIMULATED METABOLISM

A First Step Toward Simulated Cells



Until biologists really understood the functioning of the genome, they could in principle recreate it *in silico*. Instead of a choreographed swirl of molecules inside a living cell, electrons inside a computer would map out all those cell processes: DNA zipping and unzipping, transporters tugging molecules across cell membranes, enzymes latching on and letting go. >

It's an entrancing dream. For one thing, the process of developing such a simulated cell would help biologists find processes they were missing and show them when they finally understood the microscopic universe inside a cell. And there would be enormous practical value. Drugs could be tested on the model cell long before a single capsule touched a tongue. Your doctor could tell you which diet would be best for you, given your own personal genome. The full impact of genetic diseases could be worked out down to the cellular level.

A fully simulated cell hasn't happened yet, but one system—metabolism—has proven to be one of the simpler systems to tackle.

Bad news, though: "We're not even close to that," says **Joel Stiles, MD, PhD**, a computational physiologist at Carnegie Mellon University and principal co-author of MCell, a simulator of cell microphysiology. Still, it's not just a dream. Researchers have made great progress decoding the functioning of the genome in particular areas. Metabolism, in particular, has proven to be one of the simpler systems to tackle.

Multiple researchers have worked since the 1950s to develop an increasingly thorough and detailed understanding of metabolism in the bacterium *Escherichia coli* (*E. coli*). And over the last fifteen years, **Bernhard Palsson, PhD**, a professor of bioengineering, and his colleagues at the University of California, San Diego have developed and continually improved a very successful *in silico* model incorporating all of that detail.

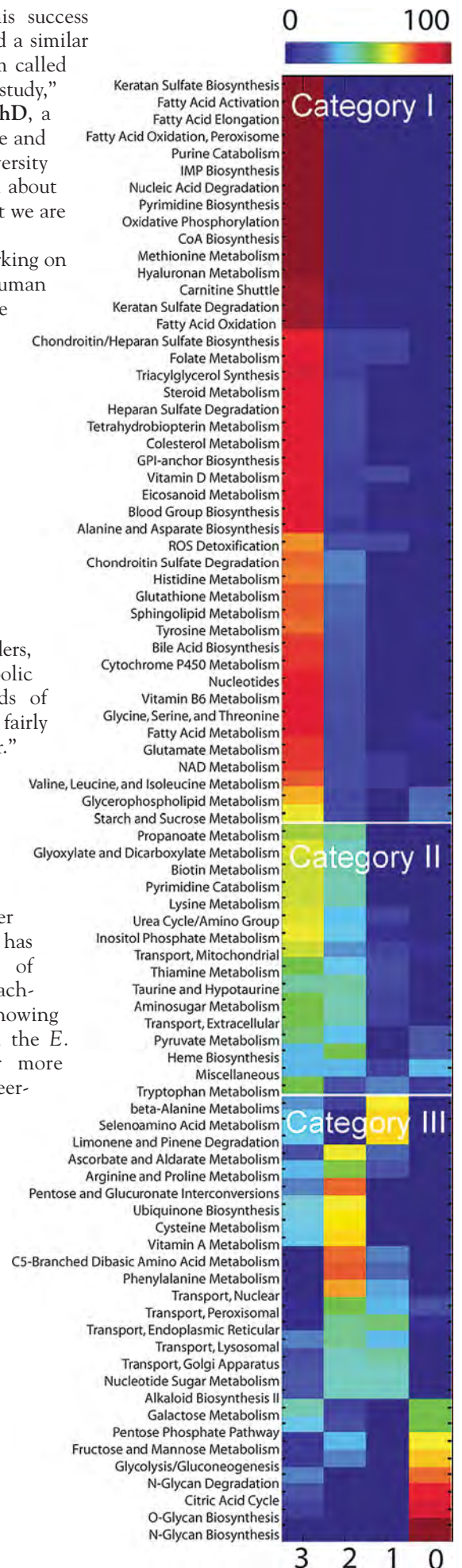
Modeling human metabolism revealed which aspects are well-understood (blue) and which need more research (red). All of the reactions in keratan sulfate biosynthesis (top line), for example, have direct biochemical or genetic evidence, while those involved in n-glycan biosynthesis (bottom line) haven't been evaluated at all. Inositol phosphate metabolism is in the middle, with many reactions having direct biochemical or genetic evidence, some supported by physiological data or evidence from a nonhuman mammalian cell, and a few with only modeling evidence. Reprinted from Duarte, NC, et al., Global reconstruction of the human metabolic network based on genomic and bibliomic data, Proceedings of the National Academy of Sciences, 104: 1777 (Feb 6 2008).

