DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

Meet the SKEPTICS

Why Some Doubt Biomedical Models—and What it Takes to Win Them Over

PLUS:

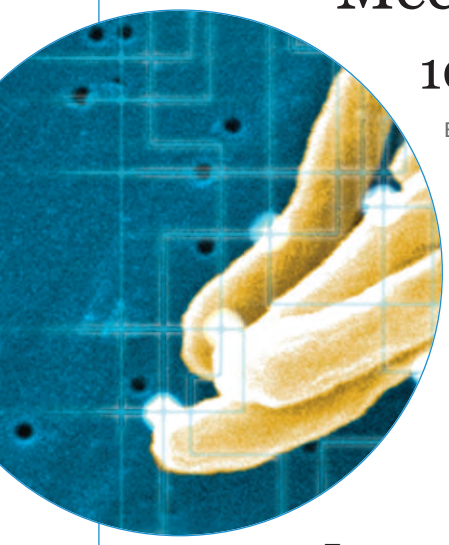
Where Tuberculosis Meets Computation: 10 Points of Intersection

INSIDE:

Summer 2012

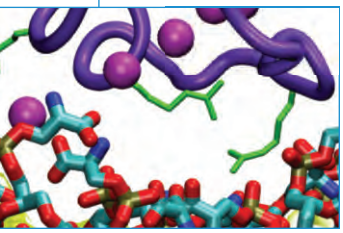
contents

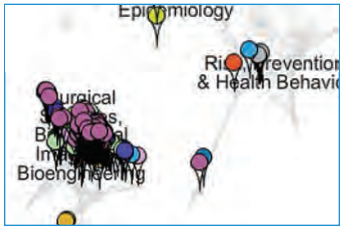
ContentsSummer 2012

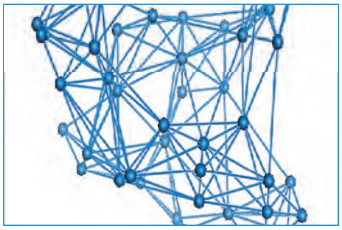


DEPARTMENTS

Cover Art: Created by Rachel Jones of Wink Design Studio using: Man in suit image © Monkey Business Images | Dreamstime.com. **Page 19:** Created by Rachel Jones of Wink Design Studio using: digital circuitry image © Siminitzki | Dreamstime.com, and TB virus image, courtesy of Centers for Disease Control, Ray Butler & Janice Carr.







Summer 2012
Volume 8, Issue 2
ISSN 1557-3192

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 by:

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

Publication is made possible through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Grace Peng, PhD (NIBIB)
Jim Gnadt, PhD (NINDS)
Peter Highnam, PhD (NCRR)
Jennie Larkin, PhD (NHLBI)
Jerry Li, MD, PhD (NIGMS)
Nancy Shinowara , PhD (NICHD)
David Thomassen, PhD (DOE)
Janna Wehrle, PhD (NIGMS)
Jane Ye, PhD (NLM)

GuestEditorial

BY JIM DELEO, PhD, NIH COMPUTER SCIENTIST

Identifying and Overcoming Skepticism about Biomedical Computing

Many collaborators[1] with whom modelers[2] work have little or no training in modeling[3] and so it is natural that they may be cautious, intimidated or disinterested—attitudes that give rise to skepticism[4]. Although, ideally, collaborators could learn more about modeling, it is understandable that they don't: They are busy keeping up with their own dynamically changing specialty fields, don't have the time, or are simply not interested. Identifying and overcoming such skepticism is important if biomedical computing is to be of greater value to society, and so I would like to suggest here that we, the modelers, take the lead in addressing and reducing this skepticism.

> We modelers also need to integrate and standardize our style of thinking as well as our terminology and nomenclature. ... Even within modeler subgroups, individuals think differently about their approaches to modeling. We need to focus on concept consilience and common ontologies!

I hope we can agree that there really is no well-established "modeling community." Typically, modelers are renegade individualists who are fuzzy members of the fuzzy subsets of different modeling disciplines such as computer science, statistics, bioinformatics, analytics and others. It would be helpful if these renegades would transcend their silos, overcome their self-oriented competitive urges and establish more cooperative relationships with one another and with their collaborators. This objective has motivated the NIH Biomedical Computing Interest Group (BCIG) since its inception 10 years ago. BCIG's mission is to encourage, support and promote good and appropriate computing methodology and technology in all aspects of biomedical research, development, and patient care; and it is open to everyone having interest in this mission. I propose that we form other geographically disparate BCIG groups and network them electronically. Are you interested? I would be happy to help facilitate this.

Conference participation, tutorial production and distribution, crowdsourcing and multi-institutional team building are examples of what we can do to improve relationships and extend computational methodology choices and accessibility. For example, BCIG is helping to formulate a panel for a workshop on "Proper Methods for Evaluating Performance of Computational Intelligence Methods and How to Encourage Use of these Evaluation Methods." This workshop has been proposed for the 2012 World Congress on Computational Intelligence. As another example, BCIG is about to put in place a mechanism for modelers who subscribe to BCIG to brainstorm on broad biomedical computing topics—a

DETAILS

Jim DeLeo has been a computer scientist for over 40 years during which he has designed, developed and implemented new and innovative computational solutions to solve medical, space exploration and defense problems. Presently at the NIH, he works collaboratively with most of the NIH institutes and centers, other government agencies, universities and industry. His current work is inspired by the NIH Roadmap translational medicine theme and is directed toward building computational, intelligent systems that have practical impact in improving patient care.

kind of local crowdsourcing operation. The first topic will be "Machine Learning and Statistics: the Interface."

We modelers also need to integrate and standardize our style of thinking as well as our terminology and nomenclature. Many other fields do this as they begin to mature. Statisticians, computer scientists and bioinformaticians think differently from one another. Even within modeler subgroups, individuals think differently about their approaches to modeling. We need to focus on concept consilience and common ontologies!

Modelers should try to convince collaborators that modeling is meaningful even when the models may be imperfect. The key is to demonstrate success in significant collaborative biomedical projects—in particular (given current priorities) in translational medicine projects, i.e., projects with results that have a direct and important positive impact on health care. Although many collaborators may not be skeptical per se, some fail to see the value of using modeling in their fields. This can be framed as a challenge for modelers. They can explore these fields and find better ways to introduce modeling. I can point to several examples where computational modeling demonstrated the potential to have significant impact on medicine and biology, particularly with respect to translational medicine. For example, I have developed methodologies that predict glucose tolerance test results, breast cancer, and adverse drug reactions with accuracies suitable for clinical use.

Unfortunately, there have been cases in which modeling has produced overhyped, misleading, or flawed outcomes. Years ago a modeler claimed that his artificial neural network (ANN) computer program could predict whether a patient presenting at an emergency room with certain symptoms and findings should be admitted to the ICU. He claimed that his ANN could do a better job than human experts faced with the same task, but his performance statistics were based only on the data used in the ANN training. He had no hold-out data for testing and validation. This is the kind of ill-designed hyped work that gives bad press to modeling. Like all good science, modeling needs good statistical oversight, which includes proper testing and validation. But modelers are often not doing this. We must correct this. When proper testing and validation are missing, it provides strong support for certain groups (e.g., certain fundamentalist, turf-protecting statisticians) who feel that these new-fangled tools from computer science are threatening their professional identity. Computer scientists and other modelers must learn to properly validate their models according to standards set by good classical statistical methodology. I know of other horror stories of modeling misuse. One example is where a physician used evolutionary computing to fit data in an application where a simple linear regression model would have been sufficient.

Like all good science, modeling needs good statistical oversight, which includes proper testing and validation. But modelers are often not doing this. We must correct this.

I have suggested here that we, the modelers, take the lead in addressing skepticism associated with biomedical computing and that we do what we can to reduce it. I have suggested several specific things we can do in this regard, namely (1) create other BCIG groups like the NIH BCIG and network them, (2) engage in conference participation, tutorial production and distribution, crowdsourcing and multi-institutional team building, (3) integrate and standardize our style of thinking, our terminology and our nomenclature, (4) demonstrate success in projects, particularly in translational medicine projects, and (5) avoid overhyping, and misleading and flawed outcomes. □

Footnotes:

1. Physicians, biologists and others who work in biomedical research and health care delivery
2. Fuzzy heterogeneous collection of individuals who work with all types of computational tools used under the general rubric "biomedical computing"
3. Developing algorithms and computer programs to solve specific problems
4. Any questioning attitude towards knowledge, facts, or opinions stated as facts, or any doubt regarding claims

BY KATHARINE MILLER

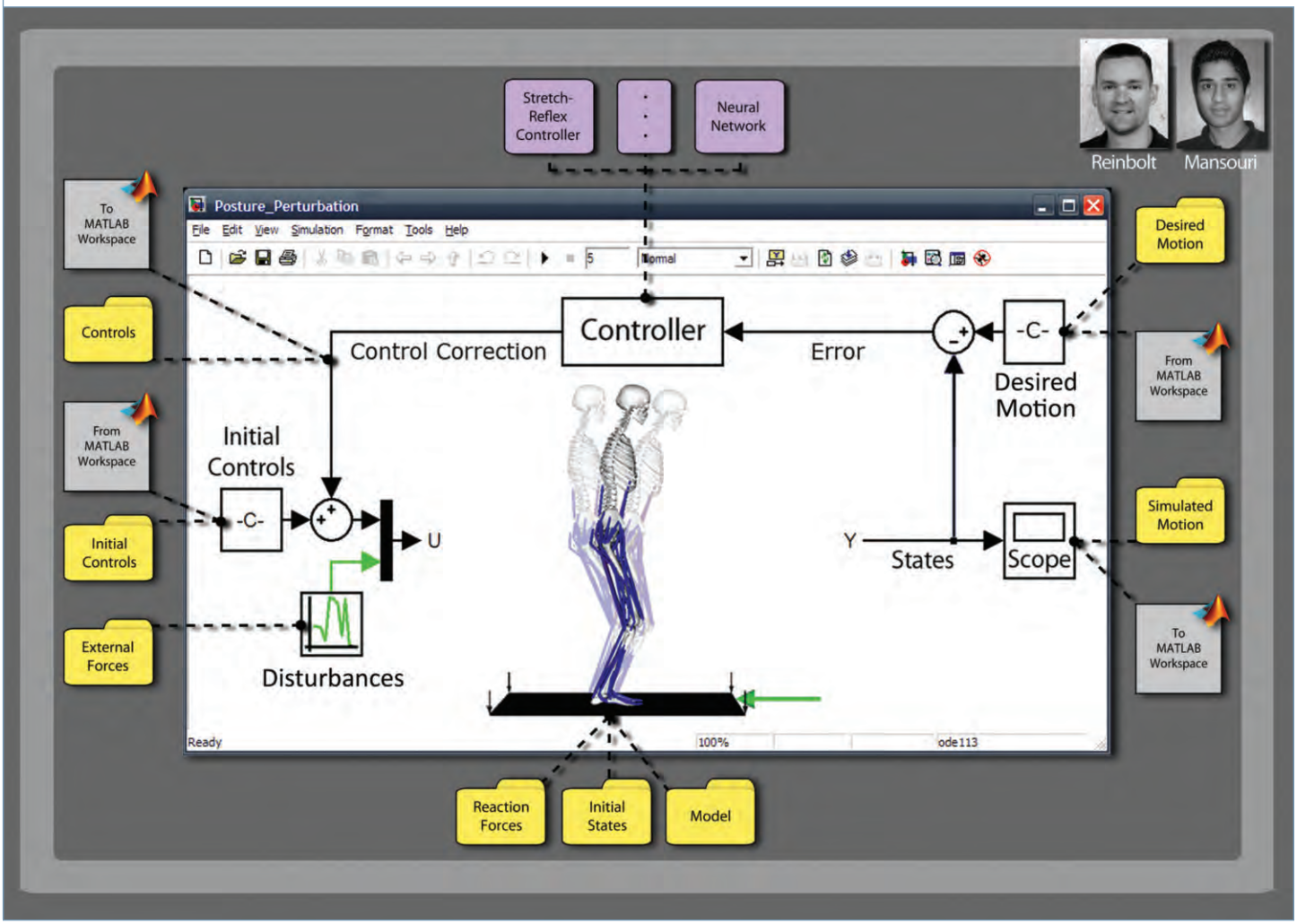
Feedback for the Brain and Body: A New Freely Available Interface Between MATLAB and OpenSim

Even when we simply stand still on two feet, our brains communicate with our muscles—firing them appropriately to keep us upright against gravity. So when scientists simulate simple or complex biomechanical movements, they need to account for that feedback between brain and body. A new, freely available, software interface now makes that possible. It connects MATLAB/Simulink, a mathematical computing and control software package, with OpenSim, a freely available musculoskeletal software package.

"MATLAB is essentially the brain and nervous system, but OpenSim is the person, their bones and muscles, the floor, and gravity pulling them down," says **Jeff Reinbolt, PhD**, assistant professor of biomedical engineering at the University of Tennessee. He and **Misagh Mansouri**, a mechanical engineering graduate student in Reinbolt's research group, created the interface under a seed grant from Simbios. They published a 2012 paper on the work in the *Journal of Biomechanics*.

OpenSim's strengths lie in its musculoskeletal models, Reinbolt says, "It has the bone geometry, muscle forces, how the joints move, and all the dynamics provided by the underlying algorithms, including Simbody." But until now, OpenSim users have had to load a file of controls that tell the model how to excite specified muscles over a certain

Reinbolt and Mansouri built an interface that lets MATLAB control an OpenSim simulation. As shown here, the reaction forces, initial states, and model (yellow files at the bottom) come from OpenSim while the controls come from MATLAB. Courtesy of Jeff Reinbolt and Misagh Mansouri.



period of time. To change controls, the user would create a new file. In addition, OpenSim could only function as an open loop, without feedback; users requiring feedback to close the loop would need to write their own software as a plugin to OpenSim. Without feedback, "you can change the model and do 'what if' scenarios, but the model has no brain to automatically compensate for it," Reinbolt says.

Meanwhile, MATLAB/Simulink provides great control options in the form of templates and blocks of code that are easily changed on the fly, but it lacks the musculoskeletal modeling capability of OpenSim.

Reinbolt and Mansouri's Simbios seed grant let them bring the two together. "The interface really allows MATLAB to access OpenSim for the muscles and bones it doesn't have; and allows OpenSim to use MATLAB to control the movement, which OpenSim doesn't really have built in," Reinbolt says. "It allows each to get the info needed to create a simulated movement." And it lets you change things on the fly.

To build the interface, Reinbolt and Mansouri took advantage of the way MATLAB calls functions to connect it with OpenSim through a Simulink block. The interface allows MATLAB to call for the OpenSim model's state derivatives—the time rate of change in joint positions and velocities of the model, as well as lengths and activations of muscles. MATLAB numerically integrates these derivatives to determine new states for the OpenSim model which allows OpenSim to tell MATLAB how gravity, muscles, and other forces are affecting the model, and then MATLAB uses that information to provide feedback: Are my joint positions at the right place? How should I adjust the controls to correct the movement?

Right now, to use the interface, researchers must join the Simtk.org project. "It's freely available but we want people to use it and give back in return," Reinbolt says. About 25 people from the US, Canada, UK, Belgium, Poland, Spain, Italy, and Taiwan have already shown interest.

Reinbolt's research group is already using the interface to simulate posture. They've added a stretch reflex controller to allow balancing on two feet. Eventually, they would like to create a controller that will reproduce someone walking with stroke gait so they can then test rehabilitation procedures to predict how the patient might walk better. "The point," Reinbolt says, "is to keep someone from falling over," which requires the feedback the interface provides. □

DETAILS

The MATLAB/Simulink interface with OpenSim is available to researchers who join the project at http://simtk.org/home/opensim_matlab. The work was published in the *Journal of Biomechanics* 45:1517-21 (2012).

Simbios (http://simbios.stanford.edu) is the National Center for Physics-Based Simulation of Biological Structures at Stanford.

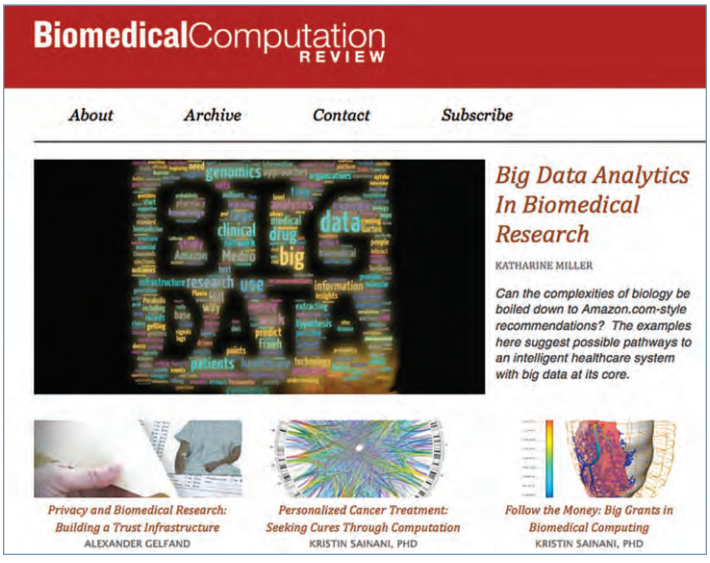
New Magazine Web Site

We are excited to announce the new website for *Biomedical Computation Review*! The site (http://biomedicalcomputationreview.org) now allows you to easily link from our stories to related web pages such as journal articles and researchers' websites; comment on and "like" stories; and find related stories.

