DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

TOP 10 RETROSPECTIVE

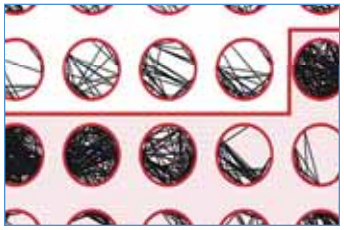
Reflections on a Decade of Biomedical Computing

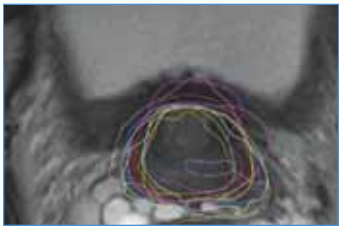
Summer 2014

ContentsSummer 2014

DEPARTMENTS

Cover and Page 23 Art: Created by Rachel Jones of Wink Design Studio using DNA art, © Anna Raspopova | Dreamstime.com.
Page 17 Art: Created by Rachel Jones of Wink Design Studio, using abstract circuit background, © Siminitzki | Dreamstime.com





Summer 2014
Volume 10, Issue 2
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
Alexander Gelfand, Esther Landuis, Katharine Miller, Kristin Sainani, Sarah C.P. Williams

Community Contributors
Megan Grove, Ruth Nussinov, PhD, Justin B. Kinney, PhD

Layout and Design
Wink Design Studio

Printing
Advanced Printing

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

Office
Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S271
Stanford, CA 94305-5444

Biomedical Computation Review
is published by:

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

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

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

GuestEditorial

BY MEGAN GROVE, MS, CGC, GENETIC COUNSELOR, STANFORD CLINICAL GENOMICS SERVICE AND STANFORD CENTER FOR INHERITED CARDIOVASCULAR DISEASE

Genomic Sequencing: Overcoming Challenges to a Bright Future

Whole genome sequencing (WGS) and whole exome sequencing (WES), which sequences only the protein-coding regions of the genome, have already begun to transform clinical medicine. They are being used to home in on the causes of rare and undiagnosed genetic diseases, determine appropriate cancer treatments for a given tumor, and match drugs and doses to an individual's genomic makeup. But as WGS takes on greater relevance in the clinic, it is increasingly important to consider the benefits and challenges of this technology.

Currently, the technology used for genome sequencing requires scientists to fragment the DNA into thousands of small pieces—"short reads"—that are then sequenced in parallel. After aligning the fragments to a human reference sequence, algorithms determine the patient's consensus sequence. Next, scientists compare the patient's DNA to the human reference sequence using a variety of computational tools that vary widely in their speed, strengths and limitations. This "variant calling" provides a list of the 3 to 3.5 million positions where individuals differ from the reference, with about 100,000 of these variants being very rare or novel.

Various aspects of each step in the process generate downstream consequences. First, short fragments can be misaligned when they exactly match more than one genomic region. Second, the ability to identify genetic variants is highly dependent on the depth of coverage, or the number of sequence reads that line up at each position in the genome. Compared with WES, WGS is generally expected to provide improved coverage of certain genomic regions, such as introns and other noncoding regions that are associated with disease risk and drug response. However, in a recent study published in *JAMA*,[1] we found that while WGS coverage is fairly high, there is still incomplete coverage of some important inherited disease genes. Finally, while variant calling algorithms often reliably identify single nucleotide variants, we and others have found that they are less consistent when it comes to identifying insertions, deletions, and larger variations (i.e., copy number variants or structural variants). This is a notable limitation, as these types of variants are often particularly important in genetic diseases.

Determining which variants matter for disease risk is also nontrivial. In the *JAMA* paper, we used automated variant annotation to help prioritize variants most likely to be impactful. We found that 50 to 100 variants per person (fewer in the undiagnosed diseases context) typically merit manual review to determine their implications for disease. Manually evaluating these candidate variants takes an average of 50 minutes of curation time. One challenge is that available information is often conflicting or limited. For instance, while specific variants may be present in variant databases, several studies have found that these databases contain high error rates, with up to 25 percent of variants incorrectly categorized as disease causing when in fact they may be common benign variants.[2]

Each of these technical, computational and interpretation challenges is currently being addressed. Advances in sequencing technology, such as long read sequencing, should allow identification of larger types of variations while also reducing errors in alignment and assembly. Incomplete coverage of specific genomic regions can be targeted with orthogonal approaches. And improved curated variant databases will greatly assist with variant assessment and the interpretation bottleneck in clinical WGS.

WGS thus has a very bright future. Clinical WES has already demonstrated a diagnostic yield of approximately 30 percent—a sensitivity higher than many routinely used genetic tests.[3] As sequencing becomes more accessible and reliable, knowledge of disease-gene relationships expand, and bioinformatics algorithms improve, our ability to interpret WGS in any context will rapidly advance. □

Note:

The information presented represents the author's own views and does not necessarily represent the views of Stanford Hospital and Clinics, Lucile Packard Children's Hospital and/or Stanford University or its affiliates.

Acknowledgments:

The author would like to thank Dr. Euan Ashley and Rachel Goldfeder for their helpful commentary on this editorial.

References:

1. Dewey, FE, Grove, MG, Pan, C, et al. Clinical interpretation and implications of whole genome sequencing. *JAMA*, 2014. 311(10):1035-1044. doi:10.1001/ jama.2014.1717

2. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med*. 2011 Jan 12;3 (65):65ra4. doi: 10.1126/scitranslmed. 3001756.

3. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;369:1502-1511.

BY KATHARINE MILLER

Curating Drugs' Potential with SWEETLEAD

Pharmaceutical research is notoriously expensive. To find safe and effective drugs cost-effectively, some researchers seek new uses for medications that have already leaped the hurdles of the FDA approval process. One systematic approach to such drug-repurposing projects involves virtual screening of molecular structures to identify compounds likely to have a particular desired effect. But these efforts have uncovered a problem: "Different databases give different chemical structures for the same drug name," says **Paul Novick**, **PhD**, who recently completed his doctorate in Vijay Pande's lab at Stanford University.

Novick decided to address that problem by creating an algorithm that automatically evaluates the structures of existing medications in various public databases. The curated database he created, called SWEETLEAD, is described in the November 2013 issue of *PLoS One*.

Virtual screening is very sensitive to precise structural information, Novick says. "A good compound might rank really low and a crappy molecule with a wrong structure might score highly," he says. "Missing out on potential active compounds is a big concern."

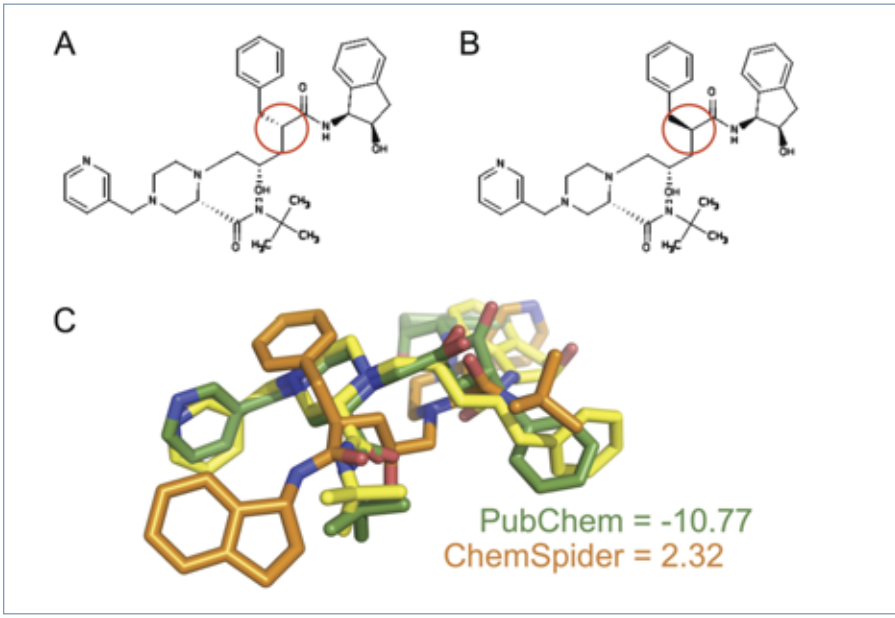
Unfortunately, the patent literature and regulatory documents that describe the molecular structure of existing medications are not currently available in downloadable form, Novick says. And reviewing that literature to re-enter all of the compounds manually would be both tedious and potentially ineffective. "It opens you up to the same kinds of errors that led to the original problem," Novick points out. "And as new drugs are approved, you want an automatic system for inclusion."

To create the SWEETLEAD database, Novick and his colleagues started by querying multiple databases (PubChem, ChemSpider, DrugBank, and others) for the chemical IDs that match a particular drug or herbal isolate's name. The algorithm then compares the structures for those IDs to see if there is a majority or consensus structure. If yes, then SWEETLEAD tags the name to that structure. For drugs with no majority or consensus structure, Novick manually reviewed the patent literature and then tagged the accurate structure.

Novick concedes that there is no *de facto* reason to trust majority structures except that they are well used by researchers who are highly motivated to correct errors. But as a final check on SWEETLEAD's accuracy, Novick compared the structures to a several other databases. "Where there were discrepancies, our structures were accurate more often than theirs," he says.

The SWEETLEAD database includes 3,600 molecules, including 2,000 approved drugs, many recreational drugs, and numerous chemical isolates from traditional and herbal medicines. "These represent a good starting point for further study by anyone doing repurposing projects," Novick says.

In addition, Novick says, SWEETLEAD can be used to explore commonalities among approved drugs. For example, the database can be used to challenge the rules-of-thumb (such as Lapinski's rule of five) that many pharmaceutical researchers use to define whether a molecule is drug-like or not. "Researchers frequently ignore compounds that violate these rules, missing out on potentially active compounds," Novick says.



The structure of indinavir, a protease inhibitor approved for treatment of HIV and AIDS, exhibits different stereochemistry (red circles) in PubChem (A) compared to ChemSpider (B). The PubChem structure was correct and received a high score for it's potential to inhibit HIV protease (C) while the incorrect structure from ChemSpider received a low score.

Novick has already used SWEETLEAD to identify several compounds that are a few steps away from clinical trials, including one for treating Chagas disease and another for Dengue fever. He's hopeful they will be effective at the same dose for which they are already approved, which would allow them to skip Phase I clinical trials.

But even if these efforts don't pan out, Novick says, "From a drug discovery perspective, any compound from our database identified as a drug candidate would definitely be a sweet lead." □

DETAILS

SWEETLEAD is publicly available at simtk.org/home/sweetlead.

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

DRILLING FOR INSIGHT:
NIH Funding for Biocomputing

By Katharine Miller

Philip Bourne's recent appointment as Associate Director for Data Science at the National Institutes of Health (NIH) signals the growing importance of bioinformatics and biomedical computing in achieving the NIH mission. Yet the NIH Institutes and Centers don't have reliable information about how much they spend on computational science. For fiscal year 2011, for example, NITRD (the Networking and Information Technology Research and Development program), reported that the NIH invested $551 million in computational science. But that report focused heavily on information technology and "high-end computing," which does not completely or accurately cover the world of scientific computing, says **Peter Lyster, PhD**, program director in the Division of Biomedical Technology, Bioinformatics and Computational Biology at the NIH's National Institute of General Medical Sciences (NIGMS).

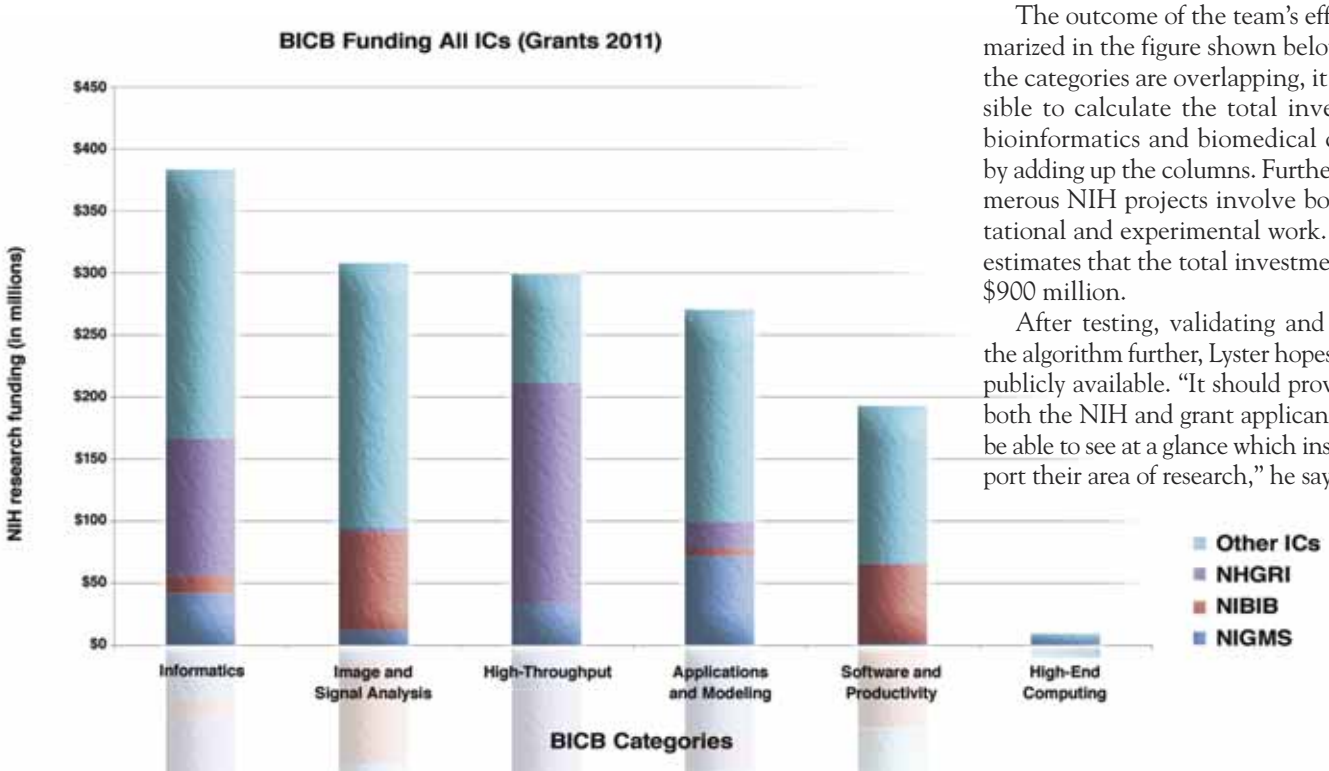
"We need a more nuanced classification," Lyster says. So a few years ago, he decided to create just that. "The main goal is to get a quantitative handle on what NIH invests in bioinformatics and biomedical computing so that we can convey this information to the public and do a good job of planning future expenditures," he says.

It is impossible to manually review thousands of annual grants to determine which ones involve computational work. "It has to be done automatically, using an algorithm that's clever enough to get around the fact that words like 'model' have different meanings in different areas of biomedical research," Lyster says.

In collaboration with Calvin Johnson and William Lau at the NIH Center for Information Technology, Lyster developed and fine-tuned a support vector machine (SVM) approach to cataloging the NIH expenditures in various subfields of bioinformatics and biomedical computing. They started by categorizing computational science into six sub-areas that are in line with NIH priorities: applications and modeling, informatics, high-throughput data-intensive scientific methods (such as next-generation sequencing, proteomics), imaging and signal analysis, high-end computing, and software and productivity. Lyster then used his expert knowledge of the field to identify a training set of about 1500 NIH projects across these areas. After training the SVM algorithm on biomedical concepts and key phrases extracted from Lyster's set of identified projects, the algorithm retrieved additional projects from the entire NIH research portfolio relevant to the six categories. Lyster reviewed a sampling of the results to confirm that the algorithm returns good hits.

The outcome of the team's effort is summarized in the figure shown below. Because the categories are overlapping, it is not possible to calculate the total investment in bioinformatics and biomedical computing by adding up the columns. Furthermore, numerous NIH projects involve both computational and experimental work. But Lyster estimates that the total investment exceeds $900 million.

After testing, validating and hardening the algorithm further, Lyster hopes to make it publicly available. "It should prove useful to both the NIH and grant applicants who will be able to see at a glance which institutes support their area of research," he says. □



Fiscal year 2011 funding for extramural (outside NIH) research into bioinformatics and biomedical computing (BICB) is shown separately for three institutes—NIGMS, the National Human Genome Research Institute (NHGRI), and the National Institute of Biomedical Imaging and Bioengineering (NIBIB)—as well as all the other NIH institutes and centers (ICs) combined. NIGMS funds a broad portfolio of computation research across all categories, including a particular focus on applications and modeling. NHGRI, on the other hand, funds quite a lot of research under informatics and high-throughput computing, which is consistent with its mission to fund basic research in genomics. And NIBIB, which has a mission that encompasses bioengineering and bioimaging, funds a significant amount of research in imaging and signal analysis.

