DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

Profiles IN Computer Science Courage

Reflections on the rewards of plunging into biomedicine

INSIDE:

Spring 2011

ContentsSpring 2011

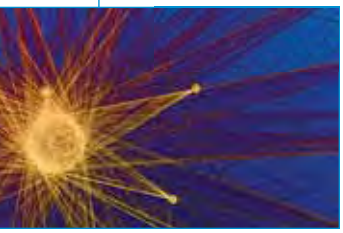
15 Profiles in Computer Science Courage:

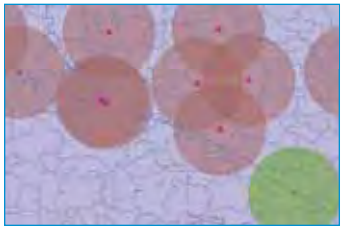
Reflections on the rewards of taking the plunge into biomedicine

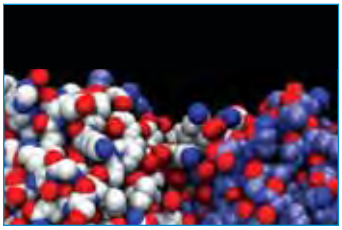
BY KATHARINE MILLER

DEPARTMENTS

Cover Art: Created by Rachel Jones of Wink Design Studio using: silhouette shapes, © Pawel Nowik | Dreamstime.com; DNA image, © Anna Raspopova | Dreamstime.com; *E. coli*, © Eraxion | Dreamstime.com.







Spring 2011
Volume 7, Issue 1
ISSN 1557-3192

Executive Editor Russ Altman, MD, PhD
Advisory Editor David Paik, PhD
Associate Editor Joy Ku, PhD
Managing Editor Katharine Miller

Science Writers
Katharine Miller
Kristin Sainani, PhD

Community Contributors
Lucila Ohno-Machado, MD, PhD,
Joy Ku, PhD,
Manuel K. Rausch

Layout and Design
Wink Design Studio

Printing
Advanced Printing

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

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

Biomedical Computation Review
is published by:

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

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

Peter Lyster, PhD (NIGMS)
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 LUCILA OHNO-MACHADO, MD, PhD

Bridging the Scientific Divide: Enabling Sharing through Biomedical Computing

Once upon a time,

a deep divide existed between scientists who did and those who did not have easy access to scientific content (journals, lectures, data), hardware (imaging devices, lab instruments, supercomputers) and people (professors, networks of peers, knowledgeable support personnel). Not surprisingly, most discoveries were limited to institutions in which these pillars were the strongest. Over time, libraries helped expand access to scientific content, but access to data, hardware and people were still constrained. Some academics preferred it this way, as they could be isolated and concentrate on their own scientific endeavors. Unfortunately, this isolation led some to completely lose perspective of the university as a place to create, propagate and disseminate meaningful scientific knowledge, not only to its own members or peers, but also to society as a whole. Research, teaching, and service became almost mutually exclusive. The achievement of prestigious academic awards essentially defined academic success. Enabling others to succeed and training the next generation of scientists remained desirable but not necessary. Translating basic scientific findings into practice remained useful but not required. For fear of change, faculty from some institutions prevented "strangers" from filling their ranks and gaining access to their resources. These "strangers" who crossed disciplinary boundaries and who were deemed collaborative were reviewed unfavorably. An unknown discipline like informatics was by definition an undeserving one.

That was then.

Then came inexpensive computers and networks and a few individuals with the vision to break the sacred circle. These individuals (librarians,[1] computer scientists, healthcare providers, engineers, biologists, physicists, and informaticists) believed that everyone should have access to knowledge, data, and tools; that everyone could start the scientific race from approximately the same baseline; and that allowing anyone to test a new idea would help the best ones prevail. With more people accessing information, discoveries would be made sooner without the hindrance of dealing with politics stemming from data ownership, copyrights, academic recognition, and fundraising.

Some could now assemble large virtual cohorts of patients while still preserving human subjects' privacy, increasing the power to study rare diseases, empower observational studies that compared effectiveness of interventions, and speed the detection of unsafe behaviors, medications, or medical devices. Others could use crowd sourcing to annotate large collections of data or construct extensive searchable knowledge bases and tools that allowed systems to interoperate. Some could build tools to generate hypotheses based on a combination of already publicly available and brand new data. Others could build resources that would help prevent the pursuit of failed ideas that were never made public because they were unpublishable. Some could use computational resources to find hidden patterns in millions of electronic medical records, images, laboratory tests, and billions of nucleotide sequences. And, enabled by computing resources, everyone would be able to discuss findings with peers across the globe, attend lectures and ask questions of world experts, and interact with peers no matter where they lived.

We can continue the bold vision started by those who challenged the establishment and dared to think beyond the ordinary. We can share knowledge, data, tools, and computational resources in a sustainable manner. While others have made this pledge here before [1-3], the current research funding climate, combined with tremendous advances in technology and social networking in the past few years makes it even more possible and imperative: let's not waste this unique opportunity to make a difference.

[1] Altman RB. Share and Share Alike: A Proposed Set of Guidelines for Both Data and Software.

[2] Musen MA. It Takes a Village: Building the Next Generation of Biomedical Ontologies.

[3] Erdemir A. Recognizing and Encouraging Timely Dissemination.

Lucila Ohno-Machado, MD, PhD, is Professor of Medicine and founding Chief of the Division of Biomedical Informatics at UC San Diego. She is the editor-in-chief of the Journal of the American Medical Informatics Association and is principal investigator of iDASH (integrating Data for Anonymization, Analysis, and Sharing), a new National Center for Biomedical Computing. **iDASH: http://idash.ucsd.edu** □

[1] *PubMed epitomizes the spirit of this enlightened era.*

SimbiosNews

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

CAMPAIGN: Expanding the Universe for Clustering Algorithms

GenBank, a repository for storing biological sequences, currently contains some 124 billion base pairs and is doubling in size every 18 months.[1] Though not a huge number compared to the billion trillion stars in the universe, a human being would clearly have difficulty making sense of it. Clustering algorithms can aid us in this effort, partitioning the data into groups. However, these algorithms can take a long time to run. A natural solution to speed up the process is to implement them on parallel architectures, but typically these do not work well with large, high-dimensional data sets. CAMPAIGN, a C++ library of **C**lustering **A**lgorithms for **M**assively **P**arallel **A**rchitectures **I**ncluding **G**PU **N**odes, was created to fill that gap.

"What would normally take an hour on a CPU, you can now do on a GPU in a second," says **Kai Kohlhoff, PhD**, a Simbios postdoctoral fellow who developed CAMPAIGN. "That's really exciting! Think about what kind of science you can do now."

For CAMPAIGN, Kohlhoff took five popular clustering algorithms—K-means, K-medoids, K-centers, hierarchical clustering, and self-organizing maps (neural networks)—and implemented them to run on both standard CPUs as well as on GPUs, or graphics processing units. The software library also includes several distance metrics, such as Euclidean and Manhattan, which can be mixed and matched with the clustering algorithms.

Kohlhoff modified the algorithms so that calculations could be run in parallel, taking advantage of the GPU architecture. CAMPAIGN's K-means algorithm can handle millions of data points with tens of thousands of dimensions (the coordinate system axes that describe the data set) and large numbers of clusters. "For a gene data set, you have a dimensionality with thousands of dimensions, but previous K-means implementations for GPUs reach at most a dimensionality of around 50 before performance starts suffering," says Kohlhoff. "That's a big limitation and that's what nobody had addressed before." By solving that problem, CAMPAIGN expands the research problems in which clustering can play a role.

Kohlhoff achieved the best results with K-means, which ran up to 2800 times faster on the GPU than on a CPU using MATLAB, a popular numerical software package, for his test case. The performance varied with the algorithm and the size of the data set. For instance, the speedup for hierarchical clustering when compared with fast C++ CPU code was five-fold, which was nevertheless noteworthy: "It is the first time, as far as I know, that you have full hierarchical clustering on a GPU, and not just the initial pair-wise distance computations," Kohlhoff says.

These faster algorithms can allow researchers to be more experimental and innovative. "Imagine if it takes you an hour to cluster your data," says Kohlhoff, "Now you can just say, 'Maybe this will work.' Then, try it out in an instance."

Kohlhoff envisions researchers using CAMPAIGN to develop novel clustering protocols, as well as to improve their clustering results. With computational time being less of an issue, researchers can explore clustering using different sets of parameters or run the algorithms with additional iterations. They could even try combining several different clustering algorithms that traditionally do not run very fast.

Clustering plays a key role in the initial steps of the Pande lab's Markov State Model software tool, MSMBuilder (see *Biomedical Computation Review*, Fall 2009), so they have started using CAMPAIGN to improve their models. "Clustering is often the rate-limiting step," says **Vijay Pande, PhD**, associate professor of chemistry at Stanford University. "The ability to speed clustering would have a fundamental impact in our ability to build better models."

CAMPAIGN currently runs on NVIDIA GPU cards using NVIDIA's programming language, CUDA. Coding algorithms in CUDA can have a big payoff in terms of performance, but it is not an easy task. That's why the availability of libraries like CAMPAIGN is significant. "I have come to appreciate how much work it is to program in CUDA, so I understand how much value there is in sharing these kinds of libraries," says Kohlhoff. □

[1] *GenBank Release Notes, February 15, 2011*

DETAILS

CAMPAIGN is open-source and is available for download from http://simtk.org/home/campaign.

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

FOLLOW THE MONEY:
Big Grants in Biomedical Computing

By Kristin Sainani, PhD

In the current economic climate, every research dollar counts. Fortunately, when it comes to biomedical computing, not everyone has been left counting change. Several big-dollar initiatives received NIH funding in late 2010, including efforts to: map all the connections in the human brain, fight diarrheal diseases with modeling, provide state-of-the-art tools for network biologists, and understand and rewire cellular stress pathways. The projects have implications for a range of human diseases from autism to cancer.

Mapping Brain Connections

In September of 2010, the NIH awarded $40 million dollars for the Human Connectome Project. The effort will create a detailed, searchable map of the neural wiring in the human brain, which is believed to include hundreds of trillions of connections.

"The overarching goal is to use cutting-edge technology to decipher as much as we can about the wiring of the brain in healthy adult humans and to probe the differences in connectivity across a large number of individuals," says **David Van Essen, PhD**, professor of anatomy and neurobiology at Washington University in St. Louis.

Van Essen is one of two principal investigators for a nine-institution consortium led by Washington University and the University of Minnesota that will acquire, analyze, and distribute the data. A smaller consortium led by Massachusetts General Hospital and the University of California, Los Angeles, will focus on alternative ways to advance the technology for imaging brain connections.

The WashU/UMinn consortium will scan the brains of 1200 healthy adults (twins and their non-twin siblings) using two types of imaging. MRI diffusion imaging traces the bundles of fibers that structurally link different regions of the brain; and functional MRI (fMRI) reveals which brain areas are linked based on functional interactions. The team will also collect extensive data on genetics and cognitive function.

"We will obtain an enormous amount of data, estimated to be about a petabyte worth," Van Essen says. (A petabyte is one quadrillion, or 10^{15}, bytes.) The data will be made publicly available, along with analysis and visualization tools. Users will be able to explore average brain connectivity, variation in brain connectivity, and the links between connectivity, cognitive function, and genetics, Van Essen says.

Eventually, they hope that future projects will enable mapping the brains of people with diseases such as autism and schizophrenia, where brain wiring goes awry, he adds.

***Brain Wiring:** Scientists from the human connectome project are using MRI diffusion imaging to visualize the structural connections in the human brain. Courtesy of: The Human Connectome Project*

Modeling Gut Immunity

In October of 2010, Virginia Tech and its collaborators received a $10.6 million grant from the National Institute of Allergy and Infectious Diseases to create and disseminate user-friendly software for modeling gut infections. The models could help scientists develop new drugs and vaccines for diarrheal diseases, a major global health problem.

"The mission of the Center for Modeling Immunity to Enteric Pathogens (MIEP) is to understand the mechanisms of action underlying immune responses to enteric [intestinal] pathogens," says principal investigator **Josep Bassaganya-Riera, PhD**, an associate professor at the Virginia Bioinformatics Institute and director of the Nutritional Immunology and Molecular Medicine Laboratory.

MIEP will design and freely distribute models and modeling tools that can be used by immunologists and infectious disease experts with minimal training, Bassaganya-Riera says. "We do not want to design models that will only be utilized by computer scientists. I think that would be a failure." The tools, COPASI (Complex Pathway Simulator) and ENISI (ENteric Immunity SImulator), will simulate specific immune responses in the gut mucosa as well as basic immunological processes, such as T cell differentiation.

Once developed, Bassaganya-Riera's team will run *in silico* experiments—such as infecting a virtual gut with diarrhea-causing *Escherichia coli* and testing various therapeutic agents. Gut pathogens cause the immune sys-

tem to overreact, leading to excessive inflammation; so Bassaganya-Riera's group is looking for drugs that can help calm this response.

The tools could eventually be extended to model chronic diseases that also involve inflammation, including obesity, diabetes, cancer, and cardiovascular disease, he says.