In addition, we will bring you occasional online-only content about the latest in biomedical computation. For example, in March we posted a 2012 Update on the National Centers for Biomedical Computing—directing readers to the March 2012 issue of the *Journal of the American Medical Informatics Association* (JAMIA), where the NCBC principal investigators and their teams highlighted their accomplishments.

Enjoy!

The BCR Editorial Team



ZOOMING IN ON BLOOD COAGULATION AND VISCOSITY:
Computation Takes On Blood Behavior

By Alexander Gelfand

Understanding blood flow and coagulation is crucial to treating blood disorders such as hemophilia and thrombosis, and to dealing with diseases such as AIDS, malaria, and diabetes that have hematologic consequences.

It's also bloody difficult. Although it behaves like a homogeneous fluid in large vessels such as arteries, human blood is really a suspension of solids (blood cells, platelets) that can alter their characteristics in response to chemical and physical provocation. In smaller vessels such as capillaries and arterioles, those particles cause blood to act like a non-Newtonian fluid, similar to ketchup, whose viscosity is subject to change. Coagulation, meanwhile, involves a complicated dance between cell membranes and biological molecules.

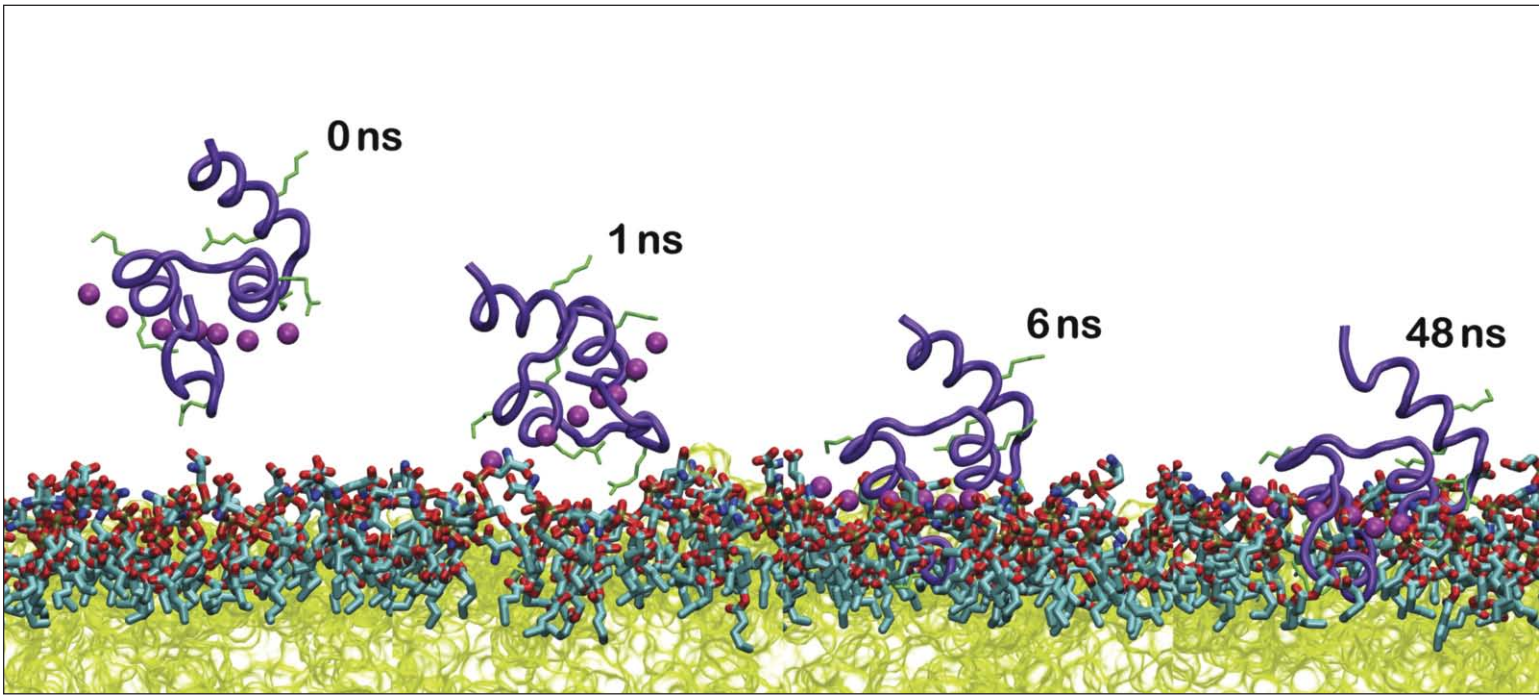
Fortunately, advances in computational modeling are helping to clarify the behavior of blood under both healthy and unhealthy conditions. Two researchers in particular are modeling blood's component parts, albeit at slightly different scales. One is trying to describe the molecular mechanisms that drive coagulation, while the other is trying to predict changes in blood viscosity by modeling individual red blood cells and their interactions. Like microscopes that offer different levels of magnification, their simulations illuminate the inner workings of blood at multiple levels.

Extreme Close-Up: Molecular Dynamics of Coagulation's Early Stages

Emad Tajkhorshid, PhD, professor of biochemistry, biophysics, and pharmacol-

Advances in computational modeling are helping to clarify the behavior of blood under both healthy and unhealthy conditions.

In unbiased full-atom simulations using a novel membrane representation, Tajkhorshid's team captured a peripheral membrane protein (purple) spontaneously binding and inserting into the platelet membrane. The model replaces the lipid tails in the membrane's hydrophobic core with an organic solvent, while preserving a full representation of lipid headgroups. This treatment enhances the lateral diffusion of lipid molecules by one to two orders of magnitude (compared to conventional full-tail membrane models) without compromising atomic details, and improves the efficiency of simulation studies of diverse membrane-associated phenomena. The work is described in Ohkubo YZ, et al., Accelerating membrane insertion of peripheral proteins with a novel membrane mimetic model. **Biophysical Journal*****, 102: 2130-2139 (2012). Image courtesy of Y. Zenmei Ohkubo, Taras V. Pogorelov, Mark J. Arcario, and Emad Tajkhorshid.***



ogy at the University of Illinois, has been using molecular dynamics (MD) modeling to simulate the earliest stages in the coagulation cascade. The process begins when blood-clotting proteins bind to the membranes of activated platelets. They do so with the help of lipid molecules that ordinarily lie buried within the membranes themselves, rising to the surface only when needed—a regulatory mechanism that prevents your blood from clotting in your veins as you read this.

Divining the mechanics of that binding process, and the specific sites on the membrane where binding occurs, could lead to the development of better anticoagulant drugs with fewer side effects. But the process is difficult to characterize experimentally because the platelets' membrane surface is a semi-fluid platform; the relevant lipids and proteins are in constant motion, and it is difficult to determine which parts of the molecules bind to one another, and where.

MD modeling, which allows scientists to simulate interactions at the molecular level, would seem to present the perfect solution. Yet the extraordinarily high resolution afforded by molecular dynamics comes at a correspondingly high cost. Tajkhorshid's models, for example, must calculate the forces between almost every pair of atoms in a system comprising hundreds of thousands of them. And they must do so at time intervals measured in quadrillionths of a second. Generating even one nanosecond's worth of simulation time requires performing those calculations millions of times.

"That's really, really expensive," says Tajkhorshid, adding that the computational burden is so high that the simulations are currently limited to timescales of hundreds of nanoseconds—"maybe a microsecond, if you really push it." The binding process itself plays out over tens of microseconds, however, which presents an obvious problem—one that Tajkhorshid's group has solved by means of an ingenious computational trick.

By replacing a portion of the virtual platelet membrane with a more fluid organic solvent, Tajkhorshid has sped up the rate at which his simulated lipids move.

"Suddenly everything is ten times faster, at least," Tajkhorshid says. "Things that usually happen at the microsecond scale are happening at the nanosecond scale." This artificial accelerant has enabled Tajkhorshid and his colleagues to simulate the interaction between platelet membrane and lipid molecules, and to work out how coagulating proteins bind to the membrane surface. Tajkhorshid is currently using his lubed-up model to pursue an even more ambitious goal: simulating how different coagulation proteins form complexes on the membrane surface in order to become fully activated, thereby driving the coagulation cascade forward.

By replacing a portion of the virtual platelet membrane with a more fluid organic solvent, Tajkhorshid has sped up the rate at which his simulated lipids move.

Medium Close-Up: Modeling Blood Viscosity

George Karniadakis, PhD, professor of applied mathematics at Brown University, also found himself bumping up against the limits of molecular dynamics when attempting to model hematological phenomena. His solution? Study blood at a coarser level of granularity. This lets him cover a larger territory in the circulatory system at a longer timescale, modeling changes in blood viscosity and simulating the kinds of abnormal red blood cell aggregation that occurs in diseases such as atherosclerosis, AIDS, myeloma, and diabetes mellitus.

To create his multiscale models, Karniadakis simulates everything from the biomechanics of individual red blood cells to their passage *en masse* through the human body's arterial tree. The method he uses, known as dissipative particle dynamics (DPD), was originally developed by a pair of Dutch chemical engineers for the purpose of modeling polymers. Sometimes called a coarse-grained molecular dynamics approach, DPD relies on virtual particles that represent clusters, or lumps, of atoms and molecules rather than delving into too much microscopic detail. "Instead of every droplet in a cloud interacting with every droplet in another cloud, we have two small clouds interacting with each other," Karniadakis says. As that metaphor suggests, DPD offers a mesoscopic or intermediate-scale tool for bridging the gap between the high-powered zoom of true MD modeling and the standard fluid models that are used to simulate blood flow writ large.

Last year, Karniadakis and his colleagues constructed a multiscale model that simulates the growth and rupture of a brain aneurysm by using DPD to capture cell-to-cell interactions within the aneurysm and classical fluid mechanics to represent the flow of blood in the brain. Now he has developed two different DPD-based models for simulating individual red blood cells and predicting their aggregate behavior.

The first model, which Karniadakis calls "cheap blood," uses only 10 DPD particles to represent each cell. (By contrast, Karniadakis says, some 30,000 points would be required to faithfully replicate the protein structure of the surface of a single red blood cell.) Yet this bare-bones model still permits accurate simulations of blood flow down to the level of capillaries. "It's not exactly the geometry you want, but it's pretty close," Karniadakis says.

The second model uses several hundred particles to accurately represent the cytoskeletal structure of a red blood cell. Though considerably more expensive, it can be used to predict the flow of blood through even the smallest vessels. By toggling between the two models, Karniadakis can tailor the degree of resolution to the

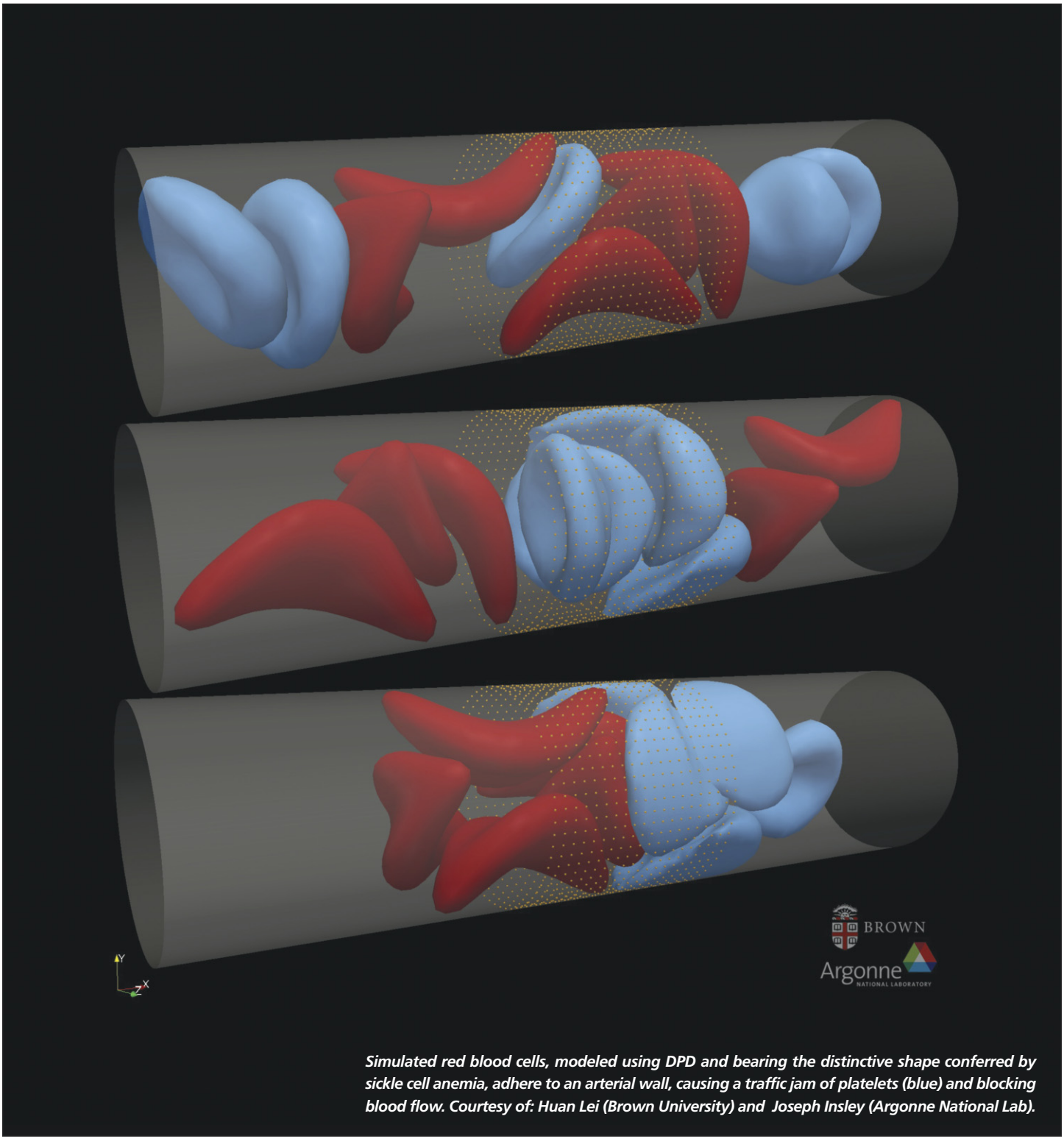
Karniadakis calibrates his models with biomechanical data gathered from experiments on individual red blood cells, then predicts the collective behavior of blood under both healthy and diseased conditions.

task at hand and avoid eating up more computational resources than necessary.

Using methods developed in collaboration with **Subra Suresh**, **PhD**, current director of the National Science Foundation, Karniadakis calibrates his models with biomechanical data gathered from experiments on individual red blood cells, then predicts the collective behavior of blood under both healthy and diseased conditions. By tweaking his models to reflect the stiffening of red blood cells infected with malaria, for example, or varying levels of a protein called fibrinogen that plays a key role in coagulation, Karniadakis has successfully predicted changes in blood viscosity—and accurately modeled, for the first time, the microscopic physical processes that cause those changes, such as the formation and destruction of "rouleaux," or stacks of red blood cells. The abnormal aggregation of red blood cells is a symptom of many diseases, and better modeling of how and why that aggregation occurs could lead to more precise diagnoses.

In addition to unpacking the physics of blood flow and coagulation, Karniadakis is also using his DPD models to figure out how diseased red blood cells interact with arterial walls and white blood cells—information that could lead to more effective treatments for both malaria and sickle cell anemia.

Now that would be bloody brilliant. □



Simulated red blood cells, modeled using DPD and bearing the distinctive shape conferred by sickle cell anemia, adhere to an arterial wall, causing a traffic jam of platelets (blue) and blocking blood flow. Courtesy of: Huan Lei (Brown University) and Joseph Insley (Argonne National Lab).

TOOLS TO UNDERSTAND THE FEDERAL RESEARCH PORTFOLIO: From Ontologies to Topic Mapping

By Katharine Miller

What biomedical research does the federal government fund? How is it allocated across important diseases? Has that changed over time? Answering these questions at any level of detail is tougher than you might expect.

The National Institutes of Health, for example, award 80,000 grants each year. But when they want to evaluate funding for a particular area of research, "It's difficult to know: Have you covered what you think you've covered?" says **Edmund Talley, PhD**, Program Director, National Institute of Neurological Disorders and Stroke (NINDS), NIH.