SEIZURES, IN THEORY:
Computational Neuroscience and Epilepsy

By Katharine Miller

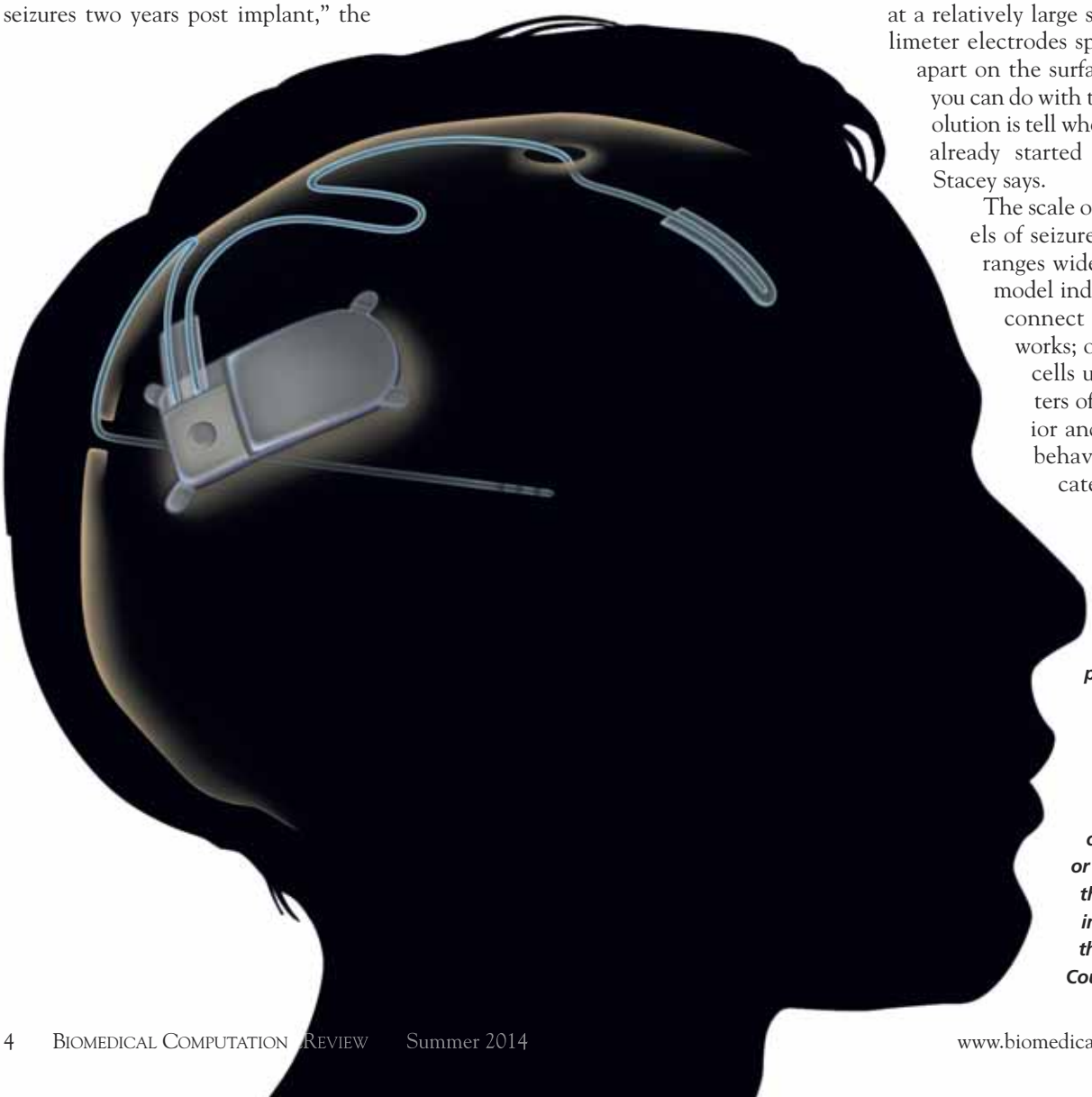
With the headline "Easing Epilepsy With Battery Power," the *New York Times* on March 24, 2014, described an implantable device for controlling epileptic seizures in patients who do not respond to medication. Developed by NeuroPace and recently approved by the FDA, the RNS® System is trained to recognize an individual patient's seizure pattern and then deliver electrical stimulation to stop seizures before they can take off.

For some patients, the device is a godsend, yet it works for only a subset of patients and even for those, its effectiveness is limited: "Fifty-five percent of patients experienced a 50 percent or greater reduction in seizures two years post implant," the company's press release declared, and most will continue to take medication.While the NeuroPace RNS® System could certainly be considered a victory for computation (it uses machine learning and could benefit an estimated 400,000 Americans), there's no question that better treatments are still needed. In recent years, even as medicines and surgical techniques have reduced seizure frequency for roughly 80 percent of patients with epilepsy, many people remain treatment-resistant.

During a seizure, voltage activity in the brain becomes synchronous. Interconnected neurons go from a state of independent processing to being connected in a massive cascade, says **William Stacey, MD, PhD**, assistant professor of neurology and biomedical engineering at the University of Michigan. It's what engineers would call a feed forward loop: Because one neuron fires, another one does until they are all firing together. "What makes a system in its normal behavior suddenly go into this self-sustaining avalanche?" Stacey asks. It's a question that has long puzzled clinicians and researchers alike.

Whether computational approaches can provide a helpful answer will require a bridging of the gap between the scales of clinical and computational research, Stacey says. Clinicians measure electrical activity at a relatively large scale—using four-millimeter electrodes spaced one centimeter apart on the surface of the brain. "All you can do with that type of spatial resolution is tell when an area of brain has already started to have a seizure," Stacey says.

The scale of computational models of seizure, on the other hand, ranges widely. Some researchers model individual cells and then connect them into small networks; others describe similar cells using lumped parameters of their average behavior and then simulate their behavior to see if it replicates reality; still others create mathematical models of dynamic networks across



Neuropace recently announced FDA approval of its RNS® System for detecting seizures and delivering deep brain stimulation (DBS) to stop them. The device is implanted in the cranium with either one or two leads for detecting the seizure and providing neurostimulation to the targeted brain areas. Courtesy of NeuroPace.

the entire brain.

Many of these models are difficult to validate experimentally. That's because there's currently no way to know if the connections in a physiological model are accurate and it's not possible to measure the network dynamics across the entire human brain, Stacey says. But that is changing. "We stand at the cusp of a very rich time in unraveling the dynamics of seizures," he says. Computational models are getting bigger and brain recordings are getting smaller. "As soon as they are at the same level—and we're close—then everything on the computer can be validated and we'll be able to play with the model to produce predictions."

The Devil in the Details

Modeling individual cells and connecting them into networks to study what makes them go haywire in epilepsy is one appealing approach, Stacey says. "It's a very intriguing problem for people interested in dynamics," he says. "And it allows us to model the brain's actual physiology, though it can be difficult to validate that the neuronal connections in such models are accurate."

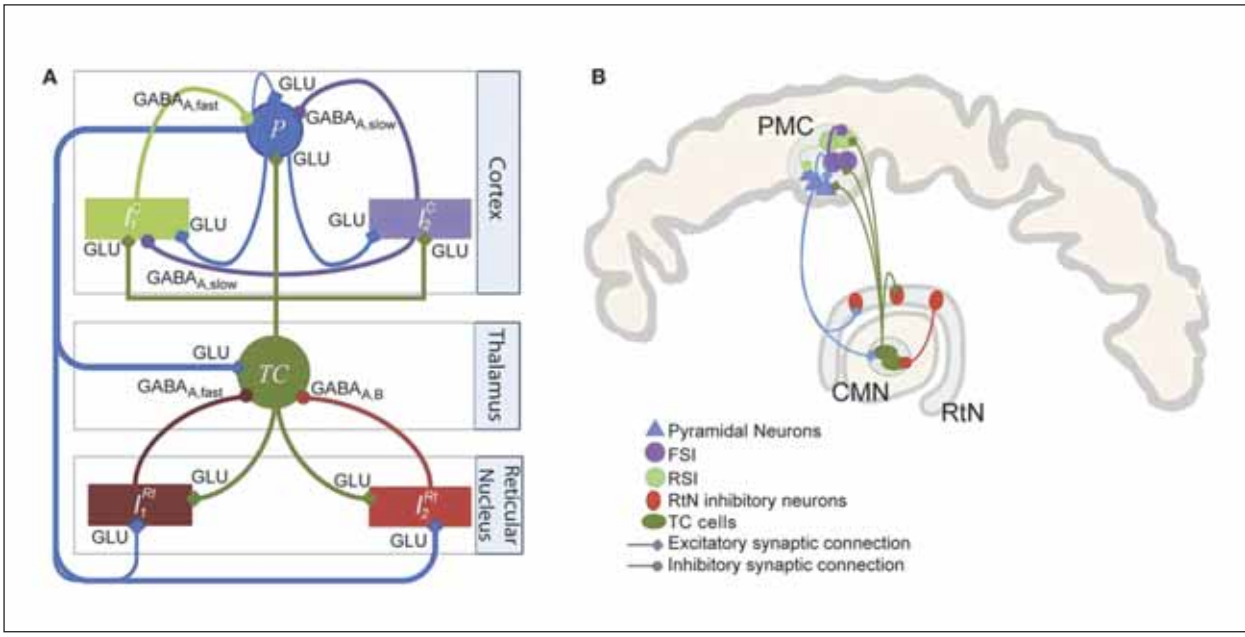
It's also very easy to make a network have a seizure using a model of a cell. In a normal brain, negative feedback keeps firing neurons from getting out of control. "It's very easy to break that feedback in a computer model," Stacey notes. "It makes you wonder why everybody doesn't have seizures."

Yet researchers who build physiologically detailed network models and simulations of epilepsy say they are valuable for generating hypotheses that get tested in the lab and then iterated back through the model. **Theoden Netoff, PhD**, associate professor of biomedical engineering at the University of Minnesota, is one such researcher. He wondered whether computer models might provide a better understanding of how and why deep brain stimulation (DBS), which is sometimes used to treat epilepsy by sending regularly scheduled electrical energy to the brain, stops or shortens some seizures but not others. The team was particularly focused on determining whether changing the frequency of DBS would shorten (or lengthen) the duration of so-called tonic-clonic seizures, in which a person first goes rigid (the tonic phase) and then starts to jerk uncontrollably (the clonic phase).

Netoff and his colleagues used a standardized computer model of an individual brain cell to build a 3,000-cell excitatory neuronal network that exhibits network statistics not unlike those in a rat visual cortex. The network is also capable of epileptic activity (it can both synchronize and desynchronize). They then added various frequency pulses of stimulation to simulate the model network's response to DBS. The result: The model predicts that DBS frequency affects the duration of the different phases of seizure in a way that is directly related to the neuron-firing rate and the level of synchronicity. For example, during the tonic phase, using a DBS frequency that matched the neuronal firing rate brought the tonic phase to a close more rapidly, while a frequency slightly below the neuronal firing rate shortened the clonic phase. Indeed, in a computer simulation, when an adaptive algorithm controlled the frequency of DBS, it was more effective in truncating seizures. Netoff is currently running experiments to test these predictions.

Lumping It

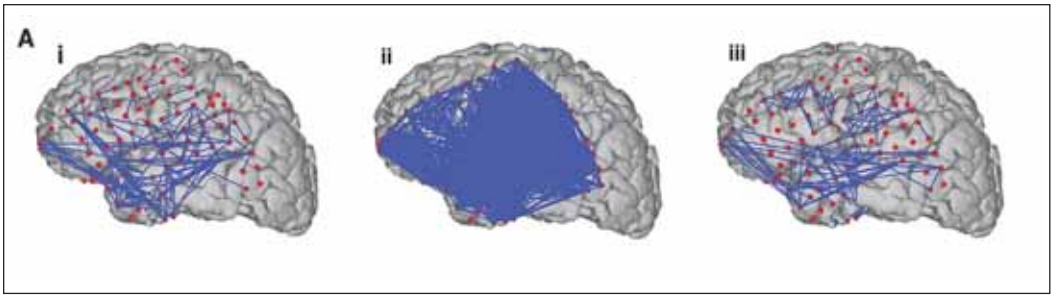
Because it is difficult to use detailed models to study the extensive brain regions involved in epilepsy, some researchers are using lumped parameter models (also known as macroscopic models or neural mass models), that use average behaviors of particular cell types. **Fabrice Wendling, PhD**, research scientist at Laboratoire Traitement du Signal et de L'Image, Université de Rennes 1, in Rennes, France, who has used this approach for some time, noticed that these models couldn't recreate one of the signatures of epilepsy: high-frequency oscillations known as fast ripples. Concerned that his macroscopic models might be missing something, Wendling set about decoding the parameters of the macroscopic model by relating them to the parameters in more detailed models. By developing a detailed model for the same system that he was modeling macroscopically, he was able to see what lay behind the macroscopic model and understand why it



Wendling uses lumped parameter models to simulate seizures in the brain. For example, in this model (A) of the thalamocortical loop, three compartments (cortical, thalamic and reticular) each contain relevant subpopulations of neurons connected in a way that is compatible with brain connectivity patterns (B) inferred from the literature (PMC = pre-motor cortex; RtN = Reticular Nucleus; CMN = centromedian nucleus of the thalamus). The model then simulates the average behavior of those regions rather than the detailed behavior of each neuron. Reprinted from Mina F, et al., Modulation of epileptic activity by deep brain stimulation: a model-based study of frequency-dependent effects,* Frontiers in Computational Neuroscience*, 7:94 (2013).

The work, which was published in *Frontiers in Neural Circuits* in February 2013, suggests that a closed-loop feedback system that can adjust DBS frequency in response to changes in the neuron-firing rate would offer greater control over seizure duration.

couldn't exhibit fast ripples. Essentially, such ripples develop in the detailed model when specific sets of pyramidal neurons are weakly synchnronized. "It makes sense that the lumped model can't see the fast ripples because it assumes the activity in each subpopulation of cells is highly synchronized," Wendling says. When the researchers increase excitability in both models, however, the same sharp epileptic spikes appear. "Once both models can generate the same type of epileptic activity (for example, epileptic spikes) then it's much easier to see which parameters at the detailed level correspond to the macroscopic parameters," Wendling says. The work was published in the *European Journal of Neuroscience* in 2012.



Kramer and his colleagues construct functional networks of brain dynamics during seizure. In part A of this graphic, the red dots represent the locations of electrodes on the brain, with blue lines showing coupled firing between various brain locations (i) just before seizure; (ii) at seizure initiation; and (iii) mid-seizure. The density of lines suggests that coupling is high at initiation but then becomes fractured in the middle of the seizure. Part B displays these same networks among 100 electrodes in one patient's brain every five seconds, with the electrodes arrayed around the edges of circles. The seizure period is shaded pink and shows coupling as the seizure starts, followed by decoupling in the middle of the seizure, and intense coupling again at the end. Reprinted from Kramer MA, et al., Coalescence and Fragmentation of Cortical Networks during Focal Seizures,* J. Neuroscience *30(30:10076-10085 (2010).

Since that time, Wendling has used his macroscopic model to help understand the relationship between DBS frequency and treatment success. For example, his team created a model of the thalamocortical network that includes multiple cell types in several compartments of the brain. They then trained the network to reproduce a particular patient's EEG recordings during seizure, and simulated various frequencies of DBS on the network. These simulations reproduced the patient's unusual and interesting response to DBS: His seizures typically stopped in response to low and high but not intermediate frequency stimulation. The work, reported in July 2013 in *Frontiers in Computational Neuroscience*, posits a possible explanation based on what happened in the model—low-frequency stimulation inhibited the feed-forward nature of the patient's seizure while high frequency stimulation inhibited thalamic output. Intermediate frequency stimulation, on the other hand, just kept the epileptic dynamics going.

Wendling says he's optimistic that DBS will prove valuable as a therapy for epilepsy once there's a better understanding of how to use it optimally. And to gain that understanding, he says both detailed and macroscopic approaches will be useful. "They are complementary and necessary," Wendling says. "What you can do with one approach you cannot do with the other and vice versa."

The Whole Enchilada

Some researchers take an even broader view of the network dynamics in epilepsy. They look at the entire system rather than one piece of it. **Mark Kramer, PhD**, assistant professor of mathematics and statistics at Boston University, for example, looks at seizure dynamics across the entire brain during the duration of the seizure. He then creates computer models to connect data to mechanisms. The goal: to help surgeons decide which part of the brain to cut out; or define optimal targets for stimulation by a device such as the one made by NeuroPace.

In work published in 2010, Kramer and his colleagues used electrocorticogram data—electrical activity measured directly on the surface of the brain's cortex—to build functional networks of the coupling and decoupling of brain areas during the course of a seizure. These networks reveal more coupling at the beginning of a seizure, less in the middle, and then more again at the end, suggesting that seizures are not simply hypersynchronous events but instead exhibit more subtle dynamics. A greater understanding of the coupling and decoupling of brain areas during seizure might suggest ways to

prevent the seizure from spreading across the brain by surgically firewalling certain connections, Kramer suggests. "Ideally, network tools could help us refine what surgeons cut out," he says. "That's one of our goals. We're not there yet."