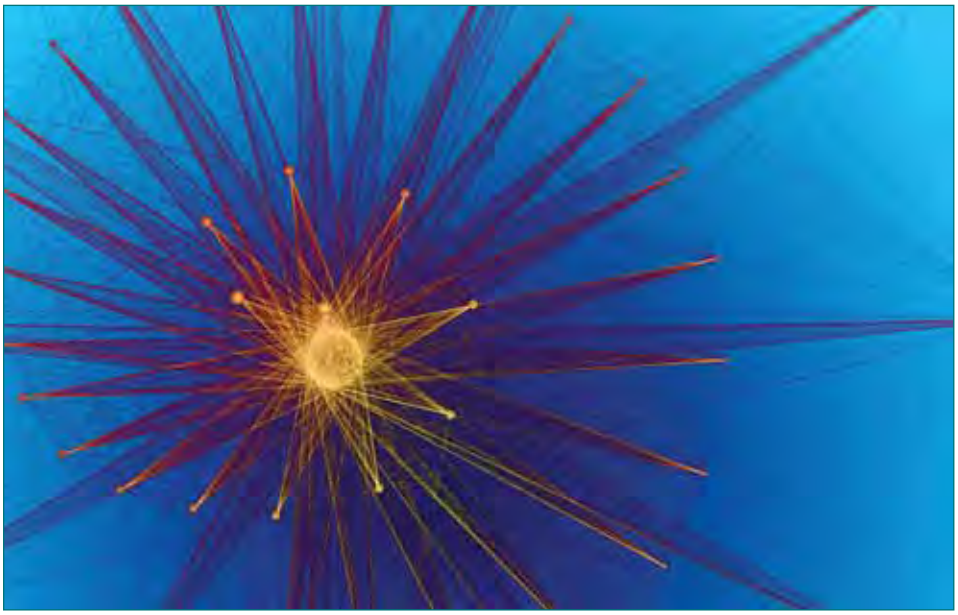
Supporting Network Biology

In October of 2010, The University of California, San Diego (UCSD), received a $6.5 million grant for a new resource center for network biologists. Genes and proteins work together in circuits and pathways; and it's these pathways that are perturbed in complex diseases, says principal investigator **Trey G. Ideker, PhD**.

"Lots and lots of data are now being produced mapping networks inside of cells," says Ideker, professor of medicine and bioengineering and chief of the division of genetics in the School of Medicine. The new center, funded by the National Center for Research Resources, will create tools for NIH-funded scientists to analyze and visualize these data. "There are quite a few centers supporting things like proteomics and genomics. But, until now, there were no centers that supported network biology. So that's our goal," he says.

Ideker's team will expand Cytoscape—their popular open-source package for network biology—in a variety of novel ways. For example, they are developing a module that allows Cytoscape to hunt for biomarkers in patient data. "Unlike other efforts we're not going to treat biomarkers as individual proteins or genes. We're treating biomarkers as networks," Ideker says. "The diagnostic of the future is going to say that some circuit or pathway module has gone awry."

"There are quite a few centers supporting things like proteomics and genomics," Ideker says. "But, until now, there were no centers that supported network biology."



***Network Biology in Action**: Picture generated by Cytoscape, a tool that network biologists can use to visualize complex molecular circuits and pathways. Courtesy of: Vuk Pavlovic and Benjamin Elliott, the University of Toronto.*

Rewiring Stress Pathways

The NIH also awarded a pair of $15.4 million grants to create two new National Centers for Systems Biology, one at UCSD and one at UCSF. UCSD researchers will explore the signaling pathways that cells use to respond and adapt to stress, such as DNA damage; and UCSF researchers will try to rewire these signaling pathways to engineer custom cells for use in medicine and biotechnology.

"Mapping out stress responses is important because there's a growing realization in biology and medicine that so many different aspects of disease and aging are interlinked by how the cell responds to stress," says Ideker, who is involved in the UCSD effort, led by principal investigator **Alexander Hoffman, PhD**. "We need a global model that links all these disease pathways together."

Cytoscape and other network biology tools developed by UCSD's National Resource for Network Biology center will play a critical role in helping the team to work out these stress pathways, Ideker says.

Once scientists understand the circuits that the cell uses to respond to stress, then they might be able to manipulate these circuits to "soup up" the cell's response. This is where researchers at the UCSF center, led by principal investigator **Wendell A. Lim, PhD**, step in. They hope to engineer synthetic circuits for therapeutic uses. For example, it might be possible to design a cell to detect and precisely kill cancer cells. □

Links to the projects:

- The Human Connectome Project: http://www.humanconnectomeproject.org/
- The Research Center for Modeling Immunity to Enteric Pathogens (MEIP): http://modelingimmunity.vbi.vt.edu/
- National Resource for Network Biology: http://www.nrnb.org/
- National Centers for Systems Biology: http://www.systemscenters.org/
- UCSF Center: http://systemsbiology.ucsf.edu/
- UCSD Center: http://sdcsb.org/

BIOSURVEILLANCE:
From Text-mining to Freakidemiology

By Katharine Miller

American officials are seeking better ways to anticipate public health crises following ten years that have seen outbreaks of SARS, avian flu, H1N1, West Nile virus, cholera and, most recently, dengue fever. There's a desire to go beyond traditional disease surveillance at local, state and regional levels and find ways to deal computationally with a fire hose of potential health data. This has led to the emergence of biosurveillance systems at the intersection of epidemiology and computational methods.

November 2010 saw the publication of a new book on the topic of biosurveillance. "It's still new to the CDC and the public health world," says **Taha Kass-Hout, MS**, one of the editors of *Biosurveillance: Methods and Case Studies* and deputy director of information science in the Division of Notifiable Diseases and Healthcare Information at the Centers for Disease Control and Prevention (CDC). "A lot of things are happening now and happening fast. It's like you're flying the plane while you're building it."

Indeed, there's so much going on, it's hard to know where to start. Researchers are expanding the types of data that can be used to predict infectious disease spread; developing novel ways to analyze that data; and trying to create systems that can help address public health problems today. The approaches range from straightforward to fairly outlandish. **Philip Polgreen, MD**, assistant professor of medicine at the University of Iowa, says he tries to be different. His infectious disease work uses futures markets, data from social media sites, and iPhone apps. "I call it freakidemiology," Polgreen says.

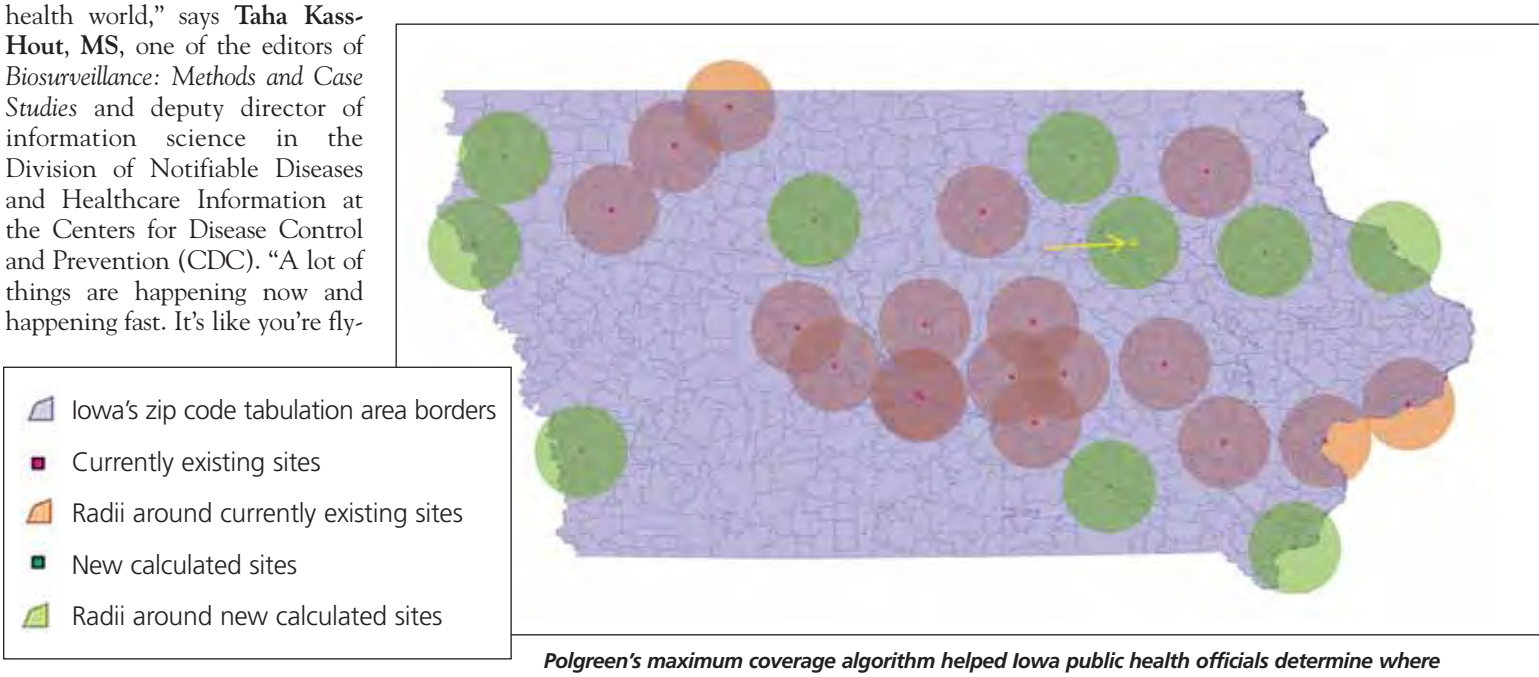
Freakish or not, biosurveillance currently provides an exciting and active niche for computational biology, with the potential to impact human health.

Expanding on Traditional Inputs

The CDC's flagship program, called BioSense, is currently computationally straightforward: It's fed with expert rules that the system uses to understand disease complaints from national, regional, and local health data sources (e.g., clinical laboratories, health departments' syndromic surveillance systems [emergency rooms, ambulatory care sites], US Department of Defense and Veterans Administration medical treatment facilities, and pharmacy chains). But according to Kass-Hout, "the next generation of biosurveillance systems is going to need to churn through massive amounts of diverse information and then present it to users in a way that can drive decisions and maintain some kind of situation awareness, either regional or national."

Some systems are already doing that by text mining online information streams to look for early warning signs of an outbreak. HealthMap, for example, mines close to 20,000 web pages to map the state of infectious diseases worldwide. BioCaster goes a little deeper, mapping layman's words for technical disease terminology in news reports in multiple languages (including several from Southeast Asia), looking to spot unusual trends, Kass-Hout says.

Kass-Hout helped develop another program, called Riff, while working at InSTEDD, a nonprofit international NGO founded by Google in 2006. This open source social networking platform scrapes information from RSS feeds or other Internet sources and uses a support vector machine behind the scenes to tag relevant items. As they are tagged, an expert gives a thumbs up or down to help the machine learn from its own mistakes.



Polgreen's maximum coverage algorithm helped Iowa public health officials determine where to put new infectious disease surveillance sites. This map shows the 20 existing sites (pink) and 10 additional sites (green) identified by the maximum coverage calculator. The radius of coverage is 20 miles. The site shown in yellow covers a population of 158,571 people. Reprinted from P.M. Polgreen, et al., Optimizing Influenza Sentinel Surveillance at the State Level,* American Journal of Epidemiology, *170 (November 2009), pp 1300–1306, by permission of Oxford University Press.

"Consensus develops and Riff does better over time," Kass-Hout says.

Social media sites also offer potentially valuable information, Polgreen says. His team used sixty million craigslist personals ads to study risk factors for sexually transmitted diseases. Because each listing is linked to a geographic location, his group can see how risky behaviors correlate with actual disease prevalence at a county level. Polgreen and his colleagues are also studying the Twitter stream to anticipate disease activity over the course of the flu season.

Remote sensing data can also help predict disease outbreaks in places where surveillance is difficult, such as in Southeast Asia. For example, researchers are using computational methods to interpret high-resolution images of environmental changes that could cause malarial outbreaks by boosting the mosquito population.

Identifying Important Signals

Regardless of data input, biosurveillance systems must identify incidents that matter. As Kass-Hout puts it, users need to be able to ask: "What's happening here that needs my decision?" To that end, he has been working with collaborators to apply change-point analysis, traditionally used in econometrics, to the problem of determining when disease activity is stable, on the rise, or going down. "An early detection algorithm can't tell you that except when it goes above a certain background level." Hopefully, he says, the approach will complement traditional detection methods by identifying which signals need to get attention. The collaborators plan to publish their results soon.

Making a Difference for Public Health

It's important to develop methods that are cutting edge but they must also be useful and understandable, Kass-Hout notes. "We don't want to build a highway to nowhere."

One system that's had some success in Illinois is called Indicator. It incorporates diverse data sources, can identify meaningful events, and has proven useful to public health officials, says **Ian Brooks, PhD**, director of the health sciences group at the National Center for Supercomputing Applications at the University of Illinois and one of the system's developers. Indicator can handle school district attendance information, hospital data, patient calls to advice nurses, and even veterinary surveillance. And it is flexible enough to allow the use of varied modeling approaches. "It's really a framework," Brooks says. So instead of committing to a particular model or way of determining whether the data is normal or aberrant, researchers can apply various algorithms to the data as it comes in, quickly creating models that are linked to that data. "You see an outbreak, and you can model how it will spread and how you would slow it down," Brooks says. During the H1N1 pandemic, he says, when school attendance data showed which schools were getting hit, the data gathered in Indicator suggested how the virus was spreading through the schools. "People then made decisions about vaccination strategies based on what we were seeing," Brooks says. "That's exciting."

> Polgreen's infectious disease work uses futures markets, data from social media sites, and iPhone apps. "I call it freakidemiology," he says.

But to be useful for biosurveillance, a computational tool can be even simpler than that. Polgreen's team has built two tools that are easy to use by various different participants in the biosurveillance enterprise. One uses a maximal coverage algorithm to help public health officials decide where to locate their outpatient surveillance systems much as a retail company decides where to locate its next outlet. The second is an iPhone app for monitoring hand hygiene in hospitals. It replaces the pen, paper and clipboard surveillance currently in place in many hospitals to check whether healthcare workers are using hand hygiene as appropriate. Polgreen says it has already been downloaded by several thousand hospitals around the country. Projects like these, he says, provide valuable experience for graduate students while also making a difference in the real world.

Screenshots from an iPhone app developed by Polgreen's team to help monitor hand hygiene in hospitals. Courtesy of Philip Polgreen.