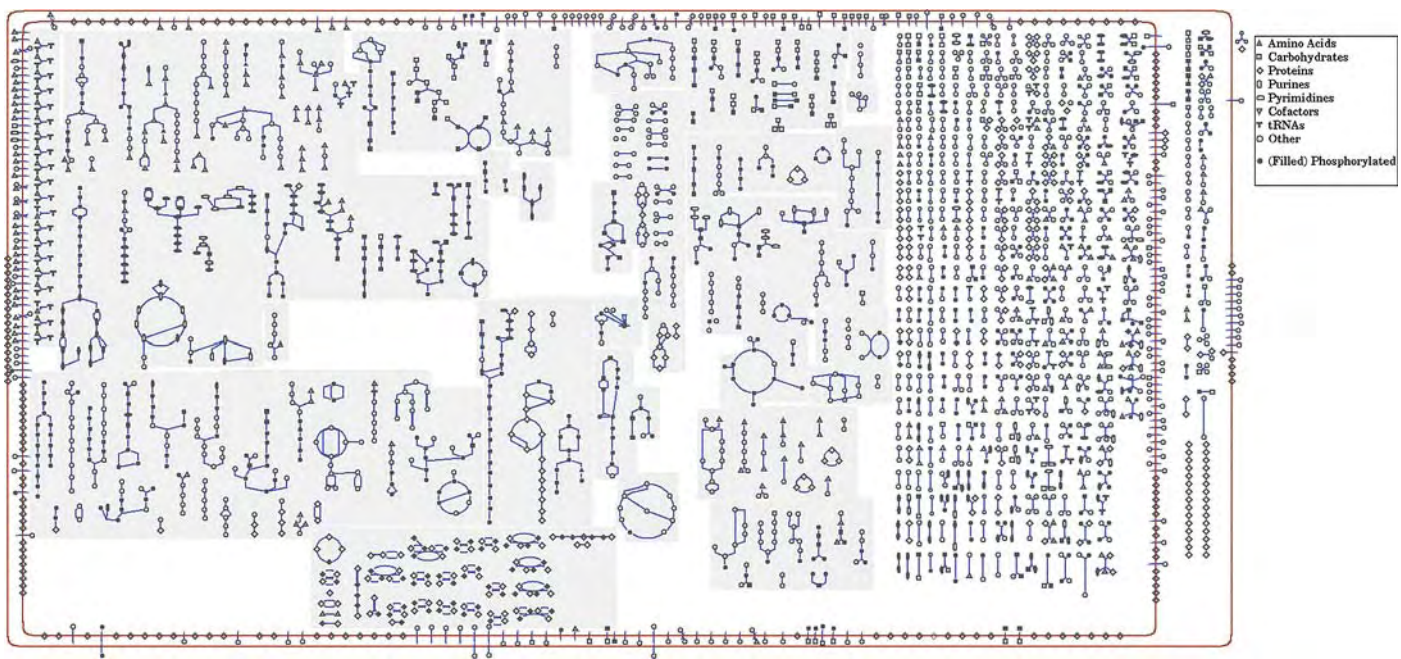
In 2007, building on his success with *E. coli*, Palsson released a similar model of human metabolism called Recon 1. "It's a landmark study," says **Eytan Ruppin, MD, PhD**, a professor of computer science and medicine at Tel Aviv University in Israel. "I'm really excited about this, but I remind myself that we are just at the beginning."

Ruppin's laboratory is working on projects using Palsson's human metabolism model. "There are many things that are really tempting to study,"

he says. "Degenerative disorders, cancer, a variety of metabolic genetic disorders. Hundreds of them could be studied in a fairly straightforward manner." Many other groups have also launched projects using the human model, but it's too soon for its potential to be clear.

On the other hand, the *E. coli* model, developed over many years, has matured and has spawned whole batches of research that are already reaching fruition. In addition to showing how the genome functions, the *E. coli* model has made far more sophisticated genetic engineering possible. Researchers are turning *E. coli* into miniature factories that pump out food additives for





manufacturing, precursors of pharmaceuticals, and even ethanol or butanol. The development of the model has also acted as a spotlight into metabolism itself, guiding lab research.

FINDING MISSING PATHWAYS

To assemble the networks for both human and *E. coli* metabolism, Palsson's group began by exhaustively combing through the existing literature, gathering a list of all known metabolic reactions and their corresponding genes. The reactions form an enormous network, with 2207 reactions in *E. coli* accounting for 1260 of

a torture test. Side by side, they immersed real *E. coli* and their virtual *E. coli* in all kinds of different media. The virtual bacterium "died" if no metabolic pathway connected the medium to all the compounds the bacterium needs to live—essentially telling the researchers "you can't get there from here" on the map. If the virtual *E. coli* died when the real one survived, the researchers knew some pathways were missing in their model. Apparently, *E. coli* was capable of performing previously unknown tricks.

The researchers then evaluated the network to identify reactions that probably needed to be added to com-

Each dot in this graphic is a metabolite. In the online version of this graphic (<http://biocyc.org/ECOLI/new-image>), clicking on a dot identifies the metabolite and the pathways it's involved in. Courtesy of Peter Karp and Suzanne Paley.

E. coli could convert succinate semi-aldehyde to succinate, they hadn't been able to find the gene that made it possible. After adding newly identified reactions to the *in silico* model, the researchers then repeated the testing of the virtual bacterium in an effort to find more missing links.

This same iterative process was used in creating Recon 1, the human model. Because the human model is

Using a combination of literature search, lab work, and iterative computational modeling, researchers have filled in missing links in models of metabolism for both *E. coli* and humans.

its 4453 genes. The human model currently contains approximately 3300 reactions, accounting for about 1500 genes, but will undoubtedly grow significantly over time.

The metabolic networks are set up like a map, with reactions forming roads connecting the metabolites. Energy flows through this network like cars do through a city. To verify this map of the "metabolic city," the researchers began

complete their map, and which genes most plausibly could make those reactions possible. Armed with these hunches, they went back to the wet lab to perform further experiments. Using this method, various teams have identified the roles of eight genes whose functions were previously unknown. They identified the gene for one reaction that had been an "orphan" for 25 years: Although researchers had known that

so much newer and because human metabolism is much more complex, researchers haven't yet pursued all the hints the model has provided about unknown reactions or gene functions. But the group created a "knowledge map," showing which metabolic functions are well understood and which ones clearly have lots missing. For example, cholesterol biosynthesis occurs in the endoplas-

mic reticulum, but the transport mechanism of its precursor from the peroxisome is still unclear, according to the reactions currently known. That's a clear indication that a transporter is missing.