Of course, federal funding agencies do analyze and report to Congress on their research portfolio. At the NIH, for example, the Research, Condition, and Disease Categorization (RCDC) Process categorizes all NIH grants according to 233 categories that it is required to report to Congress and the public. This categorization is transparently available at the NIH RePORTER website, an online searchable database of NIH grants. But, Talley says, congressional reporting categories don't necessarily cover the entire realm of research. And currently, NIH is the only agency using this system, so it can't be used to assess research across funding organizations.

Now researchers have developed two very different yet complementary computational approaches to dig deeply into the federal research portfolio. The first, developed at Stanford, relies on ontologies—structured, hierarchical categorizations of research—to answer specific questions about the federal research portfolio across all funding agencies. The second approach, NIH Map Viewer developed by Talley and a diverse team of computer scientists, uses text mining to cluster topic words from NIH grant abstracts and then visualize the results.

Both tools can help program officers—as well as grant applicants—evaluate the nature of the NIH research portfolio in ways that were previously very difficult, if not impossible.

Ontologies Get Real

Nigam Shah, PhD, assistant professor of medicine at Stanford University School of Medicine, would like to see federal funding agencies categorize their grants using a common ontology, at least for disease research. To make the case for that idea, **Yi Liu**, a graduate student in Shah's lab at Stanford, set out to demonstrate that existing ontologies could be used—in an automated way—to discover interesting information and trends in research activity across all federal agencies.

He used a decade of grants data (1997 to 2007) from the Research Crossroads database, which covers 33 different funding institutions, including the NIH, National Science Foundation, Health Resource and Service Administration, Centers for Disease Control, Food and Drug Administration and others. That database can now be searched on Bioportal, the website of ontologies created by the National Center for Biomedical Ontology. Liu also created a workflow to annotate both the grants data and a decade's worth of PubMed journal articles associated with US institutions using ontology terms from the human disease ontology (DO).

Up to this point in Liu's research, Shah says, "anyone can do this at the BioPortal." Indeed, a simple search of the Research Crosswords database provides any user with counts of grants in any disease category. But Liu went several steps further, looking at three measures of funding.

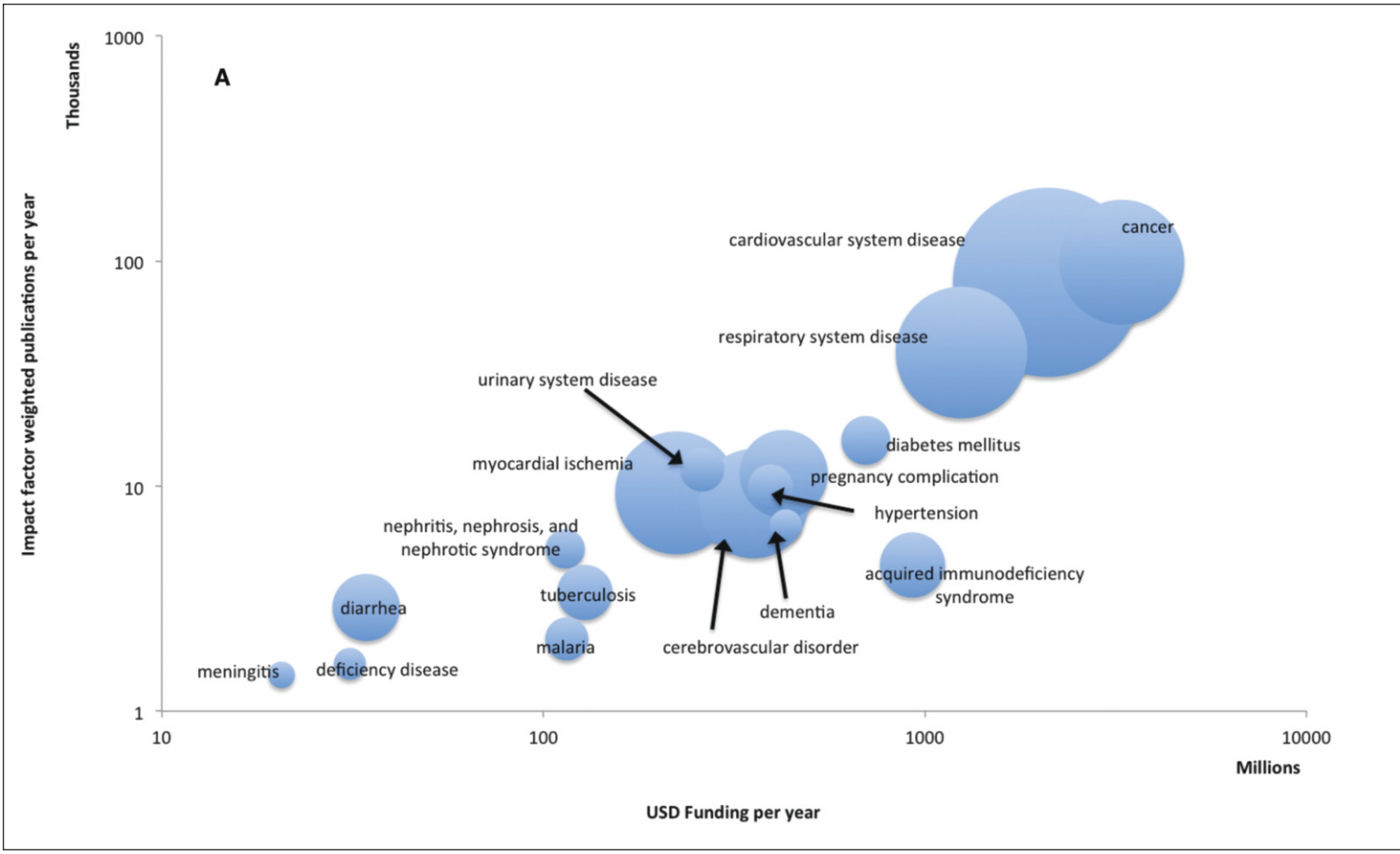
First he looked at sponsorship—the level of funding for a particular disease topic relative to the impact factor–weighted count of publications in that topic area. So, for example, Liu found that drug abuse and Alzheimer's disease are highly sponsored but are less commonly represented in high impact journals compared with cancer or heart disease. Liu's analysis can't explain this discrepancy—which could have many causes, including how expensive the research is; whether the topic is a new research area; and whether it's been hard to produce results with a significant impact on the disease—but his work makes it easier to spot the differences.

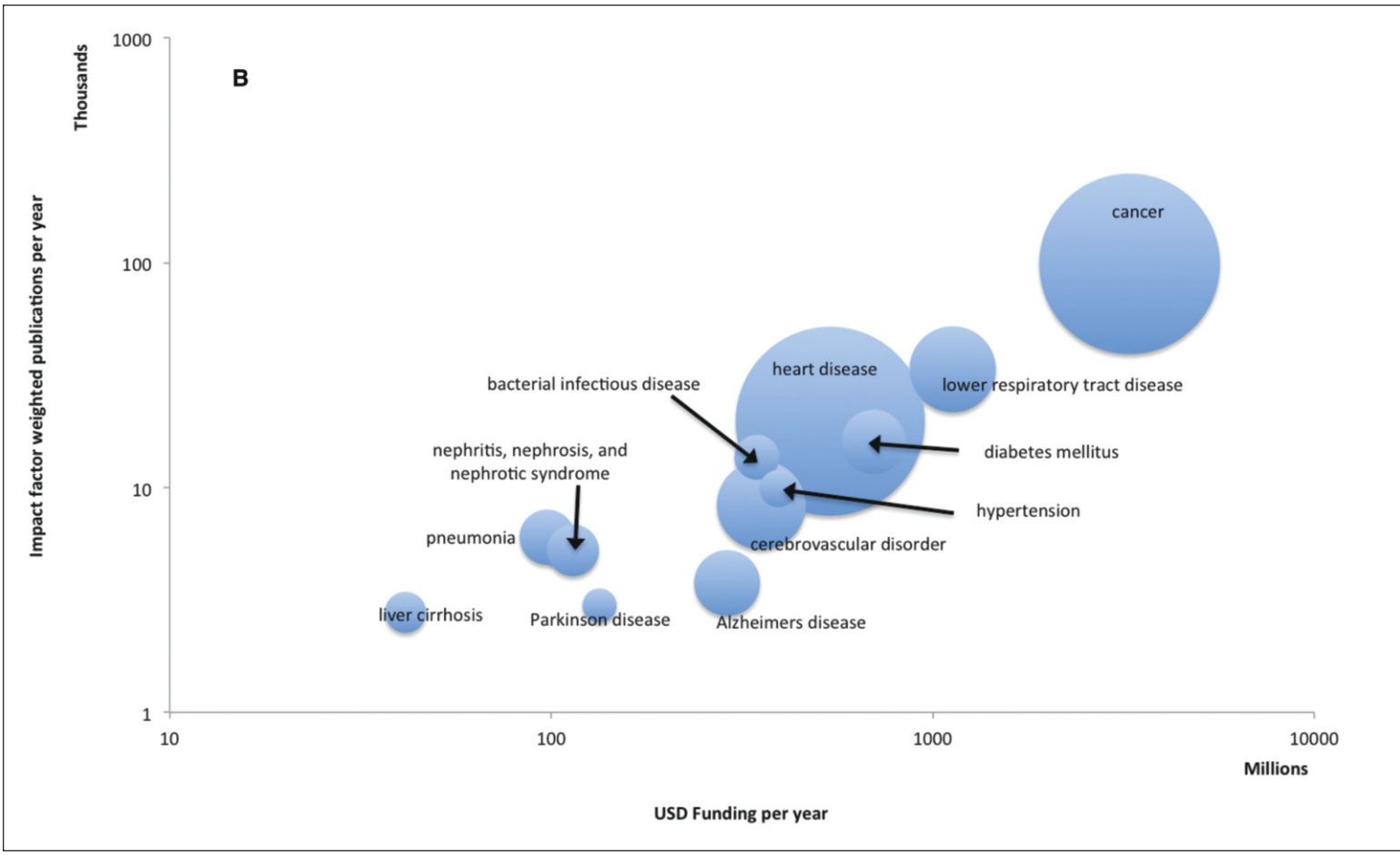
Liu also studied allocation—the level of support for a disease area as a function of mortality rates, which he used as an imperfect surrogate measure of disease burden (other measures are possible). "Allocation looks at sponsorship in the context of the size of the problem," Shah says. "Are we spending enough? Overspending? Underspending?" For example, the work showed higher funding for cancer than for heart disease, which has higher mortality rates.

And finally, Liu looked at trends across time. He determined whether, for a given disease, funding has reached a plateau,

> When the NIH wants to evaluate funding for a particular area of research, "It's difficult to know: Have you covered what you think you've covered?" says Talley.

***Opposite Page:** These panels show allocation of federal funding relative to annual impact factor–weighted publications on a per-disease basis. The sizes of the bubbles correspond to the relative disease burdens as characterized by worldwide mortality statistics for 2004 (A) or in the US in 2007 (B). Reproduced from Yi Liu, et al., Using ontology-based annotation to profile disease research,* Journal of the American Medical Informatics Association *doi:10.1136/amiajnl-2011-000631 (2011) with permission from BMJ Publishing Group Ltd.*


A
Thousands
1000
100
10
1
Impact factor weighted publications per year
cardiovascular system disease
cancer
respiratory system disease
urinary system disease
diabetes mellitus
myocardial ischemia
pregnancy complication
hypertension
nephritis, nephrosis, and
nephrotic syndrome
acquired immunodeficiency
syndrome
tuberculosis
diarrhea
dementia
malaria
cerebrovascular disorder
meningitis
deficiency disease
10
100
1000
10000
Millions
USD Funding per year


B
Thousands
1000
100
10
1
Impact factor weighted publications per year
cancer
heart disease
lower respiratory tract disease
bacterial infectious disease
diabetes mellitus
nephritis, nephrosis, and
nephrotic syndrome
hypertension
pneumonia
cerebrovascular disorder
liver cirrhosis
Parkinson disease
Alzheimers disease
10
100
1000
10000
Millions
USD Funding per year

dropped off, or increased over the years—useful information for agencies hoping to make smart funding decisions.

In the end, Liu says, "I was pretty happy about being able to see the big picture from a pretty granular database."

It's a proof of concept—a demonstration that the various agencies that fund biomedical research should switch from ad hoc categories to an existing, shared ontology. "We don't really care which one," Shah says. "But why not use an ontology based on the Unified Medical Language System (UMLS) that the National Library of Medicine is building and funding?"

It's a proof of concept—a demonstration that the various agencies that fund biomedical research should switch from ad hoc categories to an existing, shared ontology. "We don't really care which one," Shah says.

Still, using ontologies has its limits, Shah concedes. "If your research interest is one for which there is not a good ontology, then this approach is simply not going to work," he says. For example, areas such as liberal arts, political science, or even basic research are difficult to classify hierarchically.

Topic Mapping

As an alternative to ontologies, Talley and his team created the NIH Map Viewer, a tool that uses text mining, topic modeling, and visualization as a way of digging deeply into the NIH portfolio. Such topic maps have two clear advantages over ontologies: They pick up phrasings and words that are not in a pre-classified hierarchy; and they cluster words together based on their shared usage. "In topic mapping, you have a bag of words and you want to learn how it's organized—to extract structure from it," Shah says.

NIH Map Viewer was built on earlier work by the team. "We had already produced nice data using abstracts from the Society for Neuroscience annual meeting, but we didn't know if the method could scale to the entire NIH, or if it would provide a coherent view at both the local and the global level," Talley says.

The text-mining method they used is called "latent Dirichlet allocation" or LDA. It had been invented a few years before, but hadn't been tested on many real-world problems, Talley says. "There were a lot of open questions about how to evaluate what makes a good topic." LDA is a kind of component analysis, and that was a problem: The components don't have to be meaningful in order to be predictive. Talley needed good topics, not just a good model. "We had to come up with a way to assess topics in an automated way," Talley says.

In the year since the work was published in *Nature Methods* in June 2011, the team has continued to tune the parameters of the algorithm so that it now does quite a good job of extracting topics from text. "That's an accomplishment for us," he says. "The new topics will be available in the next few months."

The visualization piece of the project starts from a layout map based on similarities between the grant abstracts. Documents are clustered based on their internal texts rather than by external labels given to them by NIH, Talley says. The baseline map resembles a web with interconnected strands that represent grants with their feet firmly planted in several fields.

On top of this static baseline map, users can query for topics as well as other categories of interest, as described in the following sidebar. The NIH Map Viewer is now available at https://app.nihmaps.org, as well as from a "Links" tab within NIH RePORTER, the NIH portfolio search tool.

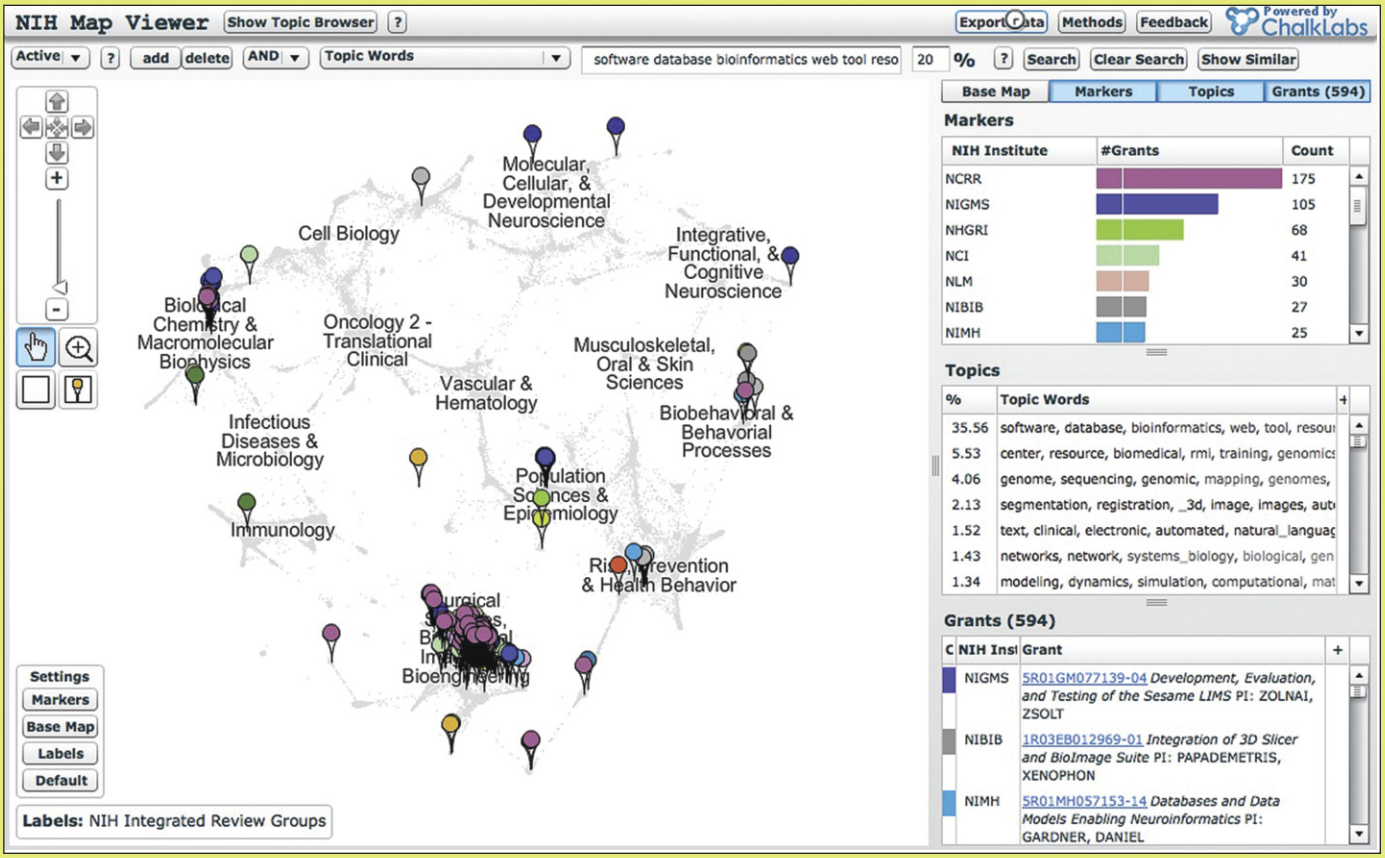
"This has been an experiment where we've said 'Let's get it out there and see where the value is,'" Talley says. "Ultimately, I think this or something like this will be valuable for policy officials. It's a new way of looking at grants."

