DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

NCBCs Take Stock AND Look Forward

Fruitful Centers Face Sunset

PLUS:
Getting It Right:
Better Validation

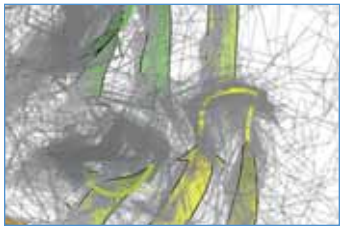
2004 – 2014 | 2014 – ?

Fall 2012

contents

ContentsFall 2012

Cover Art: Created by Rachel Jones of Wink Design Studio using:
Tree image © Jan Martin Will | Dreamstime.com, and in-house photography.



Fall 2012
Volume 8, Issue 3
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, Katharine Miller, Kristin Sainani, PhD

Community Contributors
Russ B. Altman, Ronald J. Nowling, Trevor M. Cickovski

Layout and Design
Wink Design Studio

Printing
Advanced Printing

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

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

Biomedical Computation Review
is published by:

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 RUSS B. ALTMAN, STANFORD UNIVERSITY

The NCBC Centers: Incubators for the Next Generation of Science and Scientists

In this issue of *Biomedical Computation Review*, we feature a look at the NIH Roadmap National Centers for Biomedical Computing (NCBC) program. The NCBC program was a response to the recommendations of a pivotal report[1] entitled Biomedical Information Science and Technology Initiative (BISTI). In that report, the authors recognized the need for NIH to support the creation of "an intellectual fusion of biomedicine and information technology" and support "ways to discover, encourage, train and support the new kinds of scientists needed for tomorrow's science. In their prescient report, they called for four interventions:

1 To establish between 5 and 25 National Programs of Excellence devoted to all facets of this emerging discipline, who will play a major role in educating biomedical-computation researchers.

2 To make the growing body of biological data available for study and use.

3 To provide resources for basic research in computational methods.

4 To foster a scalable national computer infrastructure to support biomedical research.

The many payoffs from the NCBC program are described in the cover story of this magazine. But I think that the program's most important legacy is its impact on human capital. Each NCBC center has created an intellectual home where a new generation of biomedical computational scientists has been created and nurtured. The centers have created a heretofore absent ecosystem that allows scientists skilled in informatics and computation to thrive. The web of people in this ecosystem includes:

- Faculty leaders striving for methodological innovation to solve big problems in biomedical science;
- Graduate students in interdisciplinary programs interacting with other students who share their passion for computation, biology or both;
- Post-doctoral fellows working in rich intellectual environments and defining the new questions and new methodological directions that will drive the field in the next 10 years;

I think that the program's most important legacy is its impact on human capital. ... The centers have created a heretofore absent ecosystem that allows scientists skilled in informatics and computation to thrive.

- Professional software engineers who've found a career path in biomedicine that offers rewards not available in more traditional areas such as finance, entertainment, social networking, and defense;
- Scientific staff who are training biologists and physicians to use powerful new software tools, and who have learned how to disseminate the fruits of their centers effectively and globally.

And, perhaps most significantly:

- NIH Program and Scientific officers who have helped lead the NCBC program and begun to learn the special features of this field—the ways in which it is similar to the other science at NIH, and the ways in which it requires special consideration because of its special technical content, its focus on methodological innovation, and its tendency to engineer artifacts (software, databases, novel hardware architectures) that require ongoing support. A well-informed and experienced set of research administrators is absolutely critical for the success of this endeavor.

As NIH leadership ponders the end of the first 10 years of the NCBC program, and considers how to evolve the NIH mission in biomedical computation, one priority must be the continued nurturing of an intellectual ecosystem for the field. It is this ecosystem that will ensure the success of biomedical research in the digital era. □

1 http://www.bisti.nih.gov/library/june_1999_Rpt.asp

SimbiosNews

BY KATHARINE MILLER

The Ease and Grace of OpenSim 3.0

OpenSim, the neuromuscular modeling and simulation software, is now available in a new digit: 3.0. The change (up from 2.4) reflects significant improvements that make this open source tool more intuitive and efficient for typical users as well as advanced developers, says **Jen Hicks, PhD**, OpenSim's research and development manager.

"We want it to do easy things easily and hard things gracefully," says **Matt DeMers**, a graduate student of mechanical engineering at Stanford University who is both a developer and user of OpenSim. With this version, that hope has become a reality, he says.

In OpenSim, users develop models of musculoskeletal structures and create dynamic simulations of movement using either a graphical user interface (GUI) or, for more advanced developers, an application programming interface (API). Version 3.0 includes numerous improvements to both the GUI and the API as well as performance improvements. "The OpenSim development team is always looking for better, faster, more accurate ways to calculate things," DeMers says.

GUI Usability

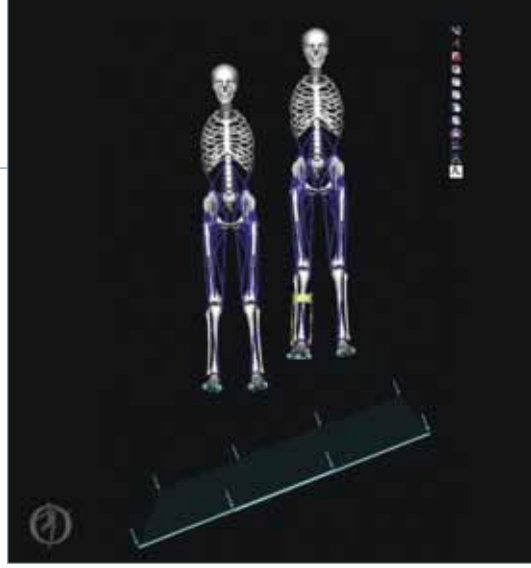
The biggest advance in OpenSim 3.0 is the ability to edit and iterate models quickly and intuitively in the GUI. "When someone is creating a model and simulation, there's a lot of tweaking that happens," Hicks says. Often, modelers are adjusting a model to better match a particular person's geometry—the length of bones and muscles, the size of the torso, the angle of the joints. In earlier versions of OpenSim, making such changes was somewhat cumbersome: Users had to close the model in OpenSim, edit and save a new text file of the model, then navigate to that file in OpenSim and re-open it. But now, right in the GUI, when users click on a muscle, bone, body or joint, they can see its properties (where it's located, what it looks like, how strong it is), change a number, hit "enter," and immediately see the updated model, providing instant feedback on whether the change is a good one. "And you can do that over and over to iterate very quickly," DeMers says. "That's the big gain with OpenSim 3.0."

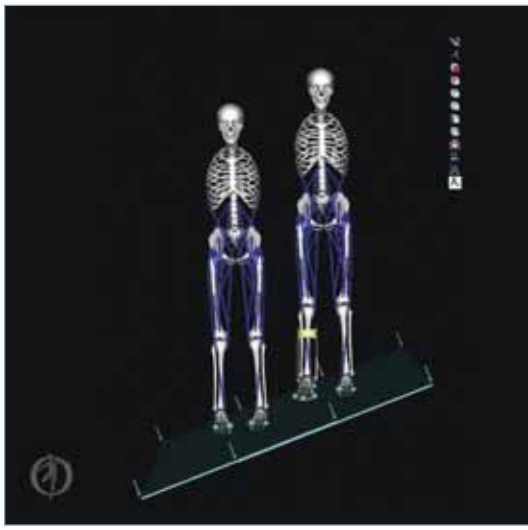
It helps that the GUI now also provides improved visualization tools as well. "You can look at forces, positions and vectors in a file all day long and not know what it really means," DeMers says. "But when you can see it in the GUI, you can tell whether it looks reasonable or resembles something physical as you'd expect it to appear."