The researchers also developed tools to automate some of the testing process in *E. coli*. "Can we use these systematic procedures to improve Recon 1? That is something the whole community is very interested in doing," Pálsson says.

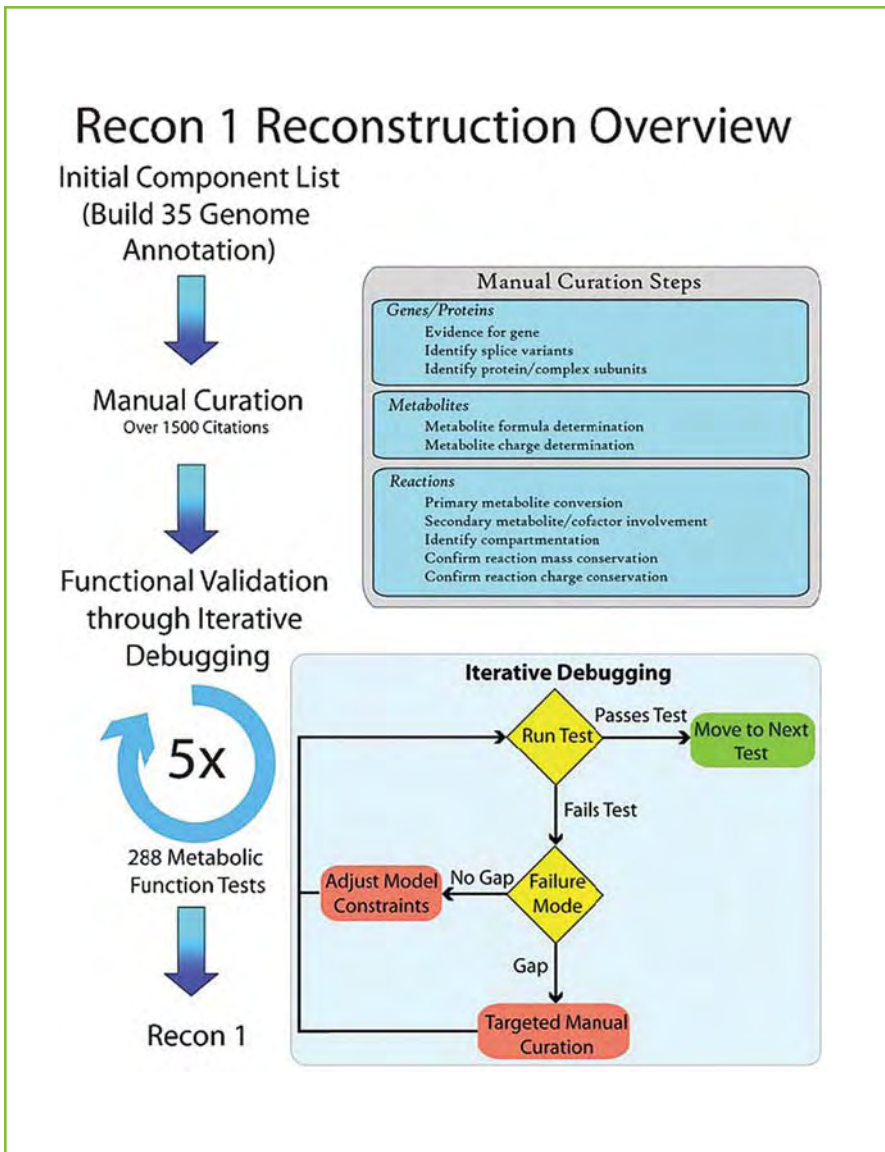
BRINGING THE NETWORK TO LIFE

Having mapped the network, the next step was to build lots of computational tools to bring the map to life. With a street map, you can figure out how to get from one spot to another, but often there are lots of different routes you might take. The same thing is true for metabolism in *E. coli*. "These models can tell you what the organism can do but not what it *will* do," says **Costas Maranas, PhD**, a professor of chemical engineering at Pennsylvania State University.

So the researchers looked for ways to deduce the pathways the bacterium is most likely to use, and to narrow the possible paths it might be taking. But they ran up against a big problem: while the reactions are known pretty well, the particular rates of the reactions aren't. Small differences in reaction rates could have a big impact on which reactions actually happen. So their map was like one that showed the layout of a city without indicating whether any particular street was a mega-highway or a dirt alley.

Masaru Tomita, PhD, a professor of bioinformatics and head of the E-Cell project at Keio University in Japan, is using high-throughput methods to identify these reaction rates and how they change in response to perturbations. This requires quantifying the rates of every different reaction in each possible circumstance—a monstrous task. Many labs have joined forces to make Tomita's project possible, and the group's work toward simulating a whole cell is ongoing.

In the meantime, and looking to simplify matters, Pálsson took a different route. He created a model of how the cell functions when in a steady state. Evolution provides a big clue. "You assume *E. coli* will evolve to use



To create the human metabolic reconstruction, the researchers first assembled a list of the components and preliminary network from the annotated human genome. They then manually reviewed more than 1500 papers to ensure that the network components and their interactions were based on direct physical evidence and reflected current knowledge. Next, they used 288 known functions of human metabolism to test the model and find missing or incorrect links. After making improvements, they repeated the tests four additional times. Reprinted from Mo, ML; Jamshidi, N and Pálsson, BØ. A genome-scale, constraint-based approach to systems biology of human metabolism. Molecular BioSystems 3: 598 (2007). Reproduced by permission of The Royal Society of Chemistry.