Eventually, Talley hopes to see a system that can provide both the accurate recall of text mining and the clarity of ontologies. This is an area of intense interest, he says. "These are complementary techniques that really need to be merged." □

AN NIH MAP VIEWER TEST CASE:
A topic search for "software"

In the NIH Map Viewer (at https://app.nihmaps.org/nih/browser), when users enter a search term in the topic window, a dropdown menu appears listing several possible topics containing that term. For example, "software" produces two possible bags of words, one of which begins with "software database bioinformatics web tool resource annotation visualization…." After selecting this topic and setting a threshold for recovering only the best topic matches (in this case the default 20 percent was used)—a search generates a list of 594 grants, all marked on the map with pushpins (as shown on the following page). Users can change the pushpin coloring to represent institute, funding level, or a number of other categories. At the same time, these categories are displayed as a bar chart in a

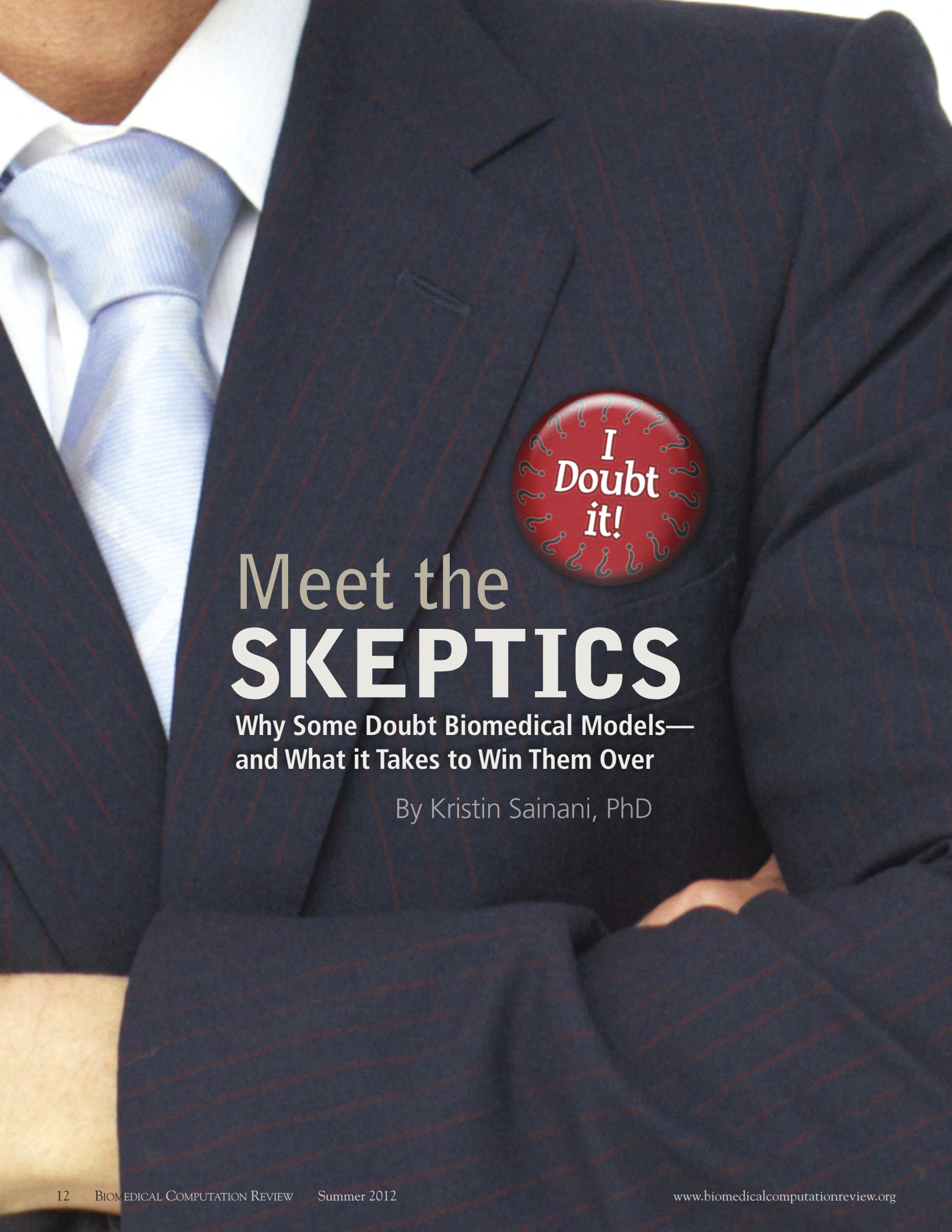
separate window. An additional window lists similar topics and users can drill down into a particular topic using the "topic info" button, which opens a separate page. There, users are given a wealth of information—including co-occurring topics, similar topics, and a list of grants—to help them evaluate whether the topic lives up to their expectations. For example, the topic info page for the map shown here helps the user ponder, for example, "Is this topic really about software development?"



By comparison, a keyword search for "software" in NIH RePORTER produces a list of 3823 grants. "You're pulling everything and there's no way to really focus it," Talley notes.

Users can also come to NIHMaps.org directly from a RePORTER search using the "Links" tab. Each grant from the search is displayed as a pushpin. Zooming in and scrolling over each pushpin identifies each grant by name. If a search produces only a handful of grants, the RePORTER's list might be adequate, Talley notes. "But when you start talking about hundreds, a list becomes intractable, and you need a way to organize the information," he says. "Our statistical analysis of this layout algorithm suggests that it's tuned to perform especially well when you start getting a hundred or more documents, which is where clustering becomes really useful."

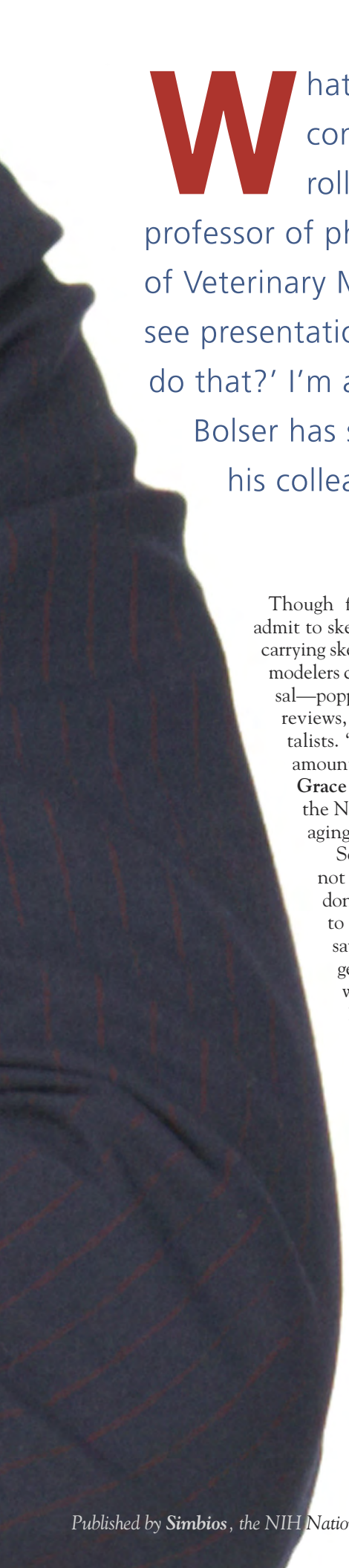
Talley and his colleagues are also continuing to improve the NIH Map Viewer. For example, it's now possible to save and share a search as a link; and in the bar chart, users can turn different categories on or off. Talley's team is also generating a map based on similarities between grants and publications that cite NIH grants. The combined map of grants and publications has higher resolution, Talley says. "The overall quantitative performance improves, and we get more clusters in places where we know we want more clusters." He hopes to release the new map in a few months.



Meet the SKEPTICS

Why Some Doubt Biomedical Models—and What it Takes to Win Them Over

By Kristin Sainani, PhD



What are the telltale signs of a modeling talk at a biology conference? Just look for the sighs, shifting, and eye-rolling in the audience, says **Donald C. Bolser, PhD**, professor of physiological sciences at the University of Florida College of Veterinary Medicine. Bolser, once a skeptic himself, says, "I would see presentations at meetings and wonder: 'Why would you want to do that?' I'm an *in vivo* person, so I never really saw the value of it." Bolser has since become an ardent fan of modeling, but many of his colleagues remain suspicious.

Though few biologists or physicians will admit to skepticism (we couldn't get any card-carrying skeptics to go on record for this story), modelers claim that skepticism is near-universal—popping up in grant evaluations, paper reviews, and interactions with experimentalists. "I have encountered a tremendous amount of skepticism for modeling," says **Grace Peng, PhD**, a program director at the National Institute of Biomedical Imaging and Bioengineering.

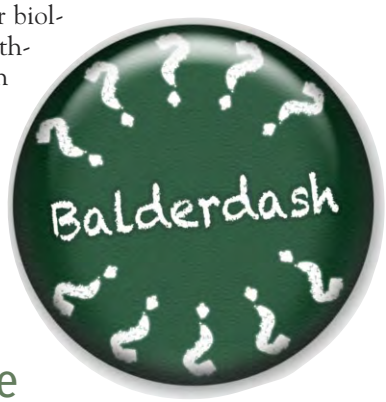
Senior-level people at the NIH may not openly oppose modeling, but they don't seem to appreciate its true power to change biomedical research, Peng says. Peng chairs IMAG (the Interagency Modeling and Analysis Group), which brings together scientists from 10 governmental agencies who manage programs in biomedical, biological, and behavioral modeling. In 2009, members convened for a two-day conference—the IMAG Futures meeting—that explored reasons for and solutions to skepticism. Threads from that meeting, as well as from interviews with modelers and biologists, form the basis for this story.

Modelers may assume that the problem of skepticism rests solely with experimentalists. But, in fact, modelers play an enabling role—in the way they treat non-modelers, present their results, and even build their models. Thus, overcoming skepticism is as much about changing the culture of modeling as it is about changing the minds of biomedical researchers.

It also turns out that skepticism is heterogeneous. The degree of skepticism varies greatly across different fields of biology and medicine; and skeptics themselves come in many different flavors. Different kinds of skepticism have diverse origins and may present unique obstacles for modelers. This article disentangles the different types of skeptics and suggests what modelers can learn from each.

The Old Guard

Some biologists are not so much skeptical of modeling as dismissive, says **Peter Sorger, PhD**, professor of systems biology at Harvard Medical School. These are mostly older biologists, who achieved success without modeling, and are stuck in their way of doing things.



The solution to this kind of skepticism is simply to "fill the place with young people," Sorger says.

When you've been doing something one way for a long time, change is hard, says **Timothy Mitchison, PhD**, also a professor of systems biology at Harvard Medical School. "That's just human nature."

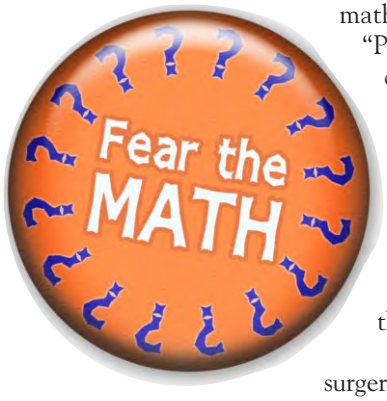
Unfortunately, there's little modelers can do to combat this kind of skepticism. "It only goes away when people die or retire; no one ever changes their mind. It's like why people vote Democrat or Republican; these things go deep," Mitchison says.

The good news is that the newer generation of biologists is much more open-minded, Sorger says. At classic biology conferences, he says, "I'm barely old enough to get a plenary talk, because everyone is in their seventies"; but at the DREAM modeling competitions, "I'm just this creaky old guy, because the mean age is about 30." So, the solution to this kind of skepticism is simply to "fill the place with young people," he says.

We also need to incorporate modeling into the curriculum of future biomedical scientists, Peng says. "I think if modeling is introduced earlier in the pipeline, even starting at the K through 12 stage, modeling will be accepted as standard practice," she says.

The Math Phobes

Some biologists are open to modeling in principle, but avoid it because they are too intimidated by the math. Modelers don't help the situation because they tend to be dismissive of people who aren't quantitatively trained. Modelers have been known to call biologists "dumb," "idiots," and "the students who weren't smart enough to go into math or physics." With this attitude, it's not hard to understand why biologists would feel intimidated and shut out.



"The models need to be as easy to use as TurboTax," Peng says.

Some modelers need a dose of humility. They also need to put more time and effort into explaining their models as simply as possible to potential users, Peng says. "The experimentalists say, 'It would be nice if I could just sit down with the modeler in front of a computer and go through the model,'" she says.

Mitchison (who counts himself among the math-phobic) recalls a math PhD student who did a stint in his lab and was able to make the math comprehensible: "Having someone who can just think that way with their eyes closed, and can explain it to you… once you have that experience of really working with someone, it makes a huge difference."

Modelers also have to be more willing to make user-friendly tools that don't require a degree in mathematics, Peng says. "The models need to be as easy to use as TurboTax," she says.

"If you think about the large number of biologists out there, the notion that everything is going to be done by making people aware of how to do all the underlying mathematics … I think is nonsense, at least in the next generation of individuals and the existing biological investigator pool," **Ron Germain, MD, PhD**, told IMAG Futures attendees; Germain is chief of the Laboratory of Systems Biology at the Center for Human Immunology and Inflammation at the NIH. Modelers worry that biologists will abuse models if they don't understand their inner workings, but Germain points out that people can accurately resize photos in Photoshop without understanding the complex math behind this operation.

Programmers in Germain's lab have developed software (called Simmune) that allows immunologists to build complex models with all the mathematics handled behind-the-scenes. "People with no computer training can do this with no assistance," Germain says. "And the response I've gotten talking to biologists—instead of the glazed over eyes, 'oh you're talking about modeling'—there's a great deal of enthusiasm." If modeling is too difficult for biologists to implement, they won't adopt it even if they think it's useful, Germain says.

Yoram Vodovotz, PhD, professor of surgery and of immunology at the University of Pittsburgh, encouraged IMAG Futures attendees to explore agent-based models—which are more intuitive than equation-based models—as a way to draw biologists into modeling. "There's an entire class of simulation platforms that already exist and that is already usable to people in high school without differential equations."

The "Modelers Are From Mars" Skeptics

Some biologists are skeptical because they feel that modelers are out of touch with the biology. Germain told attendees at IMAG Futures that he has biology colleagues who are initially excited to read a paper in the *Journal of Theoretical Biology* (or similar journals). But "the first thing they read is

'for reasons of computational complexity we decided to make the following assumptions…' And they basically throw out the three or four most important things to the biologist before they go on. At which point the biologist will stop reading."

Sorger agrees. "There was this notion, maybe 5 to 10 years ago, that basically one was going to take a series of tools that had been developed in another discipline—either computer science, chemical engineering, or control theory—and those things would whole-hog be applied to biology and that would solve the problem," Sorger says. This was not only an arrogant point of view; it was simply wrong, he says. It also "created a whole series of straw men for the people who are skeptical of modeling to hang onto."

For modelers to be successful and advance the field, "they must either understand the biology themselves or be joined at the hip with a biologist," Heetderks says.

The solution to this kind of skepticism is for modelers to become immersed in the biology. "I'm very much of the opinion you can come from either direction [biology or modeling] and become an effective computational biologist, but that ultimately you have to be a biologist," Sorger says. "I don't think biomedicine is going to be taken over by physicists and computer scientists working half-time."

Modelers need to develop a deep understanding of the biological data, agrees **William J. Heetderks, MD, PhD**, director of extramural science programs at the National Institute of Biomedical Imaging and Bioengineering. "If you just take data that was published in the literature and plug it into your model and don't understand the domain that it was acquired in, you can be badly misled," Heetderks says.