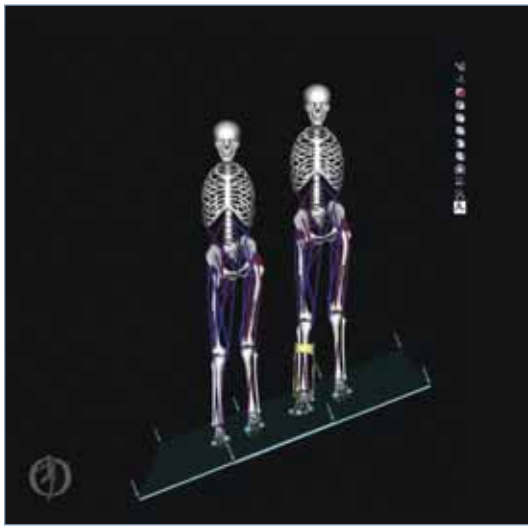
It's also now possible to fine tune functions—the math behind how a joint or muscle moves—right in the GUI. "Our bodies have messy geometry; they translate and rotate at the same time; they move in strange ways," DeMers says. It's handy to be able to adjust a function to fit a particular scenario, right in the GUI.

The GUI also features a number of other usability improvements, Hicks says, such as the ability to drag and drop models and motions; access a help button relevant to the tool you're currently using; and show a list of recently opened files. "These small usability improvements are big productivity boosters," she says.

The importance of these GUI changes shouldn't be underestimated, DeMers says. "At the end of the day, you'll send your model through a fancy numerical pipeline, but the stuff





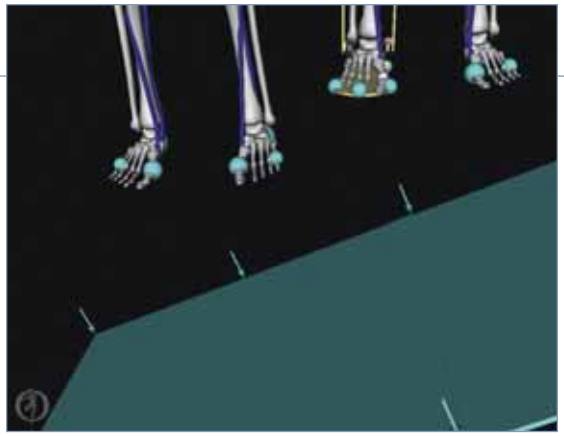


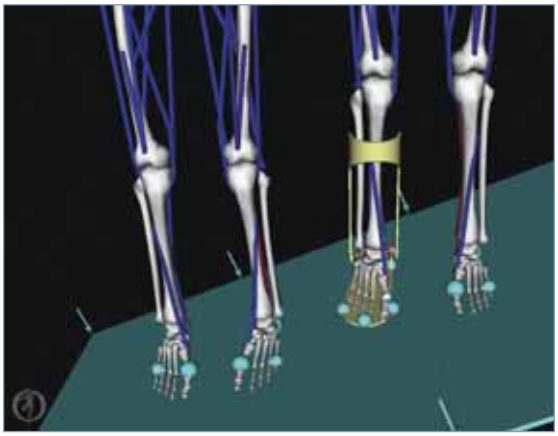
OpenSim 3.0 makes it easier for users to add new model components, such as the muscle reflex controller used in this simulation of a drop landing with an ankle brace. The controller activates muscles based on how fast they are lengthening. The 3.0 GUI also lets users quickly edit parts of a model such as the weight of a backpack, the strength of the model's muscles, or the stiffness of the ankle brace. Courtesy of Ajay Seth, Matt DeMers, and John Rogers.

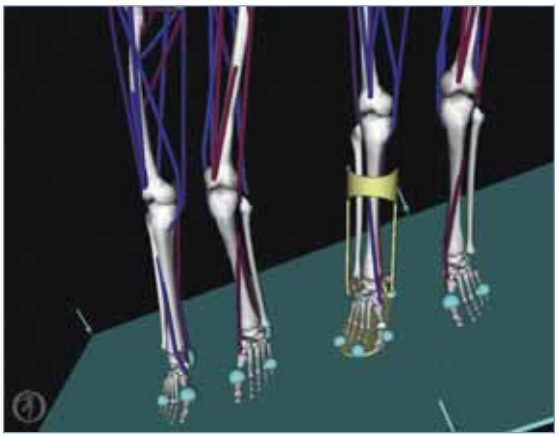
DETAILS

OpenSim 3.0 is available for download at https://simtk.org/home/opensim). It is funded by Simbios; the National Center for Simulation in Rehabilitation Research; and the DARPA Warrior Web Program.

The lead OpenSim application architect is Ayman Habib, PhD. The lead API architect is Ajay Seth, PhD. See the full OpenSim team at http://opensim.stanford.edu/about/people.html.







that takes a lot of time is the nitty-gritty stuff—changes and modifications and checking that you've configured everything correctly. Those front-end tasks all happen in the GUI, and in 3.0 it's a much nicer experience."

MATLAB and GUI Scripting

OpenSim provides ample core functionality and out-of-the-box tools for creating models of bodies, joints, and muscles. "And that's what most people use," Hicks says. But researchers might want to do something a little bit different that they can't do with the core tools, such as make a new type of muscle. Or they might want to do the same set of operations over and over again (batch operations). Until now, such things could only be done with tedious parsing of text files or using C++ programming in the OpenSim API. But C++ has a steep and often frustrating learning curve, Hicks says.

To address that limitation, OpenSim 3.0 now offers two other options: writing scripts with a Python interface in the GUI or writing scripts that call the OpenSim tools and edit models from MATLAB. Because most engineering students are familiar with using MATLAB or Python, this will be a welcome change for many OpenSim users.

A Sleeker API

For advanced developers and researchers who want to work in C++, OpenSim 3.0 also offers a cleaner, more intuitive API where they can add new features that know how to talk to OpenSim. "By providing a library of muscles, joints, and other components," DeMers says, "OpenSim's API gives other researchers a jumping off point to create new and interesting components."

But the challenge in creating an API is to make the interface clean and intuitive to use. That's where OpenSim 3.0 has made a big leap. "Every iteration of OpenSim has tried to refine that interface and now it's finally crossing a threshold where you don't have to be an expert in OpenSim to program anything," DeMers says. "So if someone wants to make a fatiguing muscle," Hicks says, "the developer has to think about the science of it and not so much the bookkeeping side of things."

Revamped Muscle Models

As part of the 3.0 upgrades, OpenSim's muscle models were revamped to improve accuracy and performance. Muscles are the primary driving force in a simulation of walking, running, or any other movement. "Having models of muscle that are robust and also easy to fine tune and extend for new uses is a huge improvement not only for developers, but also every OpenSim user," Hicks says.