"These models can tell you what the organism can do but not what it *will* do," says Costas Maranas.

the resources to grow in the most efficient way," Pálsson says. "Most of the time, that's what it does." So his model calculates the pathways that most efficiently turn a particular compound into all the different compounds needed for growth. Pálsson reasons that those are the ones the bacterium would most likely use.

Experimental evidence supports this approach when the cell's circumstances aren't changing, but researchers have developed other optimization methods that the bacterium may use under dif-

ferent conditions. For example, **George Church, PhD**, a professor of genetics at Harvard Medical School, proposed that after a knockout, cells choose the metabolic pathways that will most quickly return the cell to a steady state.

Unfortunately, Palsson's observation doesn't apply so neatly to human cells, since they don't normally grow boundlessly. Nevertheless, the team developed a description of 288 known metabolic functions in humans—such as the production of the hormone melatonin—to establish restrictions for how metabolites are likely to be processed through the map. Because those restrictions only narrow the possibilities, rather than providing a unique pathway for the metabolism of a particular compound, more work remains to be done.

THE MODEL APPLIED: FROM BIOFUELS TO DISEASE

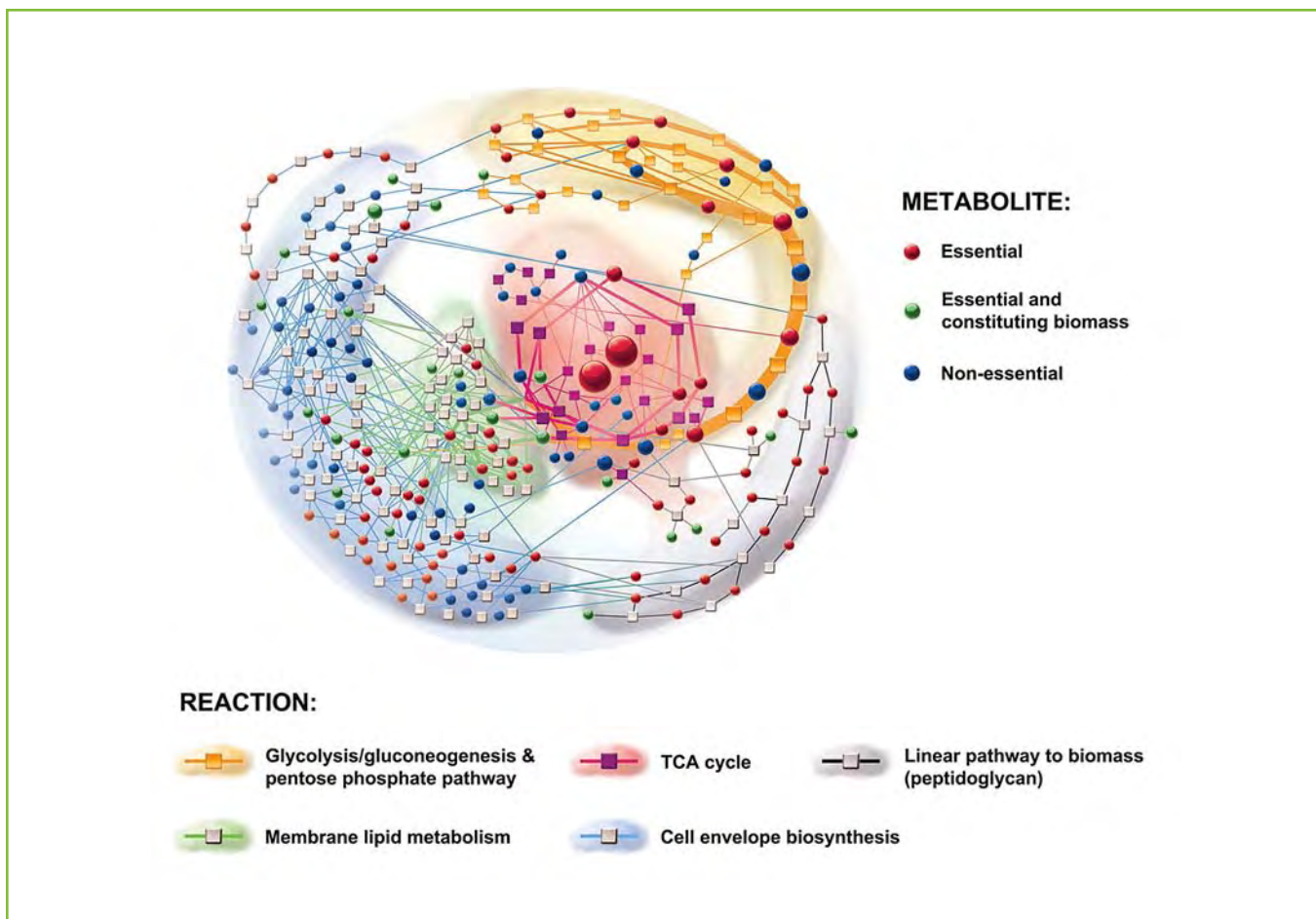
Palsson's group has made their models freely available, and the *E. coli*

model alone has been used in more than 100 research papers by groups around the world. The most obvious way to understand the functioning of the genome is to perturb it and see what happens. This has been done systematically in the lab, but with Palsson's model, researchers can knock out a gene with a keystroke, rather than spending hours or days or weeks creating a genetically modified bacterium. The model also makes it possible to systematically study *E. coli* with multiple gene deletions, or to predict the impact of adding a gene.

"So far, the effect of the model on other peoples' research has been subtle," says **Adam Feist**, a graduate student in Palsson's lab, "but going forward, it's going to be huge. People have emailed me from around the world. They are really starting to catch on."