For modelers to be successful and advance the field, "they must either understand the biology themselves or be joined at the hip with a biologist," he says.

The "I'm Not Ready" Skeptics

Some biologists think modeling is fine for others, but it's premature for their biological niche. They may think that their biological problem is too complicated to pin down with a model or that they don't have enough data yet to build an accurate model.

This type of skepticism stems from misconceptions about the role of modeling, Peng says. "People still think that models are just a tool to maybe fit the data at the end of the experiment," she says. But models are actually a platform for designing experiments. Models can help biologists organize and archive the data they do have; systematically figure out what new data are needed; and design more efficient and more informative experiments, Peng says. To turn experimentalists around, modelers should interface with them during the planning stages of grants, she says.

For example, Vodovotz told IMAG Futures attendees how models could be used to design better clinical trials. His group ran simulated trials of anti-TNF drugs for treating sepsis and predicted that the compounds would help certain types of people and harm others (with no net benefit). Had pharmaceutical companies used these simulations, they could have targeted the correct group for treatment and avoided treating those who might be harmed.

Models also don't have to be perfect to be useful. Modelers can drive home this point by highlighting the shortcomings of the alternatives to modeling, **David M. Eddy, MD, PhD**, told IMAG Futures attendees; Eddy is founder and medical director of Archimedes, a healthcare modeling company located in San Francisco.

"The only alternative in the clinical field is to use clinical judgment or what we'd call the art of medicine. And when you think about all the factors that are involved in clinical decision making, it's out of the question," Eddy says. "That's where models come in—because they're better than the alternative. They're not perfect; they're not as good as clinical trials. But we would argue that they're better than the alternative."

Models don't have to be perfect to be useful, says Eddy.

The "Show Me the Beef" Skeptics

Some biologists see modeling as esoteric because they can't point to a concrete example of how biomedical modeling has impacted human health.

Combating this kind of skepticism requires better salesmanship, Peng says. "People who develop models are so into the details of their models that they forget the bigger picture of why their model is so useful and what the model is helping them to do that they couldn't do otherwise," Peng says.

It's often difficult to read a computational paper and figure out what the breakthrough was, Mitchison says. Biologists come from a discovery-oriented tradition; if you've been looking for the receptor to a particular hormone, and now you found it, the impact is easy to see, he says. But "often successful modelers are not people who come from this tradition of telling a story. They come from different traditions, and they may not see the value of that," he says.

Combating this kind of skepticism requires better salesmanship, Peng says.

Modelers need to do a better job of communicating what they learned from a piece of work and what it lets them do that they couldn't do before, he says.

Modelers can also help convince skeptics by pointing to specific success stories. "What we as program people are always looking for is a killer app or a success story that we can tell our higher-ups: 'Look, they couldn't have made this discovery without this model,'" Peng says.

One of the goals of the IMAG Futures meeting was to compile some of these success stories. **Marco Viceconti, PhD**, professor of biomechanics at the University of Sheffield, in the United Kingdom, presented several examples of how modeling is already being used in patient care. For example, aneurIST (http://www.aneurist.org/) is a program that predicts the rupture of incidentally detected cerebral aneurysms using patient-specific information. And companies such as Phillips and Siemens are integrating similar simulations into their imaging tools for aneurysms, Viceconti says. Modeling also has a practical utility for many companies because the FDA is now allowing *in silico* simulations as part of the approval process for certain devices.

The Converted

Many biologists who once were skeptical now count themselves as enthusiastic converts to modeling, including Bolser and Mitchison. Their stories offer lessons for modelers.

Bolser studies the neurological circuits that control airway behaviors such as coughing and swallowing; few investigators in this field use computational modeling. Before his conversion, he thought that models were just about fitting data after an experiment, he says.

But he began working more closely with modelers as part of submitting a grant application. At one point, he and his modeling collaborator ran some simulations and generated a baffling result. Bolser realized that no one had ever done that particular experiment *in vivo*, so he did it. At first, he didn't believe the result and suspected a technical error.

But then came the 'aha' moment. He realized that the simulation had predicted the outcome exactly. "And it just blew me away. I was just stunned by that. I mean modelers know this, they live by this; but for me it was a long road to go before I understood internally what a model could do. What you're really aiming at with a computational model is prediction," Bolser says. "For me, once I got it, I was totally sold on it."

He's also shifted from thinking that the airway problem was too complex to model to realizing that the behavior is so complex that "there is no way we could understand it without computational modeling."

You can't sell experimentalists on modeling simply by talking about it and showing it at seminars and conferences, Bolser advises. They need hands-on experience, such as a tutorial or workshop, or directly sitting down with a modeler and working on their biological system.

Mitchison says he is also a convert, but it was a more gradual transition. He was a skeptic of both the "I'm not ready" and "math phobe" varieties described above. But he's slowly come around, helped by people like the mathematician in his lab who was able to explain the math to him clearly. He's seen a similar gradual transition in his field. For example, the American Society of Cell Biology annual meetings have gone from no one even thinking about modeling 20 years ago to having several lively, well-attended modeling sessions today. "And cell biology is a pretty tra-

ditional, conservative field of biology," he says.

Modeling has even become state of the art for some niche areas of cell biology. In 2008, a couple of modeling papers on yeast cell polarity were published in high-profile journals. Many people disputed the papers (including a new PhD student in his department who went on to do work in the area), Mitchison says. In fact, it angered them: "Here's a high-profile paper published in *Nature* on theory and it's all wrong and I'm annoyed by that, so I'm going to do better." This stimulated a number of groups to work on the problem and publish better models. It's been a healthy progression and has resulted in many hard-core biologists adopting theory collaborators for the first time, Mitchison says. So getting theory papers on hot biological topics in high-profile journals can spark interest in modeling, he says.

Bolser was converted when he saw a simulation predict an experimental outcome exactly. "It just blew me away .... modelers know this, they live by this; but .... for me, once I got it, I was totally sold on it."

The "Healthy" Skeptics

Some biomedical researchers are proponents of modeling in theory, but are skeptical about specific tools and approaches. This is a legitimate and healthy form of skepticism, Sorger says. "The skeptics rightly point out that even in the hands of people being fairly careful, the promise has run way ahead of the actual tools and knowledge needed to apply them correctly. Therefore, there are probably quite a lot of errors out there." (See "Error: What Biomedical Computing Can Learn from Its Mistakes," in *Biomedical Computation Review* online.)

The solution in this case is not to challenge the skeptics, but to fix the problems that drive their skepticism. This will require more investment into the methodology of modeling. "This general assumption that the methodology is good enough I think is really wrong," Sorger says.

We need better ways of validating models, Eddy told IMAG Futures attendees. Validation means different things to different people and some of what is passed off as validation is garbage. "I can certainly understand why potential users of models are frustrated and confused. And they don't know what they can believe and can't believe," Eddy says. Modelers need to devise a recognized standard so that users know that they can trust a model when it has the seal of validation for a particular application, he says.

Competitions and benchmark problems can demonstrate the reproducibility of models (or reveal errors in need of fixing). Standards in the reporting of models and simulations (such as MIRIAM and MIASE, Minimal Information Required In the Annotation of Models and Minimum Information About a Simulation Experiment, respectively) also help improve the reproducibility and testability of models, participants at IMAG Futures noted.

To win over healthy skeptics, modelers also need to be more humble in how they present models. In molecular and cell biology, researchers typically draw cartoon models with arrows and boxes; and it's understood that these are just working hypotheses, not to be taken too seriously, Mitchison says. Computational models are also just provisional, but they often aren't presented this way. "There's a way that computational stuff is written up that sort of implies a rigor and absolute truth that experimentalists who don't use quantitative methods have deliberately shied away from," Mitchison says. This kind of overconfidence can drive skepticism, because experimentalists know that biology always involves hidden assumptions.

Modelers need to be more upfront about the limitations and potential pitfalls of their models and to make these issues more understandable for non-mathematicians, says Kobilka.

Modelers need to be more upfront about the limitations and potential pitfalls of their models and to

make these issues more understandable for non-mathematicians, says **Brian K. Kobilka, MD**, professor of molecular and cellular physiology and medicine at Stanford University.

The "Insider" Skeptics

Some of biomedical modeling's biggest skeptics are actually modeling insiders. They may challenge specific models and applications, or even whole paradigms of how modeling is done.

When it comes to reviewing grants with a modeling component, engineers are often the worst critics, Peng says. Engineers tend to have a critical mindset; and they can actually do a disservice to modeling by being too nitpicky in their reviews, says Peng (who is herself an engineer). We also face the "grumpy Russian mathematician problem," Sorger says. He says pure mathematicians tend to give his papers unfavorable reviews because it's "1950s, engineering mathematics" rather than cutting-edge modern math.

Some modeling insiders go even further, saying that almost all biomedical modeling is done incorrectly. Awareness of these skeptics' point of view is important, because their critiques may ultimately explain some of the intuitive discomfort that biologists feel toward models.

For example, **James Bower, PhD**, professor of computational neurobiology at the University of Texas Health Science Center in San Antonio, says, "I would argue that, at present, the majority of mathematically based models in biology are not in fact useful in advancing the field." Most biomedical modelers are building models simply to explain or convince others of what they already believe; but the purpose of modeling should be to discover new features of a system. "If you don't know anything more about the system after you build the model than you did before, it is of little use," Bower says. "We are just endlessly misled by what are basically Ptolemaic or religious models that are designed to enforce a particular doctrine."

He advocates the use of anatomically and structurally realistic models that don't have built-in assumptions about function, such as embodied by the GENESIS simulation toolkit for neuronal modeling (http://www.genesis-sim.org/GENESIS/). He also promotes the notion of community models where everyone uses and freely shares the same base models. "There is a way forward and it's slowly starting to happen," Bower says.

John C. Criscione, MD, PhD, associate professor of biomedical engineering at Texas A&M, is another advocate of sweeping change in biomedical modeling. In his field of multiscale tissue modeling, he says he has shown that the current framework of modeling yields an infinity of solutions. Modelers are essentially trying to solve for three variables with two equations, he says. "If a freshman algebra student did this, I would flunk them," he says. "Everybody is going, 'but it fits the data.' Well, yeah, the sun moving around the earth fits the observational data too," he says. He says we need to get back to basics and figure out models that solve the simplest problems—like a perfectly homogenous elastic cylinder.

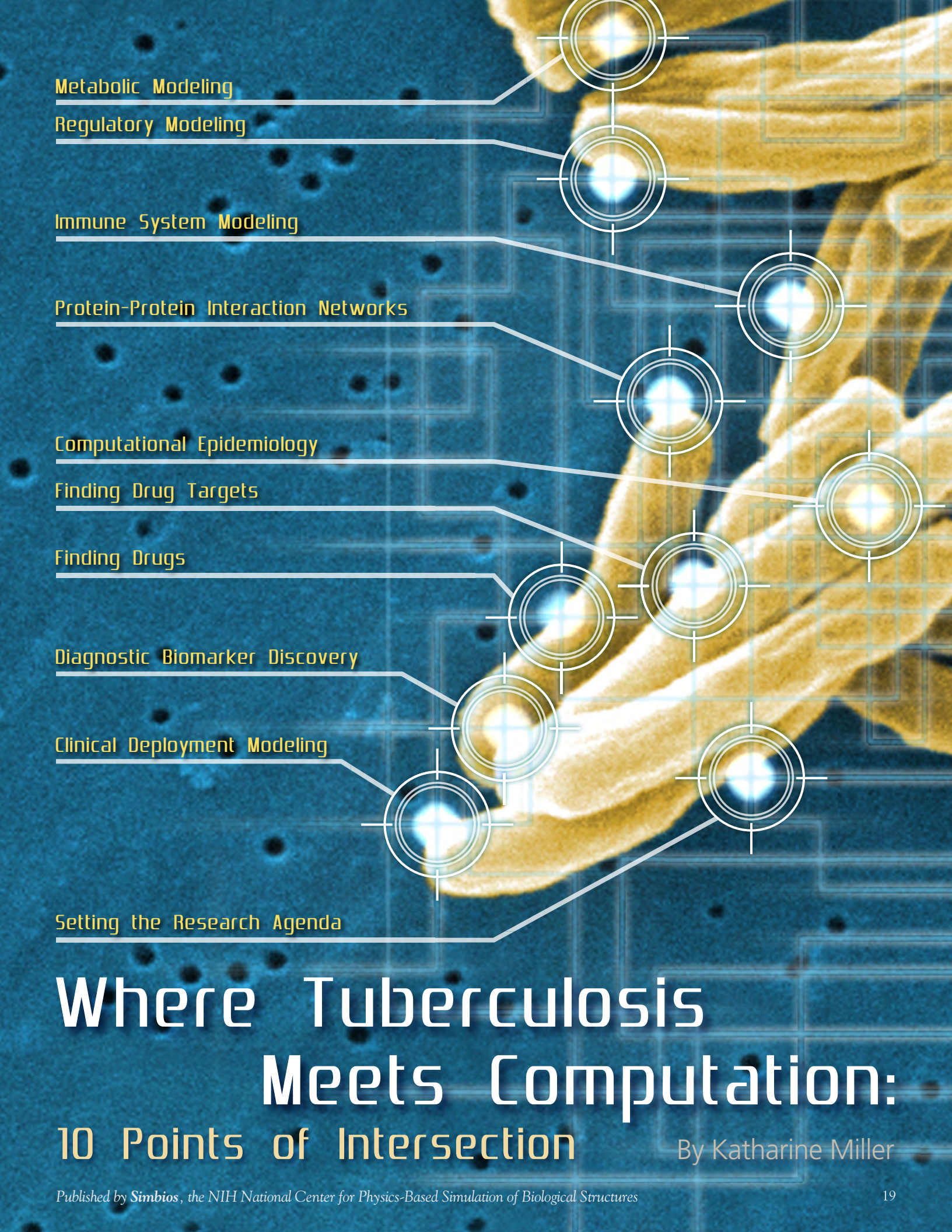
"We absolutely need modeling. I'm not saying we shouldn't do multiscale modeling. I love it. It's great stuff," Criscione says. "What we don't need is to spend money doing modeling where we'll never get a right solution."

Interestingly, both Bower and Criscione are increasingly pessimistic about convincing their colleagues. Both are therefore independently focused on exposing the next generation of engineers to modeling technology. Bower's efforts are based in Whyville.net, which he founded as the first simulation-based educational virtual world 13 years ago, which now has more than 7.2 million subscribers worldwide.

"We absolutely need modeling .... I love it. It's great stuff," Criscione says. "What we don't need is to spend money doing modeling where we'll never get a right solution."

The Demise of Skepticism?

Skepticism may be pervasive, but it's also on the decline. If researchers talked about modeling and mathematics at biology meetings 30 or 40 years ago, "we were going to be lynched almost," Eddy told IMAG Futures attendees. Acceptance of modeling will continue to grow, because modeling in biomedicine is inevitable. In the future, everyone will use models and appreciate the use of models, Peng says. But what's at stake is how quickly this transition will occur, Peng says. "How fast we get there is what I'm trying to address." □



Where Tuberculosis Meets Computation:

10 Points of Intersection

By Katharine Miller

Computational approaches to tuberculosis are unavoidable.

The growing threats of multi-drug resistant (MDR) and extensively drug resistant (XDR) tuberculosis (TB) are spurring worldwide interest in faster and more innovative research approaches, such as computation offers. And, as in other areas of biomedicine, high-throughput experiments are yielding a data deluge: The bug's bacterial genome was sequenced a decade ago and more than 26 public databases are now accumulating vast and varied information about the disease—all of it ripe for analysis.

In addition, computation makes an appealing complement to experimentation. In the lab, because the bacterium (*Mycobacterium tuberculosis* or *Mtb*) grows slowly (replicating only once a day), one experiment might require months to complete. By contrast, a virtual experiment might take seconds—and doesn't require rigorous safety precautions.

Computation also has the capacity to address important questions in TB research. Simply flipping through the TB Research Roadmap (published by the World Health Organization's Stop TB Partnership in 2011) reveals the many ways computation can contribute to developing TB drugs, diagnostics and vaccines.

In addition, notes the Roadmap, systems biologists are uniquely situated to study one of *Mtb*'s big mysteries: how it can survive inside the human lung for years—seemingly and inexplicably protected by the very immune system that should wipe it out. Only about 10 percent of the 2 billion people infected worldwide develop active disease from the get-go (and will die if not treated); the rest develop latent disease, which they control but cannot clear. And about 10 percent of latent infections will transition to active disease later in a person's life. By studying *Mtb* as a whole—rather than by looking at its individual parts—systems biologists can tease out how the bug manages these stunts.

But is TB research really benefiting from computation's promise?

Here we've highlighted 10 ways computation is currently making a difference to this problem of global significance. It's a non-exhaustive sampler, designed to whet your appetite. But the exercise of finding key points of intersection between an important infectious disease and computation is instructive in its own way: It provides a window into a disease often described as a black box and suggests novel ways to gain insight about this mysterious pathogen.