TOP 10 Improvements in OpenSim 3.0

1. Users can edit and iterate models in the GUI
2. Improved visualization tools
3. New and improved muscle models
4. New probe model component
5. Performance upgrades
6. Function editing in the GUI
7. Drag & Drop models and motions
8. MATLAB and GUI scripting
9. An intuitive, more efficient API
10. Apache 2.0 open-source license for the API

A New License

Earlier versions of the OpenSim GUI were freely available to any researcher or nonprofit user; and the source code was accessible to any user who asked for it. Now, the GUI and API are also available to commercial users and the API has a certified open-source license. "Now people can get it without having to ask," Hicks says. "This further consolidates OpenSim as a common biomechanics platform for all." □

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

SMART EMBEDDED DEVICES:
Here They Come

By Katharine Miller

Embedded medical devices that both detect symptoms and treat them have existed for decades. Take, for example, the heart pacemaker. But a new generation of implants could soon emerge to do something far more useful and daunting: These devices will learn from and adapt to changing human physiology and behavior. Projects as diverse as the artificial pancreas and closed-loop systems for deep brain stimulation are developing cutting-edge treatments for diabetes, epilepsy, and Parkinson's Disease (PD). At the heart of these efforts are machine-learning approaches: Computational algorithms that learn from patient-specific data.

Through it all, researchers are grappling with several fundamental questions: How smart do these devices need to be? Should we rely on simpler algorithms (which may be more predictable) or complex ones (which may be more robust)? And how do we make these devices failsafe?

The Artificial Pancreas: How Smart is Smart Enough?

Medtronic has run feasibility trials of an artificial pancreas in which sensor information is used in real time to modulate insulin delivery as blood sugars vary up and down. Their closed-loop system is pretty straightforward: The algorithm uses what's called a proportional–integral–derivative (PID) controller that looks for deviations from a set point and then makes adjustments to bring glucose levels back to that point. "Most of the commercially available control systems use that approach," says **John Mastrototaro, PhD**, chief technology officer at Medtronic Diabetes. The PID algorithm is quite robust over a wide range of insulin needs for a patient, Mastrototaro says. "It doesn't have to learn so quickly." But it does gradually learn. "We program in various parameters to model the sensitivity of the patient for our algorithm and then the machine will learn and change and drift as the patient's needs change and drift over time," he says. For example, pregnant women and growing children have changes in insulin needs.

Initially, Mastrototaro says, the learning can all be done manually and offline. As the patient wears the system, it gathers data that he or she can upload to online software provided by Medtronic, which recalculates and optimizes the parameter settings and feeds them back to the embedded system for use the next week. "And you can repeat that iterative process on an ongoing basis so it's analogous to machine learning in the device," Mastrototaro says. In the future, data will be uploaded and parameters adjusted automatically while the patient wears the device, he says.

Mastrototaro also envisions doing the machine learning outside the device. "With wireless technology, the device can talk to central computers like a cell phone can, so it doesn't really have to be in the device to behave as if it is," he says. Moreover, the patient can benefit from software modifications and updates without having to buy a new device.

As simple and effective as the PID controller seems to be, other groups are instead exploring a model predictive controller (MPC). **Frank Doyle, PhD**, professor of engineering at the University of California, San Diego is one of the pioneers of the MPC school. MPC controllers are used in everything from flight controllers to automobile controllers to the control of petroleum refineries. "Basically the high priority control loops in industry use more sophisticated algorithms, like MPC," he says.

The Doyle group's MPC algorithm doesn't target a single set point or number. Instead, it creates multiple zones along the continuum from hypo- to hyperglycemia. Within each zone, all measurements are considered equally good—allowing the algorithm to ignore sensor noise. "It's consistent with how a doctor analyzes data," Doyle says.

MPC acts on a time scale of minutes. "Every 5 minutes or so the forecast is for the next 30 to 60 minutes," Doyle says. This means the controller can respond quickly to changes in glucose. For example, Doyle's group developed a meal detection algorithm that can spot a sharp rise in glucose (such as might occur during a meal) and take appropriate action (provide fast-acting insulin). "Most other groups let patients interact with the pump to give a priming bolus [large injection] of insulin at lunchtime." Because children sometimes forget to bolus, he says, "we've sought results that don't require that." Other algorithms might detect and respond to exercise or illness.

"Most other groups let patients mess with the pump to give a priming bolus [large injection] of insulin at lunchtime," Doyle says. Because children sometimes forget to bolus, he says, "we've sought results that don't require that."

Doyle's group is also adding a layer of iterative learning control (ILC) that would learn over a longer time scale. He compares MPC to cruise control in a car, which works well on a scale of fractions of a mile but struggles a bit when it comes to a hill that it didn't expect. "If you know the hill is coming you can anticipate," he says. In diabetes, the hills might be exercise days or days when a patient is sick or anxious. ILC can anticipate these on a longer time horizon. "It's our hypothesis that we will get information to inform a long-term control program," he says.

"It could be punched in (by the mom of a sick kid) or it could be something the algorithm could learn from." For example, a woman's monthly menstrual period might change sensitivity and could be learned.

To make it possible for the algorithm to function in an embedded device, Doyle's group reduces the equation to an analytic solution. "We enumerate all the possible solutions and store it as a memory table," he says, "So it's a memory operation rather than a calculation."

No clear winner has yet emerged between PID and MPC. "It's a bit of a controversy," Mastrototaro says. Currently , based on the clinical data produced using different algorithms, "the PID algorithm is performing every bit as well if not better than the MPCs," he says.

But Doyle says the two algorithms have yet to be compared head-to-head in a clinical trial under the same conditions. In the year ahead, his group has been funded by the Juvenile Diabetes Research Foundation to conduct that very study in collaboration with the Sansum Diabetes Research Institute.

Mastrototaro concedes that as the MPC folks add more parameters and learning to their algorithm it will get better and better. "At the end of the day, to be quite frank, I think both of them will do a good job. The goal ultimately is to have phenomenal outcomes in managing diabetes."

In Europe, Medtronic currently sells this device, called a Paradigm Veo, which is a first baby step toward an artificial pancreas. It automatically suspends delivery of insulin for a period of time when a continuous glucose monitor starts to read lower than a certain level. In June 2012, Medtronic applied for approval to use the device in the United States. Courtesy of Medtronic.

Closing the Loop on the Brain

Deep brain stimulation (DBS), which uses electrodes implanted in the brain, has been approved to treat Parkinson's disease (PD) tremors since 2002 and epileptic seizures since 2010. Current DBS devices have fixed settings that are manually adjusted by medical experts during physical exams, and the stimulation is continuous.

Ideally, say researchers, DBS would be a bi-directional system: Electrodes would sense an imminent seizure or the onset of a PD symptom, which would signal the same electrodes to provide an appropriate level of stimulation. But because the brain's signals are complex and vary among patients, machine learning could play a role in making such systems a reality.