Forecasting with Freakidemiology

Polgreen also applies ideas from economics to forecast infectious disease spread. This approach was conceived by Polgreen after meeting the creators of the Iowa Electronic Markets (IEM), which have been forecasting election outcomes by using real money to trade in political futures since 1988. Polgreen and his colleagues then launched the Iowa Electronic Health Markets (IEHM), which have been used to forecast aspects of seasonal flu, avian flu and H1N1. For example, for H1N1, the market asked what the mortality rate will be or how many states the disease will spread to in the next month. People with prior experience of the disease are invited to participate in the market. They answer the question using their personal experience in the clinic and what they know about the disease. "It's like a survey on steroids," Polgreen says. "Ordinarily, it's very difficult to quantify subjective information but prediction markets help do that."

So far, he says, markets are producing pretty good predictions of flu spread. And although futures markets won't replace traditional forms of surveillance anytime soon, Polgreen says, when people have to make decisions with very little information, perhaps a little bit of freakidemiology can help. □

COMPUTATIONAL DRUG DESIGN:
New Tricks for Old Drugs

By Kristin Sainani, PhD

When cheap drugs are needed fast, researchers and drug companies are increasingly turning to an interesting short-cut: repurposing existing drugs for new uses. Because drugs exert multiple actions in the body, the same drug may be able to treat disparate diseases.

"There's a lot of evidence now that the so-called 'magic bullet'—one-drug to specifically bind to one receptor to treat one condition—is the exception rather than the rule," says **Philip E. Bourne, PhD**, a professor of pharmacology at the University of California, San Diego (UCSD).

Just as it speeds up the time needed to identify new drugs (see "Dock This" in http://biomedicalcomputationreview.org/3/3/6.pdf), computation can speed up the time it takes to find new binding partners for old drugs. Some algorithms focus on the receptors: They virtually "dock" drugs into receptors' 3-D structures to predict how snugly they will bind. Other algorithms focus on identifying drugs that resemble one another in structure or effect, in order to determine whether they bind with the same receptors. Besides repositioning old drugs, these same *in silico* strategies can be used to predict or explain a drug's side effects. The algorithms are already yielding some surprising and promising leads.

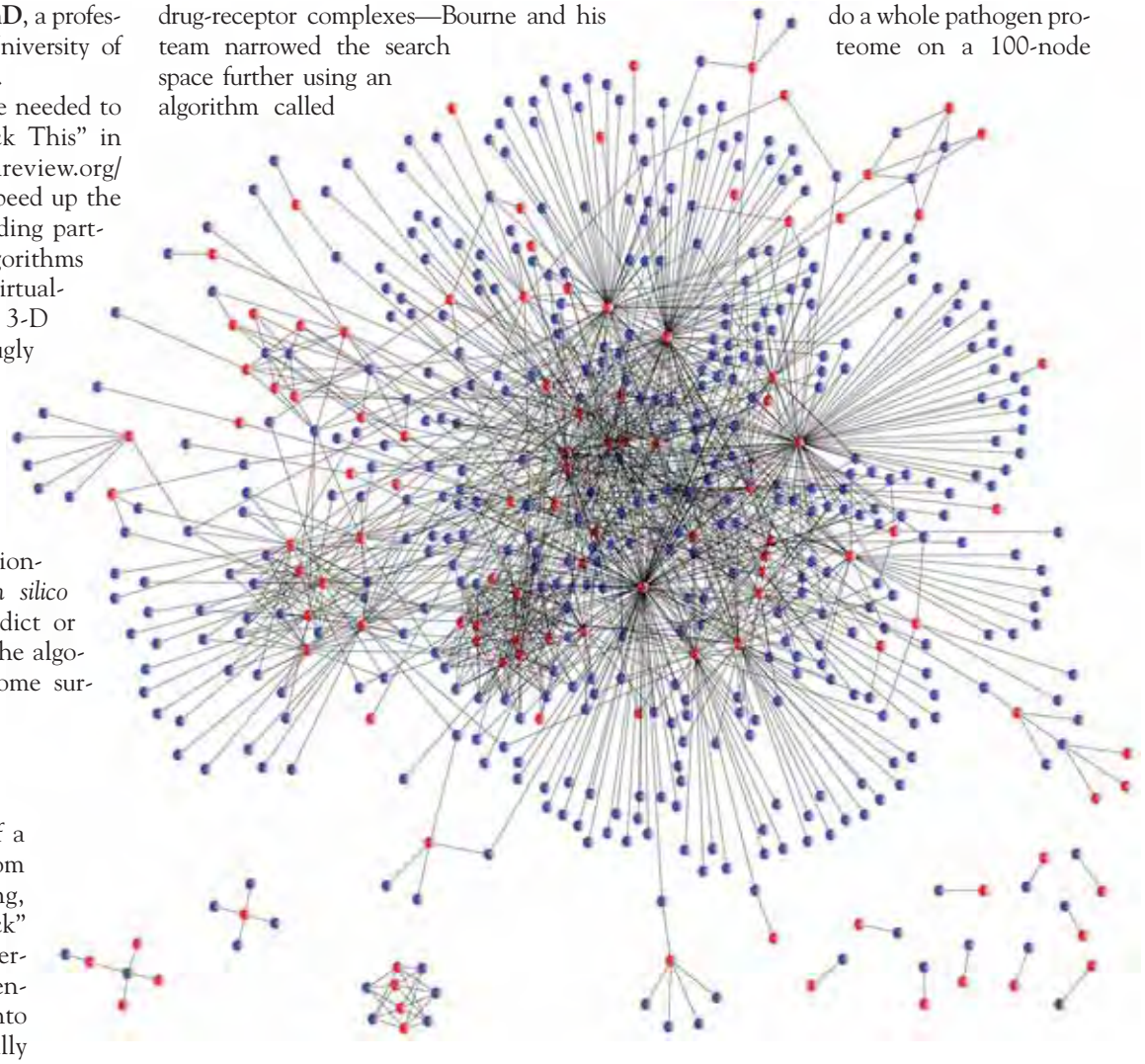
Virtual Docking: Focusing on Proteins

When the 3-D structure of a receptor is known, either from crystallography or modeling, researchers can virtually "dock" drugs into the receptor to determine if they fit, like a key opening a lock. Docking all drugs into all proteins is computationally costly, however, so Bourne and his team focused on proteins from *mycobacterium tuberculosis* (the bug that causes tuberculosis or TB) which, in many parts of the world, is still rapidly evolving resistance to all known therapies, making urgent the need for new treatments.

Bourne and his colleagues screened 274 drugs approved for human use against 1730 TB proteins with known structures (about 40 percent of the proteome). They reported their results in the November 4 issue of *PLoS Computational Biology*.

Starting with 274 approved drugs that had been co-crystallized with at least one human or animal protein—a total of 962 drug-receptor complexes—Bourne and his team narrowed the search space further using an algorithm called SOIPPA (Sequence Order Independent Profile-Profile Alignment Algorithm), which was developed by Bourne and **Lei Xie, PhD**, research scientist at UCSD. For each drug-receptor complex, SOIPPA searched for TB proteins with structurally similar binding sites. Then Bourne's team virtually docked the drug into these TB proteins to look for matches.

"There's a computer cost associated with our method, but you can do a whole pathogen proteome on a 100-node



In this network representation of the TB-drugome, red nodes are FDA approved drugs and blue nodes are binding sites on TB protein receptors. An edge is drawn between drug and receptor if the drug is believed to interact with the protein. Bourne's group hypothesizes that the highly connected drugs present the most opportunity for disrupting the normal functioning of TB. Courtesy of Philip Bourne. Reprinted from Kinnings, SL, et al., 2010, The* Mycobacterium tuberculosis *Drugome and Its Polypharmacological Implications,* PLoS Computational Biology *6(11): e1000976. doi:10.1371/journal.pcbi.1000976.

cluster in a couple of weeks," Bourne says.

They connected 123 drugs to 447 TB proteins, and validated some of the most interesting associations experimentally. "Many of these drugs have never been looked at in the TB medicinal chemistry world. No one's ever thought of trying them," Bourne says. For example, two drugs for Parkinson's disease were unexpectedly found to bind to an important TB enzyme.

Some drugs bound as many as 18 different TB proteins, which is a potential boon for preventing drug resistance. If a drug only disrupts one protein, TB can easily develop a mutation that escapes the drug, Bourne says. "So what you really want is something that's going to bind several sites and disrupt multiple pathways."

The findings also greatly expand the drug-target space for TB—previously only nine TB proteins had been investigated as potential therapeutic targets, Bourne says.

And the algorithm can be used to discover potential drugs for other diseases as well as to predict drug side effects. For example, in an earlier study, the researchers identified secondary binding sites for the breast cancer drug tamoxifen that help explain why it can cause cardiac abnormalities and blood clots.

Guilt by Association: Focusing on Ligands

Protein-centric methods are limited by the need for 3-D protein structures, says **Michael J. Keiser, PhD**. Keiser is President & COO of SeaChange Pharmaceuticals, Inc, a company that is using drug repurposing to search for new treatment options for orphan diseases (among other goals). His team builds structure-free profiles of the receptors. "We forget everything we know about the targets except one single thing: what its known ligands are," Keiser says. Using an algorithm called Similarity Ensemble Approach, or SEA, these ligands are combined into a composite based on the similarities of their chemical structures.

In a 2009 paper in *Nature*, his team screened 3665 drugs (FDA-approved or investigational) against a panel of 1400 human protein targets looking for novel matches. They compared each drug with the ligand composite developed for each protein target. "The new idea that we brought in was to compare a drug to an entire set [of ligands] rather than on a one-by-one basis," Keiser says. They identified 6928 drug-target pairs that were statistically likely to bind. When they tested 30 previously unknown associations experimentally, 23 bound with high affinity.

"There are quite a few examples that we're very excited about," Keiser says. For example, Doralese—an antihypertensive drug—was predicted to bind to an off-target receptor, dopamine D4. In experiments, it bound to this receptor 10 times more strongly than to its primary target (alpha-1 adrenergic receptor).

Another surprise: they discovered that the antidepressants Prozac and Paxil (which act on neurons) also act as weak beta blockers; they bind to the beta-1 adrenergic receptor, located in heart muscle and blood vessels. The finding may explain why some people who stop taking these drugs experience changes in heart rate and blood pressure ("SSRI discontinuation syndrome").

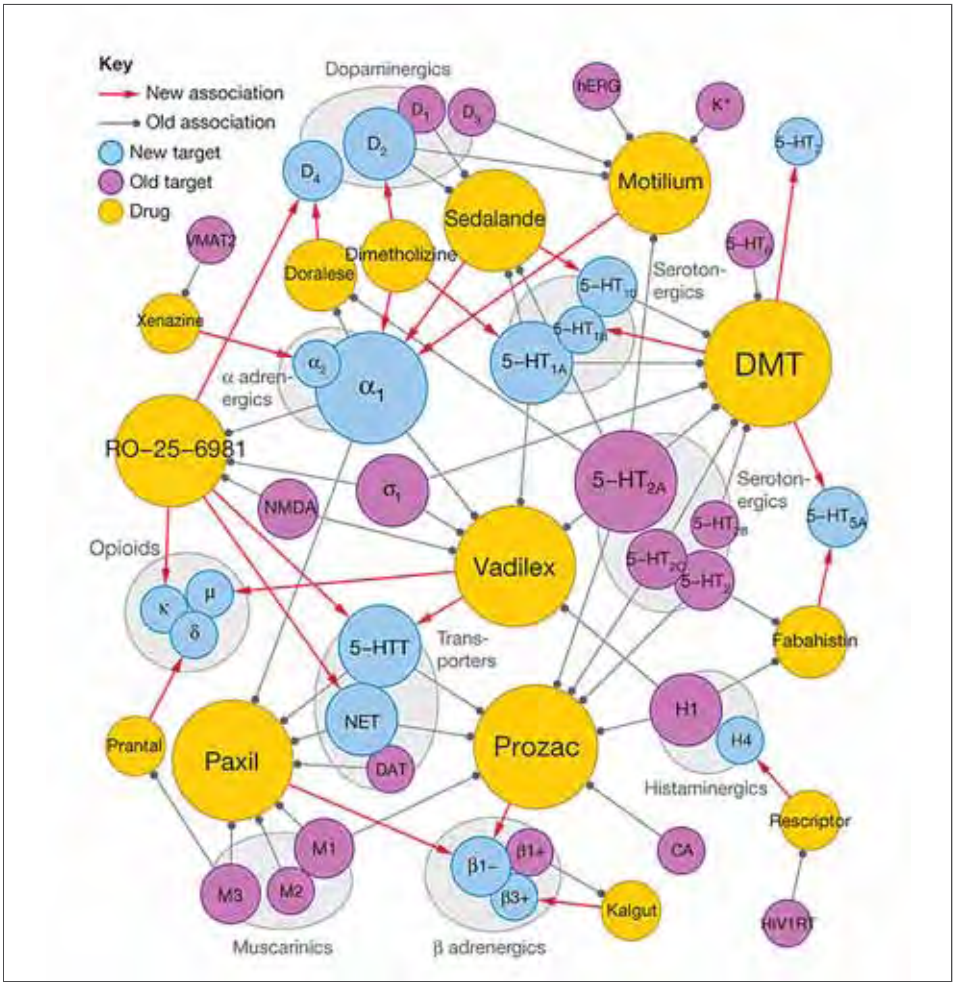
Other ligand-centric approaches connect drugs based on their phenotypic similarities, for example gene expression or side effect profiles. In a 2008 paper in *Science*, researchers used text-mining to compare 746 marketed drugs solely based on the side effects listed on their inserts. They found more than 1000 pairs of side-effect related drugs, including a couple hundred pairs that were otherwise unalike. In tests of 20 of these, they verified 13 novel drug-target interactions. For example, rabeprazole, a proton pump inhibitor used to treat ulcers, was found to bind neurologic targets, including dopamine and serotonin receptors.

Going Forward: Complementary Approaches

Each of these approaches complements the other, Keiser says. Though ligand-based approaches don't require 3-D structures, they are limited to protein targets that are already known to bind to drugs.