1 Systems Biology of TB Metabolism

To kill a bacterium, researchers try to determine what genes it needs to survive. In the wet lab, experimental biologists identify essential genes by knocking them out one at a time and then observing the result: Does the bug thrive or die? Systems biologists do this same exercise *in silico*, building metabolic models and using them to identify essential genes.

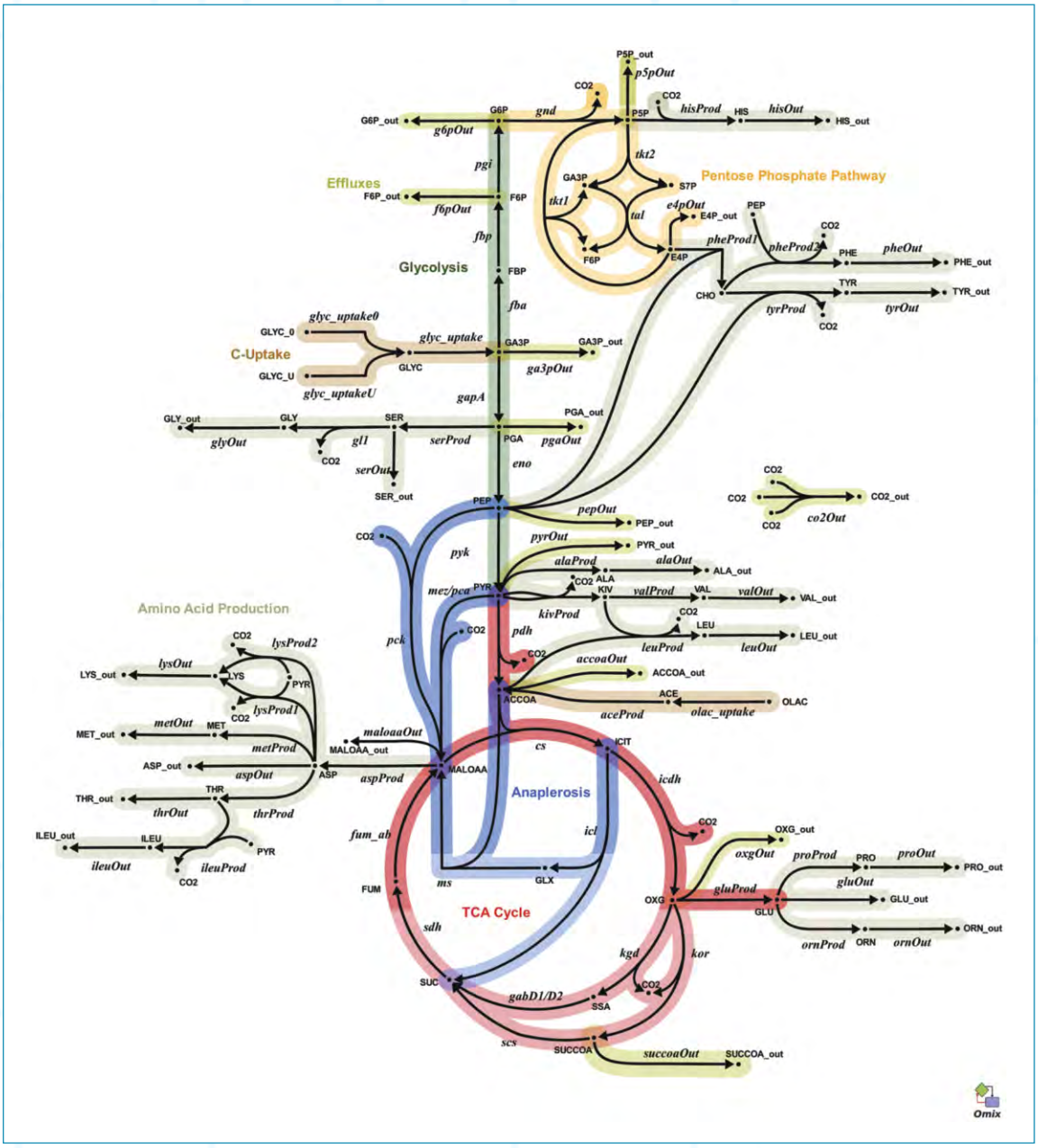
In 2007, both **Bernard Palsson**'s systems biology lab at the University of California, San Diego, and **Johnjoe McFadden**'s molecular genetics lab at the University of Surrey published genome-scale network models of *Mtb* metabolism. The McFadden team's model involved 726 genes, 849 reactions, and 739 metabolites and was calibrated by growing *Mycobacterium bovis*, a close relative of *Mtb*, in a steady state. The researchers then used metabolic flux analysis to simulate the flow of metabolites through the network.

> "TB is a black box," Galagan says. "But we're starting to open it up. We're collecting the data to map the innards of TB and using that to create predictive models."

McFadden says he thinks of fluxes as traffic through an island road network where the bacillus is the island—the United Kingdom, say—and substrates enter at the ports and are transported through cities (various chemical reactions). "If the rate of traffic going through the networks is steady," he says, "you can go to a port at Plymouth where some product is being produced, measure the rate of production and infer the fluxes inside the network using linear algebra."

The researchers then looked at what happens to the fluxes when various genes are knocked out. "Once you have the model, it becomes a virtual cell," McFadden says. "So you can do experiments instantaneously that would take months and months in the lab." If a gene is essential, the fluxes through the network change, creating blockades that make it impossible for the bacterium to survive. McFadden's lab's analysis of which *Mtb* genes were essential found a 75 to 80 percent matchup between model predictions and lab results.

McFadden concedes that the metabolic models remain incomplete, and that it's still the early days for TB systems biology. But, look-



Metabolic network of the central metabolism of* Mycobacterium bovis *BCG. Reprinted from Beste DJ, et al., ^{13}C metabolic flux analysis identifies an unusual route for pyruvate dissimilation in mycobacteria which requires isocitrate lyase and carbon dioxide fixation.* PLoS Pathog *(2011) 7(7):e1002091.

ing under the hood of TB has the potential to give researchers a more accurate and predictive view of how TB works.

"TB is a black box," says **James Galagan, PhD**, associate director of microbial genome analysis at the Broad Institute of MIT and Harvard. "But we're starting to open it up. We're collecting the data to map the innards of TB and using that to create predictive models."

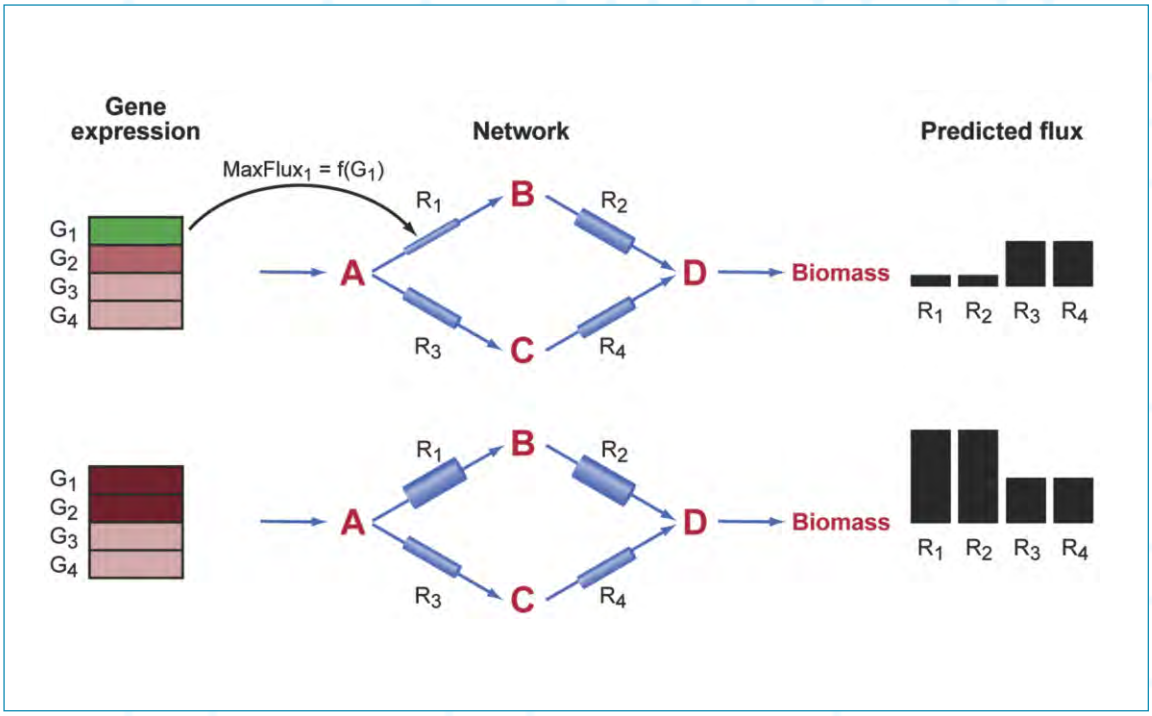
2 Combining Metabolic Models with Gene Regulatory Models to Get At Latency

Researchers would like to understand how TB survives its veiled existence in the lung. "If we can better understand latency, it might change our minds about how to treat latency," Galagan says.

To understand latency, researchers must combine gene regulatory models with metabolic network models. "There are probably 1,000 papers on how to model regulatory networks and probably 1,000 on metabolic networks" says **Nathan Price, PhD**, of the Institute for Systems Biology in Seattle. "But there are very few that do them together in an integrated way."

In a paper published in *PLoS Computational Biology* in June 2011, McFadden's team explored changes in metabolites when *Mtb* is grown inside a host cell (a macrophage). Their method, called differential producibility analysis (DPA), uses a metabolic network to extract metabolic signals from transcriptome data, allowing a glimpse at which pathways *Mtb* is using inside the host cell, what substrates it is eating, and what products it is generating.

One of the most interesting results: According to the network model, growing the bacillus in a macrophage caused it to go quiet. It stopped making DNA, RNA and amino acids and focused on one job: rebuilding its cell wall. "It seems that the TB bacillus has realized it's inside a host immune cell whose job is to kill it," McFadden says. "So the *Mtb* hunkers down, shuts



Galagan E-Flux. *E-flux uses gene expression to set maximum flux constraints on individual metabolic reactions. This can be illustrated as pipes of different widths around each reaction as illustrated here in a simple model of four metabolites (A–D), four internal reactions, an uptake reaction for A, and a reaction converting D to biomass. On the left are simulated gene expression data for four genes whose enzymes catalyze the four internal reactions (green = lower expression; red = higher expression). Where G1 is poorly expressed (top panel), a thin pipe is illustrated around reaction 1. In the bottom panel, G1 and G2 are highly expressed, corresponding to a wider pipe for these reactions. Under conditions in which uptake of A is not limiting, we would predict more flux through R1 and R2 in the bottom panel relative to the top panel and R3 and R4, as shown by the bars on the right. Reprinted from Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, et al. (2009) Interpreting Expression Data with Metabolic Flux Models: Predicting* Mycobacterium tuberculosis *Mycolic Acid Production.* PLoS Comput Biol *5(8): e1000489. doi:10.1371/journal.pcbi.1000489.*

down central processes and creates a more effective barrier for itself." Although this is not a novel insight—researchers already knew that *Mtb* shifts to cell wall building during latency—seeing it in a model was both impressive and instructive. Some of the details of the model might help researchers develop drugs to kill *Mtb* more easily, McFadden says.

Previously, when researchers have looked at transcriptomes, McFadden says, they've picked a favorite gene and looked at what that gene does. "It's like throwing a thousand stones in a pond and producing a thousand ripples but looking at just one of those ripples in an attempt to understand what's going on," he says. "What we do with DPA is look at all the ripples and put them together to get a picture of what's going on throughout the entire system."

Price and graduate student **Sriram Chandrasekaran** at the University of Illinois took a different approach to creating a unified model of *Mtb* gene regulation and metabolism. Their model, called PROM (probabilistic regulation of metabolism), was published in *Proceedings of the National Academy of Sciences* (*PNAS*) in 2010 and integrates work from Palsson's lab on metabolic networks.

PROM calculates the probability that expression of a particular transcription factor will result in the expression of a particular metabolic enzyme. These probabilities act as constraints on the metabolic model, like a dimmer switch. For example, researchers can ask how a knockout of a particular transcription factor will affect the abundance of metabolic enzymes and, in turn, the flux through a particular reaction. "We're trying to link changes in transcriptional regulation to what's going to happen in terms of a metabolic phenotype," Price says. And, like the metabolic model on its own, PROM was able to identify essential genes. "PROM picked up really well which transcription factors are essential to optimal growth in tuberculosis," Price says.

Another approach, called E-flux, brings gene expression together with metabolic models in yet another way. Developed by Galagan and his colleagues at the Broad Institute, E-flux uses gene expression data to constrain the metabolic model—essentially setting the width of the pipes leading to and from particular reactions in the metabolic network. Because the gene expression data comes from *Mtb* grown under a variety of conditions—including under exposure to 75 different substances and conditions such as hypoxia (which is akin to what the bug experiences in latency)—the results could help researchers understand how existing drugs work and identify other drugs that might also be effective.

In a 2009 publication in *PLoS Computational Biology*, Galagan's team applied E-flux to existing metabolic models of the *Mtb* mycolic acid pathway. Mycolic acids are good drug targets because they are critical components of the *Mtb* cell wall, do not exist in humans, and are the target of several existing antibiotics used to treat TB. E-flux predicted seven of eight known inhibitors of the mycolic acid pathway and identified several novel compounds not previously known to inhibit mycolic acid biosynthesis. The model also mimicked the ineffectiveness of first-line TB drugs against dormant tuberculosis.

But Galagan and his team have more ambitious goals for E-Flux: They want to improve the method so that it can make more refined interpretations of what *Mtb* is doing in latency—or at other key points in the disease process, Galagan says. His plans include building a metabolic model that isn't constrained to a steady state, by using mass spectrometry to measure metabolites directly.

In recent work with E-flux, his team observed—as McFadden did—that when TB goes dormant, most of the genes are ramped down. "But if you turn off the lights, things could go haywire," Galagan notes. "TB has to handle the process in a way that doesn't kill itself." His team observed that as *Mtb* scales down its activities, the bacterium makes a series of adaptations to the toxins that build up. "Those could be a weak link," he says. "Perhaps we could muck with that—target those processes that are important for *Mtb* not dying as it goes to sleep," Galagan says.

Ultimately, McFadden says, researchers need to link *Mtb* metabolic and regulatory models together with models that look at how *Mtb* and host cells interact. "That's really the challenge for the future," he says.

Ultimately, McFadden says, researchers need to link *Mtb* metabolic and regulatory models together with models that look at how *Mtb* and host cells interact. "That's really the challenge for the future," he says.

3 Multiscale Models of *Mtb*-Host Interactions

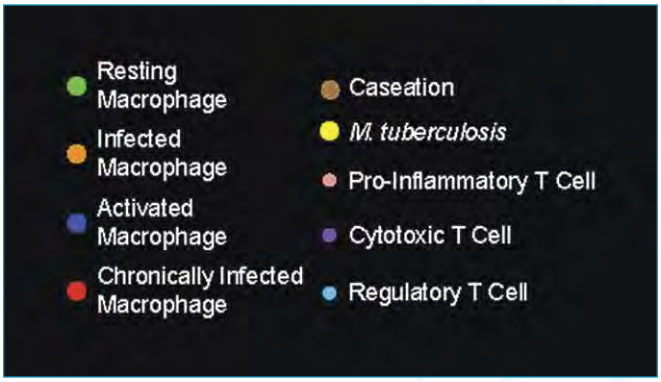
The granuloma, a spherical conglomeration of immune cells, bacteria, and tissue that walls off *Mtb* bacteria inside the human lung, offers another piece of the TB latency puzzle. How does the immune system force TB into an inactive state? Why do some people contain and wall off TB infections better than others? And why do

granulomas sometimes break down to reactivate TB infection?