A big reason for the excitement is that these capabilities have made genetic engineering of *E. coli* dramatically more efficient. **Stephen Fong, PhD**, an assistant professor of chemical

...With Palsson's model, researchers can knock out a gene with a keystroke, rather than spending hours or days or weeks creating a genetically modified bacterium.



A map of *E. coli* metabolism, based on Palsson's models. Metabolites marked in red or green are essential, those in green constitute biomass, and those in blue aren't essential. Reprinted

from Kim, PJ, et al., *Metabolite essentiality elucidates robustness of Escherichia coli metabolism*. Proceedings of the National Academy of Sciences 104: 13638-13642 (Aug 21 2007).

and life science engineering at Virginia Commonwealth University, says that before these models existed, his bio-engineering work was vastly more time-consuming. He would make one modification based on his best guess of what would work, and then test the result. Based on what he found out, he'd make another, and another and another. "Each cycle takes several months," he says. "Simulations literally take less than a second to do." That has revolutionized the process, he says. "You have a way of screening through all the things that seem like they have the highest probability for success before you do any experiments at all."

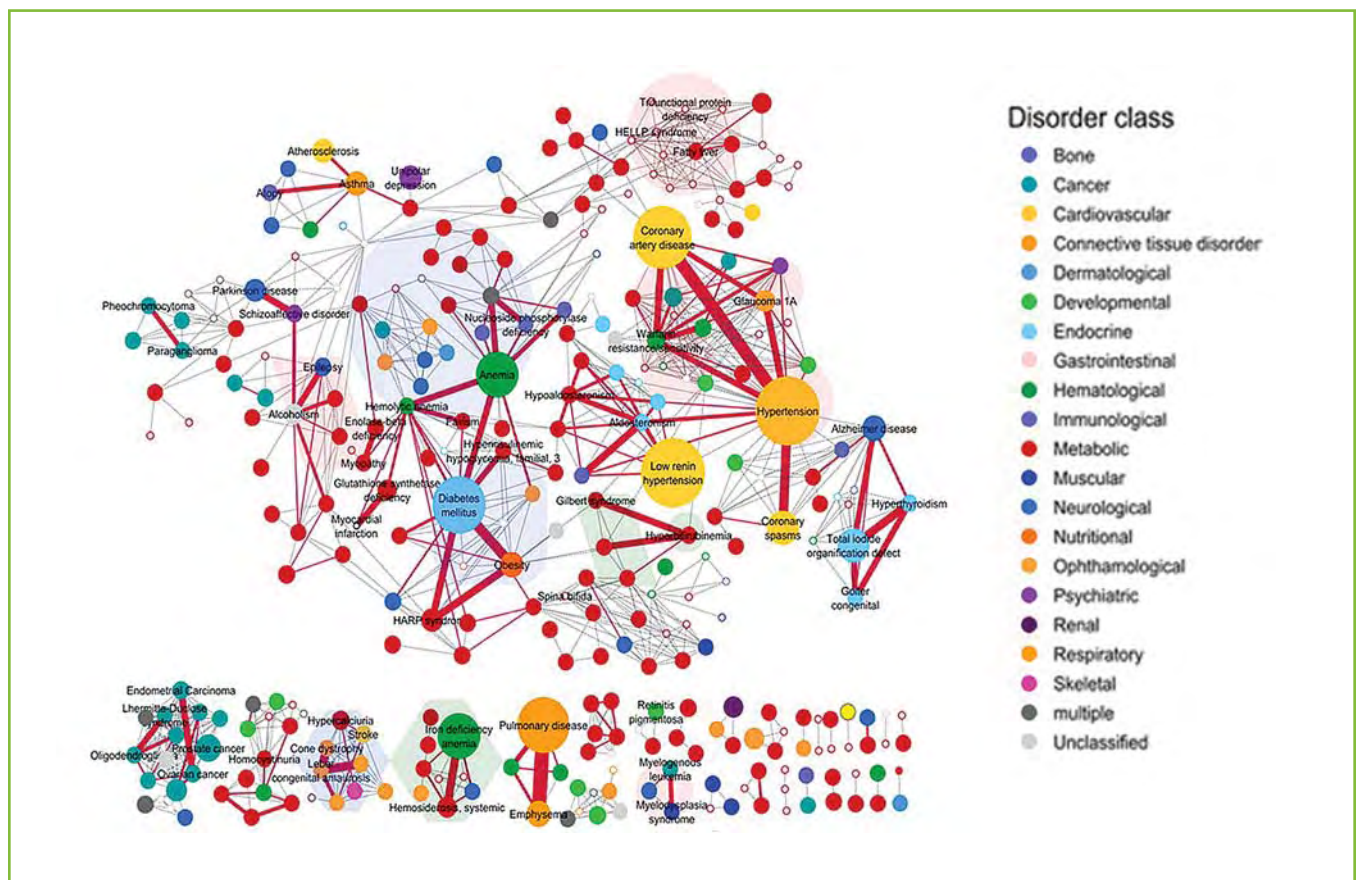
In 2005, Fong pioneered the use of Palsson's model in genetic engineering, creating *E. coli* that produce lactic acid, which is used as a food additive and to create scaffolding for tissue implants. His strategy was to essentially cripple the bacterium by eliminating genes so that its metabolism would be less efficient at turning its food into the compounds it needs to grow. Instead, the bacterium would convert some of its food into a waste product—lactic acid. Fong used Palsson's model to identify the most promising genes to knock out to achieve this, and then he experimentally modified the bacterium to confirm the model's predictions.