Finding drugs and proteins that bind is only the first step in drug repurposing. Researchers then have to show that a given drug actually has a therapeutic effect at a reasonable dose.

Drug researchers now have thousands of new drug-target pairs to explore. If even just a few prove effective, these could provide life-saving alternatives for devastating diseases such as drug-resistant TB. □



Ignoring structures and focusing only on ligand binding, Keiser and his colleagues identified new targets (blue) to existing drugs (gold) The drugs' known targets (violet) connected to the drugs by gray lines. Node sizes increase with number of incident edges. Reprinted with permission from MacMillan Publishers Ltd, Keiser, M, et al., Predicting new molecular targets for known drugs,* Nature *(2009).

BENCH-SIDE COMPUTATION:
New Tools to Accelerate Experimental Research

By Katharine Miller

Many experimental researchers rely on computational tools to push the pace and productivity of laboratory research. It's impossible to predict what the hottest new tools will be, but this column describes a few gems that have recently caught our attention: a new computer vision tool to analyze mouse behavior captured on video; a new protein simulation video database; an Excel program that makes it easier for the average bench scientist to do his or her own bioinformatics work; and a data-mining algorithm that explores complex temporal interactions among genes.

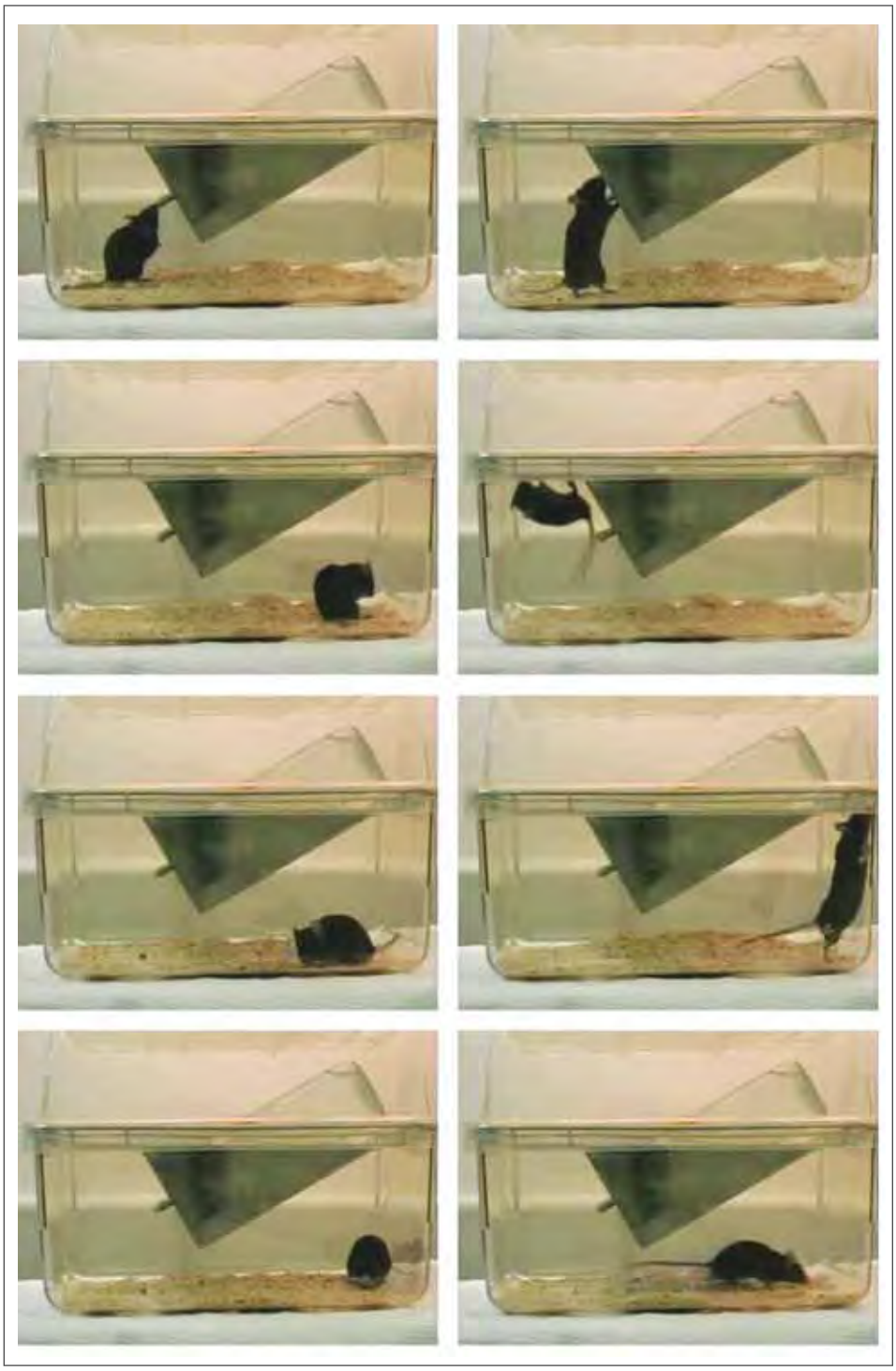
In future issues of *Biomedical Computation Review* we plan to describe other interesting new tools in this column. So send us ideas for a tool you'd like to have, a tool you're using, or a tool you've developed that you think merits coverage here. We'll stack it up against other ideas and write about those that catch our eye.

Evaluating Mouse Behavior

To conduct experiments that track mice as they eat, rest, play, and sleep, graduate students spend hours of mind-numbing time viewing video footage and categorizing actions. But a new computer model of the visual system can do that job just as well as humans. In addition to freeing researchers to do more engaging activities, it should provide more objective data and lead to more reproducible results.

"One real prospect is that you can use it over long periods of time to track the time course of disease," says **Thomas Serre, PhD**, assistant professor of cognitive and linguistic sciences at Brown University, who developed this system in collaboration with a team of colleagues at the McGovern Institute for Brain Research at Massachusetts Institute of Technology and the California Institute of Technology. "Addition of a second camera would allow computer systems to do better than a single human observer," he notes.

Serre's model of the visual system simulates what scientists know about the receptivity of neurons in various parts of the



Serre's model of the visual system was trained to identify the eight mouse behaviors shown here: Drink, eat, groom, hang, micromovement, rear, rest and walk. Reprinted by permission from MacMillan Publishers, Ltd., Jhuang, H. et al., Automated home-cage behavioural phenotyping of mice,* Nature Communications *(2010).

brain. "Neurons in some areas are sensitive to movement. Others are tuned to edges and boundaries," Serre says. "This program is trying to mimic that." The model looks at shapes and directions of motion and learns what combinations constitute certain behaviors. Using video sequences (essentially pixels changing in intensity over time), it performed as well as people in identifying eight standard mouse behaviors. His team is now extending the system to watch social behaviors of animals housed in groups. The work was published in *Nature Communications* in September, 2010 and available online at : http://serre-lab.clps.brown.edu/projects/mouse_behavior/index.html.

"Vision is far from a solved problem," Serre says. "But little by little we are getting there."

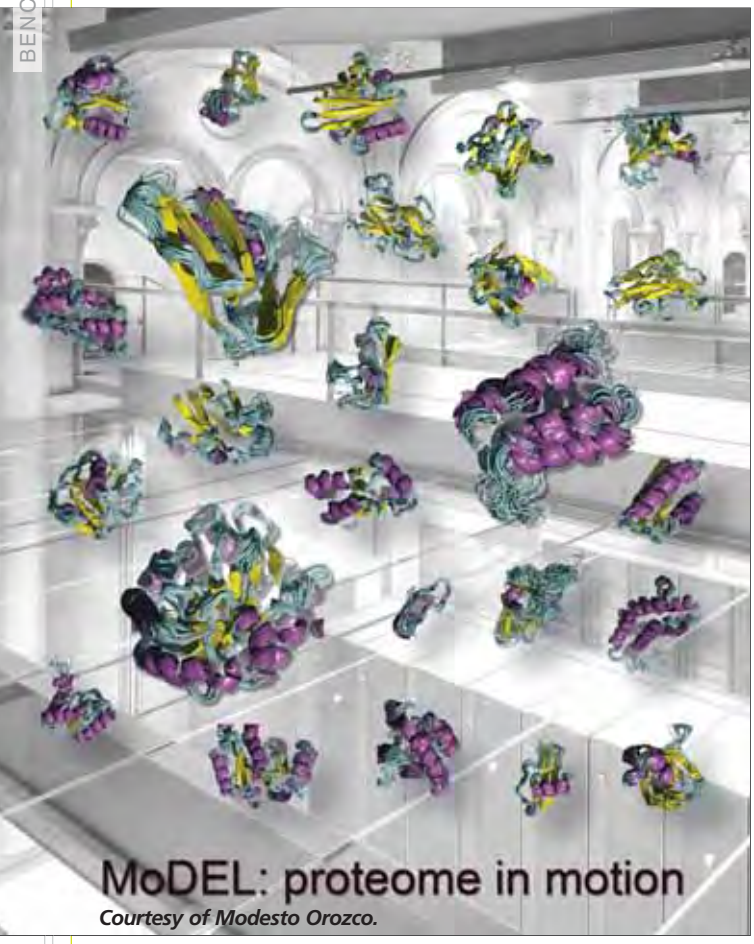
Videos of Protein Dynamics

Proteins are machines that move. Nevertheless, the pharmaceutical industry uses static protein structures when doing rational drug design. "Everybody knows that is a huge limitation," says **Modesto Orozco, PhD**, professor of biochemistry and molecular biology at the Institute for Research in Biomedicine in Barcelona, Spain. So he and his colleagues assembled a large database of proteins in motion—including proteins that are pharmaceutical targets, such as kinases and membrane proteins.

Traditionally, the pharmaceutical industry virtually screens millions of compounds against a single static structure to identify perhaps 100 compounds for testing while missing 1000 that might bind to the structure in a different configuration, Orozco says. "They live with that." But when people access the simulation results in Orozco's Molecular Dynamics Extended Library (MoDEL), they can dock compounds with 10,000 structures instead of one. "This increases the possibility of detecting potential ligands," Orozco says.

Running the simulations to build MoDEL took almost four years and nearly several hundred years of CPU time with jobs running in parallel. "It was very computationally intensive," Orozco says.

Orozco estimates that MoDEL's simulations will only increase the probability of finding a successful drug by 5 to 10 percent (since drug development is complicated by many factors beyond ligand binding, such as toxicology and patent issues). Still, given that MoDEL is open to the community (at http://mmb.pcb.ub.es/MoDEL/), why not make use of that potentially valuable 5 to 10 percent?



MoDEL: proteome in motion

Courtesy of Modesto Orozco.

Excel Anyone?

For some biologists, hiring a bioinformatician to do analyses can be prohibitively expensive. Even the software can be out of reach—as well as hard to learn. That's why **Robin Hallett**, a third-year PhD student at McMaster University in Toronto, Canada, decided to create his gene expression analysis tool in a ubiquitous program: Excel.

"There are lots of biologists sitting on data and they don't have the knowledge or means to extract useful information from it," Hallett says. "My goal was to make something for myself using programs everyone knows how to use."

Hallett's tool takes gene expression data and uses basic statistics to identify predictive gene signatures for diseases such as cancer. To test his algorithm, Hallett used data on 295 breast cancer patients. From half the data (the training set) he identified genes whose expression levels correlated with survival. When tested on the remaining patients' data (the test set), the highly ranked genes properly segregated the breast cancer patients by prognosis. The work was published in the *Journal of Experimental & Clinical Cancer Research* in September 2010.

Hallett notes that bioinformaticians haven't been terribly impressed by his paper (it was rejected by a computer science journal), but they're not his target audience. The tool was designed for graduate students with the least access to advanced computing resources. And it's meant for the learning/discovery phase of research rather than at the clinical end, when researchers might want something more powerful.

"It works well, not excellently," he says. "But more advanced machine learning algorithms are not universally accessible like Excel. To someone unfamiliar with those algorithms, it's fine." And perhaps his work will lead to other Excel-based bioinformatics algorithms, making computational biology truly available on any desktop.

Data-Mining for Transcriptional Changes Over Time

Biologists frequently have to grapple with changes in gene expression over time. Some algorithms track how gene expression peaks and what the shape of the peak is. Others try to find oscillatory patterns (such as circadian rhythms). Still others look at gene expression over an entire life cycle to understand different stages of the organism's life. But all these options only consider changes at the level of single genes.

Naren Ramakrishnan, PhD, professor of computer science at Virginia Tech, and colleagues have taken a new approach, creating a software tool that focuses on groups of genes. Different groups of active genes organize, break up, and coalesce when the cell transitions from one stage to the next, he says. By identifying the timepoints when groups of genes dynamically reorganize, his algorithm automatically determines the stages the cell goes through. "It's very unsupervised," he says. "The algorithm is not given any information about where transitions might happen."

The software, called GOALIE, is available for use by any biologist with time course data. "They can load time series data into our software and explore transitional boundaries," Ramakrishnan says. □

COMPUTATION FOR THE BEDSIDE:
Optimizing Patient Care

By Katharine Miller

Medical decision-making is often more art than science, requiring physicians to exercise judgment in the face of complex factual circumstances. But now a few tools offer the opportunity to computationally optimize patient care. Here, we present several recent projects that have begun making a difference in patients' lives. They span a range of medical scenarios affecting those with chronic illnesses, including HIV and kidney disease, as well as those undergoing CT scans or facial bone surgery.

Optimizing AIDS Treatment Protocols

People living with HIV often develop drug-resistant forms of the virus after a period of treatment. To select effective new drug cocktails, their doctors must filter through a mass of information about the patients' viral mutations and load, past drug regimens, immune counts, and symptoms. This problem cries out for a computational solution. One such tool, HIV-TrePS, launched as a free online service in October 2010 and is already being used in over 40 countries.