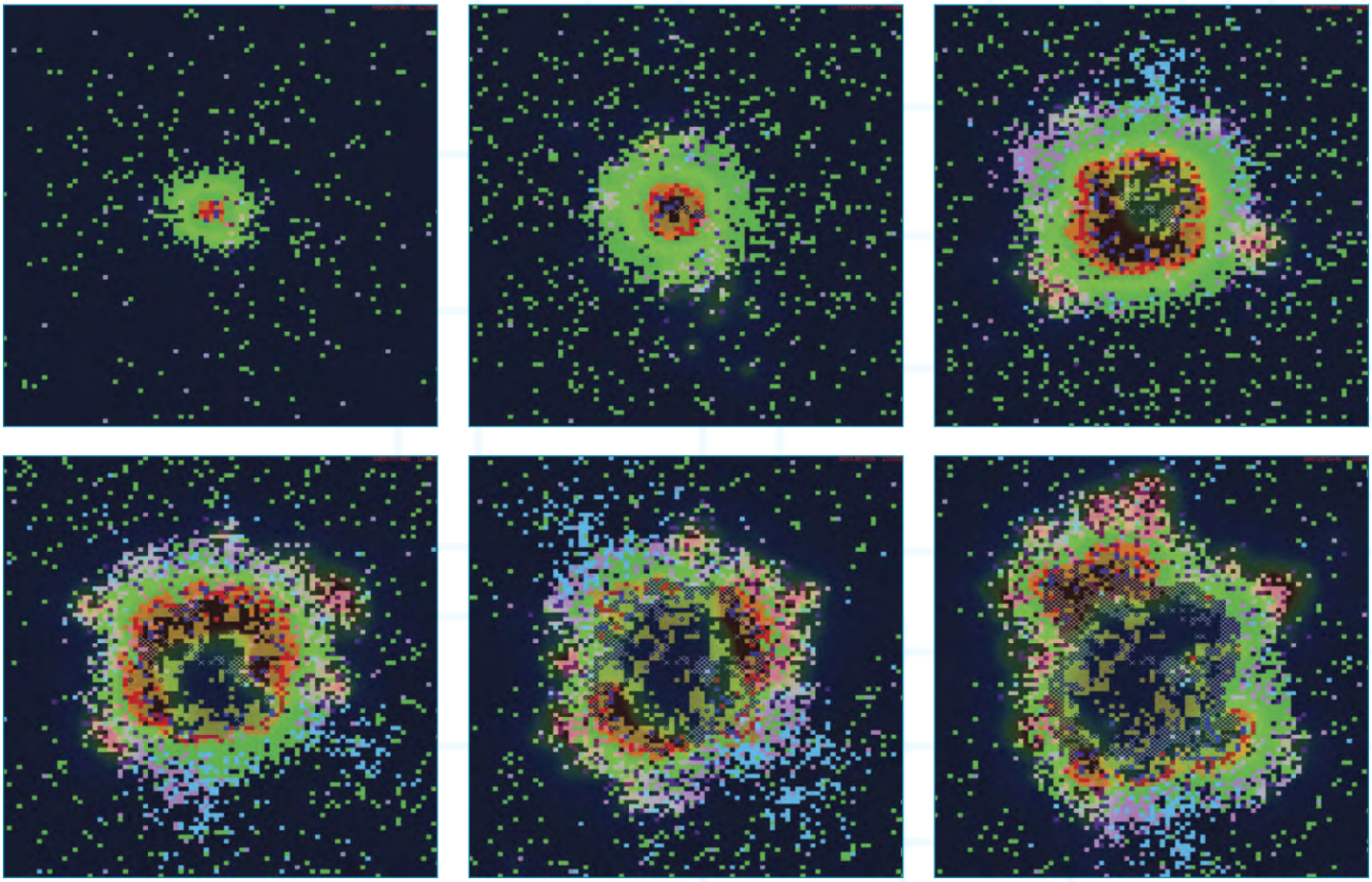
Denise Kirschner, PhD, professor of microbiology and immunology at the University of Michigan Medical School, has been modeling the granuloma for about 11 years. Her team uses a multitude of host-pathogen response data gathered from granulomas in monkeys to build models at multiple scales: Ordinary differential equation models at the molecular scale link to an agent-based model at the cellular scale that reads out at the tissue scale.

Kirschner's models are stochastic, meaning they contain probabilities that certain events will or will not occur: One hundred simulations will generate one hundred different answers, simulating the sorts of variability one would find in human hosts. Kirschner refines the agent-based models until they produce outcomes that are relatively stable—akin to TB in its latent state. "Even with slight perturbations, the models still go to the same place in the end," she says. It then becomes possible to run *in silico* experiments that perturb the models. "We can look at the finest scale and say how it's impacting at the largest scale and vice versa," Kirschner says. And they can virtually "knock out" various parts of the host immune system instantaneously to see the effect on the granuloma. Ultimately, Kirschner would like to understand what brings *Mtb* out of the stable state—transitioning the disease from latency to reactivated disease. But for now, her model predictions are helping to focus the next series of animal experiments.

Since 2004 when she published her first model of the granuloma, Kirschner's team



These six snapshots from a time-lapse simulation of a granuloma forming after infection with* Mycobacterium tuberculosis *show a 2 mm by 2 mm slice of lung tissue. The simulation covers 200 days of infection dynamics (days 0, 50, 75, 100, 150, 200). Once the infected macrophage cell takes up bacteria and initiates infection, it begins to recruit additional immune cells that arrive via vascular sources distributed on the parenchyma. This simulation represents a controlled infection where the granuloma is able to physically contain and immunologically restrain the bacteria. Courtesy of Denise Kirschner.



has refined and improved the models. They can now be run in 3-D and incorporate compartments outside the lungs, including the lymph nodes and blood. She's also taking the model to a new scale: populations. Working with a team in Italy, she says, they now have an agent-based model of TB epidemiology. The people in the population-scale model each have an immune-scale model of a granuloma running inside them, Kirschner says. These models predict how events at the smallest scales can influence epidemic outcomes, and thus can be used to test vaccine and treatment strategies.

Others studying the systems biology of TB latency and reactivation in the host are still at the early stages of their projects. For example, **Henry Boom, MD**, vice chair for research and director of the tuberculosis research unit at Case Western University Medical School, is looking at whether patients at different stages of progression from latent TB to active TB can be distinguished by looking at protein-protein interaction sub-networks inside certain host immune cells. Watch for results in the future.

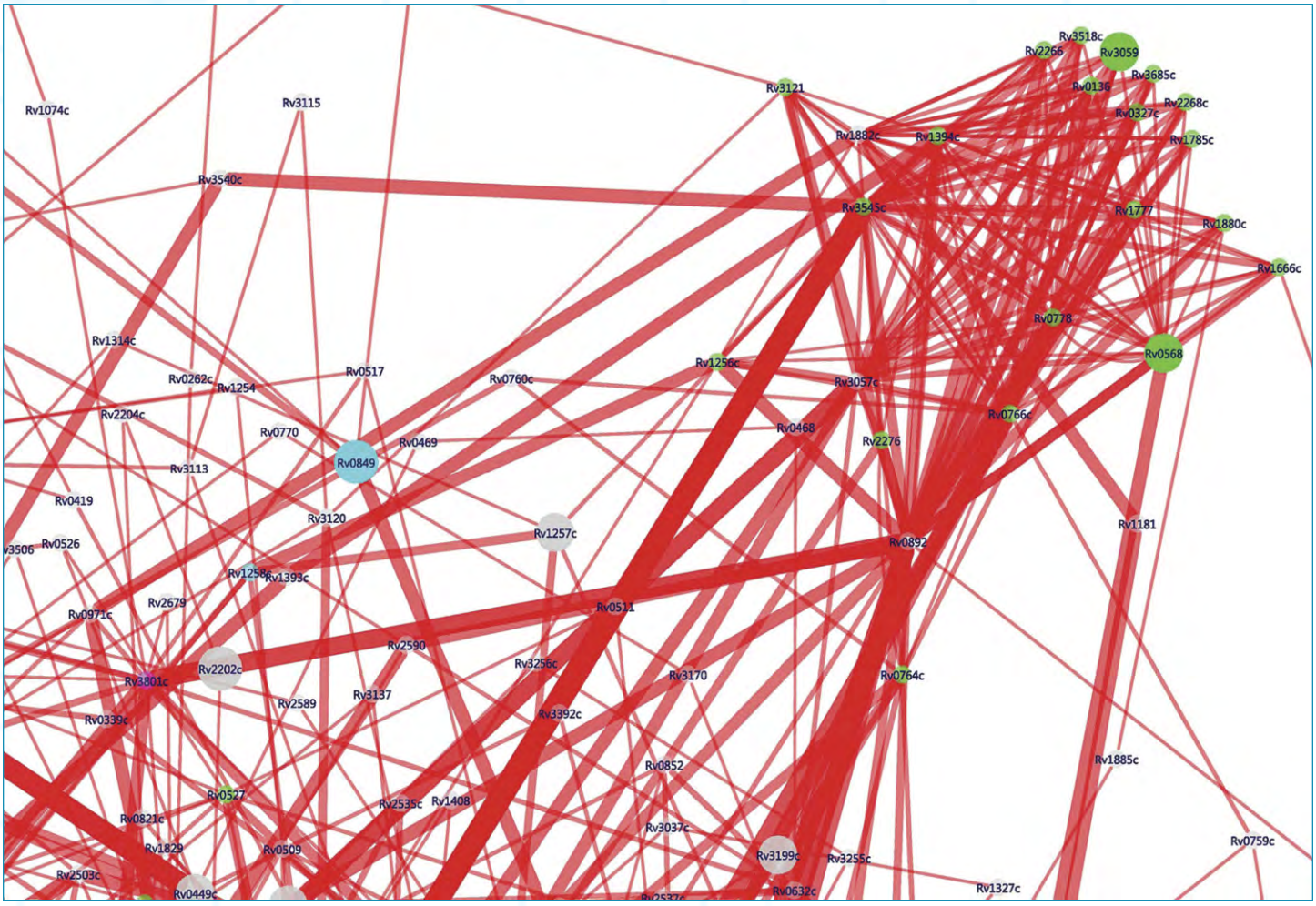
4 Protein-Protein Interactions and TB Drug Resistance

Emergence of drug-resistant TB strains is the biggest health challenge facing TB researchers. "Each time you administer a drug you are selecting for the organisms that are drug resistant," notes **Karthik Raman, PhD**, assistant professor of biotechnology at the Indian Institute of Technology Madras in Chennai, India.

While he was a graduate student in the lab of Dr. Nagasuma Chandra at the Indian Institute of Science, Raman and his colleagues used a systems biology approach to unravel the different mechanisms by which TB drugs trigger resistance. His team created an *Mtb* protein-protein interaction network (from the STRING database of protein-protein interactions). They then merged that network with *Mtb* gene expression data gathered under exposure to seven different TB drugs, allowing an analysis of the possible routes leading to resistance.

One of their key aims was to identify what Raman calls "co-targets"—proteins that could be inhibited along with a primary drug target to reduce the likelihood that the *Mtb* bacillus will become drug resistant. "Some proteins were more important than others in terms of their strategic location in the network," Raman says. "Our hypothesis is that one could try to disable these proteins and not just the target proteins and it could probably reduce resistance."

Further work in this area is needed, Raman says. "It's probably an issue we'll never conquer." Yet developing novel strategies, such as the co-target concept, could help.



This portion of the TB protein-protein interaction network involved in drug resistance shows a tight cluster of cytochrome proteins (green nodes) and Rv0892 (blue node), a potential co-target that links the cytochrome clusters to proteins in the mycolic acid pathway. The nodes (individual proteins) are sized in proportion to the number of MAP drugs that induce their upregulation. The thickness of an edge is proportional to the number of times a shortest path is traversed through that edge. Reprinted with permission from Karthik Raman and Nagasuma Chandra, **Mycobacterium tuberculosis** ***interactome analysis unravels potential pathways to drug resistance,*** **BMC Microbiology** ***8:234 (2008).***

5 Computational Epidemiology and the Emergence of Drug Resistant TB

Computational models are also proving useful in exploring the population-level causes of drug resistant TB. Surprisingly, a recent statistical model shows that TB multiple drug resistance can evolve spontaneously—it is not necessarily caused by mono-therapy or by patients failing to complete a course of antibiotic treatment. This is a fundamental shift in our understanding of how combination drug resistance can emerge, says **Ted Cohen, MD, MPH, DPH**, assistant professor in epidemiology at Harvard University. "And it may help explain how highly drug resistant forms of TB have independently emerged in many settings."

Cohen is also working out the order in which drug resistance mutations occur and their probabilities of occurring. Similar work has proven helpful in understanding and treating HIV. But exploring the order of mutations over time typically requires longitudinal genotypic data—which is often not available for *Mtb*. So Cohen and his colleagues decided to determine whether this information could be inferred from phenotypic data (e.g., *in vitro* tests of drug resistance) gathered at one point in time. Using branching trees, a special kind of Bayesian network that makes it possible to infer past and future events, they were able to infer some possible patterns in which TB drug resistance phenotypes arise, Cohen says. "It's a promising approach and should be even more useful as genetic and genomic data becomes available and we can look not only at drug resistance phenotypes but also at the actual resistance-conferring mutations."

Cohen is also doing work in South Africa to try to understand the phenomenon of complex TB infections—multi-strain infections that arise either from multiple infections by unrelated strains or from within-host evolution. He and his colleagues developed a modeling framework to investigate mechanisms of strain competition within hosts and to assess the long-term effects of such competition on the ecology of strains in a population. His initial modeling efforts suggest that the presence of mixed strains in a single host can increase the likelihood that drug-resistant strains will persist and potentially evolve.

"For me, modeling is a cycle," Cohen says. "We've used models to identify the gaps in our understanding of TB that most limit our ability to project trends or design effective interventions. Then we try to conduct studies that reduce this uncertainty so we can refine the models and improve our understanding of how best to intervene."

"For me, modeling is a cycle," Cohen says. "We've used models to identify the gaps in our understanding of TB that most limit our ability to project trends or design effective interventions. Then we try to conduct studies that reduce this uncertainty so we can refine the models and improve our understanding of how best to intervene."

6 Using Computation to Find TB Drug Targets

The National Institute of Allergy and Infectious Diseases (NIAID), part of the National Institutes of Health, supports a portfolio of computational approaches for modeling TB drug targets and determining how promising drugs bind to these targets, says **Karen Lacourciere, PhD**, program officer for tuberculosis and other mycobacterial diseases at NIAID.

Computation can simplify the entire drug discovery process. For example, Raman developed a pipeline (TargetTB) for pinpointing which essential *Mtb* genes (and their protein products) are also good drug targets. The pipeline starts by identifying essential genes as predicted by both existing metabolic models (including McFadden's and Palsson's) and by experiments. Next, the pipeline filters out proteins with structural similarities to human proteins because targeting such proteins can produce negative side effects. This computationally intensive step involves exhaustive pairwise comparisons of several thousand pockets on more than 750 *Mtb* proteins with more than 70,000 sites on more than 15,000 human proteins. The pipeline filters the resulting short list using additional criteria and also prioritizes the proteins' importance based on their level of expression during TB latency. The work, published in *BMC Systems Biology* in 2008, also examines several known

"Ten to twelve years back [TB] was considered a third-world disease, but with the appearance of drug resistant forms, pharmaceutical companies are showing more interest," Raman says.

and predicted drug targets based on their filters, and postulates why many known targets may produce adverse drug reactions.

About 400 potential drug targets have emerged from Raman's pipeline. "It's a beautiful place to start, those 400," Raman says. And pharmaceutical companies have shown some preliminary interest in pursuing these targets. "Ten to twelve years back it was considered a third-world disease, but with the appearance of drug resistant forms, pharmaceutical companies are showing more interest," he says.

McFadden's metabolic model is also being used to better understand known drug targets, says **Desmond Lun, PhD**, associate professor of computer science at Rutgers University. For example, Harvey Rubin at the University of Pennsylvania asked Lun to work out the mechanism of action of NDH-2. Rubin hypothesized that NDH-2 could be a lethal knockout in *Mtb* because removing the gene from a related bacterium leaves a non-viable organism. Using a metabolic model that consolidates two others, Lun identified a possible mechanism by which knocking out NDH-2 would be lethal in *Mtb*. "It takes a long time to develop drugs," Lun says, "so you want to know in significant detail what the mechanism is and the possible effect on the host and related organisms."

7 Using Computation to Find TB Drugs

Some chemists are screening for TB drugs using computational approaches, says **Sean Ekins, PhD**, vice president of science at Collaborative Drug Discovery and senior consultant for Collaborations in Chemistry. "People started by thinking we'd do structure-based screening but the molecules that looked good against the target didn't really have the right physical properties to get into the cell," he says. "TB is a pretty tough cookie in terms of the types of molecules it will let in."