Now, he and many other groups are after a more exciting quarry than lactic acid. They want to revolutionize the creation of biofuels using the same process. *E. coli* has been engineered to produce ethanol, as well as fuels like butanol or alkanes that are better substitutes for gasoline.

Church is associated with four different companies that are using these approaches to develop biofuels, and he says the only challenge is to scale up the manufacturing process to create the fuels inexpensively enough. "Inevitably as it scales up, it'll be able to beat petroleum out of the ground," he says.

Church's dreams extend beyond bio-

"Simulations literally take less than a second to do." That has revolutionized the process, Fong says. "You have a way of screening through all the things that seem like they have the highest probability for success before you do any experiments at all."



Barabasi and his colleagues analyzed a Medicare dataset to create this map showing the relationship between diseases. Diseases in the network are connected if mutated genes associated with them catalyze metabolic reactions that are closely related. Diseases that occur more frequently are depicted with larger

dots, and two diseases that tend to occur in the same person are connected with a heavier line. Reprinted with modifications from Less, DS, et al., The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences 105: 9880-9885 (July 22 2008).



fuel as well. “This kind of biochemical engineering is pleasantly easy these days,” he says, because of the rise of these computational methods. “It’s really interesting and fun to use these cellular models to think of all the different products you can make that are currently fairly expensive.”

One of his ideas is to create *E. coli* that he can feed off agricultural waste to produce non-biodegradable precursors for plastics. “If you pull carbon dioxide out of the air and make a wax or a plastic, and don’t burn it and don’t let it degrade, then you’ve had a net loss of carbon dioxide from the atmosphere. Rather than sequestering carbon at the bottom of the ocean, why not sequester it into roads and schools?”

Palsson’s lab is working on biofuels as well, but they’re also pushing to make further improvements in the *E. coli* and human models. They’re inching the *E. coli* model closer to the vision of a fully functioning cell inside a computer by integrating the metabolism model with models of gene regulation and transcription. At the same time, they’re applying the knowledge they’ve gained from *E. coli* to human

E. coli has been engineered to produce ethanol, as well as fuels like butanol or alkanes that are better substitutes for gasoline.

pathogens like salmonella.

The human cell model is developing quickly, both because of Palsson’s work and that of others. Currently, the model includes all metabolic reactions known to happen in any human cell. But only a portion of those reactions occurs in a specific cell type, say, a liver cell or a brain cell or a heart cell. A model identifying which reactions occur in which types of cell will allow

for the study of specific cell types, and that is expected to come out soon.

Then there are applications of Recon 1. “There are so many different ones that it’s hard to choose,” Palsson says. “It’s become clear in recent years that metabolism is involved in all of the major human diseases, either as a consequence or a cause.”

Already, the model is beginning to be used. A team led by **Albert-László Barabási, PhD**, of University of Notre Dame, USA and **Zoltán Oltvai, MD**, of the University of Pittsburgh School of Medicine, used the Recon 1 metabolic network to discover relationships among various metabolic diseases in a Medicare dataset of 13 million patients and 30 million hospital visits. They found that if two genetic diseases were caused by mutated genes whose associated enzymes were close to one another in Palsson’s network, then odds were increased that someone with one of the diseases would have the other as well.

“My expectation is that the human applications will develop much faster than in *E. coli* because of the interest,” Palsson says. “The momentum is enormous.” □

BY JOY KU, PhD

Enhanced Function Recognition in Protein Trajectories over Space and Time

If a picture's worth a thousand words, then a motion picture, such as that provided by molecular dynamics (MD) simulations, must contain a wealth of information. It's this potential payoff that has Simbios investing in the infrastructure to speed up these simulations (see last issue's article on OpenMM) and to store and share the trajectories that are generated.

Research, such as that of **Dariya Glazer**, a graduate student in genetics at Stanford University whose work is partially supported by Simbios, illustrates the insights that MD simulations can provide. In recent work, she demonstrated that simulating how a molecule moves over time can lead to more accurate predictions of molecular functional sites, such as active enzymatic or drug-binding sites. Her poster based on this work received an Outstanding Poster Award at the Intelligent Systems for Molecular Biology (ISMB) conference in Toronto in July 2008.

"Dariya has shown that simulating the motion of these proteins using molecular dynamics can markedly improve the ability to detect function and should probably be routinely employed by these algorithms," says **Russ Altman, MD, PhD**, Glazer's advisor who is also a professor of bioengineering at Stanford University and a principal investigator for Simbios.

Most function prediction algorithms use data from experimental techniques, such as X-ray crystallography, which

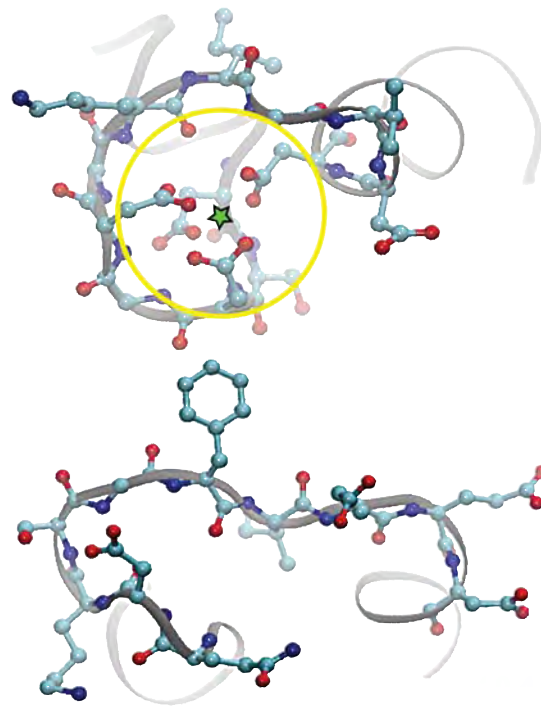
show a molecule's structure at one particular moment in time. But the molecule may not be in a functional configuration at that instant, resulting in incorrect predictions.