Kramer is also interested in how seizures end. Recent research suggests that synchrony increases just before the seizure ends. "It gets more and more similar and then the brain shuts down," Kramer says. He hypothesizes that seizures end because they cross some kind of tipping point or critical transition. Moreover, perhaps when seizures keep going and going (a condition called status epilepticus), the brain's rhythmic activity tries to slow but then speeds up again, repeatedly approaching an ending but not quite making it. "What was nice about the hypothesis was that it led to specific testable measures," Kramer says. The model simulation of the tipping point theory replicated the expected brain dynamics, with the same features of rhythmic slowing, increased coupling, and flickering between seizure and non-seizure states that had been observed in functional networks during the transition. The work was published in *Proceedings of the National Academy of Sciences* (*PNAS*) in 2012.

"It's a different way to think about seizure termination, focusing on the mathematical mechanisms rather than biophysiology," Kramer says. It's possible, for example, that the mathematical constraints might help rule out other models that don't fit the predicted pattern.

Stacey took an even broader approach to the tipping point question in a recent collaboration with **Viktor Jirsa** (physics) and **Christophe Bernard** (neuroscience), both at the Université de Marseille in France. They found that seizure dynamics in any species can be described by a common set of abstract mathematical equations. They validated the equations with data from humans, monkeys, rats, mice, zebrafish, and flies. This work, to be published in the journal *Brain* in 2014 (in press), suggests that seizures are, in fact, among "the normal repertoire of brain activities," Stacey says. Moreover, they suggest that treatments should be directed toward altering dynamical properties of the brain rather than specific pathways.

A Question of Control

Some researchers are betting that work like Kramer's and Stacey's will yield a greater understanding of seizure dynamics that could eventually lead not only to better treatments for epilepsy, but even to a cure. **Paul Carney, MD**, professor of pediatric neurology at the University of Florida College of Medicine and director of the University's Center Of Excellence for Epilepsy Research and Comprehensive Pediatric Epilepsy is especially optimistic about control theory, an approach borrowed from finance, weather, and airplane cruise control or autopilot. "The airplane makes subtle adjustments as you fly," Carney says. In the brain, he says, there's also a controller that

THE PROBLEM OF PREDICTION

The NeuroPace RNS® device relies on seizure detection—spotting a seizure just as it's starting, typically only seconds before onset. By that point, Stacey says, the seizure is already underway. Prediction, along with the possibility of true prevention, has to occur sooner. "Is something already burning or is there just heat and smoke?" Stacey says. "It's a lot easier to put out a fire before the flames start."

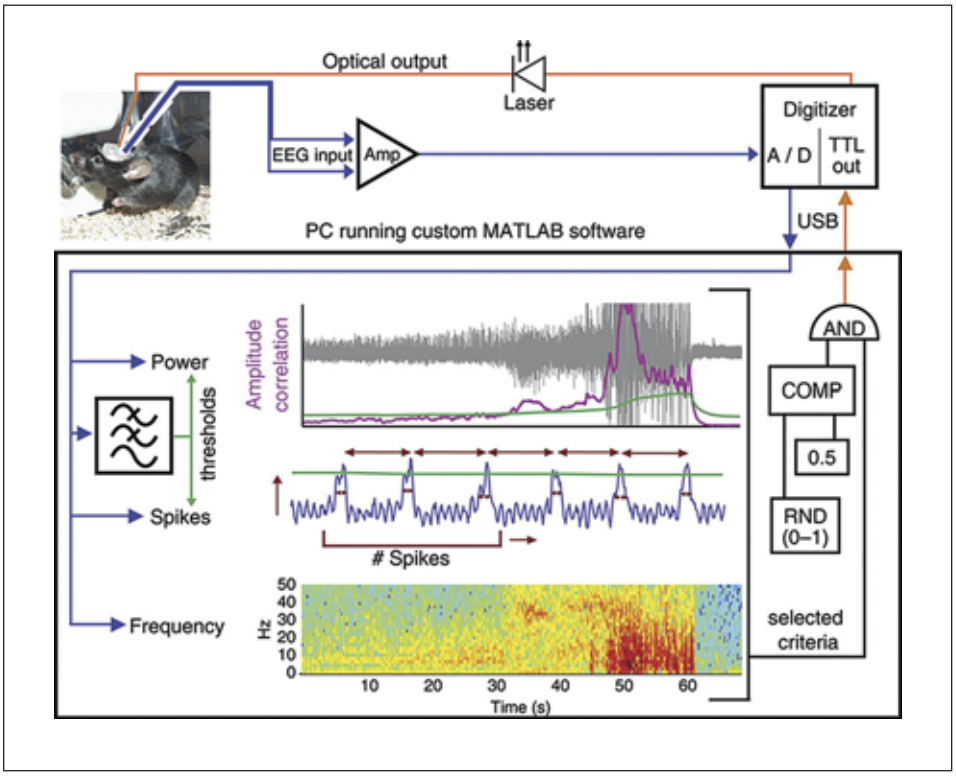
People with epilepsy would welcome a way to know when a seizure is coming, Carney says. "If I told you that you'd have a seizure today at 1 p.m., you could arrange your life around it."

About 20 years ago, researchers got very excited about using computation to predict seizures well before they start. They set about looking for patterns in the electroencephalograms (EEGs) of people having seizures to see if they could predict seizure onset well before it actually starts, and at least a minute ahead of time. An initial flurry of promising algorithms didn't pan out because they used flawed statistical methods.

Then in 2007, a company called Neurovista published an appropriate statistical framework for prediction. They went on to develop an implantable prediction device that can show patients, on a handheld device, whether likelihood of a seizure is low (blue light), moderate (white light) or high (red light) in the next few hours. A 2013 paper in *Lancet Neurology* reported that for 8 out of 11 patients tested, the device predicted seizures accurately between 56 and 100 percent of the time. For investors and the FDA, that apparently wasn't enough of a home run, Stacey says. Funding dried up and the future of the prediction device is uncertain.

At this point, Carney says, "The field is back to trying to understand what's going on rather than trying to predict seizures based on what we know now."

applies gentle adjustments to keep things within a certain dynamical range. "Can we take advantage of those intrinsic mechanisms to prevent seizures?" he wonders. Perhaps as a seizure is ramping up, there might be a point when intervention (turning on a stimulator or taking a medication) would keep the brain out of the danger zone. "Rather than responding to the hurricane, you break it up in advance."



In an optogenetic closed-loop system for stopping seizures, Krook-Magnusson and her colleagues fed EEG signals coming from the mouse brain (blue arrows) into real-time seizure detection software containing several possible algorithms for recognizing changes in features such as signal power, spikes, or frequency. The software was tuned to recognize certain thresholds for seizure in each mouse. Once detected, the experimental protocol called for administration of light (orange arrows) in half of the events in random fashion. The result: Optogenetic control reduced the frequency and duration of seizures in the mice. Reprinted with permission from Krook-Magnuson E, et al., On-demand optogenetic control of spontaneous seizures in temporal lobe epilepsy, **Nature Communications** ***4:1376 (2013).***

Unlike DBS, which Carney describes as a black box, control theorists would start by figuring out what features in the brain can be acted on to provide the necessary control.

One approach that is already showing great potential is optogenetics: Using a pulse of light to activate genes involved in epilepsy. A device for detecting and then automatically and optogenetically stopping spontaneous temporal lobe seizures recently proved effective in transgenic mice. The research team, led by **Esther Krook-Magnuson, PhD**, postdoctoral fellow in the department of anatomy and neurobiology at the University of California, Irvine, used two breeds of mice, each designed to express light-sensitive proteins that would either inhibit certain excitatory brain cells or activate the power of inhibitory (GABAergic) cells. They then implanted the mice with electrodes for detecting seizures and an optical fiber for delivering light to the target cells. First, the detector had to be trained on the specific mouse's seizure data, a not insignificant hurdle because temporal lobe seizures are tricky to detect. Detection also had to be fast, because it would occur only seconds before a seizure would otherwise start. "Computations have to be done efficiently and at an appropriate time scale," Krook-Magnuson says.

For both breeds of mice, the device reduced seizures and seizure duration with no obvious side effects. "Since it is 'on demand' rather than continuous treatment, we're not interrupting good network activity," Krook-Magnuson says.

The work offers a tool for understanding the roles of specific cell types in causing and stopping seizures, and might lead to new pharmacologic approaches, she says. There's also the possibility of using optogenetics to treat humans, although currently the idea of transfecting a human brain with a virus carrying the necessary genes is out of favor, Carney notes. "We have not been able to convince reviewers that optogenetics has a clinical future," he says.

But the epilepsy field's interest in control theory goes beyond optogenetics. In his 2012 book, *Neural Control Engineering*, **Steven Schiff, MD, PhD**, director of the Penn State Center for Neural Engineering, and a pioneer of using computational neuroscience to study epilepsy, proposes applying non-linear control theory to models of epilepsy at all scales—neuronal, lumped, and whole-brain—and paints a picture of where control theory could take the field.

According to Carney, Schiff's interest in control theory reflects a shift in computational neuroscience away from a signal processing approach to epilepsy and toward more advanced dynamical modeling. "Ultimately we want prevention and cure," Carney says. "We have treatments right now. But computational neuroscience lets you take experiments or results to the next level." □

PROSTATE CANCER: Crunching the Numbers

By Alexander Gelfand

The numbers tell the story: Prostate cancer is a killer. It's the second most common form of cancer and the second leading cause of cancer death among American men. More than 230,000 new cases are expected by the end of this year alone, and nearly 30,000 men are expected to die from it in the same period.

Early diagnosis is vital, but current methods are far from perfect; and once it metastasizes to the bones, the disease is incurable. Some researchers and clinicians are embracing computation as a means of improving both diagnosis and treatment: They're finding better ways of detecting and treating the cancer with robots and magnetic resonance imaging, figuring out how prostate cancer evolves, and homing in on the genes that regulate the disease.

Image Enhancement

Anant Madabhushi, PhD, wants to use computers to get a better picture of prostate cancer.

Doctors typically use manual ultrasound scans to guide their biopsy needles into patients' prostates or to implant radioactive seeds in the gland during a treatment called brachytherapy—an approach that can destroy tumors before they spread. But studies show that ultrasound-guided biopsy fails to detect prostate cancer in at least 20 percent of patients who have it.

MRI images, with their high resolution and excellent soft-tissue contrast, can do a better job of enabling both targeted biopsy and more narrowly focused treatment of prostate tumors. They could also fuel computer-aided detection and diagnostic algorithms, and be combined with other kinds of imaging data to improve computer-assisted surgical navigation and radiotherapy. But manually segmenting MRI images to identify a tumor's borders within the prostate is not currently standard practice, partially because it is difficult and time consuming. Segmentation algorithms would seem to offer a faster and better way of parsing MR data, but they aren't yet reliable enough for clinical use—especially since the scans themselves tend to be highly idiosyncratic, with differences amongst scanners generating lots of variability in image appearance and quality.

Madabhushi, who is associate professor of biomedical engineering and director of the Center for Computational Imaging and Personalized Diagnostics at Case Western Reserve University, has therefore been trying to draw more algorithm developers into the fray. In 2012, he was the lead organizer for the Prostate MR Image Segmentation challenge (aka PROMISE12), which had 11 teams from industry and academia compete to see whose algorithms—some fully automated, some highly interactive—could best segment scores of MR images provided by imaging centers in the United States and Europe.

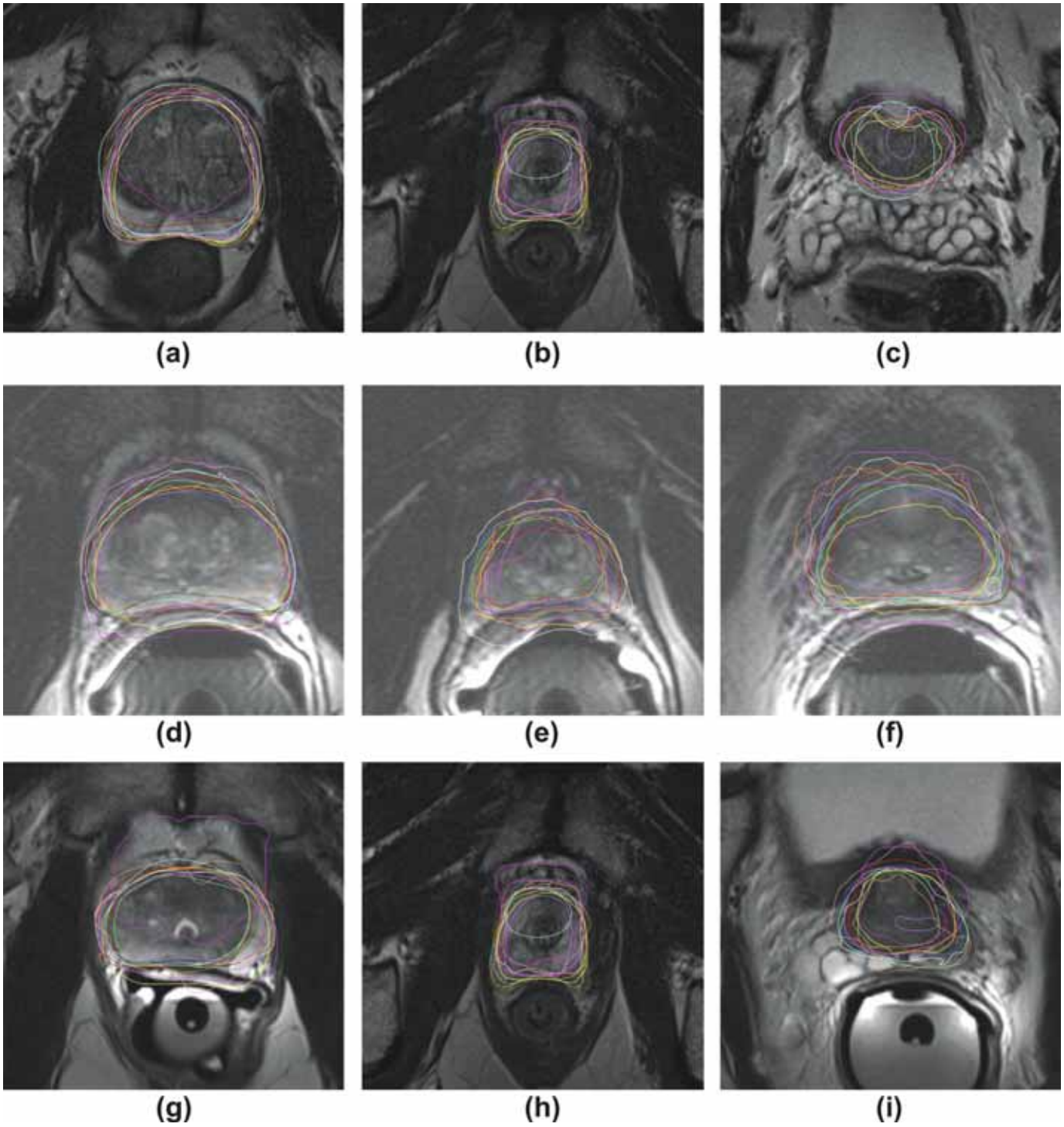
After tuning their algorithms on a training data set that included a reference standard comprised of manual segmentations by expert human annotators, the teams downloaded and segmented one test set that did not include such a benchmark, and were handed yet another at a live workshop in Nice, France. The organizers ranked the algorithms based on how closely they approached the reference standard, and on how well they did compared to an inexperienced human annotator.

In some particularly tricky cases, the interactive algorithms, which relied on human users to digitally paint large parts of the images, outperformed their fully automated counterparts. But much to Madabhushi's surprise, the two algorithms that scored best overall were completely automated. Both employed active appearance models, which use large data sets to construct models of the shape and appearance of the prostate; and on at least some measures, both managed to outdo even the inexperienced human annotator. (In a second challenge organized in 2013, Madabhushi's own team from Case Western won with a semi-automated algorithm.)

Madabhushi hopes that the algorithms will continue to improve, but he doesn't think they'll ever completely replace expert human annotators. "You have the autopilot, and you can go on cruise. But you still want the pilot there when you're taking off or landing," he says, pointing out that the inexperienced annotator still outshone most of the algorithms in the 2012 challenge.

Madabhushi hopes that the algorithms will continue to improve, but he doesn't think they'll ever completely replace expert human annotators. "You have the autopilot, and you can go on cruise. But you still want the pilot there when you're taking off or landing," he says, pointing out that the inexperienced annotator still outshone most of the algorithms in the 2012 challenge. Which begs the question: what would happen if a whole bunch of different algorithms, all using different approaches, joined forced

Three sets of images representing three different cases from the PROMISE12 challenge. Different colors are used to illustrate prostate segmentations by different teams; on average, case 3 (images a, b, and c) had the best algorithm scores, case 10 (images d, e, and f) had reasonable scores, and case 25 (images g, h, and i) had the worst scores. The two algorithms with the best overall scores in the contest were fully automated; but in case 25, a large area of fat around the gland caused most of the algorithms to make large errors in prostate volume, and a more interactive algorithm did best. Reprinted from Litjens G, Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge, **Medical Image Analysis** ***18:359-373 (2014), with permission from Elsevier.***



with a whole bunch of inexpert humans?