So far, there's been a lot of work on algorithms that detect or attempt to predict epileptic seizures, and some pilot work is now exploring similar detection schemes for PD and other disease states. In general, bi-directional systems are still in an investigational stage of maturity, says **Tim Denison, PhD**, engineering director in Medtronic's Neuromodulation business. "We are still on the journey, not yet to the destination. Detection systems with high sensitivity and specificity are challenging; prediction systems are even more difficult."

Among those working on predicting seizures are **Mushfiq Saleheen** and **Homa Alemzadeh**, graduate students in **Ravi Iyer**'s engineering group at the University of Illinois at Urbana Champaign. They used neural networks to train a device to predict seizures. The device relies on multiple parameters—not only electro-encephalogram (EEG) readings, but also oxygen saturation and body movements. And it's a flexible device that can also work for detection of traumatic brain injury, cognitive decline, and heart attack prediction. "The math underlying making these predictions is not that different, but the device would be configured differently for each disease," Iyer says. Iyer and Alemzadeh's goal is to design a flexible device that can predict the onset of a traumatic event such as seizure or heart attack several minutes in advance, but the intention is to set off an alarm rather than provide treatment. "With an alarm, patients can take action to prevent the worst of the consequences."

Medtronic hopes to use similar types of machine learning to distinguish seizure from non-seizure events but with an eye to using DBS as a closed-loop treatment. This presents unique challenges. For example, when DBS starts (say in response to brain signals suggesting a seizure is imminent), the large stimulation pulses could immediately drown out the brain signals needed to assess the patient's state. Denison and his colleagues have found a way to teach the algorithm to distinguish this noise from the valuable background signal. It's a key step toward making a bi-directional system a practical reality.

Creating a bi-directional system for DBS treatment of PD presents additional challenges. PD signals aren't very strong compared to seizures. "They're about 1000 times smaller than what cardio pacemakers

detect today, and usually about 10 to 100 times smaller than a seizure," Denison says. So it can be tough to get robust measurements. And unlike seizures, many of PD's brain signals appear to be embedded as variations in normal rhythms. "The challenge is to discover what's normal and what's disease related when designing a robust classifier," Denison says.

A stimulator and leads used for deep brain stimulation surgery. Courtesy of Medtronic.

Eduard Bakstein, a PhD student at the department of cybernetics at Czech Technical University in Prague, is using neural networks to discover features of PD tremor. "I look at the signal and I know when tremor was present or not, and then I extract different features from the data," he says. This process involves applying various transformations to the data—Fourier transform; wavelet transform; standard deviation of the signal—to observe how the features behave during the on- and off-tremor periods. The goal is to identify the features that change most when the tremor is starting. "Then I use these in the machine learning to identify the on and off periods," he says. In a small pilot study of his model, it worked well for some patients and not so well for others.

Medtronic has access to more patient data than Bakstein, and has therefore applied its machine learning approach to large datasets of EEG data both across subjects and over time. Although there's a particular signal that they believe seems to correlate with the presence of symptoms in an animal model, Medtronic researchers are still trying to clarify when it appears and whether it can be used to titrate stimulation as part of a closed-loop therapy. "We might really want something that gently coaxes the brain, rather than responding as it would to a large-scale event," Denison says. Previous thinking about responsive stimulation in the brain evolved from a defibrillation mindset—applying a very strong stimulus as is done with cardiac devices. "That might not be the right approach," Denison says, "And we are now exploring other schemes more in the spirit of a circadian or homeostatic feedback concept, based on first-principles measurements of physiology."

Medtronic is also trying to simplify its machine learning approaches so that they don't require too much power. "The therapy today to provide DBS is on the order of 100 microwatts—about a million times less than an incandescent light bulb," Denison says. This limits the amount of power available for sensors and detection algorithms. "We can only get a budget of 10 percent of the therapy power," he says. Using what's called the reduced sets method and other schemes, Medtronic systematically seeks to simplify its detectors and algorithms to reduce the energy needed. "Frankly, a lot of detectors draw too much power for the performance that is achieved," Denison says. "You're not doing anyone any favor if you can't implement the technology practically in an implant."

Denison's team spends a lot of time optimizing algorithm methods that are simultaneously accurate and low power. As an example, Medtronic uses a posture response algorithm in its RestoreSensor device, which uses stimulation to treat chronic pain. "We customized an accelerometer and algorithm to build a reflex into the device, drawing only microwatts of power," Denison says.

Safety is Everything

For an embedded medical device to succeed, it must not only do what it's designed to do but also do it in a failsafe way. That's one of Medtronic's big concerns now with the artificial pancreas. Because an incorrect dose of insulin can be deadly, putting decisions in the hand of a tiny device with a detector and a learning algorithm is a bit scary. One option, says Mastrototaro, is for the machine to learn the patient's normal cycle of blood sugar variation and then use that information to send an alarm or suspend closed loop control when there's something unusual going on—when the pump or detector aren't working correctly. "That's where our focus is now," he says. □

"Frankly, a lot of detectors draw too much power for the performance that is achieved," Denison says. "You're not doing anyone any favor if you can't implement the technology practically in an implant."