"Unfortunately, most prediction methods forget that the beautiful crystal structure of a protein is nothing but a snapshot of what the actual protein looks like *in vivo*," says **Marco Punta, PhD**, a research scientist in biochemistry and molecular biophysics at Columbia University and the posters committee chair at ISMB this year.

To increase her odds of making accurate predictions, Glazer used MD simulations to model molecular motion. From the resulting molecular trajectories, she extracted hundreds of frames and then applied traditional function prediction methods to each of those frames to identify potential binding sites for calcium.

For one protein she examined, the static structure offered no hints of a binding site. "It was only during the simulations that the binding site adopted an appropriate conformation so that the algorithm could identify it," she says.

After running the MD simulations, Glazer faced a new problem: how to combine the information from the tens to hundreds of structures that presented



Glazer's study showed that molecular dynamics simulations can increase the accuracy of predictions about a molecule's functional sites, as compared with a static image. Shown above are two frames from a simulation. (top) A molecular configuration with a likely calcium-binding site, shown by the green star. (bottom) The same molecule from a different time point in the simulation that is unlikely to have any calcium-binding sites. A static image would show the molecule in only one of the configurations.

with potential binding sites. "It wasn't always obvious to us whether the data represented a single binding site or several." So she came up with a multi-tiered clustering scheme to identify the independent sites.

With the MD simulations and her new clustering scheme, Glazer was able to identify upwards of 60 percent more true calcium binding sites. In fact, one of the prediction methods could not identify any binding sites at all without the MD.

Glazer is looking forward to experimenting with MD simulations on graphics processing units (GPUs), something that Simbios is making possible. "I'll be able to investigate more complicated functions in larger systems," she says. "The possibilities are exciting." □

DETAILS

Glazer tested her approach using two different prediction algorithms: **FEATURE** (<http://simtk.org/home/feature>), which looks at about 80 different properties to make its prediction; and a valence method, which uses a molecule's local charge to determine where it might bind.

Her test cases consisted of both a calcium-bound and a non-calcium-bound version of five different proteins. The simulation trajectories generated for these molecules will be made available at <http://simtk.org/home/mdfxnpredict>.



BY MATHIAS BROCHHAUSEN, PhD



How Upper Level Ontologies Deal With Functions and Other Realizable Entities

Before categorizing things, you have to decide on the categories. For material “things” (e.g., molecules, organs, etc.) or entities, the task is relatively straightforward. But often in biomedicine, you need to also categorize abstract aspects of these material entities, such as their function or role. To tackle that task, ontologists create what are called “upper level” ontologies. Such ontologies (e.g., DOLCE, BFO) provide a basic classification of reality without addressing domain-specific entities (such as heart, platelet, or patient). Upper ontologies support the process of ontology development by providing a first framework. Furthermore, they foster harmonization among ontologies by representing the root classes.

One upper level task that has been neglected to date by the most widely used terminology resources is the coherent representation of terms like function, role, dis-

position and tendency. These abstract aspects of material entities may encode more knowledge than the entities themselves. For example, the function of red blood cells to transport oxygen, the function of the heart to pump blood, and the function of sexual reproduction to generate genetic variability, are supremely important concepts in biomedicine, and understanding them can help us determine how to fight disease. Yet they are not easily described using today’s biomedical ontologies.

So, how should functions, roles, tendencies and dispositions be represented in biomedical ontologies? They are what we can call realizable entities. They are marked by their realizations: functions are realized by their bearers being active in a specific process, roles are realized by the processes being performed in the corresponding contexts (examples: the student role is realized when a person studies; the pathogen role is realized when a bacteria infects).

Ontologies need to represent realizable entities cor-

Function, role, disposition and tendency: These abstract aspects of material entities may encode more knowledge than the entities themselves.

position and tendency. These abstract aspects of material entities may encode more knowledge than the entities themselves. For example, the function of red blood cells to transport oxygen, the function of the heart to pump blood, and the function of sexual reproduction to generate genetic variability, are supremely important concepts in biomedicine, and understanding them can help us determine how to fight disease. Yet they are not easily described using today’s biomedical ontologies.

As another example, people and other entities can act in specific ways, what we can think of as roles—the role of my mother as a patient, the role of an electrode array as a prosthesis, the role of belladonna as a drug, or the

role of a bacterium as an infectious agent. There are also dispositions and tendencies—for example the tendency of smokers to develop a cancer, and the disposition of a zygote to develop into a morula. These, too, are entities that do not exist independently of their bearers. And all of these are critical to understanding biological processes.

From a theoretical point of view, scientists’ neglect of functions in ontologies and terminologies might reflect an aversion to introducing an unwelcome teleological element into the domain of biological reality. We scientists exploring functions talk as if biological systems worked towards aims in contradiction to commonly accepted interpretations of Darwinian biology. I am strongly convinced, however, that we should not shy away from talking about functional aspects of organisms and their behaviour. This should not arouse objections from true Darwinians since it is functionality, after all, that explains why organisms can be viewed as survival machines. It is functionality, too, which offers our most coherent understanding of what clinical medicine is all about. □

rectly for several reasons. First, they really are different in nature from the material entities with which they are associated. For example, a bacterium is not the same thing as its role as an “infectious agent.” Second, an object might have multiple functions, such as a chemical substance that could be used as either a drug or a poison. Likewise, a function may apply to several things. Think about the function to pump blood. It might be carried out by either a heart or a machine.