Madabhushi describes one possible scenario in which large numbers of non-experts (e.g., high school and college students) segment batches of MR images, and automated algorithms check their annotations for accuracy. The algorithms could identify the best of the inexpert human annotators for future reference; segment the harder cases themselves; and send the ones that even they can't handle on to expert human annotators, who could then pass their own properly segmented images back down to the algorithms for training purposes. Whether such a system would adequately preserve patient privacy or gain FDA approval remains to be seen. But the ensuing virtuous circle of data-sharing and analysis amongst experts, amateurs, and algorithms could, says Madabhushi, yield something "more enriched, and perhaps more accurate, than any individual source of information."

Open-Source Revolution

If Madhabhushi envisions a future where crowdsourcing and computer automation enhance prostate cancer diagnosis and treatment, **Gabor Fichtinger**, **PhD**, a professor in the School of Computing at Queen's University in Kingston, Ontario, and adjunct professor of computer science and radiology at Johns Hopkins University, sees one dominated by open-source software.

In collaboration with the National Alliance for Medical Imaging Computing (NA-MIC), Fichtinger led the development of Prostate Nav, a prostate-specific module within 3D Slicer (www.slicer.org), the Alliance's free, open-source platform for visualization and image analysis. Prostate Nav allows researchers and clinicians to use medical robots to perform biopsies and brachytherapy. It can create an interface between a robot and the rest of the equipment (scanners, navigation systems) in the medical suite; register the robot to the same coordinate system that 3D Slicer uses to pinpoint the location of any other tracked surgical instrument; issue commands to the device; and even cause an animated model of it to appear on the operator's screen. Fichtinger and his colleagues have used Prostate Nav to support an entire family of MR-compatible robots that can function inside the bore of an MRI scanner, guiding needles into patients with far greater accuracy than the standard manual ultrasound-guided method can achieve.

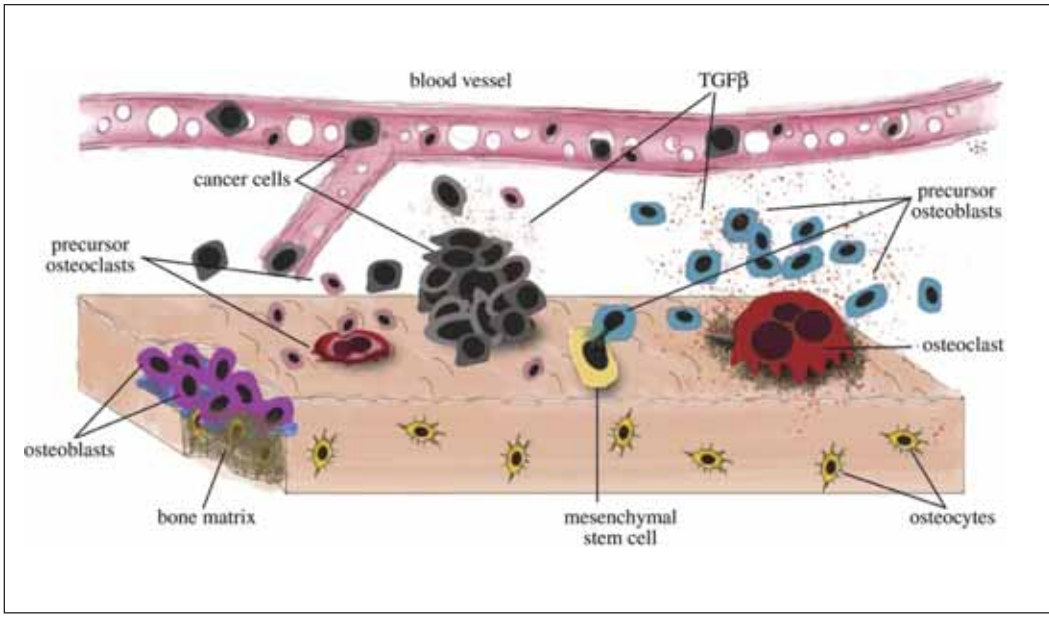
Now Fichtinger is exploiting open-source software to more quickly and efficiently build systems that combine robots and tracked surgical tools with preoperative MRI, intraoperative ultrasound, and other imaging modalities. For example, he and his colleagues recently added color stereo optical imaging to their software platform to accommodate researchers who are interested in laparoscopic prostate surgery. And last year, Fichtinger's team developed a custom system for MRI- and ultrasound-guided prostate intervention research at Harvard's Brigham and Women's Hospital in just eight weeks—and it only took that long, Fichtinger says, because of some "funky requests" by the clinicians, such as simultaneous image acquisition from multiple ultrasound transducers. Within the next decade, Fichtinger predicts that the technology will have matured to the point where it will be possible to derive working, clinical-grade applications from open-source platforms such as 3D Slicer in a matter of days—though getting FDA approval for them "will still take a good bit of time."

An Ecological Approach

Identifying and treating tumors in the prostate is critical. But prostate cancer becomes truly lethal when it migrates elsewhere. **David Basanta**, **PhD**, and **Arturo Araujo**, **PhD**, in the Integrated Mathematical Oncology department at the Moffitt Cancer Center in Tampa, Florida, have therefore built a computational model that combines agent-based techniques with conventional mathematical modeling methods to simulate how prostate cancer metastasizes to bone in order to better understand, and hopefully foil, the process.

Previously, Basanta used another hybrid model to investigate how the protein TGF-beta affects tumor growth. He has also employed evolutionary game theory to explain how interactions between prostate cancer cells, normal cells, and their shared microenvironment influence cancer progression, comparing tumor cells to invasive species that disrupt the ecosystem of healthy tissue. His latest work, carried out in conjunction with a group led by molecular biologist **Conor Lynch**, **PhD**, and reported in 2014 in the journal *Cancer Research*, builds on those earlier efforts, using a hybrid cellular automaton model to illustrate how metastatic prostate cancer cells are able to exploit elements of the bone ecosystem, including TGF-beta and another signaling molecule called RANKL, to their own advantage.

The agents in Basanta's model include not only prostate cancer cells, but also the osteoclasts and osteoblasts that break down and build up bone tissue during the course of normal bone maintenance. Partial differential equations, meanwhile, are used to mimic the production, diffusion, and decay of TGF-beta, RANKL, and other molecules that coordinate normal bone maintenance yet also facilitate the proliferation of cancer cells. In simulations



The ecosystem in a prostate to bone metastasis comprises several types of cancer cells interacting with other cellular populations such as osteoblasts, osteoclasts, osteocytes and stem cells. Tumor cells compete and cooperate for resources such as nutrients, space and growth factors. Reprinted with permission from Basanta D, Anderson A, Exploiting ecological principles to better understand cancer progression and treatment,* Interface Focus, *3, 20130020 (2013).

that ran for 240 virtual days, Basanta's model demonstrated how prostate cancer cells manipulated levels of TGF-beta and RANKL to create a vicious cycle of aggressive tumor growth and abnormal bone formation and resorption. The model was also able to predict the efficacy of two types of drugs that are commonly used to slow the progress of bone metastasis, and offered some insight into how one of them—an anti-RANKL inhibitor—might be used more effectively in the clinic.

Basanta and Lynch are now testing the efficacy of TGF-beta inhibitors using both *in silico* and *in vivo* tools, and they have joined forces with several clinicians to develop a computationally and mathematically enhanced method of personalizing treatments for patients with metastatic prostate cancer. Basanta hopes to use models to predict how tumors with particular mutations might evolve and grow in response to different drugs, then use that information to optimize the sequence of treatments a patient receives "in order to reduce the tumor burden in the bone and, presumably, extend quality of life—and improve their chances of coming out of this alive."

The Root of the Problem

Andrea Califano, **PhD**, professor of chemical systems biology and chair of the department of Systems Biology at Columbia University, is pursuing the same goals with a different set of computational tools. Ultimately, Califano wants to personalize cancer treatments by reconstructing the regulatory networks, or interactomes, that control different kinds of tumors. This approach would allow researchers to look beyond the bewildering array of genetic mutations that accompany the various tumor types and focus instead on the master regulators of the disease: those genes that are necessary for the survival of a given form of cancer. Because they rarely harbor genetic mutations, these master regulators cannot be identified through standard genetic sequencing. "But you can find them by analyzing these networks," Califano says. And once found, they may be inhibited by existing drugs.

That was the case in a study that Califano and his colleagues, **Cory Abate-Shen**, **PhD**, and **Michael Shen**, **PhD**, recently published in *Cancer Cell*. They began by using an algorithm called ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) to reconstruct two interactomes: one responsible for producing prostate cancer tumors in human beings, and one responsible for producing them in mice. Reverse engineering each network required sifting through hundreds of thousands of possible interactions between thousands of transcription factors and their target genes. The team then ran a different algorithm to determine which cancer-related transcription factors controlled genetic programs that were conserved between mice and people, and were therefore most likely to be significant.

By combining the genes discovered in the indolent tumor and master regulator studies, Califano and Abate-Shen hope to develop a comprehensive five-gene panel that can give prostate cancer sufferers a "complete report" on the aggressive potential of their tumors.

Califano and his collaborators then used an algorithm called MARINa (Master Regulator Inference Algorithm) to identify the transcription factors that were most likely to induce the genetic signature observed in aggressive prostate tumors. The seven conserved master regulators that emerged were then computationally analyzed for potential synergistic interactions among themselves, and a single pair of synergistic master regulators—the genes FOXM1 and CENPF—were found to drive aggressive prostate cancer in both mice and humans. Silencing one gene slowed cancer growth in the mouse models; silencing both shut it down completely. And protein expression analysis of prostate tissue samples taken from more than 900 prostate cancer patients at Memorial Sloan-Kettering Cancer Center revealed that patients with elevated expression levels of both genes experienced by far the worst outcomes—including shortest time to metastasis, and death. Abate-Shen and Califano have already identified two drugs that can inhibit these master regulators.

In addition to identifying the master regulators that induce aggressive prostate cancer, Califano and his colleagues have found a cluster of genes that can be used to predict whether tumors that seem indolent, or slow-growing, are destined to stay that way. It's a crucial task, since overtreatment of prostate cancer is both costly and potentially risky, yet the only thing worse than unnecessarily treating a person with an indolent tumor is failing to treat one whose tumor only appears to be so.

Califano's indolent tumor work, published last year in *Science Translational Medicine*, began with a manually curated list of 377 genes associated with the tumor-inhibiting processes of cellular aging and senescence. He and colleagues used Gene Set Enrichment Analysis (GSEA), which ranks genes on a spectrum from most to least expressed, to identify 17 senescence genes that were over-expressed in indolent mouse tumors and under-expressed in aggressive human ones; then applied a decision-tree algorithm to prune them down to a trio of genes with the greatest predictive power. All three were validated in the lab by Abate-Shen and were found to be underexpressed at the protein level in biopsies taken from prostate cancer patients whose tumors initially appeared to be indolent, but nonetheless became aggressive.

By combining the genes discovered in the indolent tumor and master regulator studies, Califano and Abate-Shen hope to develop a comprehensive five-gene panel that can give prostate cancer sufferers a "complete report" on the aggressive potential of their tumors.

More accurate diagnoses, new biomarkers, and improved therapies: add it all up, and computation just might make prostate cancer's numbers look a little less menacing after all. □

PROBING HUNTINGTON'S ORIGINS:
Computational Approaches May Lead to Earlier Interventions

By Esther Landhuis

Uncontrolled writhing and jerking. Poor judgment. Depression and irritability. It's hard to imagine how this unnerving mix of movement, cognitive and psychiatric problems arises from a single genetic blip—one that plops unusually long stretches of the amino acid glutamine in the culprit protein for Huntington's disease (HD). Researchers who study this brain disorder are still puzzling over how the rogue molecule causes so much to go awry.

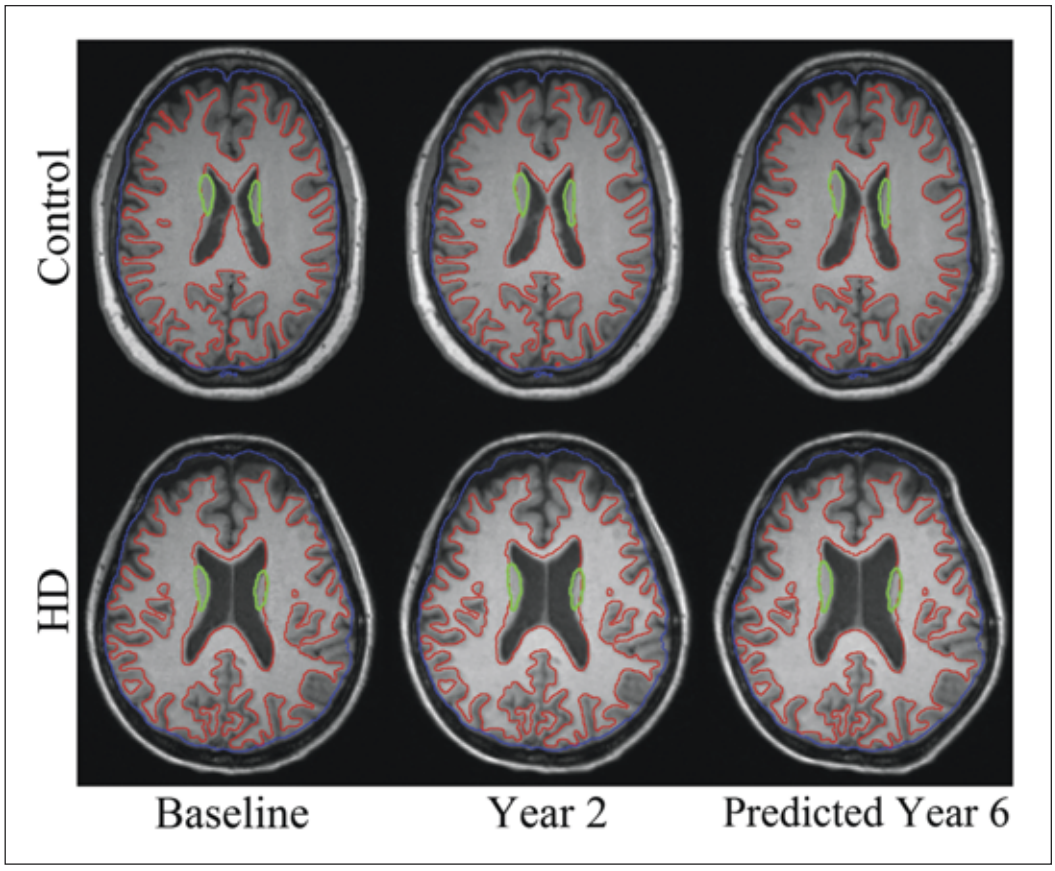
What they do know is that people who inherit the huntingtin gene mutation are sure to develop the disease and die of it. In Western nations, the disease strikes one in every 10,000 to 20,000 people, destroying neurons in areas at the base of the brain known as the basal ganglia. On magnetic resonance imaging (MRI) scans that measure brain volume, regions of the basal ganglia appear heavily shrunken in Huntington's patients, relative to normal adults. By the time symptoms appear, "they've already lost a tremendous amount of brain structure. It's hard to regain that," says **Hans Johnson, PhD**, assistant professor of psychiatry and biomedical engineering at the University of Iowa Carver College of Medicine in Iowa City.

While structural MRI measures of brain volume can be useful for understanding the biological degradation that has occurred at the time of diagnosis, Johnson and other researchers are now using state-of-the-art neuroimaging and computational approaches to look much earlier in the disease process. Churning out four- or five-dimensional data, these newer methods burn through 1,000 to 2,000 times as much mathematical and computational power as volumetric MRI. But they are yielding valuable clues—subtle changes in circuitry and function that seem to lurk within the brain for years, perhaps even decades, before symptoms become serious enough to prompt a doctor's visit. The researchers hope that teasing out these early abnormalities will guide them toward new strategies to protect the brain from impending damage. Other scientists are taking a different tack, using bioinformatics to uncover other genes and pathways that may be linked to Huntington's pathogenesis.

Neuroimaging: Structural and Functional Brain Changes

One of the field's biggest surprises—which also may explain why Huntington's symptoms remain under the clinician's radar for so many years—is the brain's incredible ability to adapt. "We can function pretty darn well by recruiting non-traditional parts of our brain to work on tasks," Johnson says. For instance, early trouble in memory circuits can trigger the brain to rewire itself in ways that bypass problem spots by activating neurons from other areas. Some researchers believe this process kicks in as a compensatory process to help people on the verge of Huntington's retain function in the face of early degeneration.

Such insights come from studies that use functional MRI (fMRI)—a technique that measures brain activity by detecting changes in blood flow. A typical functional neuroimaging session produces 20 gigabytes of raw data, Johnson notes. "In addition to drawing a ruler on the screen and counting the number of voxels in a region, we need to correlate the task being run with other



The brains of HD patients (bottom) show progressive expansion of the ventricle (large dark area in the middle) and thinning of the caudate (green outline) from baseline (left) until two years later (middle). And researchers have extrapolated that change six years into the future as well (right). In contrast, a control subject does not show much change at all (top). Animations of the series shown here are particularly compelling and available online at http://www.cs.utah.edu/~jfishbau/docs/ctrl_and_hd.gif. Courtesy of James Fishbaugh.