THE MICROBIOME:
Dealing with the Data Deluge

By Alexander Gelfand

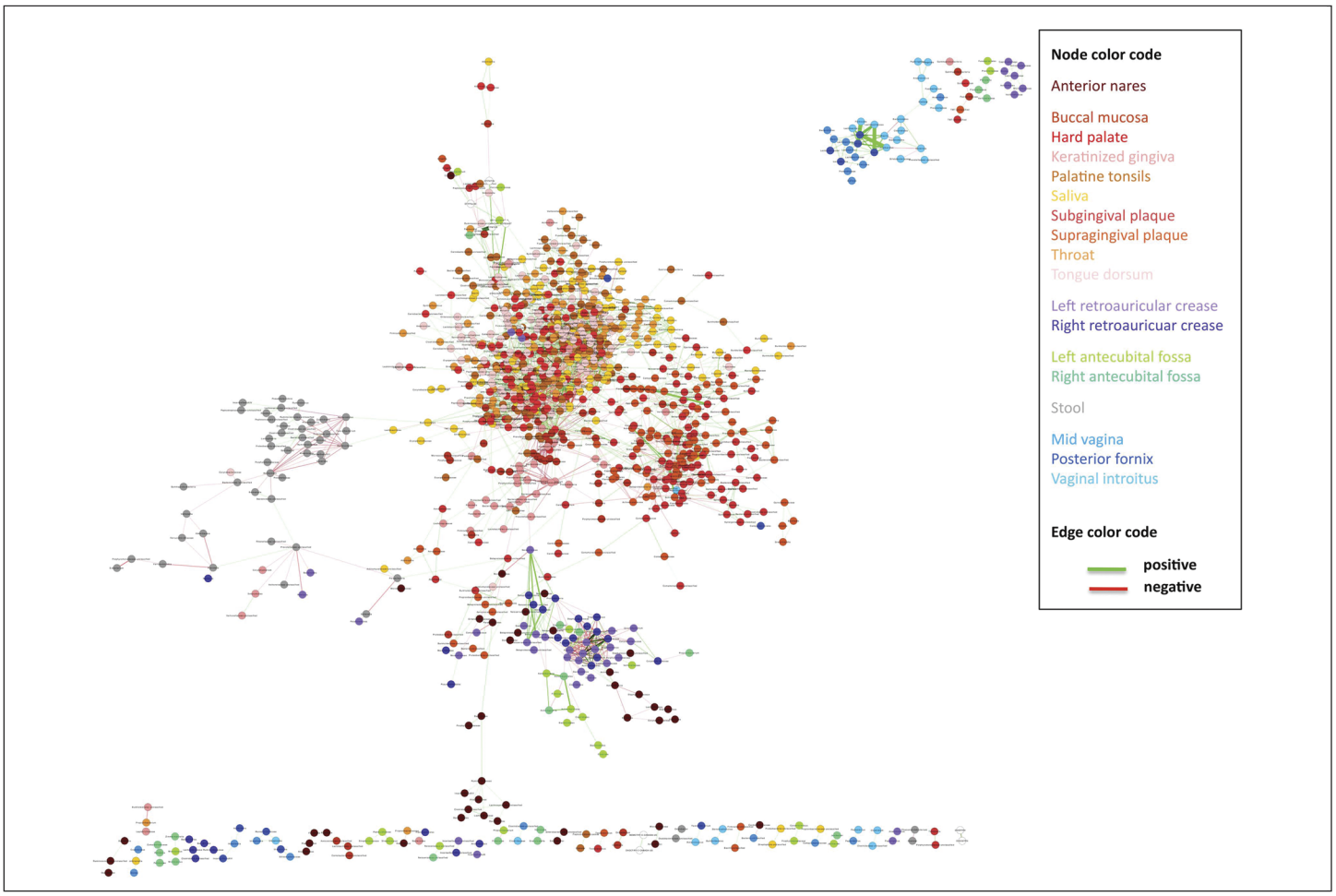
This past June, 200 members of the NIH-funded Human Microbiome Project (HMP) Consortium published a slew of papers offering fresh insights into the role microbial communities play in the human body—including how changes in the vaginal bacteria of pregnant women affect the health of their babies, and how gut microbes influence inflammatory bowel disease.

But the research was only possible thanks to a team of experts in computational biology and bioinformatics.

HMP participants sequenced thousands of metagenome samples—which contain genetic material from hundreds of diverse microbes—from up to 18 body sites in 242 healthy individuals. State-of-the-art tools for analyzing individual genomes aren't well suited to analyzing metagenomes, as the data are much more massive and messy. So a team of researchers from the Department of Energy's Joint Genome Initiative (JGI) joined forces with software engineers and computer scientists from the Biological Data Management and Technology Center at Lawrence Berkeley National Laboratory to develop and maintain a suite of novel tools, including a quality control filter, a curation and annotation pipeline, and methods for analyzing and integrating the data.

These efforts culminated in a one-stop shopping data management and analysis system for microbial metagenomic studies, the Integrated Microbial Genomes and Metagenomes (IMG/M) system; and an HMP-specific web interface known as IMG/M-HMP that supports comparative analysis of HMP genomes and metagenomes against the vast pool of microbial data in IMG/M.

HMP scientists can come to the IMG/M-HMP—which is neck-deep in genomics tools and annotated microbiome data—knowing that they will find much of what they need. The IMG/M tools can do a range of analyses, including identifying microbes and genes within a metagenome; predicting gene function; and comparing populations of microbes across metagenomes. According



Using data that was processed by the HMP team, researchers created this global interaction network showing the associations among phylotypes (the nodes in this plot) within and across 18 body sites (colors), with edges representing significant relationships between the phylotypes, whether positive (green—co-occurence) or negative (red—co-exclusion). The network shows significant niche specialization. Reprinted from Faust, K, et al., Microbial co-occurrence relationships in the human microbiome, **PLoS Comput Biol.** ***2012 Jul;8(7):e1002606. Epub 2012 Jul 12.***

to JGI functional annotation group leader **Natalia Ivanova, PhD**, HMP researchers assemble their sequences and perform structural annotation, or gene prediction; while JGI scientists perform functional annotation—assigning the predicted genes to conserved protein families—and provide data integration and the user interface.

Finding Known Genes and Microbes

HMP researchers must find protein-coding genes and determine what bugs they come from. The process typically starts with comparing a set of metagenomic data to millions of genes in the IMG reference database in hopes of finding a match. Using this process, the HMP has harvested approximately 200 million genes from its metagenomic samples. "And that requires a lot of computations," says JGI computational genomics group leader **Konstantinos Mavrommatis, PhD**.

For decades, researchers have used an algorithm called BLAST (Basic Local Alignment Search Tool) to search for similarities between nucleotide sequences. But BLAST alone is too slow and computationally expensive to handle the metagenomic sifting required by the HMP. So Mavrommatis and his colleagues incorporated novel computational approaches into IMG/M.

At first, the team at JGI investigated alternatives to similarity-based pattern searches, but those produced too many false results. So they turned instead to new similarity-based algorithms, such as USEARCH, capable of producing results similar to those of BLAST, only faster and more efficiently. USEARCH looks for a small number of good matches rather than trying to identify all homologous sequences, cutting down on search time without affecting sensitivity.

Finding Novel Genes and Predicting their Functions

Matching a microbe buried inside a metagenome to a genome in a reference database is akin to finding a needle in a haystack. But identifying the genes from a microbe that hasn't previously been sequenced and figuring out what those genes actually do is even more challenging.

To enable researchers to find novel genes, the IMG/M includes gene-predicting algorithms that rely on generic features of nucleotide sequences rather than relying on comparison to known sequences (as similarity-based algorithms do). The mathematical methods used in gene prediction, such as hidden Markov models, "work quite well," says Ivanova. As a result, even when they are fed radically new content, their error rate remains below 10 percent. "It's still not perfect," says Mavrommatis, "but considering all the other sources of error, it's not the worst."

These algorithms in the IMG/M were the basis for characterizing the diversity of the microbiome in many of the papers published by the HMP.

Teasing out the function of a novel gene is similarly demanding. In general, Ivanova says, gene function is determined by comparing unknown genes to ones whose function has been verified experimentally. Function can be confirmed by analyzing the distribution of similar genes in known genomes, or by looking at a gene's chromosomal neighborhood, since "genes that are next to each other are more likely to be functionally related."

Few genes in the HMP database have been characterized to the point where scientists can say precisely what they do. And while perhaps 75 percent of the genes in IMG have been at least broadly characterized, that figure falls by half for genes within the HMP pool.

Yet considering the number of genes involved, that's still an awful lot of information. And the methods that Ivanova describes can be used to create clusters of microbial sequences that might be worth examining in the lab, where researchers can learn more about them through experimentation. The gene prediction and annotation pipeline developed by JGI has already led to the creation of the HMP Gene Index, a collection of 690 annotated sequences from 15 different body sites. And this past September, a group of users attending a Microbial Genomics & Metagenomics (MGM) workshop run by the JGI used the data in an attempt to identify potential antibiotic-resistance genes in different metagenome samples.