Physicians can log in, input their patients' baseline parameters (about 80 variables), and they will receive a list of drug combinations and an assessment of the probability that they will reduce the virus to undetectable levels, explains **Brendan Larder, PhD**, scientific chair of the HIV Resistance Response Database Initiative (RDI), a nonprofit research group in the United Kingdom which, along with several collaborators, created HIV-TrePS.

Larder and his colleagues trained the HIV-TrePS system using data from 60,000 real patients and a computational learning method called random forests, which builds decision trees and detects patterns. During testing and validation, the tool correctly predicted whether a combination therapy will lead to response or failure approximately 80 percent of the time. And physicians changed their treatment plans about one-third of the time based on its recommendations, Larder's team has found.

In a few months, they will launch a similar system that doesn't require viral genotyping data, which is unavailable in most developing countries. "The accuracy drops to around 70 to 75 percent," Larder says, "but that's still pretty good."

Optimizing CT Images to Reduce Radiation Exposure

X-ray computed tomography (CT) scans provide useful, clear 2-D and 3-D images, but require using low doses of radiation. The question is: Can computers help radiologists get sufficiently clear images at the lowest necessary radiation dose? Already, computers can reduce radiation 20 to 30 percent by automatically adjusting to a patient's body size and the thickness of the body part being scanned. But recent research at the Mayo Clinic in Rochester, Minnesota goes further.

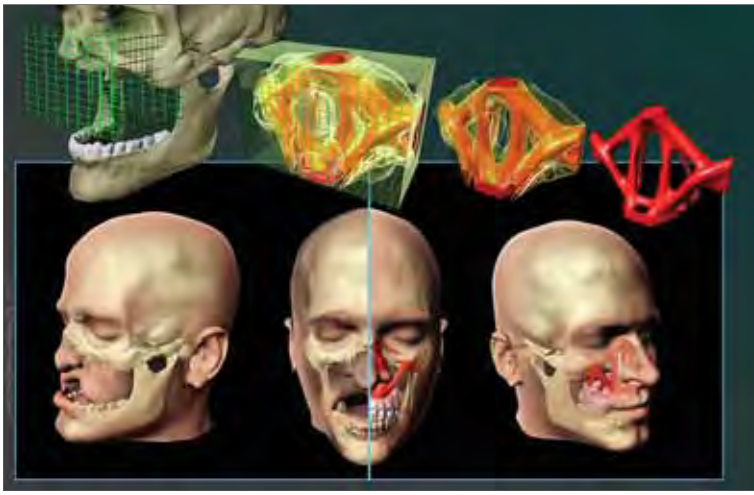
"We showed you can use almost half the radiation dose with a lower tube energy and get diagnostically acceptable images," says **Joel Fletcher, MD**, a radiologist at the Mayo Clinic. Lowering the tube energy makes the iodine contrast (which is injected into the patient) brighter, but it also makes the image noisy. To deal with that, Mayo has developed a novel computational way to de-noise the data. The work was published in *Academic Radiology* in October 2010.

Mayo's novel method—projection-space de-noising—reads how the x-ray tube current is modulating to adapt to the patient size and shape, predicts where the noise will be, and corrects for it before the data have been reconstructed into an image for interpretation. The resulting images compared favorably with routine images reconstructed using a higher radiation dose.

It may be possible to lower the dose further if radiologists can increase their tolerance for noise. "There is evidence that radiologists are just as good at looking at noisy images, even without fancy de-noising software," Fletcher says. It's a tougher case to prove that with higher noise, diagnoses are still optimal. But that may be the optimization of the future.

Optimizing Facial Bone Replacements

Patients with cancer or injuries to the face sometimes need bones replaced. Today, surgeons fashion solutions based on what has worked before in other patients, not knowing for sure if the result will be strong enough for important functions, such as chewing, which can exert up to 700 Newtons of force. In recent work, researchers at Ohio State University used topological optimization to improve the design of these bone grafts.



Beginning with a prism shape (left), Sutradhar's team's algorithm gradually refines and optimizas the shape of a facial bone replacement, ensuring that it can handle the forces imposed by chewing and other actions of the face. Courtesy of Alok Sutradhar.

"For each specific function, you can optimize the structure," says **Alok Sutradhar, PhD**, a post-doctoral researcher and lead author on the paper published in *Proceedings of the National Academy of Sciences* in July 2010.

Topological optimization does not try to mimic structure that was there before.

Starting with a rectangular prism filled with material, as well as boundary conditions that fix the height, the algorithm iteratively designs the optimal shape to withstand the loads (the direction and size of the forces) exerted by, say, chewing and swallowing. Videos accompanying the research show the gradual evolution of the optimal shape. "The algorithm is such that it will take you to the correct shape in less than an hour," Sutradhar says.

Sutradhar's group is now testing the optimized shape in a skeletal model, confirming that it does actually withstand the required forces. In the future, scientists might be able use tissue engineering to grow bones on a topologically optimized scaffold, he adds.

Nationwide Optimization of Live Kidney Donor/Recipient Matches

People who need kidney transplants often have friends and relatives who are willing to donate. They will even donate to a different person as long as their loved one gets a kidney out of the transaction. This might happen either in a "cycle" of multiple donor/recipient pairs (pair A gives to pair B who gives to pair C who gives to pair A) or in a chain (where an altruistic donor sets off a chain reaction of donation from one pair to the next and the next and so on).

Organ centers are therefore faced with the constant problem of matching multiple potential donors with multiple potential recipients. This scenario created a challenging computational problem that caught the attention of **Tuomas Sandholm**, **PhD**, professor of computer science at Carnegie Mellon University. He created an optimization algorithm that, following a successful pilot program, began being implemented nationwide in October 2010.

Unlike predecessors who have tackled the problem, Sandholm's algorithm solves the problem optimally and scales to larger populations without any simplifications. All the possible combinations of cycles would be "more than the number of atoms in the universe," Sandoholm says. "The algorithm has to prove that there are combinations that aren't worth trying. Otherwise you're dead in the water." His algorithm identifies combinations of cycles and chains that you shouldn't even try. The problem has interesting computational bits and pieces, Sandholm says. For example, he had to constrain the algorithm such that no donor gives out more than one kidney; assign weights to maximize better quality over lower quality matches; deal with the fact that computer memory was the bottleneck rather than speed; and design the rules of the exchange.

Optimizing X

The work described here suggests additional opportunities: optimizing drug protocols for hepatitis; optimizing hip replacement designs; and optimizing liver and bone marrow transplants. And there may be other low-hanging fruit out there just waiting for someone to pluck them and find amazing satisfaction. As Sandholm says, "It's very rewarding and unusual for a computer scientist to be able to save lives like this." □

LIFE IS CROWDED: Modeling the Cell's Interior

By Kristin Sainani, PhD

Molecules in cells behave like people in crowded subway cars. Because they can barely budge or stretch out without bumping into a neighbor, they move more slowly, smush themselves into more compact forms, and coalesce into aggregates more often than in a less congested setting, says **Allen Minton**, **PhD**, a physical chemist at National Institute of Diabetes and Digestive and Kidney Diseases, who coined the term "macromolecular crowding" in 1981. In addition, short distances separate crowded molecules, so they may also exert forces on one another, sometimes altering the effects of limited space.

In the past, intracellular crowding was routinely ignored in both experiments (which are typically run in uncrowded solutions) and computer models. As a result, scientists' understanding of intracellular biology might be inaccurate. But in studies during the past five or six years experimentalists have added crowding agents—complex polysaccharides that take up space—to their test tubes to get a better picture of crowding effects. And modelers are using recent gains in computational power to consider the complex interactions of hundreds or thousands of macromolecules at once. 2010 saw these computer models begin to yield surprising insights about molecular diffusion as well as protein folding and function.

> ...Modelers are using recent gains in computational power to consider the complex interactions of hundreds or thousands of macromolecules at once.

Slow going: Modeling diffusion

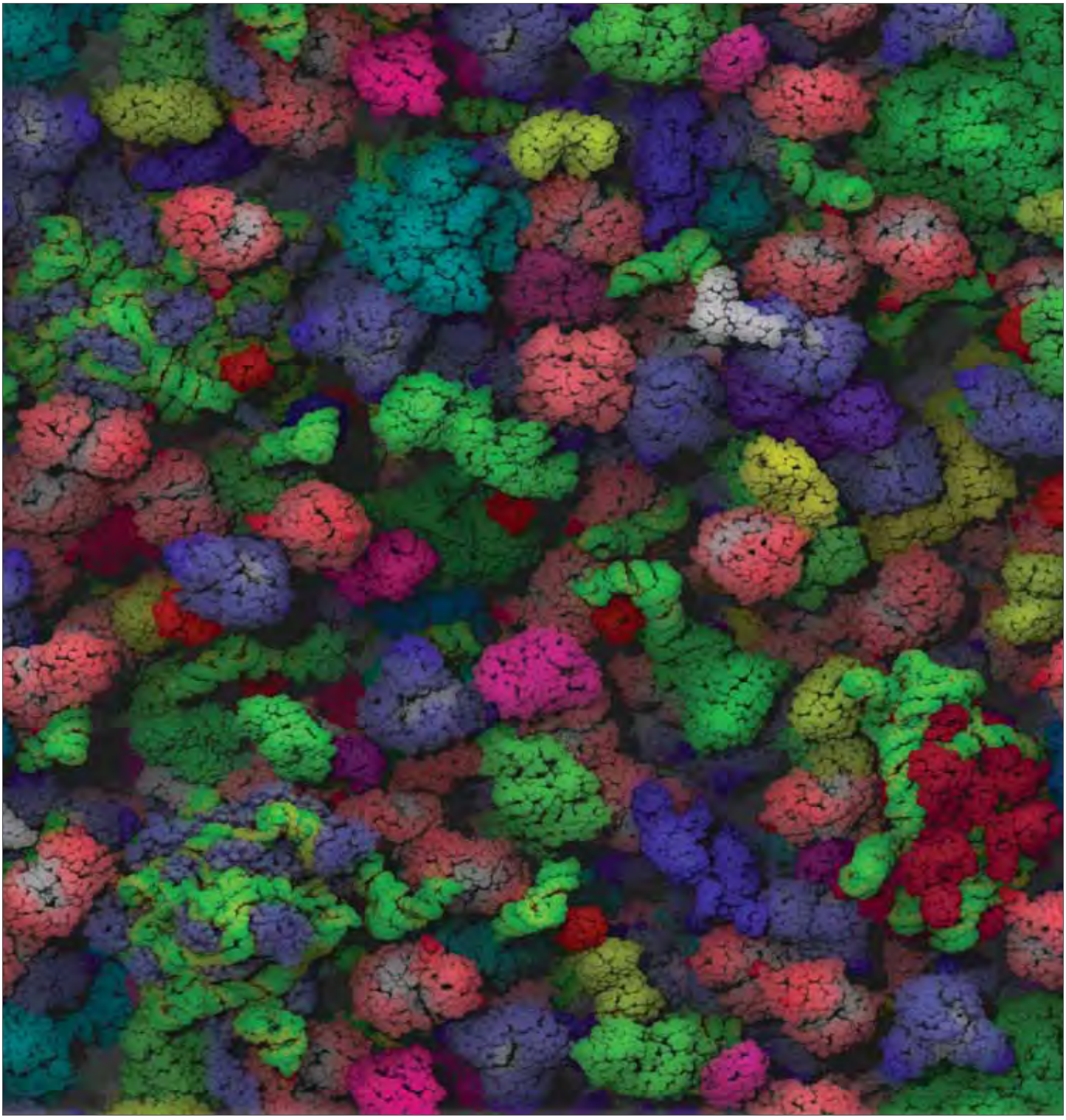
Fluorescent-tagged proteins move 10 to 15 times more slowly inside an *E. coli* bacterium than in a test tube, says **Jeffrey Skolnick**, **PhD**, professor of systems biol-

ogy and director of the Center for the Study of Systems Biology at Georgia Tech. To try to work out the exact causes of this slow down, his team ran Brownian dynamics simulations of a virtual *E. coli* packed with more than 1000 macromolecules (including 15 unique types). They reported their results in October 2010 in the *Proceedings of the National Academy of Sciences*.

Crowding alone—just molecules taking up space, or "excluding volume"—explained only about one third of the reduction in diffusion speed. But the combination of excluded volume plus hydrodynamic interactions—molecules creating wakes like sailboats in a lake—achieved the 10 to 15 percent reduction.

"If you have a whole bunch of sailboats, your behavior is going to be modified by the presence of the wakes created by all the other sailboats," Skolnick says. "In the same way, when one molecule starts to move it creates an eddy in the

The Crowded Cell: ***This picture shows an atomically detailed model of the crowded*** **E. coli** ***cytoplasm, including the 50 most abundant macromolecules. RNA is shown as green and yellow. Reprinted from: McGuffee SR, Elcock AH (2010) Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm.*** **PLoS Comput Biol** ***6(3): e1000694.***



solvent which perturbs the flow around other molecules."

Hydrodynamic interactions had largely been ignored in previous cell simulations, because they act over a long range and time frame and thus are computationally expensive to implement. "I'd rather throw it away if I could," Skolnick says. His team had to reduce the total number of molecules in the simulation to about 400 to keep it computationally tractable.

Skolnick's team also considered weak attractive interactions, such as van der Waals forces. If you make proteins sufficiently "sticky," you can slow diffusion to any speed—even zero, Skolnick says. But his team showed that these forces are much more dependent on particle size, when stickiness dominates, as compared to hydrodynamic interactions. "So it seems that crowding and hydrodynamic interactions are the dominant effects," Skolnick says.

Squished together: Modeling protein folding and function

When space is at a premium, proteins are driven to fold and compact. But accounting for crowding in simulations of protein folding takes enormous computing power.