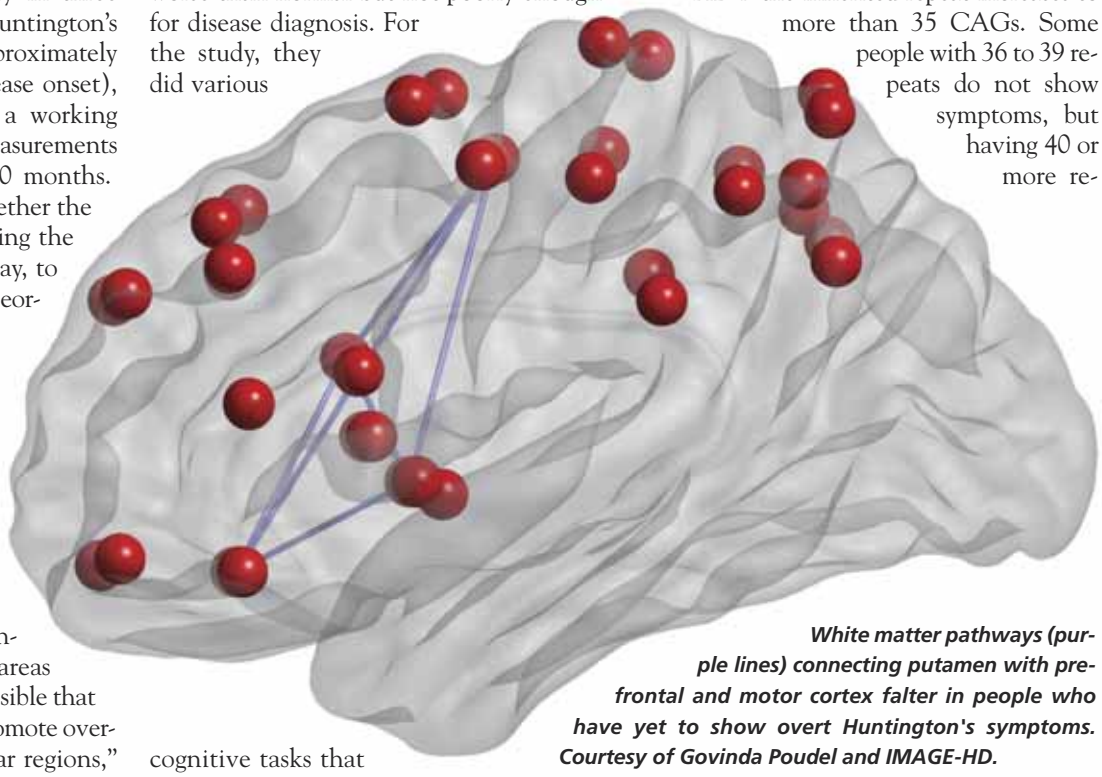
variables such as heartbeat and breathing rate, in order to get to the signal that we then have to extract mathematically with the task being performed."

Prior fMRI analyses support the idea that functional reorganization occurs early in disease, but results have been mixed—some showed a boost in brain activity while others portrayed a decrease. Moreover, past studies have been cross-sectional, analyzing data on patient subgroups at just one point in time. Researchers led by **Nellie Georgiou-Karistianis, PhD**, a cognitive neuroscience professor at Monash University in Melbourne, Australia, investigated the compensation issue more rigorously in a longitudinal study known as IMAGE-HD. The team used fMRI to monitor brain activity in three groups of people—those with Huntington's symptoms, mutation carriers (approximately 15 years prior to estimated disease onset), and healthy controls—during a working memory task. They made the measurements at baseline, 18 months, and 30 months. "Our focus was to determine whether the brain is able to compensate during the 'pre-manifest' stage—in some way, to help individuals stay on task," Georgiou-Karistianis says.

As her team reported in a November 16, 2013, *Brain Structure & Function* paper, the answer seems to be yes. Over the 30-month study, the fMRI scans showed subcortical and cortical areas firing more intensively in people on the verge of symptom onset, compared with the other groups. In addition, connections between those brain areas appeared to be faltering. "It is possible that the reduced connectivity may promote overcompensation in these particular regions," Georgiou-Karistianis says, noting that these areas could be used as targets for future pharmacological intervention. "The overcompensation might also be a response to the structural brain changes that are also happening very early during the disease."

Some of those abnormalities may show up with diffusion-weighted imaging (DWI), a newer type of MRI that requires massive computing power to visualize tissue architecture. Instead of a single picture of a brain structure, DWI generates some 80 high-resolution sub-pictures that help scientists model the vibration of water molecules. Those vibrations give insight into the integrity of the underlying tissue. Using the analogy of an electric circuit, fMRI measures the strength of signals whereas DWI gives insight into the integrity of the wires that carry the signals. For example, is the insulating shell around the wires intact, or does it have holes where it's rubbed against something else? Does the insulation work properly? The "insulation" that surrounds the axonal projections of neurons is known as white matter. Diffusion MRI helps researchers determine how the white matter bundles are organized, whether they're packed densely or loosely, for example.

A team of researchers led by **Jane Paulsen, PhD**, a professor of psychiatry at the University of Iowa, Iowa City, used diffusion-weighted imaging to measure white matter changes in the prefrontal cortex of people with prodromal Huntington's disease. These individuals perform slightly worse than normal but not poorly enough for disease diagnosis. For the study, they did various cognitive tasks that required them to identify words or colors, or link numbers together. In a paper published in April 2013 in *Human Brain Mapping*, Paulsen and colleagues reported that myelin sheath integrity and other white matter measures seemed to track with cognitive readouts in those areas. "For example, in the word-finding task, white matter deficits showed up in the part of the brain needed for word processing," noted Johnson, who was a coauthor on the research.



White matter pathways (purple lines) connecting putamen with prefrontal and motor cortex falter in people who have yet to show overt Huntington's symptoms. Courtesy of Govinda Poudel and IMAGE-HD.

Bioinformatics: Other Genes and Pathways

While neuroimaging afficionados strive to understand the earliest stages of Huntington's by probing the brain's inner workings, others are coming at the issue using bioinformatics. Their computational strategies are scouring massive gene-expression datasets for pathways that are altered by mutant huntingtin.

The huntingtin protein is expressed in many cell types, but scientists don't understand quite what it does. It interacts with a huge number of proteins and has structural features found in organizers of molecular complexes.

The huntingtin (*HTT*) mutation is also intriguing. All of us have CAG (cytosine-adenine-guanine) trinucleotide repeats in our huntingtin gene. In fact, we can have up to 35 *HTT* CAG repeats and still be considered normal. In rare cases, people with 27 to 35 *HTT* CAG repeats can have children with HD if the inherited repeat increases to more than 35 CAGs. Some people with 36 to 39 repeats do not show symptoms, but having 40 or more repeats virtually guarantees disease. The size of the *HTT* CAG repeat mutation correlates inversely with age of symptom onset—in general, the more *HTT* CAG repeats, the earlier a person will develop disease.

The prevailing view holds that the *huntingtin* mutation acts through full-length expression of the gene. "We're not dealing with a protein that loses function. It's either gaining function or getting dysregulated," says **James Gusella, PhD**, who directs the Center for Human Genetic Research at Massachusetts General Hospital in Boston.

Until recently, researchers doing gene expression studies to understand the effect of the *HTT* CAG repeat assumed they would get the strongest signal by analyzing extreme populations—that is, genomes with

strong disease mutations and those devoid of mutations. Hence, they have compared two groups: patients with HD symptoms and normal patients (controls). These dichotomous analyses failed to take into account differences in CAG repeat length.

Gusella and his colleagues therefore took a different analytical approach: They correlated gene expression across the continuum of *HTT* CAG repeats from low to high (15 to 92) in 97 lymphoblastoid cell lines (the training set). They used the results to mathematically predict CAG repeat numbers in a set of 10 cell lines (the test set). The proof-of-concept study showed that differences in transcript levels can detect the continuous effects of increasing *HTT* CAG repeat length and provide an approach to discovering factors contributing to the pathogenic process, which also increases with *HTT* CAG repeat length. The expression changes appeared in genes involved in chromatin remodeling, energy metabolism and axonal transport, suggesting that CAG repeats have downstream consequences on molecules involved in these pathways. However, how these systems connect is not clear, Gusella says. "The big picture hasn't yet come together." The work appeared in *Human Molecular Genetics* in April 2013.

Future: Bigger Datasets

In the long run, gaining a more complete picture of Huntington's disease progression will require a pooling of many different types of studies—gene expression, fMRI and DWI and others, Johnson says. "Investigating it jointly rather than independently is really where the future promise is." □

DESIGNING LIFE'S LAYERED CIRCUITS: Tools of the Trade

By Sarah C.P. Williams

In synthetic biology labs around the world, brainstorming has often begun at the same place: in front of a whiteboard. Marker in hand, researchers jot down the parts needed to form a new circuit, draw lines and arrows to show how they interact, and scrawl notes about how to assemble the parts into an appropriate whole.

"It's usually based on intuition, and what we know has worked in the past," says **Timothy Lu, MD, PhD**, who heads up the Synthetic Biology Group at the Massachusetts Institute of Technology.

The whiteboard has been used to design many novel genetic programs—whether aimed at turning bacteria into biosensors or forming networks of enzymes to churn out a particular product. But the way of the whiteboard might be fading. As circuits become more and more complex, and researchers move toward the design of larger networks and whole-cell programs, it's becoming harder to manage all the required parts for a new project in hand-written dry erase.

"When I was looking at a simple circuit with two inputs, I could by hand iterate through all the possible states of the system," Lu says. "Now, I'm interested in things with six or eight inputs, and intuition starts to fail."

Costas D. Maranas, PhD, professor of chemical engineering at Penn State University, concurs. While synthetic circuits of a decade ago had a single switch and just a few inputs to alter genes, Maranas is trying to reconstruct and regulate the entire repertoire of pathways involved in a microbe's metabolism.

And it's not just the number of switches that adds complexity. Adding new enzyme activity into a bacterium is more complicated than just adding the enzyme. Take nitrogenase, for example, which Maranas and collaborators at Washington University would like to be able to control within a cyanobacterium. Getting the right levels of nitrogenase activity, he adds, doesn't just mean having the right levels of gene and protein expression, but also accurately reproducing the light to dark transitions and providing sufficient energy in the form of ATP to power the nitrogenase.

To help manage this complexity, researchers are developing, refining, and applying computerized design programs that track the parts involved in their systems and pinpoint the best method to assemble a new circuit. There's not yet one program that fulfils the dream of "plug and play" biology—where a few simple clicks choose the parts for the essential biological circuits and, voilà, synthetic life! But several programs are emerging as crucial to the field.

Inspired by Engineering

In the mid-2000s, **Jean Peccoud, PhD**, a computational synthetic biology researcher at Virginia Tech's Bioinformatics Institute, was working on recreating the genetic networks that control cell division in yeast. Like other synthetic biologists, Peccoud viewed the components of the network—genes, promoters, ribosome binding sites, and terminators, to name a few—as discrete parts, with defined functions, that could be shuffled around between networks. But he realized that no software existed that could track which parts worked together, guide how the parts could be plugged into genetic circuits, and model how a proposed circuit would function.

"It seemed reasonable to assume that synthetic biology would need some computer-aided-design tools just like any other engineering discipline," says Peccoud. CAD programs are heavily relied on by electrical and mechanical engineers, for example, to design electrical circuits or structures on the computer before they're created and tested.

So his lab began developing such a program for biology. The result: GenoCAD, an open-source, synthetic biology CAD software. GenoCAD manages lists of genetic parts and gives users an interface where they can set design rules, apply them to their system and then assemble genetic parts into plasmids. It also includes a simulation engine to test new circuits.

In the December 2013 issue of *ACS Synthetic Biology*, Peccoud and two collaborators describe using GenoCAD to create a set of

grammatical rules for building novel synthetic transcription factors from seven different types of parts. The program was able to generate eight possible designs that met all the rules governing what parts were required and what order they should fall in. The rules, which were derived from experimental information, can be revised and updated over time. As new synthetic circuits are tested in living cells, their success or failure can help guide the design of future circuits.

A Growing Toolbox

In addition to GenoCAD, there are a rapidly growing number of synthetic biology tools, Peccoud says. He adds that his ultimate goal isn't for GenoCAD to beat out other tools. "I don't think it should be a goal to converge to one tool," he says. "Our field is so new that it is necessary that people explore different avenues."

When designing DNA to characterize new promoters, **George McArthur, PhD**, a chemical engineer at Virginia Commonwealth University, turns to a different software program for nearly every step of the process. Aside from GenoCAD, he uses a ribosome binding site (RBS) calculator that develops an RBS of whatever binding rate he needs; a tool that produces inert spacer sequences; and the automated DNA assembly program J5 that gives him a list of primers for use in assembling the sequences he designs in GenoCAD.

"As a user, I'd love to have everything in one place," McArthur says. "And already it's great that a lot of these tools adhere to the same file standards. I think that eventually we'll have different collections of software that aggregate together."

One effort to encourage the consolidation of tools—or at least the development of a start-to-finish synthetic biology design protocol—is DARPA's "Living Foundries: 1000 Molecules" program. Approximately $110 million in grants will be doled out by the end of 2014 to scientists who aim to build infrastructures for engineering biological molecules. The proof of principle for any infrastructure will be the design and production of 1,000 new molecules. But following through to such an outcome will likely require a strong set of software tools and provide an example for the rest of the field to follow.

Predictive Power Still to Come

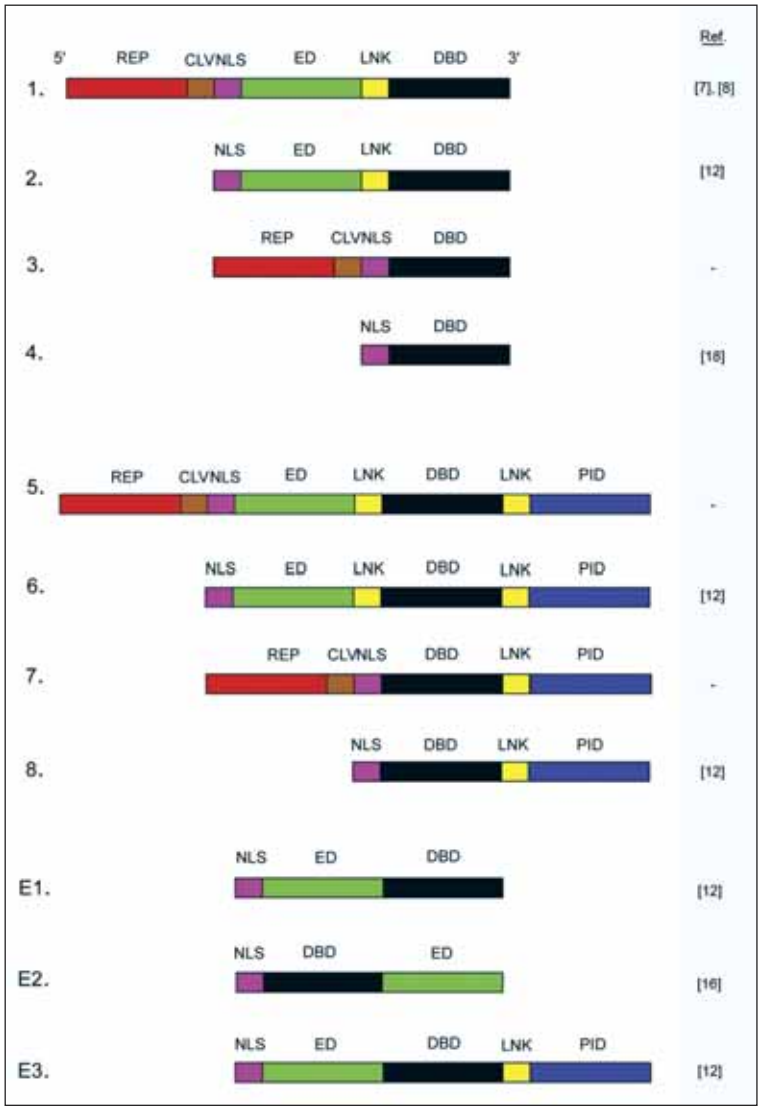
Peccoud admits that the weakest part of GenoCAD is the newest addition to the program—the simulation engine. "Being able to run simulations of the behavior of a synthetic genetic system before making it is the holy grail of synthetic biology," he says. "The science is not there yet but it is our hope that a tool like GenoCAD can help support the research necessary to understand gene expression better."

Lu says that getting more accurate simulations of biological circuits will require more data on how different organisms interact differently with the various parts that make up circuits.