Human Health

One of the grand challenges of the microbiome project is to discover new information that could help to diagnose or treat disease. This is a huge challenge computationally for several reasons. First, differences in the diversity and complexity of the microbial communities found in different body sites (e.g., the skin, the mouth, the gut) make it difficult to do comparisons between them, Ivanova says. As a result, researchers tend to focus on comparisons of populations at the same body sites but in different individuals.

Perhaps even more significantly, due to privacy restrictions, the HMP metagenome datasets themselves come with very little metadata attached. Such metadata, which might describe the sex or dietary preferences of the human donor, is crucial to determining which metagenomic datasets might be of interest. Scientists at JGI have manually applied their own five-tiered classification scheme to the data, moving from the general (e.g., "host-associated" versus "engineered") to the specific ("respiratory system," "digestive system," "skin and appendages"), but the approach has its limitations.

"It has to be much more granular," says Ivanova. "There are some scientific questions that you won't be able to answer because of the lack of metadata."

Information Overload

The IMG/M helps researchers access and manipulate microbiome data in useful ways, but the sheer volume of data continues to present challenges.

For example, standard methods of storing and retrieving data from relational databases are no longer sufficient. The JGI-BDMTC team is exploring options such as nonrelational or NoSQL databases, and while they have yet to find a one-size-fits-all solution, they continue to explore alternatives.

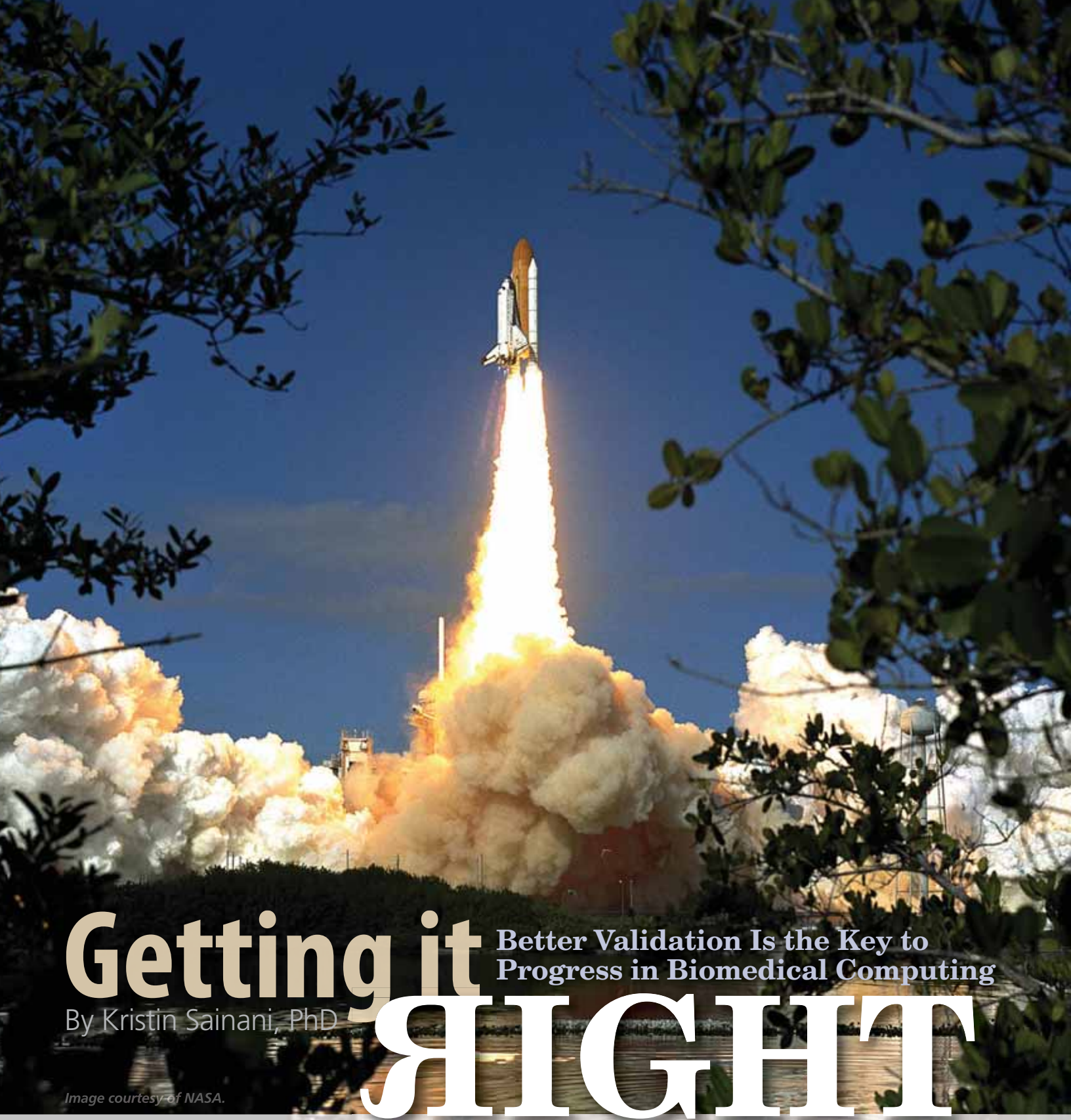
And then there's the question of how to provide access to the data and distribute the information. "We are struggling with the challenge of devising tools that don't overwhelm our scientific users," says **Victor M. Markowitz, DSc**, head of the Biological Data Management and Technology Center.

> "We are struggling with the challenge of devising tools that don't overwhelm our scientific users," says Markowitz.

Giving scientists tools that are easy and efficient to use is critical because these tools drive research as much as they support it.

"In our experience most researchers don't have a clear idea of what they really want and how to achieve it until they start getting the data," Mavrommatis says. "For good or for ill, there is frequently no prior design of the analysis; we generate the data, and then the researcher starts trying to address questions based on what tools are available."

Which is all the more reason to ensure that those tools are the best possible ones for the job. □



Getting it RIGHT

Better Validation Is the Key to Progress in Biomedical Computing

By Kristin Sainani, PhD

Image courtesy of NASA.

When the ill-fated space shuttle *Columbia* launched on January 16, 2003, a large piece of foam fell off and hit the left wing. Alerted of the impact, NASA engineers used a computer model to predict the possible consequences. Their conclusion: It will likely be okay. But, in fact, the foam had catastrophically exposed the shuttle's thermal protection system, causing *Columbia* to disintegrate during reentry and killing all seven crew members.

Investigators later concluded that the disaster might have been averted. One of the key failures: The computer model got it wrong. The model had been validated for small pieces of foam, not "huge hunking pieces," says **Jerry Myers, PhD**, chief of the Bio-Science and Technology branch at NASA's Glenn Research Center. "Because it had been well-validated down in the low end in that operational scheme, everybody took it at face value that it would work in the upper scheme." (In fact, one simulation did predict catastrophic failure—but engineers distrusted that particular simulation.) The assumptions and uncertainties of the model were never fully presented to higher-ups, who consequently made the wrong decisions. This chain of failures led NASA to implement a comprehensive standard (NASA 7009) for vetting models and simulations.