DETAILS

Mathias Brochhausen, PhD, is a researcher at the Institute of Formal Ontology and Medical Information Science (IFOMIS), Saarbrücken, Germany and Executive Director of the European Consortium of Ontological Research.

For further information, see Realizable entities in Basic Formal Ontology (BFO) <http://www.ifomis.org/bfo>, & Robert Arp and Barry Smith: Function, Role; and Disposition in Basic Formal Ontology, in Proceedings of Bio-Ontologies Workshop (ISMB 2008), Toronto, 45-48, <http://bio-ontologies.org.uk/download/Bio-Ontologies2008.pdf>

Biomedical Computation Review

Symbios A NATIONAL CENTER FOR BIOMEDICAL COMPUTING

Stanford University

318 Campus Drive

Clark Center Room S231

Stanford, CA 94305-5444

seeing science

SeeingScience

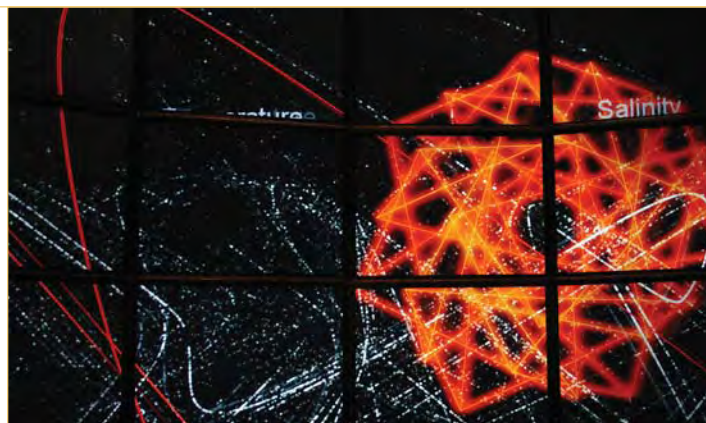
BY KRISTIN SAINANI, PhD

Sensational Sequences

What's it like to be immersed in a dataset of millions of DNA sequences? Audiences of *ATLAS in silico*—a new media artwork that explores novel ways to represent and intuitively understand nature in the metagenomic era—are about to find out. The installation, which is a hybrid of art, science, and technology, was displayed in Cleveland in July 2008 and will be displayed in Los Angeles in November. It is expected to reach more than 100,000 people.

ATLAS in silico transforms raw metagenomics data—predicted protein sequences derived from millions of ocean-dwelling microbes collected by the Global Ocean Survey—into haunting digital sounds and luminous 3D geometric forms that appear in a virtual world. Head- and hand-tracking systems allow users to “push and pull” on objects in 3D to see more detail. “The experience is one of being immersed in something that is flowing, as if you’re in a kind of fluid,” says **Ruth West**, who leads the collaborative project. West is director of interactive technologies at the University of California, Los Angeles, Center for Embedded Networked Sensing, and artist-research associate with the University of California, San Diego, Center for Research in Computing and the Arts.

“The whole idea is you are able to get some sense of the internal structure or patterns within the highly abstract data that you can viscerally relate to,” West says. More information is at: <http://www.atlasinsilico.net/>. □



ATLAS in silico uses a custom algorithm to translate genomic data, as well as social and environmental data from regions where the biological samples were collected, into unique 3D shapes, which are displayed on a room-sized, 100-million pixel semi-circular tiled display. The installation is equipped with computer vision, which allows users to interact with the data through simple hand and head movements.

Biomedical Computation Review

Simbios A NATIONAL CENTER FOR BIOMEDICAL COMPUTING

Stanford University

318 Campus Drive

Clark Center Room S231

Stanford, CA 94305-5444

what *you* think

What YOUThink

Controversy in Biomedical Computing

In our next issue, we'll introduce a column featuring topical debates between leaders in the field of biomedical computing. You'll read what prominent researchers think about controversial topics facing biomedical computing—and have the chance to share your own opinions.

BUT FIRST:

Go to our Web site (<http://biomedicalcomputationreview.org>) to vote on which of the following topics you're most interested in, take our survey, and enter your name as a possible debater. **You could even win an iPod.**

TOPICS:

To Fund or Not To Fund: Should grant applications for the development and maintenance of software and infrastructure compete against basic research applications or should there be a separate mechanism?

To Mine or Not To Mine: Are clinical data repositories useful sources of untapped discoveries awaiting data-mining algorithms or are they too noisy and messy?

Too many tools in the toolbox?: Is the massive proliferation of analytical tools for biomedical informatics diluting the best ones and limiting their visibility and usage?

Open source vs. proprietary research: Is open source unfair to scientists whose primary work product is computer software? Do other engineers just give away their inventions?

Technology Transfer: Is technology transfer in biomedical computing occurring sufficiently?

GPUs vs. ASICs: Will GPU computing dominate high performance computing in the next decade? Are ASICs (application-specific integrated circuits) the next big thing?

Dry vs. Wet: Should the computational tool developers of tomorrow be spending significant time doing wet lab work for exposure?

Cloud Computing—Here to Stay or Gone Tomorrow?: Is cloud computing a great, relatively untapped resource for the scientific community or is it over-hyped?

Journal Requirements or Individual Choice?: Should journals require that software and data be made available in public repositories before a related publication is accepted?

XXX vs. XXX: What do YOU think is the hottest topic in biomedical computing?