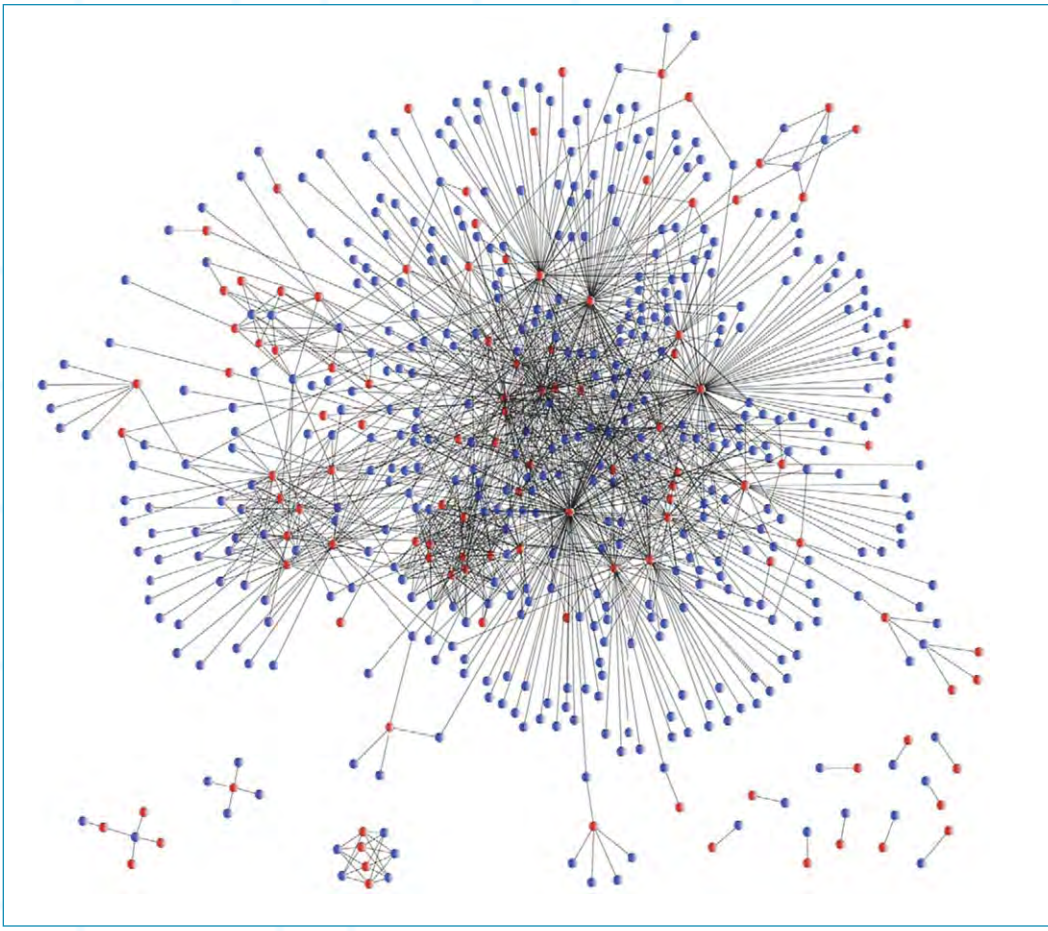
Then a few years ago, researchers started whole cell screening—testing lots of compounds to see what would kill *Mtb*. This produced lists of thousands of lethal compounds but gave researchers no hints as to which ones merited further research, Ekins says. So Ekins decided to try a data-mining approach on an NIH database of more than 200,000 compounds with known activity against *Mtb*. The approach uses Bayesian machine learning to cherry-pick the compounds that have a high likelihood of whole cell activity and a low likelihood of human toxicity, he says. "These models are a step in the right direction: We're leveraging all this data, using it to build models, and using the models to make future experimental decisions for our collaborators."

Indeed, Ekins has used the model to pick groups of compounds for various researchers to test. "That's sort of the acid test, really: backing up your predictions experimentally," he says. "I don't think we've got the magic bullet, but we're giving it a good try."

Using a different approach, **Lei Xie, PhD**, associate professor of computer science at Hunter College, the City University of New York, and **Philip Bourne, PhD**, professor of pharmacology at the University of California, San Diego, sought TB drugs among the pool of currently FDA-approved drugs by creating a genome-scale drug-target network they call the TB drugome. The network compares binding sites on *Mtb* proteins against the known binding targets of FDA-approved drugs (there are only 250). "If two proteins have a similar ligand-binding site, then our assumption is that they can potentially bind a similar drug," says Xie. Next, they did protein-ligand docking to predict the binding affinity of the drugs to the *Mtb* proteins. The network revealed that about a third of the drugs examined have the potential to be repositioned to treat tuberculosis and that many currently unexploited *Mtb* receptors may be chemically druggable and could serve as novel anti-tubercular targets. The researchers are currently seeking collaborators to validate these findings.

8 Computation to Find Diagnostic Biomarkers

Since the 1940s, the primary diagnostic test for active TB has been the same: Take a sample of sputum (i.e., coughed up mucus) and look at it under a microscope. But the test is insensitive, producing many false-negative results. In addition it takes two



The TB Drugome, a protein-drug interaction network, shows* Mtb *proteins (blue circles) connected to drugs (red circles), a single connection indicates binding site similarity between any of the structures of the connected* Mtb *protein, and any of the binding sites of the connected drug. This Drugome is highly connected, indicating that many binding site similarities were observed between* Mtb *proteins and drug targets, even though those proteins had different overall structures. Reprinted from Kinnings SL, et al., The Mycobacterium tuberculosis Drugome and Its Polypharmacological Implications,* PLoS Comput Biol *6(11): e1000976 (2010).

weeks to get results to tests for drug resistance, risking further spread of TB's most virulent forms.

Although newer, more rapid TB diagnostic tests are now hitting the market, many of them require specialized labs and skills not available in areas most stricken by the disease. In addition, the rapid diagnosis of drug resistance remains elusive with one exception: A new test now being deployed in parts of Africa uses the GeneXpert system, a molecular assay that can identify strains of TB that are resistant to treatment with rifampin, a commonly used TB antibiotic. This test became possible because, for rifampin, we know the affected gene and the majority of mutations that result in resistance, says **James Posey, PhD**, a research microbiologist in the Center for Disease Control's Division of Tuberculosis Elimination. But for many of the other first- and second-line TB drugs, this information is unknown. Computational approaches are helping fill this gap; for example, Posey's lab is using whole genome sequencing to identify the genes and mutations that cause resistance to additional TB drugs.

Another group of researchers is searching for metabolic changes that could be used to rapidly flag drug-resistant mutants. Lun is working with **Greg Bisson, MD**, assistant professor of medicine at the University of Pennsylvania Medical School, on the novel hypothesis that, just as *Mtb*'s ancestors—non-pathogenic soil bacteria—respond to assault by making new metabolites, perhaps TB does the same when confronted with drug treatment. Lun is perturbing a consolidated metabolic model using protein abundance data as a way to study that possibility. "This is something we can do to help develop a cheap and easy diagnostic test to work out whether someone has a drug resistant strain," Lun says. "This can make a major difference in public health outcomes."

9 Planning for the Clinic

Even after effective drugs and diagnostics for TB have been developed, deploying them in the clinic can be tricky. For example, when putting a new diagnostic test into action, doctors need to know who should be tested and whether the new tools should replace or complement existing ones. And the answers to those questions might depend on the setting—both the incidence of TB and the availability of resources. To explore these questions, Cohen and his colleagues combined an epidemiological model of TB spread with a health system model. The work shows that, in concept, one can begin to evaluate the operational impact of a diagnostic tool using information not only about how the bug spreads but also about the logistical characteristics of the healthcare system. "That's something many people have simplified out of the problem," he says. "There is a role for modeling to inform local, immediate, real world–type decisions while taking into account detailed knowledge of local conditions." He's now working with others to build on this initial work to look at the potential effects of specific diagnostic tools such as the Gene Xpert system.

"There is a role for modeling to inform local, immediate, real world-type decisions while taking into account detailed knowledge of local conditions," Cohen says.

There are now at least nine drugs and 12 vaccines in the clinical research pipeline. If they prove medically effective, similar sorts of deployment models will be essential to their ultimate impact.

10 Setting the Clinical Research Agenda

About 10 years ago, the Bill and Melinda Gates Foundation began investing heavily in a tuberculosis research portfolio that included the development of drugs, vaccines and diagnostic tests. Having set that agenda, they then wondered: If we achieve our goals, what will TB morbidity and mortality look like in 2050? So they hired a team headed by **Elizabeth Halloran, MD, DSc**, professor of biostatistics at the University of Washington and the Hutchinson Research Center, to create a model showing the long-term effect of a successful program.

The work, published in *PNAS* in 2009, produced several interesting insights that could affect funding decisions. For example, introducing a new vaccine for infants would have very little effect by 2050 because people get TB when they are older. But a program of mass vaccinations is much more effective. "So one has to rethink vaccination strategies and clinical trials," she says. The model also showed that finding and curing latent infections—which we currently don't know how to do—would have a very large effect. Halloran notes: "That might affect how you allocate research resources." For example, this might suggest the wisdom of funding systems biology studies of TB latency, which brings us full circle.

TB The Opportunities Are Many

Although these intersections between computation and TB might suggest the field is pretty well picked over, that is not at all the case. The systems models need refinement and must be layered together with other models to gain a multiscale picture of the bug. Host-pathogen interaction research is really in its infancy. Researchers don't know the extent of MDR and XDR TB, let alone how to deal with it. And though there are multiple new diagnostics, drugs and vaccines in the pipeline, no one really knows how to implement them so that they will have the greatest impact.

Thus, each of these intersections between TB and computation suggests more that can be done. Perhaps because of its complexity, modelers haven't flocked to TB research, Cohen says, but he believes that will change: "If you want to ask questions that have global impact to improve the lot of humanity," Cohen says, "I think TB is a great thing to choose to work on." □

"If you want to ask questions that have global impact to improve the lot of humanity," Cohen says, "I think TB is a great thing to choose to work on."

Under TheHood

BY JENELLE BRAY, PhD

Normal Mode Analysis: Calculation of the Natural Motions of Proteins

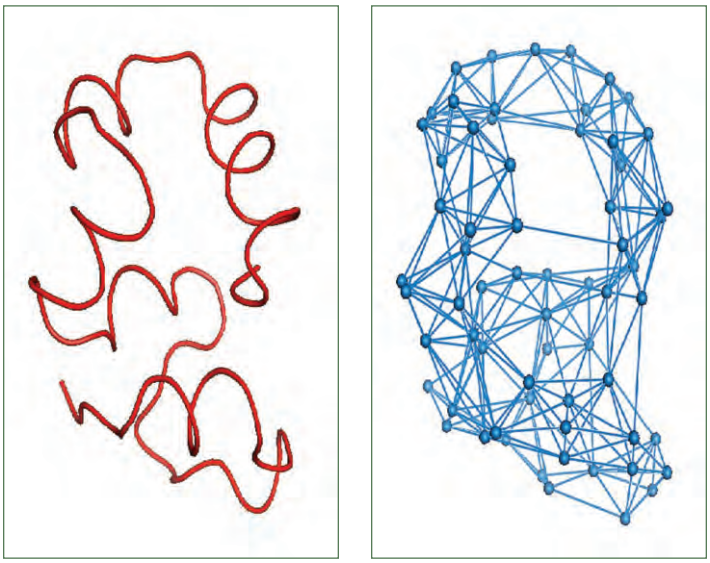
Advances in computational power and algorithms have led to longer and more accurate molecular dynamics simulations of protein folding. But these approaches, because they are computationally intensive, cannot yet be used to model conformational changes of large, already-folded proteins at biologically relevant time scales. Yet these kinds of movements are often biologically interesting: For example, understanding and predicting the conformational change a protein undergoes upon the binding of a small molecule—such as a drug—can lead to better rational drug design.

Normal mode analysis fills the gap: It can quickly reveal the overall change in the conformation of large proteins, without the need to calculate the specific molecular mechanism, such as the motion of specific bonds.

Normal mode analysis, also known as harmonic analysis, existed before the advent of computational biology and is applicable to many fields. It identifies the natural, resonant movements of a physical object, such as a building, bridge or molecule. In acoustics, a guitar string's harmonics (or overtones) are its normal modes, and the modes' corresponding frequencies are their pitches; in geology, normal mode analysis of low frequency normal seismic waves generated by large earthquakes leads to greater understanding of the deep substructures of the earth.

For proteins, normal mode analysis represents each amino acid as a bead. All pairs of beads less than a specified distance away from each other in 3-D space are connected by springs. The force constants of the springs are determined by fitting to experimental data. Once the spring constants and the masses of the atoms have been tabulated, the movement of the beads and springs can be described by a matrix version of Newton's second law of motion. Analytically solving this equation gives the protein's natural motions, called normal modes, and their associated frequencies.

A combination of the normal modes describes the motion of the protein. The biologically relevant modes are the low-frequency modes because they describe the large-scale, overall motion of the protein. Normal mode analysis does not specify which of the low-frequency modes describe most of the conformational change. However, with a small amount of experimental data, such as the change in pairwise distances between a few pairs of residues (derived from fluorescence resonance energy transfer or FRET) or in overall shape (derived from cryo-electron microscopy), the relevant modes can be determined. Then the conformational change of the protein can be predicted.



A protein shown on the left as a simple ribbon and at right in a beads and springs representation for normal mode analysis. Courtesy of Jenelle Bray.

In addition to predicting conformational changes of proteins, normal modes can be used to help solve x-ray crystallography structures or to improve protein-ligand docking calculations. Additionally, if there are two known conformations of a protein, the normal modes that contribute most to the conformational change can be calculated. Then we can use the modes to understand the pathway and to describe the conformational change in only a few degrees of freedom.

The application of normal mode analysis to proteins has given researchers an important tool for solving problems in computational biology. □

DETAILS

Jenelle Bray is a Simbios Distinguished Postdoctoral Fellow in Russ Altman's and Michael Levitt's labs. She works on the development and application of torsion angle normal mode analysis.

For more information about normal mode analysis, go to http://www.igs.cnrs-mrs.fr/elnemo

or check out a recent paper by Jenelle Bray and her colleagues: Optimized torsion-angle normal modes reproduce conformational changes more accurately than cartesian modes. Bray JK, Weiss DR, Levitt M. *Biophys J.* 2011; 101 (12): 2966-9

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

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

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

seeing science

SeeingScience

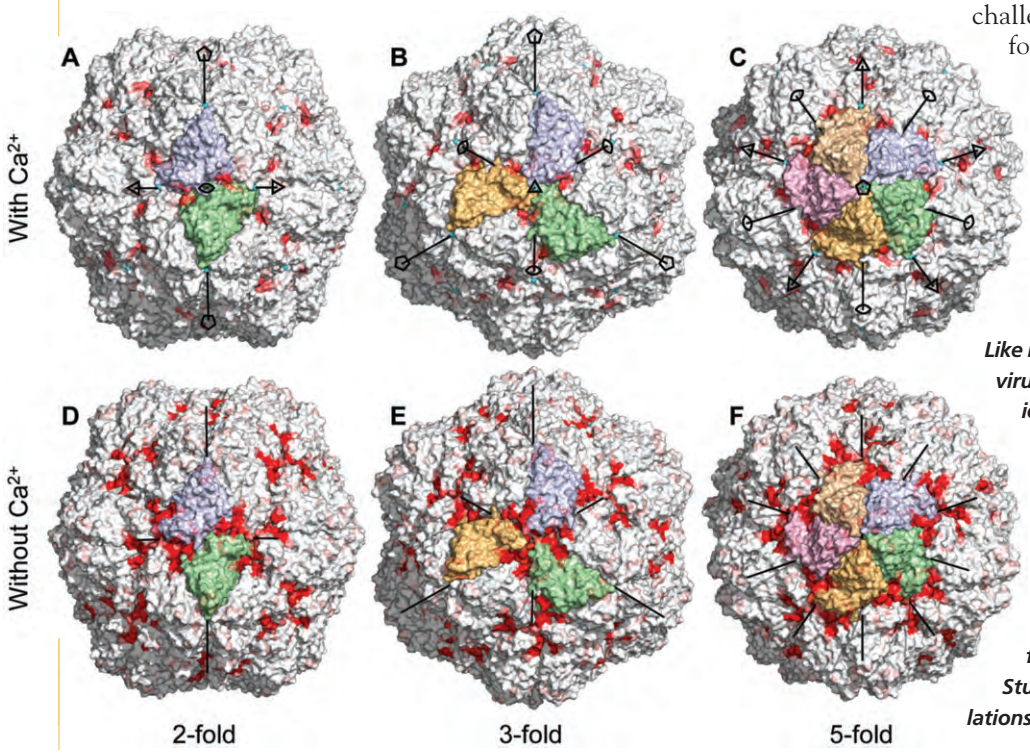
BY KATHARINE MILLER

Dissolving a Viral Capsid

After a satellite tobacco necrosis virus particle infects a cell, it sheds the calcium ions that hold the capsid proteins together. Next, the proteins start to repel each other, the capsid swells and water begins to enter. It's a process that hasn't been observed directly, but can now be seen in the longest and biggest virus simulation to date—a one-microsecond long, full-atom, molecular dynamics simulation by **David van der Spoel, PhD**, professor of biology in the department of cell and molecular biology at Uppsala University, Uppsala, Sweden and his graduate student, **Daniel Larsson**.

"We are seeing the beginning of the infection process as the capsid starts to open up," van der Spoel says. Next, his lab plans to add the genome to the simulation—a challenge because there is no structure for the genome.

"Even though it's not a virus that attacks humans, most viruses have a similar protein shell that protects the genome," van der Spoel notes. "If you can tinker with the shell, then you can use it as an additional route to combat viruses." □



Like many virus particles, the satellite tobacco necrosis virus (the smallest known virus) has multiple lines of icosahedral symmetry—two-fold (A,D), three-fold (B,E) and five-fold (C,F). Larsson and van der Spoel's simulation reveals areas where water can permeate the capsid (red) with (A,B,C) and without (D,E,F) bound calcium ions. The nearly symmetrical water-permeable zones suggest where the capsid is least stable and most likely to open up to release the genome. Reprinted from Larsson DSD, et al., Virus Capsid Dissolution Studied by Microsecond Molecular Dynamics Simulations.* PLoS Comput Biol *8(5) (2012).