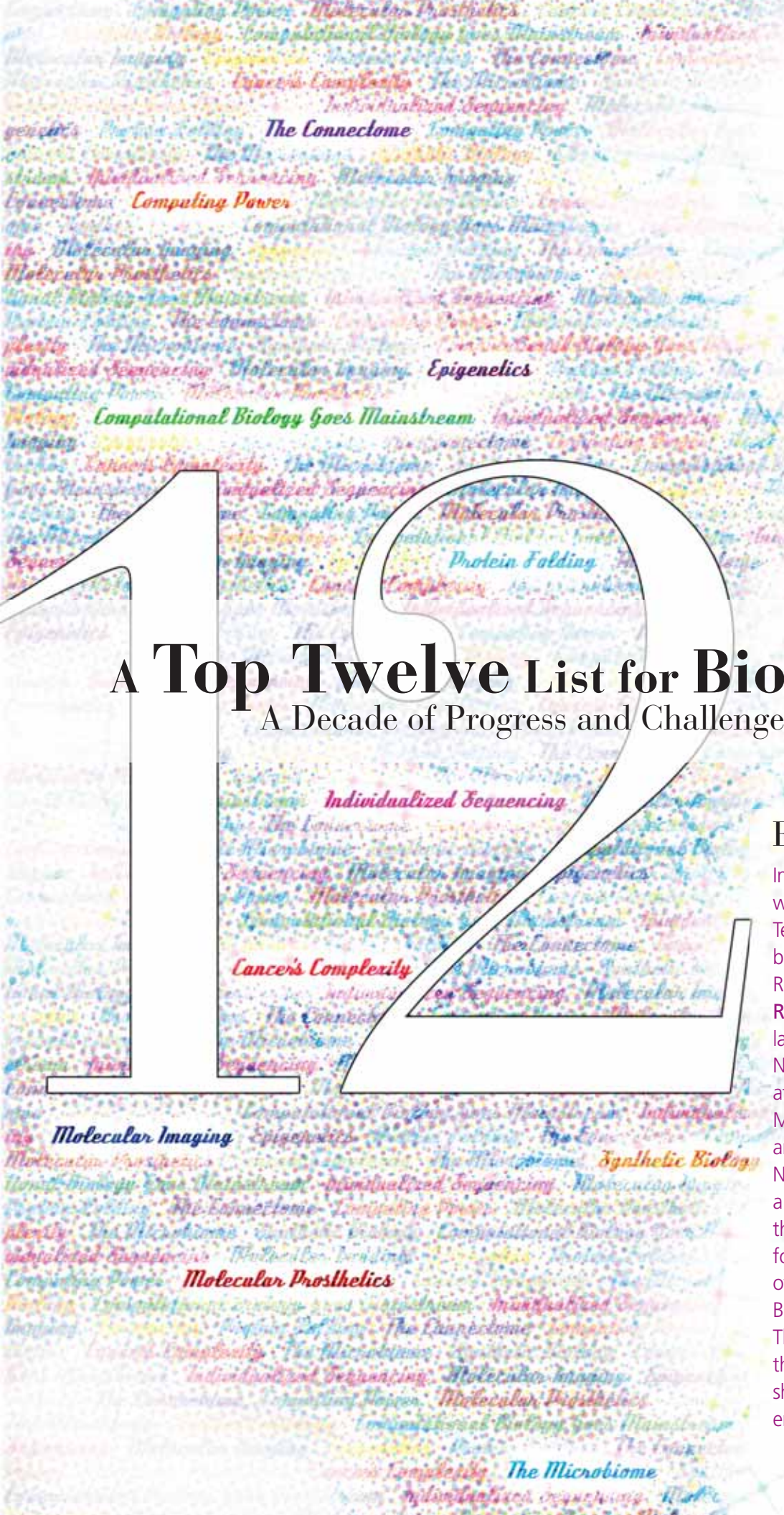
"In other engineering disciplines, the manufacturer of a system will give you parameters that define that particular system," Lu says. Those parameters can be entered into CAD software to make your computer models accurate. "But in biology these days, no one has defined what, say, the *E. coli* parameter set is," he points out. Even if a circuit is completely worked out in one strain of *E. coli*, he says, moving it to a new strain can drastically change how it functions.

Recently, Stanford scientists created a whole-cell computational model of the circuits within *Mycoplasma genitalium*, a human pathogen. Lu has collaborated with the Stanford team to start putting synthetic circuits into the organism. With the whole-cell model at his disposal, he hopes to start predicting how new circuits will work in the organism. But even that has been slow going, he says, and it's just one bacterium.

Even without full predictive power though, programs like GenoCAD are pushing the boundaries of synthetic biology, offering a more modern "whiteboard" to sketch out complex circuit designs and organize growing libraries of biological parts. By giving users a place to organize the grammatical rules that govern their design process, and the parts that they want to use, it makes the design step of the standard engineering "design, build, test" cycle that much easier. □



Seven molecular building blocks, each shown here in a different color, can be arranged in numerous ways to form a functional transcription factor. But rules dictate which parts come first and which are required, similar to grammatical rules regarding sentence structure. Using the design program GenoCAD, researchers can determine which grammatical arrangement of parts will work best for their system, using the molecules they already have. Reprinted with permission from Purcell, O, Peccoud, J, Lu, T, Rule-Based Design of Synthetic Transcription Factors in Eukaryotes,* ACS Synth. Biol.*, DOI: 10.1021/sb400134k. Copyright 2013, American Chemical Society.



By Ruth Nussinov, PhD

A Top Twelve List for Biocomputing:

A Decade of Progress and Challenges Ahead

Editor's Note:

In addition to asking 10 experts to weigh in on Eric Jakobsson's 2005 Top Ten Challenges for the field of biomedical computing (Top Ten Retrospective, in this issue), *BCR* asked **Ruth Nussinov, PhD**, to reflect on the last decade and offer a new list. Nussinov, who is a professor of genetics at Tel Aviv University School of Medicine and senior principal scientist and principal investigator at the National Cancer Institute, came up with a Top Twelve that melds advances of the last decade with related challenges for the future. Not surprisingly, this list overlaps somewhat with Jakobsson's. But the two stories make a nice pairing: Their similarities clarify the continuity in the field, while their differences let the shifting nature of the scientific enterprise shine through. >

When asked to choose the most significant computational biology advances of the last decade, I welcomed the opportunity to reflect on the field. But soon the daunting nature of the task became clear: The field has come a long way in ten years and is so broad (covering the entire range of biomedicine) that the possible choices are numerous. Moreover, there may well be entire categories of research that haven't come to my attention. On top of that, there's the question of subjectivity: The significance of any research advance is a matter of opinion.

To make the job tractable, I opted not to identify specific research papers but instead to focus on broad topic areas where computation has made, and will continue to make, a major contribution.

It is my sense that, these days, computational biology is closely tied to experimental research. And this is an advance in itself (see #1). Thus, the progress described here is not purely computational in nature; it is tied to biomedicine. And that's as it should be.

1. Mainstreaming Computational Biology

Ten years ago, as the National Institutes of Health were preparing to fund the National Centers for Biomedical Computing, we lived in a different world. The field of computational biology, though established, was dispersed and not entirely trusted by mainstream biological and medical researchers. Today, by contrast, computational biology is intimately connected to the rest of biomedicine. It's easier to collaborate across disciplines. And laboratory researchers have a better understanding of the value of using computational models for hypothesis generation, as well as the need to iterate through a cycle of modeling and laboratory testing. In sum, computational biology has become more closely integrated with experimental work, to the betterment of biomedicine as a whole.

It's a change that's also reflected in societal expectations about computation's potential generally. This attitude shift is felt throughout the field: Computational biologists now get the respect they are due.

Looking to the future, I see computational biology increasingly taking the lead in the medical sciences. I expect experimentalists will frequently find themselves testing hypotheses generated by computational analyses of massive datasets.

2. Individualizing Sequencing

Over the past decade, sequencing technologies have changed the face of genomics at an unprecedented pace, spurring new opportunities in computational biology. The one billion dollar cost of the first two human genome projects has dropped by a factor of a million. Soon a completely different and novel approach called single molecule sequencing may become viable. The method involves pulling a DNA strand through a membrane with a nanopore designed to measure tiny base-specific changes in impedance across the pore.

This technique has great promise. For example, it will be possible to compare corresponding sequences in diseased and healthy tissue of a single human being. Such methods may lead to early screening of ailments and epidemics, and suggest more effective personalized treatment decisions.

Eventually, single molecule sequencing may also permit delivery of drugs (attached to DNA via sticky ends) to specific locations using complementary sequences. Such an approach has already been proposed and could be used to target malignant genomic aberrations.

3. Imaging Molecules in Action

A vast improvement in our ability to image living systems at all levels has been the key to many crucial developments and is likely to remain so into the future. Over the last decade, X-ray crystallographic imaging has proven its efficacy by revealing structures of thousands of proteins and by successfully decoding the structure of a whole RNA-protein complex—the ribosome.

Recent advances in nuclear magnetic resonance (NMR) imaging have led to gains of a complementary but more dynamic nature. Whereas X-ray crystallography captures the single dominant native structure of a protein, NMR can often verify the existence of several and even dozens of transient alternate forms of a protein structure. Understanding the range of possible shapes a protein can take will help in the development of drugs.

There have also been remarkable advances in the imaging of single cells, individual organelles, and even single molecules using fluorescence, electron microscopy, atomic probes and other techniques. These allow researchers to follow morphological and dynamical changes over reasonably short timescales.

In sum, molecular imaging tools now at our disposal not only allow the reconstruction of static protein structures, they can track transitions between structures, as well as reveal the kinetics and dynamics of such things as gene transcription and splicing. As we move forward, such tools, combined with computational approaches, may be able to track processes in the living cell across space, time and environment, ultimately revealing cross-scale relationships such as those between cellular outcomes and morphogens (signaling molecules involved in tissue development), genetic mutations, post-translational modifications or pathogenic proteins.

4. Going Beyond the Genome

In the nature versus nurture debate, there's been a surprising shift toward nurture over the last ten years, with computational analyses playing a role in discoveries on both sides.

In support of nature, an ever-increasing number of human traits—including even the tendency for happiness—have been found to have a genetic component. And epigenetic activity affecting the dynamics of gene expression has been found in the areas of DNA that don't code for specific proteins (areas previously considered "junk DNA"). Indeed sections of DNA that are physically far away from a protein-coding location can critically affect function.

On the side of nurture, recent developments show the importance of DNA methylation and post-translational modifications of proteins as well as other changes that aren't built into the genome but instead develop during the course of an organism's life. For example, prion disease and type A diabetes directly stem from the misfolding of proteins even in the absence of purely genetic causes. This undermines the notion that the sequence of amino acids in a protein uniquely determines its consequent shape and function, and weakens the linkage between genetic information and phenotypic traits exhibited by a cell or organism. Nurture—which causes cellular changes during an organism's life—matters.

Looking forward, computational tools will continue to play a role in pursuing an understanding of both nature and nurture and how they interact.

5. Approaching Protein Folding Sideways

Researchers would love to be able to accurately predict a protein's shape simply by knowing its amino acid sequence. That's because a protein's shape can yield an understanding of the protein's function and reveal the likelihood it will bind to other molecules, including drugs.

Naturally occurring proteins typically fold quickly *in vivo* into the correct, native, functional form. Yet predicting the complete, tertiary structure of proteins by using only sequence information and a reliable force field has challenging chemical, geometric and combinatorial aspects. It may be an intractable NP hard problem that requires approximate solutions.

Rather than attacking the problem head-on, researchers have spent the last decade developing alternative strategies—using known fold motifs, sequence homologies and the many existing structures in the Protein Data Bank (PDB)—to piece together a reasonable guess for the folded form that can be further refined and improved by additional information and calculations.

Though we've come a long way toward efficient handling and quick access to protein structural information via a large variety of different types of queries of massive structure/sequence data, there remains an ongoing need to efficiently and reliably exploit these data in order to predict accurate shapes.

Having determined a protein's accurate three-dimensional structure, computational researchers can move on to the task of determining its function in the cell, including how it cooperates with larger assemblies in biological systems. In the past decade, for example, coarse-grained molecular dynamics simulations have revealed how various proteins interact with the cell membrane. Moving forward, as greater computational power comes on line, we can expect larger and ever more detailed simulations, which will reveal valuable clues to the workings of proteins and cells.

6. Untangling Networks

Over the past decade, computational work has helped highlight the extent to which the programming of life relies on complex biological networks. Such networks include the basic metabolic cycles that have been remarkably conserved throughout evolution, as well as more complex networks that regulate various cell functions. The onset of many diseases, including cancer, is often related to a malfunction of these intra- and inter-cellular communication networks. Going forward, computational modeling and simulation, together with experimental efforts to decode complex feedback loops (the hallmark of regulation), are likely to play a major role in generating a deeper understanding of biological networks at all levels.

7. Tackling the Brain: From Artifical Intelligence to the Connectome

Computer programs inspired by the central nervous system and designed to "learn" from their environments have been around for some time now. They can steer robots through diverse terrain and are being increasingly successful at tasks such as computer vision and speech recognition. In recent years, some researchers have even attempted to replicate the behavior of neurons and the brain on a computer, with the intention to both gain an understanding of the brain and build more powerful computers.

Advances in computationally intensive imaging modalities, such as functional MRI and diffusion tensor imaging, have also increased our understanding of the brain's anatomy in health and disease. Unlike computers, which store data and perform operations at specific locations within their chips, we now understand that both memory/data and operations in our brain seem to be distributed over many neurons and synapses. It is this richness that may underlie the phenomena of associative memory and thinking. Patterns of closed loop sympathetic multi-neuron firing may eventually prove to be the basis of consciousness.

Among the most exciting developments of the last decade are the emerging efforts to map the human "connectome"—the connections among all of the neurons of the brain. Since the human brain network is more complex than that of the entire worldwide web, its complete mapping will likely be a challenge for the next decade or two. Like the human genome project of the 1990s, such a connectome project may mobilize a concerted effort and accelerate its achievement. Though we will likely face many challenges in taking on such a project, the potential benefits to be gained from a deep understanding of the (mis)function of the human brain and mind are extraordinary.

Since the human brain network is more complex than that of the entire worldwide web, its complete mapping will likely be a challenge for the next decade or two.

8. An Explosion of Computing Power: More Is Not Enough

Large-scale genome sequencing (and the shotgun method in particular) would be impossible were it not for massive computer power and associated efficient, fast algorithms for sequence alignment. Likewise, molecular dynamics simulations of large biological molecules, recently honored with a Nobel Prize in chemistry, depend on vast computational resources. So too does the dawning age of big data, with its need for efficient storage and quick access in order to probe for patterns and clues and gain a more comprehensive understanding of biological molecules, cells, tissues and organisms.

Thanks to increasing miniaturization, computing power has increased exponentially. However, the variety and complexity of biological information is growing much faster than computational power. Many researchers deal with computational limits by varying the computational resolution of their models and simulations as the scale of the task increases. But the desire to understand biological complexity in all its glorious details makes that approach less than satisfying.

Barring a true breakthrough that provides massive amounts of computational power more efficiently, we will soon be inundated with un-analyzable and therefore largely useless amounts of information. Most potential solutions remain, at this point, more hypothetical than practical—including quantum computers, light-based computers, or an unforeseeable but profound theoretical insight that can vastly improve algorithmic speed. I am somewhat more hopeful that the machinery of DNA/RNA/proteins may prove useful in computing. An all-purpose universal DNA computer likely remains too far in the future, but researchers are already turning to less ambitious lower-level utilizations, such as having complementary DNA probes efficiently search large DNA libraries. When and if any of these possibilities materializes, biomedicine will be ready with the data to take advantage of it.

9. Moving Forward with Molecular Prosthetics: From Synthetic Biology to Nanobiology

Just as computation has been instrumental in designing large-scale prosthetics, such as artificial hip bones, so too is it proving valuable for designing and synthesizing potential molecular bio-prostheses. For example, synthetic biologists are designing robust substitutes for certain amino acids, which can still be integrated into living proteins. The advent of nanotechnology and nanobiology may further close the enormous gap between large- and small- scale prosthetics by allowing useful intimate interfacing of biological and hardware components at many intermediate scales.

Imagine, for example, the creation of nanocircuits using sticky-end unpaired single stranded portions of DNA that can spontaneously attach with high specificity to complementary ends of other double helical chains; or the induced self-assembly of multi-cellular biological structures that can in turn control the assembly of various other nano-hardware pieces designed to interact with it; or novel materials such as the magical graphene, bucky-balls and carbon nano-tubes being integrated into tissues to collect electronic impulses. It is even conceivable that bionic ears could be designed to increase the range of audible acoustic waves.

Such possibilities raise societal questions as well: For example, to what extent do we want nanomachines roaming our bodies? But it's clear that as these issues are being explored, computation will play a role.

10. Confronting the Complexity of Cancer

No single field of medical research manifests more clearly the diverse, distributed and multifaceted nature of biological information than the study of cancer and cancer therapy, making it an excellent testbed for computational approaches. Indeed, the last decade has seen computational work contribute toward a better understanding of cancer on numerous fronts—from basic biology to diagnosis and treatment.

No single field of medical research manifests more clearly the diverse, distributed and multi-faceted nature of biological information than the study of cancer and cancer therapy, making it an excellent test-bed for computational approaches.

Computational researchers have worked hand in hand with experimentalists to map the pathways

and biological networks associated with the initiation, growth and spreading of cancer, as well as the critical junctions where these can be blocked via appropriate drugs or drug cocktails. High-end molecular dynamics simulations have started to reveal the mechanisms by which several oncogenic mutations in key cancer-related proteins (such as Ras and p53) wreak havoc in the cell.