High-Stakes Consequences. *Streaks of burning debris from the U.S. space shuttle orbiter* **Columbia** *as it broke up over Texas on February 1, 2003. The accident killed all seven astronauts aboard the craft. A poorly validated model played a role. Credit: Dr. Scott Lieberman—AP Photo/Tyler Morning Telegraph.*

Models can be extremely valuable: They complement experimental studies by providing additional insights in a cost-effective way. But the value of a model depends on rigorous validation, as the Columbia accident tragically shows. Modelers in high-stakes fields—aeronautics, nuclear physics, bomb making, and weather prediction, for example—understand this well. But "in the biomedical sciences, there hasn't been such a culture of holding people to the fire of validation," says **Peter Lyster, PhD**, program director in the Division of Biomedical Technology, Bioinformatics and Computational Biology at the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH).

This laissez-faire ethos is going to have to change. Biomedical modeling has now entered a high-stakes era: Models are increasingly being used to make direct clinical decisions, with life-and-death consequences, such as choosing between cancer drugs. At the same time, there is a brewing crisis of confidence in bioinformatics and biomedical computing (see: *Meet the Skeptics*, in the Summer 2012 issue of this magazine). Scores of papers have been published claiming "success"—for everything from disease signatures to drug targets—but practical applications have been few, and some models have been debunked (see: *Errors in Biomedical Computing*, Fall 2011 issue of this magazine). These factors are fueling an intense discussion on validation in biomedical modeling circles.

The point of validation is to help modelers and model consumers decide: Does the model get close enough to reality so that they can use it with confidence in a particular scenario? Perfect validation isn't always the goal; sometimes a less costly validation might suffice if the costs of making a mistake are low.

The point of validation is to help modelers and model consumers decide: Does the model get close enough to reality so that they can use it with confidence in a particular scenario?

The problem is that current validation schemes for biomedical models are often inadequate given the stakes. This article describes several common pitfalls of current practices, as well as several efforts to remedy these issues by innovating or standardizing validation for biomedical models.

The Status Quo

When biomedical modelers talk about "validation" currently, they may mean many different things. Some researchers may confuse verification—checking that the code does what it's supposed to—with validation; but verification is only a prerequisite to validating a model. Some researchers also confuse peer review with validation. "We had a long discussion with a couple of researchers a while back as to what constituted validation. And their response was, 'publication in the general literature,'" Myers says.

At most, peer review provides a very low-level, "do-my-concepts-look-good" validation, Myers says.

"But that is just not right." At most, peer review provides a very low-level, "do-my-concepts-look-good" validation, he says. Peer review is simply not equipped to vet high-throughput data and complex models in a meaningful way.

Researchers who go beyond verification and peer review will typically validate their models against existing data. Using a kind of statistical validation, they fit the model on one set of data while holding out some of the data for subsequent "independent" testing. For example, in the old days of predicting protein structure from sequence, people used to fit an algorithm to one set of known structures and then test it on a separate set of known structures, says **John Moult, PhD**, professor of cell biology and molecular genetics at the University of Maryland. In theory, this could provide reasonable validation—but in practice, there's good evidence that it simply doesn't work. "In practice, we're all rather fallible. It's very hard if you know the answer not to be unconsciously biased by it," Moult says.

"I don't mean that people deliberately cheat," Moult says. "I think it's a lot subtler than that. In the field that I'm familiar with, there are a lot of very, very smart people and they're very honest people by and large. But somehow we fool ourselves." Information inevitably "leaks" from the training set to the test set; for example, if the model doesn't fit quite right on the test set, researchers go back and tweak the algorithm a little, Moult says. Or the training and test set may contain such similar samples that the algorithm works well on both, but does not generalize to other problems.

Researchers do better when they get beyond statistical validation and benchmark their algorithms against truly new experimental data (or data that they were blinded to during algorithm development). However, even in this situation biases slip in. Researchers may selectively report the most optimistic validation results, for example. "We call it the self-assessment trap," says **Gustavo Stolovitzky, PhD**, manager of functional genomics and systems biology at the IBM Computational Biology Center. "You want to publish your paper and, therefore, at the end of the day, some of the objectivity of the scientific enterprise is lost."

Researchers may selectively report the most optimistic validation results. "We call it the self-assessment trap," says Stolovitzky. "You want to publish your paper and, therefore, at the end of the day, some of the objectivity of the scientific enterprise is lost."

In a 2011 paper in *Molecular Systems Biology*, Stolovitzky and colleagues surveyed 57 modeling papers—within a few specific areas—in which authors assessed their own methods. Sixty-eight percent of authors reported that their method was best for all metrics and all datasets; and 100 percent reported that their method was among the best. But, of course, this is impossible—all these methods cannot be the best.

Another problem with the status quo is that most researchers view validation as a one-time, one-size-fits-all endeavor. "The word 'validated' can get slipped in very, very easily," says **David M. Eddy, MD, PhD**, founder and medical director of Archimedes, a healthcare modeling company in San Francisco. "A team can validate the model in one population for one outcome for one treatment, and then they'll attach the word 'validated' to the model as though it's a property of the model, that goes with the model wherever the model goes—to any treatment, to any outcome, to any population, to any time period," he says. This, of course, leads to the kind of dangerous extrapolation that happened with the Columbia disaster. Plus, if you only validate a model once, that model is going to be out of date in a few years, Eddy says.

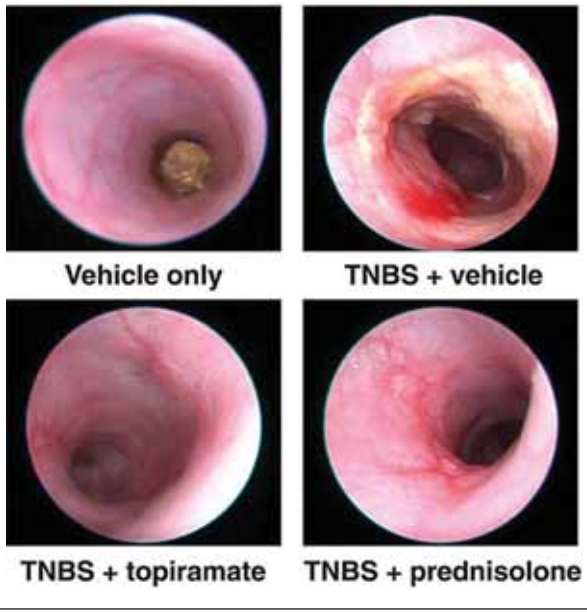
Finally, most prevailing validation efforts omit a critical element: error bars. Since a model can never match reality perfectly, "validation is mostly about knowing what the errors are and accounting for them," Lyster says. The uncertainties in the model and data need to be quantified by putting error bars around model predictions. "It's not just a matter of

having a forecast; it's a matter of knowing accurately how fat the error bars are," Lyster says. "You've got to know that you've got good error bars so that people can go to the bank with them."

"It's not just a matter of having a prediction or forecast; it's a matter of knowing accurately how fat the error bars are," Lyster says. "You've got to know that you've got good error bars so that people can go to the bank with them."