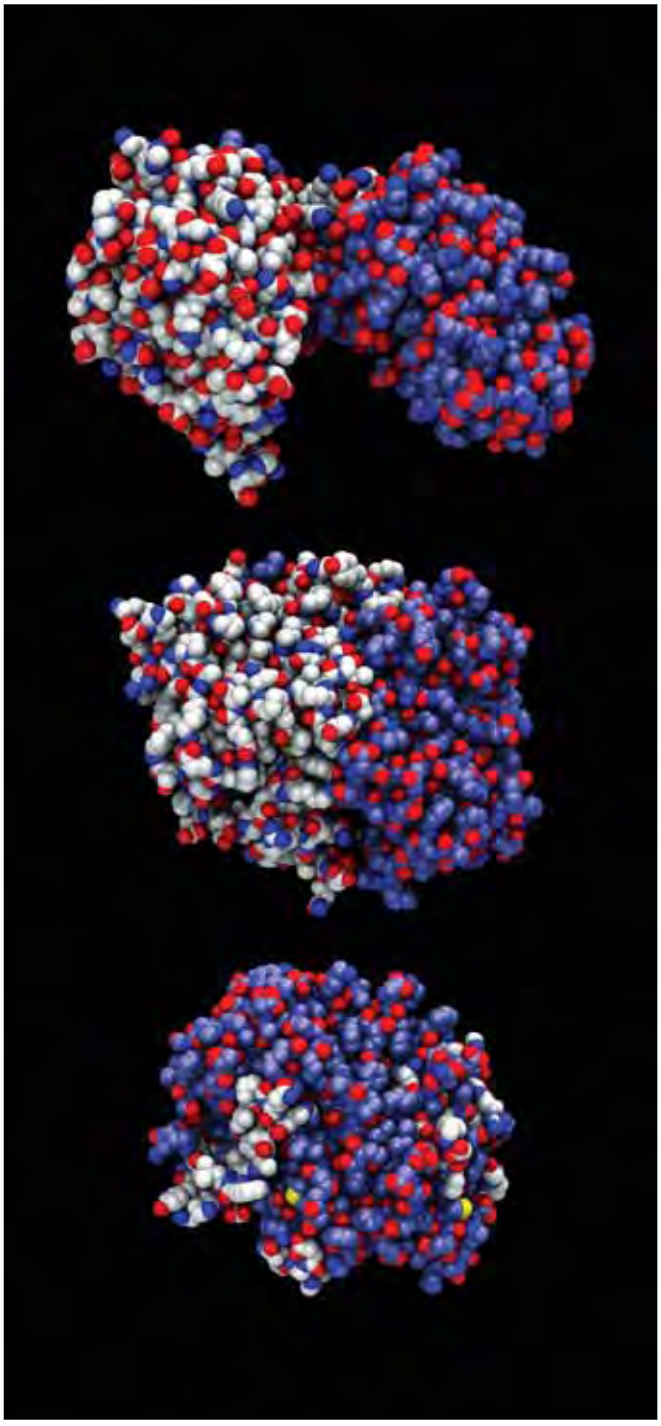
"It's a very intimidating task to think about not only just one protein, but many, many proteins," says **Margaret Shun Cheung, PhD**, assistant professor of physics at the University of Texas, Houston. In 2005, she and her mentor—**Devarajan Thirumalai, PhD**, professor of chemistry and biochemistry at the University of Maryland—published some of the first simulations of protein folding in crowded conditions.

In an October 2010 paper in *PNAS*, Cheung and her team reported the effects of crowding on PGK, an enzyme involved in glycolysis (the breakdown of sugar). In its native state, PGK is shaped like PacMan—it has two subunits where substrates bind, connected by an open hinge. Researchers thought that substrate binding caused PacMan to close his jaws, bringing the substrates together and igniting the reaction.

But using coarse-grained models, Cheung found that the enzyme actually remains in a closed, non-native state in the crowded cell (see video at: http://vimeo.com/15969373). Cheung's experimental collaborators attached fluorescent tags to PGK's two subunits and confirmed the finding.

The closed conformation keeps the binding sites near each other, allowing the substrates to bind one another quickly. PGK can therefore act 15 times faster *in vivo* than in dilute solution. "This indicates that protein function inside a cell may be very different than in a test tube," Cheung says.

Cheung, like many others, models crowding agents as simple spheres, to save computing power. But such models may miss important protein-macromolecule interactions, says **Adrian Elcock, PhD**, associate professor of biochemistry at the University of Iowa.



***Protein Compactor:** The crowded conditions in a cell cause the enzyme PGK to compact. The protein folds into three states depending on the level of crowding—from a more open conformation (top) to the most closed conformation (bottom). Courtesy of: Margaret Shun Cheung, University of Texas, Houston.*

In a March 2010 paper in *PLoS Computational Biology*, his team described an atomically detailed model of *E. coli* cytoplasm, including about 1000 instances of the 50 most abundant macromolecules. The molecules were modeled as "hollowed out" rigid shells, with atomic details only on the surface. It took a year to run the simulations.

Crowding is expected to stabilize protein folding. But when Elcock's team considered the thermodynamics of two particular *E. coli* proteins, they found that folding was actually less stable *in vivo* than *in vitro*. The reason: electrostatic and hydrophobic forces actually countered the excluded volume effects.

So just as Skolnick's work showed the importance of hydrodynamic interactions in a crowded environment, their work showed the importance of electrostatic and hydrophobic interactions.

"I think both studies are first generation models. Second generation models will have to take aspects of both," Elcock says.

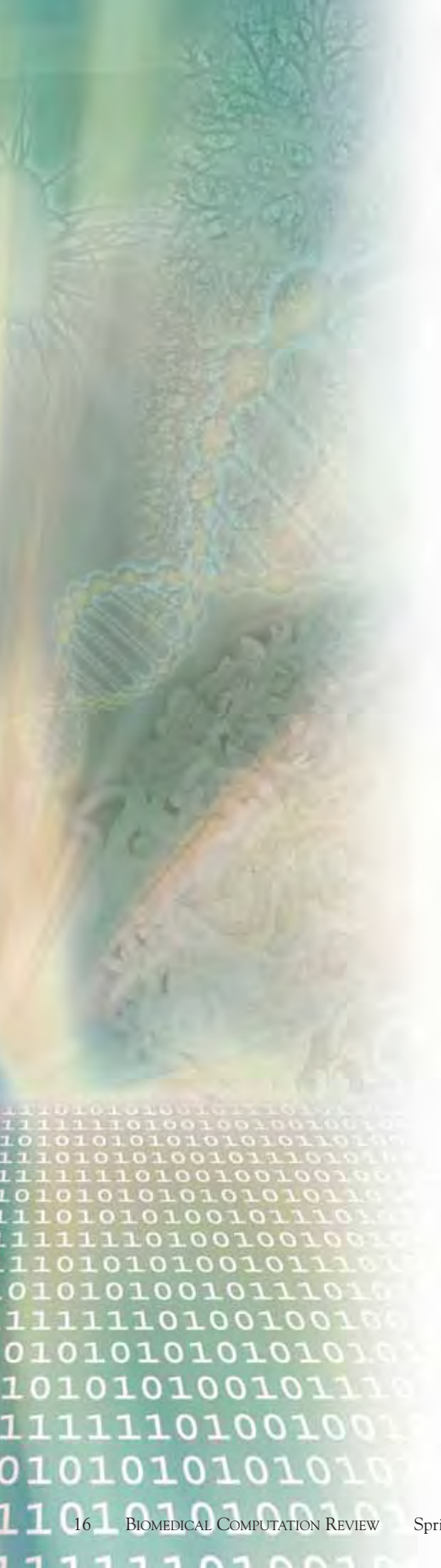
The Future: Modeling the cell and beyond

Understanding the effects of crowding on macromolecules is a necessary first step toward whole cell simulation, Skolnick concludes. "And now given the algorithms and the computational resources, it's not a preposterous question to begin to look at these things." □

By Katharine Miller

Profiles IN Computer Science Courage

Reflections on the rewards of plunging into biomedicine



To a computer scientist, the fields of biology and medicine can seem like the vast Pacific Ocean, says **Leonidas Guibas**, **PhD**, professor of computer science at Stanford University. "You go to the edge and stare out for thousands of miles. How do you know where to go in? It's scary."

And Guibas is talking about how it feels today. That vast sea must have seemed quite daunting thirty years ago when the field of computational biology barely existed. But a few pioneers from computer science saw an opportunity to bring their skills and intuitions to bear in a new arena—an arena that could impact human health while also advancing the field of computer science. So they dove in: They learned the language of biomedicine; adjusted to a different peer-review and publishing system; and successfully developed a new field.

Today, many universities offer not only graduate degrees in computational biology, but undergraduate majors as well. Yet the field of biomedicine still presents tremendous opportunities to the pure computer scientist who knows little about the area.

The people profiled here provide a sampling of those opportunities. Some were pioneers thirty years ago; others are relative newcomers. Some now dedicate their careers to biomedicine while others still maintain a computer science focus. Their skills span a variety of computational techniques including computational algorithms, imaging, knowledge representation, robotics, machine learning, and computer vision. And they are applying their skills to biomedicine's vast sea: genomic sequencing, molecular biology, phenotyping, drug design, epidemiology, neuroscience and more.

For researchers contemplating following in these scientists' footsteps, it's clear that each person must find his or her own path. Yet the stories and advice of these role models should prove reassuring: "The challenges in this space are never-ending and there's always a need for smart people to look at the data and figure out how to extract the most information from them," says **Daphne Koller**, **PhD**, professor of computer science at Stanford University.

Ron Shamir: A Gradual Transition

During the early years of his academic career, Ron Shamir had no idea that he would ever develop an interest in biology. He never took biology in school; his PhD thesis in operations research at the University of California, Berkeley in 1984 was completely theoretical; and his research after grad school focused on graph algorithms and optimization, with no biological applications. But around 1990, after presenting some research on temporal reasoning—an area of artificial intelligence—at Rutgers, an audience member (the late Gene Lawler from Berkeley) commented that it was a beautiful model for the physical mapping of DNA. The same approach—which determines whether the time periods for a group of overlapping events can be arranged to satisfy a set of constraints—could be used to study overlapping clones along the chromosome.

The comment sparked Shamir's curiosity and he started reading biology texts. His wife, a biologist, helped him with the basics, and the newly launched human genome project fanned his fascination. "It was just serendipity," he says. He counts himself lucky to have been involved when things really took off. "It was evident that there would be a need for a good deal of computing. Otherwise the human genome project wouldn't fly."

In the late 1980s, 100 percent of Shamir's work involved optimization and graph algorithms. By the mid-1990s, 50 percent of that had been replaced by biologically motivated problems, and more recently, the vast majority of his work became driven by biology. Making the shift was pretty risky, he says. "I was moving into a discipline that had no name and none of my colleagues knew what I was talking about," he says. He also had to bridge a large cultural gap and language barrier, which have both shrunk a lot since. "But I didn't make a 90-degree turn. You do it gradually and build your confidence in the field over time."

Ron Shamir, PhD, professor in the School of Computer Science, Raymond and Beverly Sackler Chair of Bioinformatics and head of Edmond J. Safra Bioinformatics Program, Tel Aviv University

Advice... ON Taking the Plunge

Find Your Passion

"Not every computer scientist will fall in love with the field like I did," says Ron Shamir. "And that's an essential part of doing research: to be fascinated, excited and enthusiastic about what you do."

It helps, Bruce Donald says, to find a great lab where people are doing computational biology that excites you. "You must develop the ability to admire the work and decide whether you'd like to do work like that."

When he was a master's student at Stanford, Michael Black told a professor he wanted to do his PhD in cognitive science and study human perception. The professor told him, "If you were my brother, I'd tell you to get a computer science PhD because you'll make more money." Black took that advice, but managed to find his way back to cognitive science through the study of computer vision. The advice he gives his students is different: "Follow your heart."

Haussler agrees: "You're limited only by your passion and commitment."

Learn Some Biology

Certainly a computer scientist who wants to work in biomedicine must learn some biology. The question is, how much? According to Shamir, "Initially, a computer scientist can pick up what he or she needs to know about a biological problem by reading chapters in one good book. To get more seriously into the field, one has to attend conferences and follow the recent literature."

Daphne Koller suggests that instead of taking introductory biology classes, which can be descriptive rather than quantitative, computer-science graduate students should start by reading a more advanced textbook and some more computationally oriented papers in the good journals. "Get a sense for the kinds of work people are doing," she says. "Find a problem that interests you and then find the background courses and reading you need."

David Haussler, PhD, professor of biomolecular engineering at the University of California, Santa Cruz

David Haussler: The Chance to Address Great Scientific Questions

David Haussler got a taste of biology when he worked in his older brother's biology lab at the University of Arizona in the 1970s and again when working with his PhD advisor, Andrzej Ehrenfeucht, in the 1980s. "He was a real polymath, interested in all aspects of science," Haussler says. It was before the field of bioinformatics really existed, Haussler says, but Ehrenfeucht led discussions of how to analyze DNA using computer algorithms.

But it wasn't until the 1990s that Haussler began using his computer expertise for biological applications. He was interested in artificial neural networks and hidden Markov models—trying to get a handle on what was learnable by a machine. Then one day, recalling his happy days with Ehrenfeucht, Haussler proposed to his postdoc, Anders Krogh (now a professor at the University of Copenhagen), that they should apply neural nets and hidden Markov models to protein and DNA sequences. "So we tried it and it worked perfectly," Haussler says. Their paper on hidden Markov models is now a mainstay of bioinformatics. "It was one of those magic moments where things took off and very rapidly we were revising the earlier work and pulling together a unified viewpoint for the field," Haussler says.

Since then, Haussler's research has gradually become completely focused on biology. After more than ten years as a professor of computer science, he became a professor of bio-molecular engineering in 2004—reflecting his shift. He now supervises both experimental and computational biological research.

What draws him to apply computer science to biomedicine? Two things, he says: First, the chance to address some of the great scientific questions. "The questions we look at are among the greatest. How did we become human? How does the cell work? How did life come to be?" he says.