On the diagnostic side, scores of biomarkers for many common forms of cancer have been identified with the help of computation and are being widely used. Since cancer is most likely induced by synergistic pathways, further work is needed to determine whether clusters of markers may jointly serve as more reliable indicators.

Insofar as treatment and therapy are concerned, novel combinations of drugs are being suggested for experimental testing based on computational screening. At the same time, physics-based techniques are helping clinicians deliver more accurate and localized radiation to cancerous cells.

And then there is the Big Data organizational and analytical effort to identify patterns of driver mutations and genes, which is beginning to tap into the vast data from cancer patient cohorts in electronic medical records.

Despite this progress, there remains a long road ahead. We need to computationally combine and integrate data types across a range of scales with the aim of predicting which mutational combinations will be oncogenic, and which drugs will benefit individual patients. The successes of the last decade have laid the groundwork for such an effort.

11. Decoding the Microbiome

The last decade saw our initial efforts to decode the microbiome—the microscopic life that resides within our own bodies. The importance of this internal ecological niche to human health is becoming increasingly clear. The microbiome includes not only well-known disease-causing microbes, but also thousands of species of bacteria, such as *E. coli*, with which we have a symbiotic relationship.

The discovery that transplanting specific bacteria from fat individuals to lean ones and vice versa can transfer tendencies toward obesity or reverse it suggest a much more subtle and deeper layer of interplay between humans and the huge variety of species inside them. Understanding how this phenomenon works at a molecular level will require intensive computation and may aid in our fight against obesity as well as other diseases. It could also provide a diagnostic tool, as enhanced populations of specific bacteria may signal disease onset as well as provide potential clues for treatment.

The time for directly channeling the vast diversity of life inside our bodies to our own medical advantage has come.

12. Building Life from Scratch: From Life's Origins to Synthetic Biology

With the various burgeoning datasets now available, computational researchers are digging into the origins of life as well as designing new forms of synthetic life capable of performing novel tasks.

Researchers are simulating computational models of various evolutionary theories to predict which are most realistic, including theories of how self-replicating molecules arose from a primordial mix of organic molecules; whether RNA or proteins—or for that matter, metabolism—came first; and how proteins have evolved. They are even exploring whether alternative evolutionary schemes are possible on Earth or elsewhere in the cosmos.

At the same time, the last decade has seen the launch of synthetic biology as a powerful field. Achievements include computational models that predict the minimum genome required to support life and validation of that prediction by inserting the genome into a cell to produce a self-replicating organism. In addition, researchers are designing engineered organisms capable of sensing environmental toxins or producing biofuels.

As these fields progress, their revolutionary implications will likely amaze us in ways I cannot begin to imagine. □

TOP 10 RETROSPECTIVE

Reflections on a Decade of Biomedical Computing

By Kristin Sainani

The first issue of this magazine (June 2005) featured a story called "Top Ten Challenges of the Next Decade" written by **Eric Jakobssen, PhD**, who had recently left his position as Director of the Center for Bioinformatics and Computational Biology in the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH).

Today, as we near the end of that decade, we've asked ten domain experts to weigh in: **How well has each of these challenges been met? And, with the benefit of hindsight, were they the right challenges in the first place?**

CHALLENGE 1

In Silico Screening of Drug Compounds

Status 10 years ago:

In 2005, Jakobsson hoped the next ten years would see researchers advance our ability to "predict the efficacy and side effects of lead compounds using computer modeling and simulation," thereby reducing the need for human testing while also saving time and money spent in the laboratory.

Update by:

Arthur Olson, PhD, professor in the Department of Integrative Structural and Computational Biology at the Scripps Research Institute

Progress made:

We've made a lot of progress in terms of how many people are doing *in silico* screening. There seems to be a larger and larger community of people doing virtual screens, many of whom are not computational chemists. The tools have improved because the toolmakers have had to respond to the demands of all these users. The chemical libraries have become larger, better characterized and more focused. The peer-reviewed science using virtual screening that has been published over the past 10 years has also been staggering. I believe that structure-based drug design has informed development of many of the new drugs that have come out in recent years. I'm guessing that this was the case with the Hepatitis C antiviral drug from Gilead, which made the news recently as a cure for the disease.

In terms of specific advancements, we've improved the ability to rank the results of screening. We do broad screens using quick docking methods and then pass the top candidates along for evaluation using more computationally intensive methods (calculating molecular dynamics-based binding free energies). While the basic theoretical framework hasn't changed that much in the past 10 years, properties that were difficult to estimate 10 years ago are now possible because computing has become so much more powerful and available. The docking algorithms have also gotten incrementally better. For example, we've improved how we model water during a docking calculation; this can make a significant difference in which poses are selected. We're also making better use of parallel computing—the fact that the analysis by molecular dynamics can be broken up into multiple runs and information exchanged between them can improve sampling and throughput.

Challenges ahead:

We still face the challenge of designing synthetic drugs that modulate protein-protein interactions. Most successful small molecule drugs at this point have been targeted to individual protein active sites. While solving this problem won't require any new physics, it will require new algorithms that can model complex interactions efficiently.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? Yes, this was the right challenge a decade ago, and it's still the right challenge for the next decade. The payoff could be very, very large in terms of human health.

CHALLENGE 2

Predicting Function From Structure of Complex Molecules

at an engineering level of precision

Status 10 years ago:

In 2005, molecular simulation and analysis methods could "capture the essence of the mechanism of biomolecular function, but could not predict that function with quantitative accuracy," Jakobsson wrote. He hoped the decade would lead to improved capability in this regard, enabling a precise understanding of the consequences of mutations and other biological variations, and the ability to design molecules for medical nanotechnology.

Update by:

Predrag Radivojac, PhD, associate professor of computer science and informatics at Indiana University

Progress made:

I think this particular challenge has not been met if we look strictly. Since 2005, we have broadened the concept of function tremendously and now understand the "breadth" a lot better. Today, we think of function in more specific terms (such as whether a residue binds to a protein or DNA) and at more levels (for example, a protein may participate in a specific reaction, in the cell cycle, or in a disease). As a result, we have not reached this goal because of the many new challenges we have discovered along the way.

Still, we have made a lot of progress in the past 10 years. We can now predict many aspects of function surprisingly accurately, such as certain metal-binding residues, catalytic residues, ligand-binding sites and protein-DNA binding sites. All these different aspects of function have some specificities in their methods; there's no silver bullet to address all of them. But each of these little sub-fields has pushed things forward. I believe that in the next 10 years we will be able to deliver on the goal of predicting the consequences of mutations and sequence variants; and we will see some fascinating discoveries.

There have also been individual success stories, where researchers were able to achieve an engineering level of precision of function prediction. For example, David Baker's group designed a protein with particular functionality *de novo* by structural modeling of an enzyme with increased catalytic activity. This is exactly what this challenge had in mind.

Challenges ahead:

The 2005 article does not talk about the fact that proteins are dynamic molecules. To predict function at an engineering level of precision, we will have to have some sort of dynamic models both at the micro and macro levels, including large irregular movements. And this will require advancements in mathematical, computational, and physical approaches. Current methods do not scale. We cannot model motions of proteins at the appropriate granularity and length of time in order to be able to extract the signatures of motion that would be predictive of function. Another important challenge is that structure data are noisy, reflecting many experimental artifacts. We have to find the right statistical and machine learning approaches to model the uncertainty in the structure data, and then integrate it with other types of data in order to be able to infer function.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? This challenge was the right one. I would have refined it slightly, to: predicting function from structure *and dynamics* of complex molecules. We discovered that we have a lot of sub-problems to solve. The stars need to be aligned for us to be able to deliver on this challenge. But I definitely think it was the right challenge.

CHALLENGE 3

Prediction of Protein Structure

Status 10 years ago:

In 2005, Jakobsson noted that there were many more known protein sequences than structures. He hoped that through a combination of accelerated experimental structure determination and improved techniques for mining known structures to determine the rules for predicting unknown structures, researchers would gain the ability to assign a structure to every sequence. Jakobsson believed this achievement would advance the field of biomedicine in many ways.

Update by:

Adam Godzik, PhD, professor and program director of bioinformatics and systems biology at the Sanford Burnham Medical Research Institute

Progress made:

I think this is the challenge where the most progress has been made. We don't have a tool that works every time; but, compared with 10 years ago, the progress has been amazing. Ten years ago, you would look at predicted structures and just cringe; now some of them are as good as real.

Much of the progress is due to David Baker's efforts with the Rosetta algorithm for energy-based predictions. The tool doesn't work for every case, but when it works, it works fantastically. The second big thing that happened is people started to realize the importance of distant homology prediction (finding sequence relationship with proteins that have already been characterized experimentally). This approach is actually much more powerful, because in addition to giving clues about 3D structure, it also tells you about what the protein does. With advancements using hidden Markov models, we can now recognize much more distant relationships than we did 10 years ago.

The CASP (Critical Assessment of Protein Structure Prediction) experiments also gave the field an enormous push, because blind tests and judges allow you to actually see what is working and what is not.

Challenges ahead:

Currently the energy-based methods work well for cases where one energy term dominates, but often get it wrong when there are multiple opposing forces. We'd like to advance this to the point where it works in every case. For distant homology prediction, we're still missing a lot. Sometimes after a structure is solved experimentally, we realize that there were homologies we missed.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? I think it was perfect. And, because of CASP, the progress has been verified every year.

CHALLENGE 4

Accurate, Efficient, and Comprehensive Dynamic Models of the Spread of Infectious Disease

Status 10 years ago:

In 2005, Jakobsson saw an opportunity for modeling to take advantage of the extensive data that had been gathered on the spread of infectious disease and the consequences of various strategies of intervention. Such models, he

hoped, would provide a basis for rational, informed, real-time decision making in combating natural epidemics and bioterrorist attacks.

Update by:

Stephen Eubank, PhD, professor in the Virginia Bioinformatics Institute and Population Health Sciences department at Virginia Tech.

Progress made:

There are three big areas where there have been some substantial changes: (1) surveillance (what feeds into the models), (2) the models themselves, and (3) the use of modeling evidence to inform decision making in government agencies.

In the past decade, the scope of surveillance has broadened from simple factors, like vaccination, to more complex social behaviors such as whether or not people stay home from work when they're sick. We are beginning to get a better handle on measuring people's behaviors and how they change during an outbreak.

The models themselves have also advanced. Models 10 years ago usually assumed homogenously mixed populations. Now, we're using high-resolution network-based models that model every single person in a large region. People come and go, and they change their behaviors in reaction to things they hear on the news or their perceptions of what's going on around them. So the system's not stationary and it's not well mixed. And the new network-based models are able to take both of those things into account.

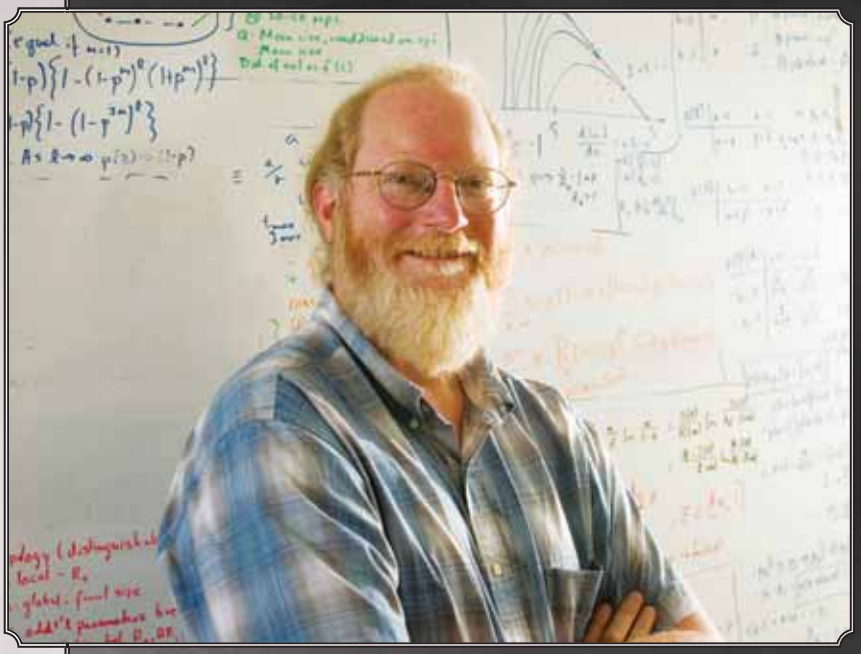
Finally, we've made a lot of progress in getting decision makers to pay attention to what the models say. One of the things that might have made people hesitate before is that the models were so generic it didn't seem as if they could really be applied to any specific circumstance. But by creating these highly resolved models that are representative of particular places and particular outbreaks, I think we've managed to convince folks that, yes indeed, these models should be taken seriously.

Challenges ahead:

I think the jury is still out on a lot of the new surveillance techniques; there's something there and there has got to be some way to use the flood of information coming at us from sources like social media, but I don't think we've perfected that art yet. We also need faster turnaround on traditional disease reporting surveillance, which hasn't been brought into the electronic era. There's still a one to two week delay in getting really good, accurate information from emergency departments and clinics up to a scale where the modelers can get hold of it.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? I think it was right on the money.



CHALLENGE 5

Intelligent Systems for Mining Biomedical Literature

Status 10 years ago:

In 2005 there existed no efficient and effective way to organize data from biomedical literature into computable databases from which accurate interpretive and predictive models could be constructed. Jakobsson hoped the ensuing decade would see better access to the abundance of information about the functioning of genes, gene products, and cells that was then buried in published papers.

Update by:

Graciela Gonzalez, PhD, associate professor of biomedical informatics at Arizona State University

Progress made:

In the past decade, there has been significant progress on this challenge. We've ad-

vanced the furthest in our ability to recognize named-entities—specific genes, diseases, chemicals, drugs, and other entities—mentioned in biomedical text. For many entities, the problem is considered pretty much solved. By retraining machine-learning-based tools such as our system, BANNER (http://banner.sourceforge.net/), or others like it, one can have an entity recognition system with little effort. We've also progressed along the next step in the pipeline: entity normalization. For example, once you find a gene, you need to know exactly which gene it is referring to out of multiple possible mappings (homologues for different species). There are different systems available for different entities. For example, the NIH's National Center for Biotechnology Information (NCBI) recently released DNorm (Disease Name Normalization), an automated tool for determining specific diseases mentioned in biomedical texts. Finding and normalizing entities are key steps towards enabling intelligent systems. There has also been significant progress toward integration of data from the literature with experimental results. For example, the NCBI links GenBank sequence data or genome data in the GEO database back to the literature in PubMed.

Challenges ahead:

Where the challenge remains is in connecting all these pieces of knowledge into larger, complex systems and hunting out causal relationships between entities. When I moved to this field in 2005, we wanted to use text-mining tools to model and make inferences from biological pathways, such as protein-protein interaction pathways; but we couldn't do that as there was no system available to extract them. This still remains a challenge. Nobody has solved the problem of pathway extraction. The NCBI links notwithstanding, the challenge remains how to coherently integrate all the experimental data being produced, such as whole genome sequencing data, with knowledge from the literature so a scientist can automatically hone in on support for or against a hypothesis or novel theory. In short, there is still a large gap from data to discovery.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? I think this was the right challenge because the literature remains a major source of valuable yet hidden knowledge. Every paper that's written represents months or even years of work by a team of scientists. But it's practically impossible to find them all without a lot of help from an automated system.

CHALLENGE 6

Complete Annotation of the Genomes

Status 10 years ago:

In 2005, so-called "complete" genomes were far from "complete," Jakobsson wrote. He hoped the research community would select eukaryotic and prokaryotic model organisms for a focused attack on complete annotation, and use all experimental, bioinformatics, and data-mining tools on these organisms. As a sequel to complete annotation, he challenged researchers to elucidate the target organisms' complete metabolic, signaling, and homeostatic pathways and networks.

Update by:

Terry Gaasterland, PhD, professor of computational biology and Genomics Director, Scripps Genome Center University of California, San Diego

Progress made:

In the past ten years, the genomes of many different species have been sequenced and the 10,000 genomes project headed by David Haussler at University of California, Santa Cruz, is making progress toward sequencing the genomes of many more. But perhaps the biggest thing that we have done in the past decade is to become capable of dealing with incomplete genomes. We've started to understand that no genome is ever truly complete. Every individual has its own genome and all these genomes are inter-related—so we have local variation, and we have large-scale variation. The best we can do are snapshots and draft genomes. As a community, we're becoming comfortable with this and even building tools to leverage this knowledge.

The single most important contribution to genomics over the last 10 years, beyond the data, is the one-stop shopping that has emerged through the UCSC genome browser project. The community needed a common way to view, manipulate, and manage genome data; and David Haussler's team built that. They're providing production-quality comparisons and calculations across many prokaryotic and eukaryotic genomes. Also, in the past decade, the development of short-read high-throughput sequencing has been of utmost importance. The community has exhibited such creativity in using the high-throughput sequencing—to sequence anything from new genomes of new organisms to RNA to DNA binding sites to nascent transcripts.

There has also been enormous progress toward elucidating target organisms' complete metabolic networks. For example, in 2007, Bernard Palsson's lab at University of California, San Diego published a detailed *in silico* model of human metabolism. Using that model and others, researchers can simulate the effect of virtual knockouts as a prelude to laboratory testing.

Challenges ahead:

Ever bigger datasets need ever faster and more efficient data storage arrays. The University of California, San Diego, Super Computer Center presents a prototype of how to provide this kind of computing power to a local community. What we've done here is we've all bought pieces of the larger system. For example I spent $20,000 to buy nodes and because of that I have access to a two million dollar computer. I'd love to see this happen over and over again at many universities. Another challenge is clinical phenotyping. For annotating the human genome in a disease-aware way, the computational biologists have to be in lock step with the physicians.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? Because we've realized that no genome is ever complete, its annotation can never be complete as well. Thus, Jacobsson's goal was per se unattainable. Nevertheless, I think it was exactly the right challenge. You lay out the ideal, you shoot for Mars, and you might get to the moon.

CHALLENGE 7

Improved Computerization of the Healthcare Delivery System

Status 10 years ago:

Jakobsson wrote in 2005, about his concern that "the relatively primitive information technology environment supporting the delivery of health care" resulted in extra expense and avoidable error. He called for "a nationally interoperable system of medical records to support transferable patient records, diagnosis and treatment based on integrating the patient record with relevant basic and clinical knowledge, and efficient patient monitoring." The deployment of personalized medicine would be, he believed, a logical consequence and extension of this computerization.

Update by:

Lucila Ohno-Machado, MD, PhD, professor of medicine, University of California, San Diego

Progress made:

Of the top-ten challenges Jakobsson listed in 2005, this is one of the ones that has made the most progress. In the past decade, electronic health

records (EHR) have been widely adopted, thanks to large investments from the government. In 2005, we were talking about institutions not even having electronic health records; now we're talking what to do with them. There are still challenges, but we are at the next level now. So, it's a very exciting time in our field.

The next major breakthrough is also on the horizon. In late December, the Patient-Centered Outcomes Research Institute (PCORI) awarded $93.5 million for the creation of PCORnet, the National Patient-Centered Clinical Research Network. The network will securely link EHR data for millions of patients, which will enable large-scale comparative effectiveness research—figuring out which types of medical care work best. Someday, EHR data may even be linked to bio-samples, such as DNA sequencing data or proteomics, with an eye toward personalized medicine. With huge numbers of patients, we will be able to correlate responses to particular therapies with very specific biomarker profiles.

Challenges ahead:

The technology for enabling preservation of privacy has evolved a lot, and that's removed many barriers. But we still need to improve data quality and standardization. We need to promote a broad understanding from patients, clinicians, administrators, and researchers, of what it takes to make these data useful.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? Yes, this was the right challenge for the past decade. But it will also remain a key challenge for the next decade. The challenge will not be as primitive, but it will still be there. It's not going to be solved overnight.

CHALLENGE 8

Integrating Computational Tools to Make Systems Biology a Reality

Status 10 years ago:

In 2005, Jakobsson observed that "many useful tools for systems biology have been created, but they are not integrated into computational environments that provide for automatic interaction of multiple programs and functionalities to address generally useful issues in biomedicine." The tools themselves also need improvement in their scope of applicability, computational efficiency, and ease of use, he wrote. The aim: a much-needed computational environment for information-based modeling of pathways, networks, cells, and tissues.

Update by:

Markus Covert, PhD, associate professor of bioengineering at Stanford University

Progress made:

There has been a lot of motion in the space particularly from pathways to cells and cells to tissues. I wouldn't say this challenge has been accomplished, but it's going well. I remember that during the first funding initiative on multiscale modeling, that term was still being defined, even at the programmatic level. But I don't think people would have that same confusion now.

In terms of tool integration, we still don't have a unified, integrated, seamless situation. Centers for systems biology are bringing different professors together, but there isn't a one-stop shop

where you can find all the tools you need for your modeling interests. People have tried to start a biomodels database, but it's challenging because you don't always know in advance what you will want to store. So it's still largely up to individual teams to make their models widely accessible.

Along these lines, we developed a comprehensive whole-cell model that predicts phenotype from genotype (*Cell*, July 2012). For this model, we've been trying to give people access at a variety of levels. We've made a knowledge base that is structured to hold all the information that you would need to run a model.

Challenges ahead:

Many problems could be solved if systems biology would reach even further outside of itself. It's already a highly interdisciplinary field, but we need to take another major step forward that would literally involve talking to people who you think you have nothing in common with. For example, systems biology tools could be greatly improved by an influx of industry talent. The best coding in the world is not happening in systems biology; it's happening at companies like Google and Facebook. For our whole-cell model, we hired a software engineer from Google for six months; and I was very impressed by how much we needed that software help. I have also realized that we have a lot of visualization tools that can be used for education, but few that can be used for exploration and discovery. To develop these more sophisticated visualization tools, we're going to need artists and graphic designers, as well as coders.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? This challenge was not the best specified one, probably out of necessity, but Jakobsson definitely did show some foresight. We're in that space now; we're progressing along the vector he outlined.

CHALLENGE 9

Tuning Biomedical Computing Software to Computer Hardware

Status 10 years ago:

In 2005, biomedicine used substantial computing resources at all levels, from the desktop to high-end supercomputing centers, but "a large fraction of these resources are not efficiently used, as the hardware and software are not tuned to each other," Jakobsson wrote. He believed that addressing this problem would allow research to advance more rapidly.

Update by:

Vijay Pande, PhD, professor of chemistry, structural biology and computer science at Stanford University

Progress made:

This challenge is an ongoing issue. People have gotten much better at tuning software to hardware, but the challenge has gotten even harder. As time goes on, the hardware is getting more heterogeneous, and getting the best performance out of it requires more effort. So, as people have advanced on this challenge, the goalposts have been pushed back as well.

On the hardware side, the key breakthroughs are advances in graphics processing units (GPUs) and in how people handle large amounts of memory. But GPUs are very specialized. And to roll out an engineering algorithm and have it run on them very quickly is a challenge. We and others have been trying to push the area of domain-specific languages, which are intentionally not general purpose and can easily be ported to GPUs. This approach has been quite powerful, allowing us and others to rapidly create code that still executes quickly. So, these languages have been a major breakthrough on the software end.

Challenges ahead:

With each generation of new GPUs, we have to

re-tune our domain-specific languages. So the constant maintenance of doing this is an ongoing challenge. Sustainable funding is also a challenge. People think software is written and then it's done. But software is like your lawn—it needs constant maintenance and upkeep to make sure it remains in good shape. With our current funding system, it's very difficult to sustain codes over long periods of time. Many people have chosen to commercialize their software. But this leads to a closed-off system that slows the community down. Compared with a decade ago, the NIH is doing much better on this issue, but I would like to see even more progress.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? It certainly should be one of them. Whether it should be in the top 5, 10, or 20, could be debated. But it's certainly a significant challenge.

CHALLENGE 10

Promoting the Use of Computational Biology Tools in Education

Status 10 years ago:

To help forestall a likely shortage of quantitatively competent researchers, Jakobsson called for the adaptation of biomedical computing tools to education at all levels in order to capture their power to motivate youngsters to pursue biomedical research careers. He believed that the same developments that were making biomedical computing tools useful to experimental researchers could also make them the basis of compelling problem-solving educational environments for students.

Update by:

Brian Athey, PhD, professor and chair of computational medicine and bioinformatics at the University of Michigan

Progress made:

I think we made good progress at getting computational tools out there, thanks largely in part to the National Centers for Biomedical Computing (NCBCs). The imaging pipeline that came out of the Center for Computational Biology NCBC, LONI, was a key to the Alzheimer's Neuroimaging Initiative. Andrea Califano's network biology tools in cancer have made a dramatic impact on our understanding of systems biology and cancer. The National Center for Biomedical Ontology put together collections of ontologies that are being used worldwide.

But it is a fundamentally different world that we're living in now compared with 2005 because of the proliferation of data. A decade ago, we were focused on computing tools and software; that focus has now been eclipsed by big data analysis. The computer is more in the background; the data and information are in the foreground.

Challenges ahead:

There's more of a need than there was even 10 years ago for training. Most biomedical researchers, from the basic to the clinical sciences, are dealing with heterogeneous digital data. They need to learn how to access and analyze these data. We need to bring forward basic exposure and instruction about data science at all levels from undergraduates through to the faculty.

20/20 Hindsight:

Given the advances of the last decade, was this challenge the right one? It was the perfect challenge, very important to put on the list. And it's important to keep it on the list for the next decade, with a new focus on data and information.

Under TheHood

BY JUSTIN B. KINNEY, PhD

Mutual Information: A Universal Measure of Statistical Dependence

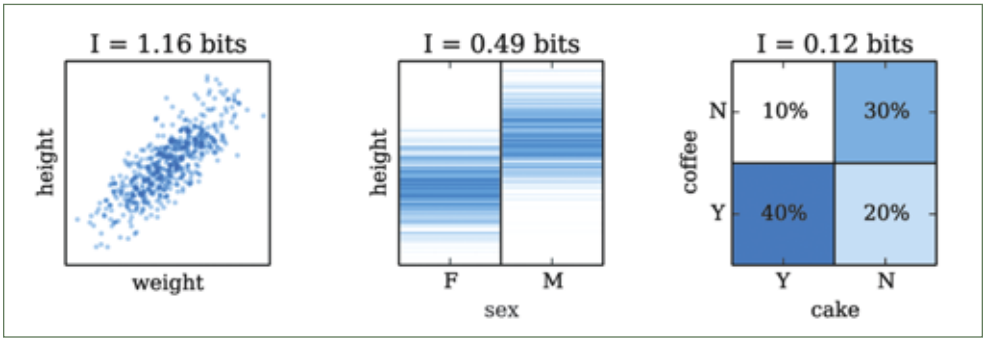
A deluge of data is transforming science and industry. Many hope that this massive flux of information will reveal new vistas of insight and understanding, but extracting knowledge from Big Data requires appropriate statistical tools. Often, very little can be assumed about the types of patterns lurking in large data sets. In these cases it is important to use statistical methods that do not make strong assumptions about the relationships one hopes to identify and measure.

In this tutorial we consider the specific problem of quantifying how strongly two variables depend on one another. Even for data sets containing thousands of different variables, assessing such pairwise relationships remains an important analysis task. Yet despite the simplicity of this problem and how frequently it is encountered in practice, the best way of actually answering it has not been settled.

One standard approach is to compute the Pearson correlation coefficient. Unfortunately, Pearson correlation has severe limitations. First, it only applies to variables that are continuous real numbers; it cannot be used when either variable represents a discrete category, such as gender. Second, the assumptions underlying Pearson correlation are violated by relationships that are nonlinear or have many outliers. Such violations can result in correlation values that conflict with more intuitive notions of dependence.

A more general way of quantifying statistical dependencies comes from the field of information theory. This branch of mathematics arose from a classic 1948 paper by Claude Shannon titled "A Mathematical Theory of Communication." Although Shannon's immediate purpose was to describe information transmission in telecommunications systems, his work illuminated deep truths that have since had a profound impact on fields as diverse as engineering, physics, neuroscience, and statistics.

Shannon argued that the concept of "information" can be formalized by a mathematical quantity now known as "mutual information." Mutual information quantifies the amount of information that the value of one variable reveals about the value of another variable. It is measured in units called "bits:" A value of zero corresponds to no dependence whatsoever, while larger values correspond to stronger relationships.



Data from three hypothetical relationships with corresponding mutual information values shown. Mutual information can quantify dependencies regardless of whether one or both of the variables in question are continuous numbers (e.g., a person's height and weight) or discrete categories (e.g., a person's gender or after-dinner food preferences).

Importantly, mutual information retains its fundamental meaning regardless of how nonlinear a relationship is. Mutual information can also be computed between variables of any type, be they continuous or discrete. Some hypothetical relationships illustrating this are shown in the accompanying figure.

Computing mutual information from data is complicated, however, by the difficulty of estimating a continuous probability distribution from a limited number of samples. Fortunately, there are algorithms that can solve this problem well enough for many practical purposes, and estimating mutual information becomes easier the more measurements one has.

Mutual information therefore provides a sensible alternative to Pearson correlation in many Big Data settings. As better ways of estimating mutual information are developed, this important concept from information theory is likely to become increasingly useful in data analysis efforts, both in science and in industry. □

DETAILS

Justin Kinney is a Quantitative Biology Fellow at Cold Spring Harbor Laboratory.

His research combines theory, computation, and experiment in an effort to better understand quantitative sequence-function relationships in molecular biology. An expanded discussion of mutual information and its merits as a statistic can be found in the recent paper, Kinney, JB and Atwal, GS (2014) Equitability, mutual information and the maximal information coefficient, *PNAS* 111(9):3354-3359.

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

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

Stanford University
318 Campus Drive
Clark Center Room S271
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

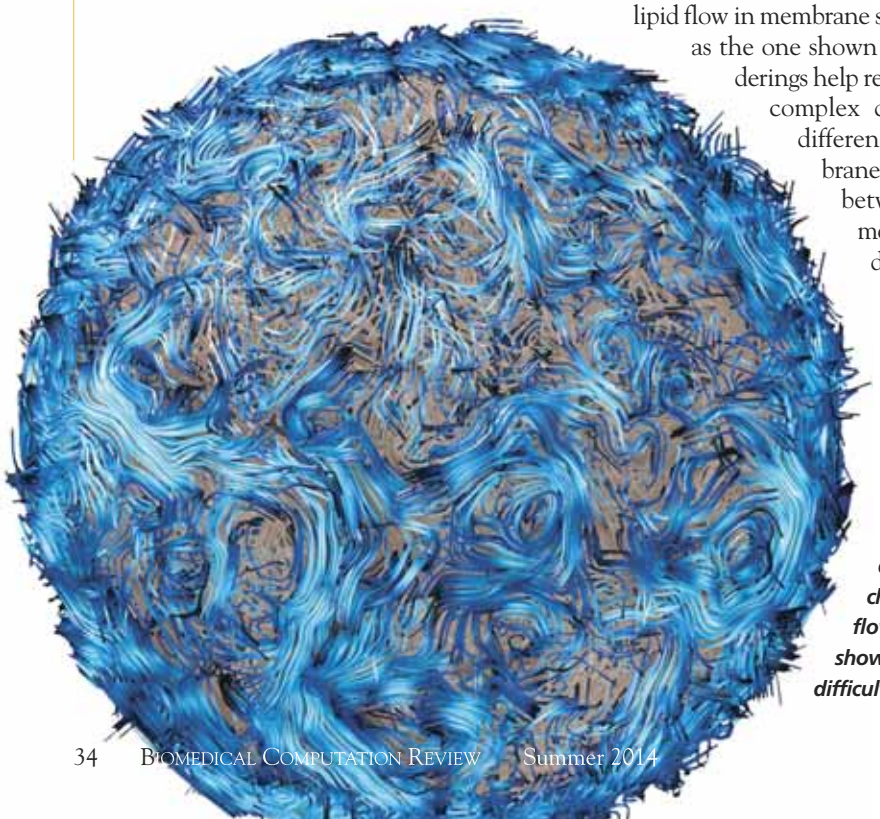
Streamlining Lipids

As computational power grows, researchers can model and simulate larger and larger molecular complexes. To visualize such systems in action, **Matthieu Chavent, PhD**, a postdoc in Mark Sansom's laboratory at the University of Oxford has turned to a physics-based strategy often used to depict atmospheric flows and oceanographic currents. It's called streamline visualization.

"It is really well-adapted to large molecular systems," says Chavent. Indeed, he has used the method to render images of lipid flow in membrane simulations, such as the one shown here. Such renderings help researchers see the complex dynamic among different types of membrane lipids as well as between lipids and membrane-embedded proteins. The insights gained could lead to a greater understanding of how drugs and viruses interact with membranes.

To create the streamlined visualizations, Chavent divides a detailed membrane model into a grid of cells. He then associates a vector to each cell, creating a vector field. The streamline approach connects the vectors together. "Instead of focusing on every lipid, which is quite complicated and may blur the view, we obtain a larger view to focus on the membrane as an ensemble," Chavent says. The approach may also prove valuable for visualizing smaller systems consisting of many molecules, such as water flow around macromolecules, Chavent says.

Chavent is currently developing a streamline visualization plugin that will work with VMD, a popular molecular visualization program. The method is freely available at the following address: http://sbcb.bioch.ox.ac.uk/flows/. □



This streamline visualization, which recently won first place in the Biophysical Society Art of Science Image Contest 2014, displays lipid movement in a spherical vesicle membrane bilayer. The outer leaflet is depicted by colored streamlines—white lines are higher velocity than blue. Brown lines represent the inner leaflet with speed not shown. The rendering is based on a lipid vesicle model developed by Syma Khalid, PhD, senior lecturer in computational chemistry at the University of Southampton. The visualization reveals flowing movements as well as vortices around the proteins (not shown) that move like rafts in a lipid sea. These circular movements were difficult to see using other methods. Courtesy of Matthieu Chavent.