Second, he says, the chance to really affect medicine. Haussler works on the cancer genomics and cancer genome atlas projects, which apply large-scale analysis to find all of the mutations in a tumor and determine which ones are driving the cancer.

Haussler also heads the Genome 10K project, which is dedicated to sequencing the genomes of 10,000 vertebrate species. The goal is to map out the evolutionary changes that produced the amazing diversity of life on this planet, he says. "The computer science challenges are nothing short of enormous. This is an incredibly exciting time to be alive."

"The questions we look at are among the greatest. How did we become human? How does the cell work? How did life come to be?"

Michael Black, PhD, professor of computer science at Brown University

Michael Black: Changing Lives with Computation

Michael Black has a longstanding interest in human perception. He contemplated a graduate degree in cognitive science (but was advised to stick with computer science because "you'll make more money") and later enjoyed hanging out with cognitive scientists at NASA/Ames while working on his PhD on optical flow estimation. Yet Black's career remained firmly rooted in computer science until he described his computer vision research to his wife's French-Canadian grandmother. According to Black, she shook her head and said, "That's all a lot of excess baggage. I've got my garden, my health, and my family. I've put away vegetables in the cellar for the winter. That's all I need." And Black thought, "She's right!"

On the flight home from Canada, Black considered whether he might be able to use his skills to help people do the most basic and important things. And he sketched out an idea for a brain-machine interface (BMI) to help paralyzed people gain back some of their independence. When initial support for these ideas evaporated at Xerox PARC where he worked at the time, he put it aside for a while. But when he landed a job at Brown University in 2000, he confided his interest in BMIs to a colleague. "I was sort of embarrassed because it sounded kind of crazy," Black says. But he was told, "that's not crazy—there's a guy here working on that." Thus was launched a successful collaboration between Black and John Donoghue, a neuroscientist at Brown.

Ten years in, Black says, patients are using the brain-machine interface systems he helped develop. As a result, he's driven less by computational elegance than by the patients' needs and what's practical for them. "It's not just a scientific question anymore; it's a usability question," he says.

Although Black still does basic computer science work, his experience with biology has changed him, he says. Computer scientists are trained to think like engineers or mathematicians rather than experimentalists, he says. "Learning to think like a biologist has made me a better computer scientist." He's also developed a drive to work on problems that could change someone's life. "I'm a little addicted to finding some of that in everything I do. It doesn't have to be a biological impact, but I want to somehow affect peoples' lives outside the academic realm."

Advice...

When Gene Myers made the leap 30 years ago, "I was lucky enough not to have to know a darn thing," he says. "I would have a conversation and do the best I could." But now, because the level of sophistication in computational biology is increasing, he thinks more is needed. "Take some biology courses or go study with somebody in the field. I think at this point that's a requirement."

Donald agrees. In the early days, he says, people felt you didn't need to know a lot of biology and biochemistry to pick a deep problem and work on it. "I'm not sure that's true anymore," he says. "I'm not sure it's good enough to learn a little."

Find Great Collaborators

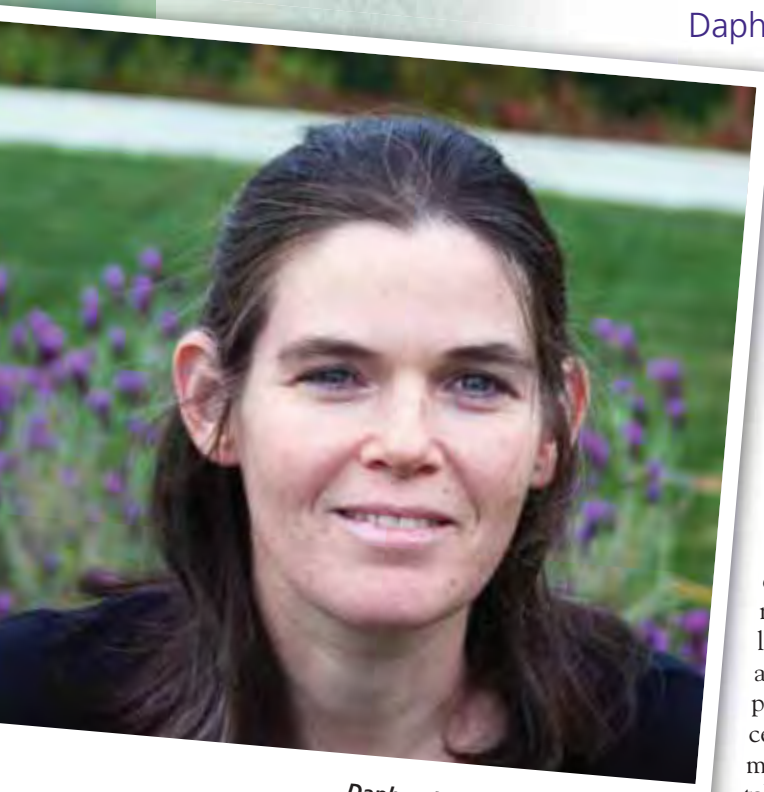
Computer scientists agree that working in biomedicine depends on personal connections with biologists you can trust. Virtually all of those profiled here say they had great collaborators early on. "You need people you can ask, 'am I doing the right thing?'" Leonidas Guibas says.

Today, because there are more biologists with a quantitative background, it's easier to find people who both understand what computer scientists can offer and speak the same language, Myers says.

But beware collaborators who have a naïve view of the computer scientist's skill set, says Black. "They might see the computer scientist as the programmer who comes in and writes some code," Black says. "Collaborators should understand that a computer science collaborator brings ideas and ways of looking at the problem and understanding the data and maybe whole new ways of thinking about what the biological system is doing." Computer scientists also shouldn't make the mistake of seeing biologists as a source of data, Black notes. "Collaborations require people to appreciate each other."

Experiment with Experiments

All of the students in David Haussler's lab have the opportunity to work in his wet lab. He doesn't expect that the computer science-oriented students will remain there, but many do a stint out of curiosity and to broaden themselves.



Daphne Koller, PhD, professor of computer science at Stanford University

Daphne Koller: Hammer Looking for a Nail (at first)

About ten years ago, Daphne Koller was working on a project to extract meaningful networks of relationships from complex heterogeneous data. She tested it on a dataset of scientific papers and authors and also on a database of movies, actors and directors, but wanted to try it on something even more complex. "I basically had a hammer and was looking for a nail," she says. And because biological datasets were rich and readily available, she decided to see if her techniques would be valuable in biological analyses. She says, "Over the course of the first few months of working on the problems, I became more interested in the nail than the hammer."

Koller's hammer was useful for studying networks both in the clinical and molecular setting. Initially, she used her tools to study networks of tuberculosis patients in San Francisco. Koller has worked partially in computational biology ever since, while still researching hard-core machine learning and other computer science problems as her mainstay.

Koller likes the fact that her biological research can have a much more direct effect on peoples' lives than can much of her computer science research. For example, she's developed a tool to evaluate a neonate's risk of developing major complications. Using only noninvasive data collected by a heart rate monitor during the first 3 hours of life, it calculates a risk score that is considerably more accurate than any other risk score previously proposed. She also developed a tool that finds pathways in cancer, the first step in identifying new drugs or personalizing cancer treatments. "I also really like the puzzle nature of trying to figure out how to take a new problem that no one has looked at computationally and thinking about how to model it, what's the right way of thinking about it, what's the right algorithmic approach. That's very satisfying."

Eran Halperin: Having an Impact

Eran Halperin began his academic career as a computer scientist working on purely mathematical problems with little regard for applications. But, while working on his PhD, he joined a bioinformatics company. It completely changed his view. Compared with designing a new theoretical algorithm, if you find a new gene or potentially new treatment, he says, "you feel the impact on society much more strongly."

> "What you learn in math and computer science is a way of thinking about a problem and how to attack it.

When he moved on to postdoctoral research, Halperin gradually changed the focus of his research to the study of applications of computer science to biology. With Halperin's background in theoretical computer science, a natural way to choose the research problems would be to look for problems that are based on computational interest. However, Halperin doesn't choose his research problem this way. "It's not the driver," he says. He chooses based on the potential impact of the project and whether his background provides some kind of advantage.

"What you learn in math and computer science is a way of thinking about a problem and how to attack it," he says. That's something he brings to the table. Halperin is perfectly willing to set his ego aside to serve the needs of biomedical research. "Everything we do is service," he says. "Eventually it serves the purpose of advancing science."

Eran Halperin, PhD, Senior Lecturer, Blavatnik School of Computer Science, and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University

Gene Myers, PhD, Group Leader,
HHMI Janelia Farm Research Campus, Ashburn, Virginia

Gene Myers: Seeking Uncharted Territory

In the early 1980s, biology was primarily a descriptive field rather than a quantitative one, Gene Myers says. "Most biologists could not compute. And this created the opportunity."

Myers was drawn to apply his computer science skills to biology—primarily in gene sequencing—because "it was a cool source of problems," he says. "I was being challenged to extract interesting variations on traditional problems. I'm a big fan of that."

He is also a fan of working in a field without much competition. "It was nice to be one of the few researchers in the field because you had a lock on a niche." In 2005, viewing the sequencing field as "crowded," and in some ways "passé," Myers made another switch—from sequencing to microscopy imaging, a relatively new niche. "It means I'm back in the 1980s," he says. "It's a really small club. I'm having a great time."

Myers believes that the most interesting computational work in molecular biology in the next 10 to 15 years will involve using microscopy to understand phenotype. "The genotype/phenotype correlation is not going to yield itself just by looking at genotype, which is what the DNA sequencing people are doing," he says. Microscopy yields rich, high-dimensional data for phenotyping, which will help researchers get the "most bang for the buck" from genomic data, he says. "So I realized that if I want to be in it, I've got to be an imaging guy."

Because computer scientists divide themselves by technique, he says, "it's very hard to get people to go from sequences to images." In addition, most academicians at his career stage are "walking this tightrope of seeking funding and managing a large group of students," he says. "They are basically supertankers, and changing the direc-

> "It was nice to be one of the few researchers in the field because you had a lock on a niche."

Advice...

"It's important to learn and understand the other person's language, concepts and worldview," Haussler says. In the end, some might find they are adept at both the pipette and the keyboard. "They can lead a complete and rich double life," Haussler says. "But not everyone has to do that to be successful in this field because we can do work in teams with people who complement each other."

Ask Lots of Questions:

"Ask the right questions and don't assume you know the biology," Paul Groth says. A little biological knowledge is a dangerous thing. "You may miss something important when helping [biologists] design new systems or designing new computational approaches to what they're doing," he says.

Be Adaptable

"If you only want to prove theorems, you will not get very far in biology," Shamir says. "You have to compromise: if you can't provide an elegant formal solution, you should be willing to sometimes work with heuristics and algorithms for which you aren't able to prove much. And you have to work with real data and interact with biomedical experts who think differently and have different goals. This requires adaptation."

Teach

"The best way to learn is to teach," Shamir says. "Teaching in a different discipline is hard, but it is also very rewarding." Preparing course materials can be a bigger commitment than just writing papers, Shamir says. That was especially true in the 1990s when he started creating bioinformatics courses and there were virtually no textbooks. "I had to create the course out of primary journal papers. It was very hard work – but it taught me a lot and was fun." And lecturing is a learning experience too. "You interact with young minds, force yourself to organize your knowledge systematically, and through the process come up with new research questions."

tion of the supertanker is hard." But Myers made the switch anyway. "I was lucky to come to Janelia Farm and really have a chance to retread myself."

When he moved to Janelia Farm, Myers says he felt like a postdoc for a few years as he got up to speed on imaging methods and developed an intuition about what techniques should work to solve computer vision problems. Now, Myers is exactly where he wants to be. He's addressing imaging problems that have requirements no one has addressed before, he says. "It's great. I'm in new territory."

Leonidas Guibas, PhD, professor of computer science at Stanford University

Leonidas Guibas: Feeding on Biology's Abstractions

Leonidas Guibas is driven to understand abstractions. He deals in the mathematics and algorithms for describing the shape and motion of things. For many years, he taught a course on geometric modeling in computer science graphics that covered only manufactured shapes such as car hoods, airplane fuselages and the like—geometric forms that people have designed. To take on the challenge of applying these same ideas to biological shapes, in 2003 he moved from Stanford University's computer science building to the Bio-X program at Stanford's Clark Center, where interdisciplinary research is encouraged.

For Guibas, biology offered the opportunity to study imprecise shapes such as protein surfaces, which have electrons floating around them. Studying proteins requires fundamentally different kinds of tools than those used to model the shape of a car or airplane, Guibas says. "That's feeding me something interesting to work on."

Guibas enjoys his interactions with biologists, but he's clear where his interests differ from theirs. "I care about computation as an object of study by itself," he says. "The biologists are interested in proteins because they are essential to life. I don't have this predilection. I study proteins as something that has geometry to it. I'm interested in something more abstract—something with shape and motion that can help me develop mathematical tools and representations that are appropriate to proteins but may also have many other uses."

"I study proteins as something that has geometry to it. I'm interested in something more abstract—something with shape and motion that can help me develop mathematical tools and representations that are appropriate to proteins but may also have many other uses."

Paul Groth, PhD, postdoc in the artificial intelligence department at the Free University in Amsterdam, the Netherlands.

Paul Groth: Biology Drives Interesting Computation

Paul Groth's career has been built around e-Science, computationally intensive science carried out over a network. "In e-Science you get really difficult computer science problems as well as really simple ones," Groth says. And many of the more complex ones come from biology. "That's the key interest for me as a computer scientist," he says.

In any kind of scientific research, it's essential to know where the data come from—i.e., their provenance—Groth says. In biology, where vast storehouses of remote data continue to grow and change, figuring out how to connect provenance information to the data themselves has become an interesting area of computer science research involving both knowledge representation (how to describe where results come from) and distributed systems (if data are coming from many different places, how do we capture and store that information?).

"Biology drove this computer science problem," Groth says, "because biologists were the first to really use publicly available data provided by web services that they didn't control and that could be updated remotely."

Often, Groth says, biologists might ask for a solution to a simple problem. "As the computer scientist, you have to ask what they would really want," he says. "You end up discovering the bigger computer science problem behind the little problem."

For example, Groth worked with a bioinformatician who wrote many different scripts but would then forget which version was the one that produced his results. "This sounds like a simple thing of being more organized," Groth says, "but in the end, the question was how to help him automatically determine what he did, which turns out to be not a simple problem." And

Advice...

Be Prepared to Slow Down

Some computer scientists find the pace of experimental science frustrating, Black says. Particularly in cognitive science, the area in which Black works, it takes time to train animals, perform required surgery, deal with governmental regulations, and obtain experimental observations. Human studies can be even more frustrating. People leave the study, patients die, "many things are out of your control," Black says. "So I've had some computer science students back away from the biology to stick with computer science."

Mind the Gap

When you work in interdisciplinary science, Guibas says, you have to decide what community you want to be part of. "There's a danger of falling in the gap between fields. Your work might be too computational for biological publications or too biologically specific for computer science publications." Computer science done for a biologist might end up in the fine print at the end of a biology paper, Guibas warns.

Be Both Bold and Careful

In computer science, Donald says, people are excited about creativity, spontaneity, innovation and boldness. "Computer science has an element of surprise," Donald says. "You're trying to make the computer do something that it couldn't obviously do before, such as redesign an enzyme to have a novel function."

Experimental scientists have a different set of values built around being careful and controlled, Donald says. They dot their i's and cross their t's. Being bold, as computer scientists are wont to do, might seem risky to them, Donald says.

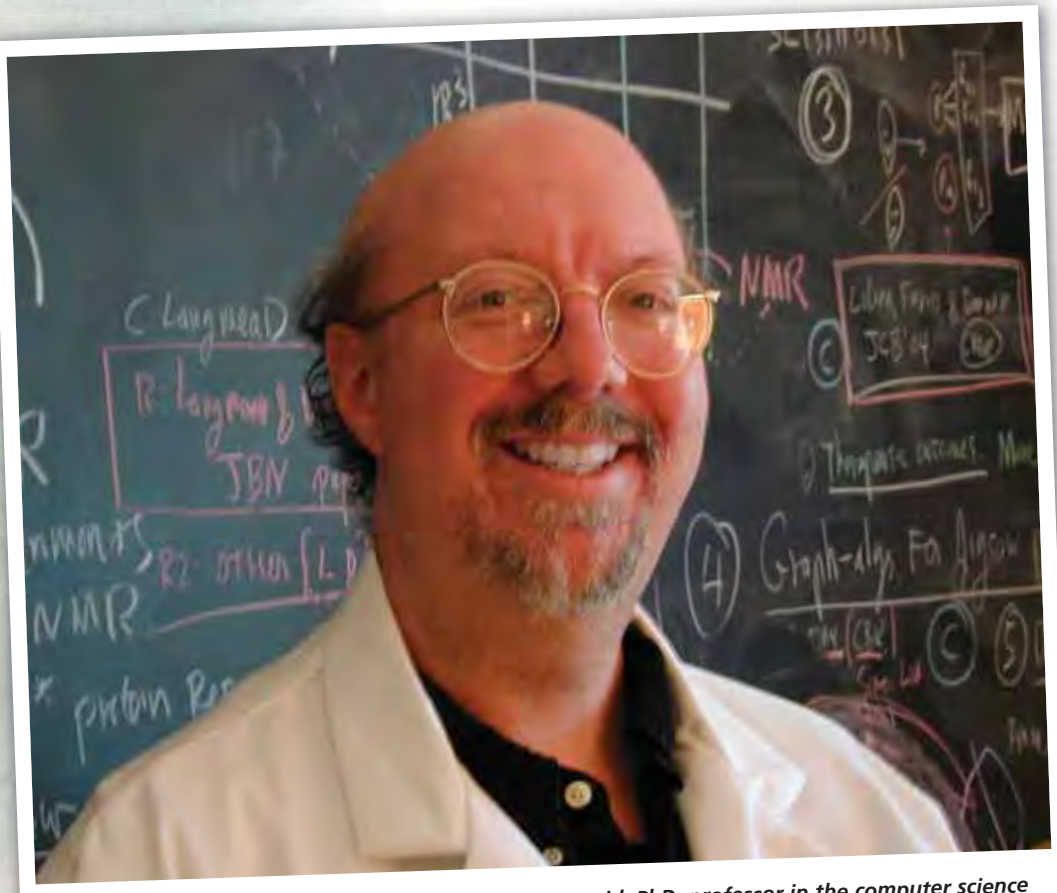
But really, the kind of care that is necessary to doing biology can be useful in computer science, Donald says. And at the same time, "the boldness and creativity of computer science in the hacker generation can be really exciting for trying new approaches in biomedical research." Computational biologists can do both: be bold and careful at the same time.

when the researcher moved from a desktop to a supercomputer, Groth had to address the more complex provenance questions raised by a supercomputer consisting of multiple machines where the mechanics could fail in many different ways. Solving this problem for one researcher led to software that could then be used by others. "The common representation we helped develop, called the open provenance model, is now becoming widely deployed," Groth says, as are several other systems developed by other groups. "Biology is a very good example of why you need this sort of provenance model."

Often, Groth says, the computer scientist has to be clear that he or she is not a programmer for the biologist. "I'm not here to design perhaps the program that helps you immediately," he says. "I might do that because I'm a nice guy and it helps me understand the collaboration. But in the end, I'm looking for the computer science research challenges that will help you eventually."

Bruce Donald: End-to-End Computational Biology

When Bruce Donald began his career in robotics in the early 1980s, he was excited about the opportunity to do what he calls "end-to-end" work. In robotics, a researcher could go from math, to algorithms, to software, to simulations, to actually making metal or silicon move.



Bruce Donald, PhD, professor in the computer science and biochemistry departments at Duke University

In 1998-99, when Donald turned his attention to using computer science for structural biology, "end-to-end" work remained his predilection. While his lab focuses on developing mathematical and highly sophisticated algorithms, they don't stop there. They will also engage in a substantial software project, implement it, and test it experimentally. That might mean, for example, performing nuclear magnetic resonance on certain proteins, developing algorithms to determine a realistic structure that captures real properties of the proteins (for example, its flexibility), predicting algorithmically how that structure would interact with a library of possible drugs, and then testing that prediction experimentally to find a drug with the desired characteristics.

To make that start-to-finish approach a reality, Donald collaborates with experimentalists. But, more recently, he gathers together the necessary experimental techniques in his own lab. "That's what I'm most excited about and proud of," he says. "A real algorithmic accomplishment is one that when applied to real data and real protein systems, really works and produces some insight in biomedical research."

And working on problems with relevance to human health has another benefit, Donald says. "I don't really have to ask myself why it's important. It's manifestly important and manifestly interesting as well." □

"A real algorithmic accomplishment is one that when applied to real data and real protein systems, really works and produces some insight in biomedical research."

Under TheHood

BY MANUEL K. RAUSCH

Continuum Mechanical Modeling of Biological Growth

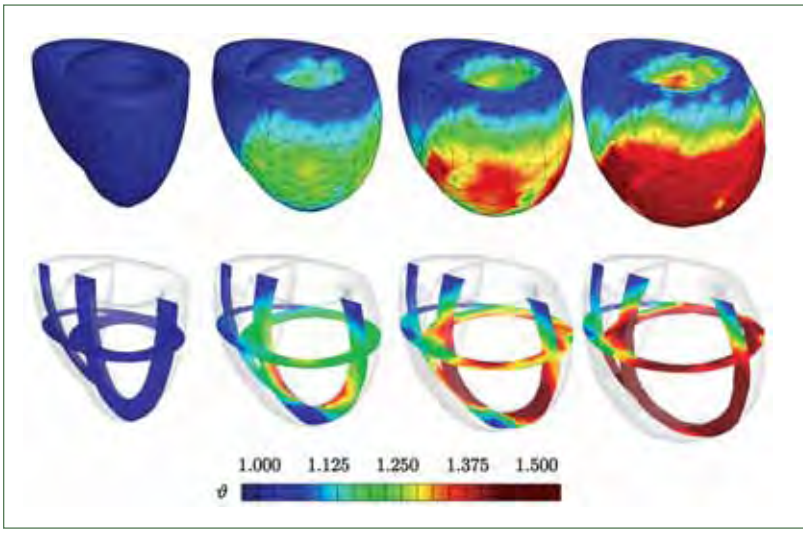
Unlike most classical engineering materials, biological tissues can adapt to external stimuli by growing in volume: Skin grows in response to wounding; muscles grow in response to exercise; cancer cells grow into tumors; and heart muscles become enlarged in response to high blood volume. To understand these adaptive processes and their role in various chronic diseases, it can be useful to study them in predictive computer models of cells, organs, organ systems and whole organisms.

As with most mechanical problems, volumetric growth can be described using continuum mechanics, a fundamental mathematical framework that describes the motion of and forces acting on a system while ignoring the discrete microscopic structure of the material under observation. Three sets of equations are used to model the system: kinematic equations, or equations of motion; constitutive equations, which model the behavior of the material; and balance laws of linear momentum, which are a generalized form of Newton's 2nd Law (force equals mass times acceleration).

But some aspects of continuum mechanics must be adjusted to address the unusual behavior of living materials. For example, unlike other materials that elastically snap back to their original shape after being stretched and released, growing tissues do not return to their original shape. The kinematic equations therefore need to account for reversible elastic and irreversible growth deformations. The precise definition of growth, and hence the exact form of this irreversible deformation, generally differs depending on tissue type and can ideally be tied to changes on the cellular or even molecular level.

For many tissues, growth may be expressed in the continuum mechanics framework as a matrix, which can be formulated in a particularly beautiful form with an intuitive interpretation of its entries. In the case of cardiac muscle growth, for example, these matrix entries can be interpreted on the molecular level as the addition of sarcomeres—the individual units that make up muscles—in series and in parallel to the existing sarcomere units. These changes result in changes on the organ level: adding sarcomeres in series results in the dilation of the cardiac muscle, while sarcomeres added in parallel results in the thickening of the cardiac muscle. Typically, the growth process is highly dynamic, and growth takes place until equilibrium between external mechanical stimuli and the growth process is reached. In its final configuration, the entries of the growth matrix will be proportional to the number of sarcomeres that were added to the tissue in series and in parallel. (Rausch MK, Dam A, Göktepe S, Abilez OJ, Kuhl E. Computational modeling of growth: Systemic and pulmonary hypertension in the heart. *Biomech Mod Mechanobio*, DOI:10.1007/s10237-010-0275-x.)

As one might imagine, the equations modeling this growth process are highly non-linear, reflecting the complex geometries and heterogeneous materials involved. As such, they typically must be solved numerically rather than analytically, which results in approximate solutions. Several numerical methods exist, such as the finite difference method (FDM) and the finite elements method (FEM). Both break down the partial differential equations describing a system into a set of algebraic equations that can then be efficiently solved using high performance computers. However, FEM has been shown to be advantageous for mechanical modeling of biological growth for a number of reasons. For instance, FEM allows us to discretize the given equations in space and time using a variety of different elements, to capture complex, ideally patient-specific, geometries. In summary, the combination of enhanced continuum theories and a powerful numerical method such as FEM now enables us to reliably predict biological growth across the different scales. This new information allows us to better understand, treat, and hopefully one day reverse the progression of pathological conditions, such as tumor growth, atherosclerosis, and heart failure. □



An idealized ventricle discretized in space and time. After spatial discretization, the model consists of 4000 elements. In this simulation, the heart undergoes growth in response to volume overload—too much blood and therefore, too much stretch in the left heart chamber. The color map denotes the amount of growth. A value of 1 (dark blue) corresponds to the normal number of sarcomeres within the heart cell. A value of 1.5 (red) indicates a 1.5-fold increase of sarcomere units.

DETAILS

Manuel K. Rausch is a PhD Student in the Mechanical Engineering department at Stanford University. He works in Ellen Kuhl's research group, studying the computational modeling of cardiac growth. To learn more about his work and the Kuhl lab, visit http://biomechanics.stanford.edu.

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

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

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

seeing science

SeeingScience

BY KATHARINE MILLER

Jackson Pollock's Protein Interaction Communities

Splashes of bold color seem to drip down the page, bringing to mind the paintings of Jackson Pollock. Spurred by the beauty of the image she had created, **Anna Lewis**,* a graduate student studying biological networks and systems biology at the University of Oxford in the United Kingdom, submitted it to the Art and Science Exhibition at the International Conference on Intelligent Systems for Molecular Biology in July 2010.

But the image is more than a pretty picture. It is actually a plot showing how functional communities in yeast protein interaction networks (horizontal axis) change in size and nature at different levels of resolution (vertical axis). Before now, the default method for community detection looked at only one horizontal slice through this image—i.e., a single level of resolution. It therefore tends to miss structures below a certain size relative to the total network size, says **Mason Porter**, **PhD**, university lecturer of applied mathematics at Oxford and one of Lewis' supervisors. Lewis' approach, Porter says, "lets you tune between the levels." If a new community is identified in this way, "that might be suggestive of further investigations one might do," he says. □

At the top of this plot, it's as though we're looking at the entire network of yeast protein interactions from a great distance, represented by the continuous solid teal color. But as we move toward the bottom of the image—getting closer and closer—various clusters resolve into smaller and smaller communities. Communities below a certain size are shown in black. Any horizontal slice through the plot provides a view of the communities in the network at a particular resolution. Work related to this image was published in* BMC Systems Biology *in 2010. Courtesy of Anna Lewis, Nick Jones, Mason Porter and Charlotte Deane, University of Oxford.

*Because Lewis is currently attempting to break a group speed record for rowing across the Atlantic Ocean she was unavailable for an interview!