Outsourcing Validation

Many modelers resort to statistical validation on existing data because they don't have the expertise, time, or resources to generate new experimental data. But it's becoming increasingly easy to outsource validation experiments, says **Atul Butte, MD, PhD**, associate professor of pediatrics at Stanford University. Outsourcing validation doesn't mean assays performed on-the-cheap in China or India. Rather, modelers can hire companies or university core facilities—experts in a particular research technique—to run the specific experiments needed to test their model predictions. "I'm a big fan of this approach," Butte says.



Outsourced Validation. *Using modeling, Atul Butte's team predicted that topiramate, an anti-seizure drug, would be effective against inflammatory bowel disease. Butte outsourced the experimental validation to companies he found through AssayDepot.com. One of the companies was able to provide colonoscopies of the rat, pictured here. Reproduced with permission from: Dudley, JT, et al. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease,* **Sci Transl Med** *17 August 2011; 3: 96.*

Using online marketplaces—such as ScienceExchange.com, AssayDepot.com, and Biomax.us—modelers can find exactly the services or samples they need. It's a lot like shopping on Amazon.com. Need high-quality serum from breast cancer patients treated with Tamoxifen? Just drop them in your shopping cart. Need to test a drug in a rat model of inflammatory bowel disease? Here are 15 companies that can do it for you. "This is the most amazing thing for us in informatics and computational biology. If we want to do this kind of translational work, all this is here waiting for us," Butte says.

In 2011, Butte's team published back-to-back papers in *Science Translational Medicine* that highlight the value of outsourcing validation. Butte's team devised an algorithm that mines publicly available gene expression data to find new uses for old drugs. The model predicted that cimetidine, an antiulcer drug, would be effective against lung cancer. Butte hired the Transgenic Mouse Research Center core facility at Stanford to test this prediction; the result: the drug indeed slowed the growth of lung cancer in mouse models. Butte's algorithm also predicted that topiramate, an anti-seizure drug, could be used to treat inflammatory bowel disease. Butte collaborated with scientists at Stanford to test the prediction in rats and additionally hired two companies that he found through AssayDepot.com to perform independent replications. All three experiments gave strong evidence of the drug's efficacy. One of the companies even provided colonoscopies of the rats, something that his Stanford collaborators couldn't do. Statistical validation doesn't resonate much with physicians and biologists, but "when you show them the colonoscopy from the rat, that's huge value-added for your model," Butte says.

Using outsourced validation, Butte has repositioned one drug—moving it from computational prediction to cell and mouse models and then to clinical trials (which are about to begin)—in the span of eight months. "It is getting to be too trivial to get just a simple bioinformatics paper published. Those kinds of papers are slowly losing their impact, especially with non-computational scientists," Butte says. "I think if you want to find and show something big, this is how you're going to do it."

Are the data trustworthy? Sure, it's a worry, Butte admits. "But that worry lasts maybe about 48 hours—the time it takes you to get the samples that you would have otherwise waited months or years to get," he says. And because companies send Butte the raw samples as well as the data (he has the formalin tissue slides on his shelf), he can look at them himself and even seek a second opinion. (For example, a pathologist might read slides for a dollar apiece, he says.) And, if the price is low enough, one can always send

the same experiment to two or more vendors, for maximum independent validation, Butte says.

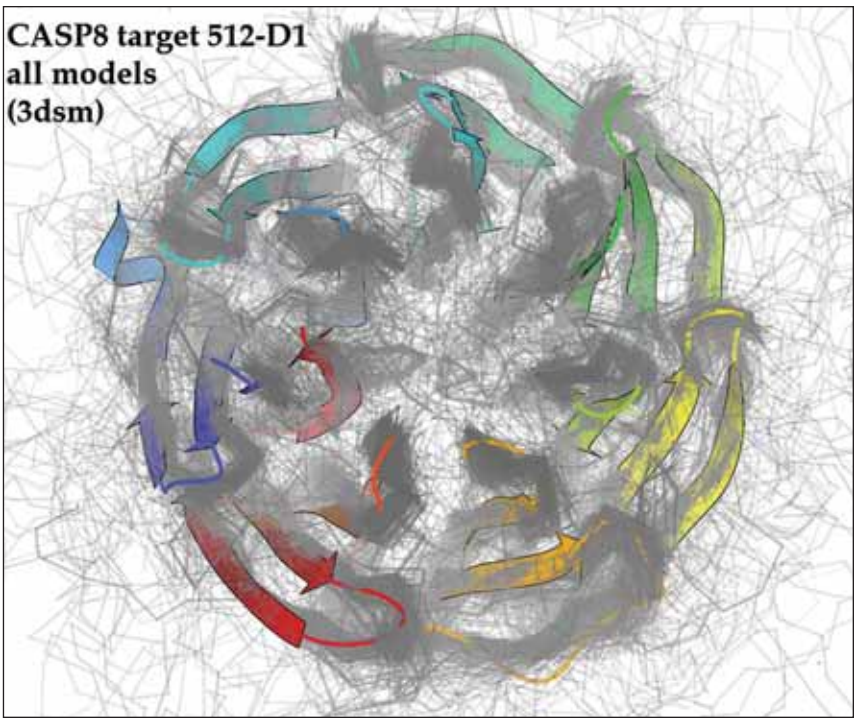
Outsourced validation experiments may actually be more robust and of higher quality than experiments done by the computational modeler, says **Elizabeth Iorns, PhD**, the cofounder and CEO of Science Exchange in Palo Alto. Science Exchange is a marketplace for university core facilities. "If you have one person who is doing all the experiments, no matter how hard they try, subconsciously they're looking at the data in a way that matches what they want it to say. So, distributing the experiments across multiple investigators is a way to eliminate the individual investigator bias." Plus, the core facilities tend to be extremely specialized in a particular experimental technique. So the quality tends to be higher than if an inexperienced postdoc or graduate student is running the experiment, she says.

To promote the cause of validation, Science Exchange recently launched the Reproducibility Initiative. Scientists may apply to have previously published research (including models) independently tested through the Science Exchange network; then they can publish the results in a special issue of *PLoS ONE*. Even if a validation study refutes a computational model, Iorns points out, it's better to publish this failing yourself rather than for someone else to discover and publicize it.

Crowdsourcing Validation

One of the most successful innovations in validation is the use of collaborative competitions. These competitions engage the community in an ongoing, cyclic model of validation that helps the field progress, Lyster says. The first of these competitions, CASP (Critical Assessment of Techniques for Protein Structure Prediction), began in 1994 and is now in its tenth round. Others quickly followed, including CAGI (Critical Assessment of Genome Interpretation), CAPRI (Critical Assessment of PRedicted Interactions), the American Society of Mechanical Engineers (ASME) Grand Challenges, and DREAM (Dialogue on Reverse Engineering Assessment and Methods)—which is now in its seventh round.

Teams work on the same challenges, so it is possible to directly compare their performance; and independent judges evaluate the methods using several well-defined metrics. These objective assessments help to "break that vicious circle of self-assessment traps and lack of sufficient rigor," Stolovitzky says. Competitors are also blinded to solutions, which further reduces bias. For example, in CASP, organizers gather unpublished data from X-ray crystallographers and NMR spectroscopers who are on the cusp of solving a structure. "The key thing about CASP is that one doesn't know the answers; one is doing genuine blinded prediction," Moult says.



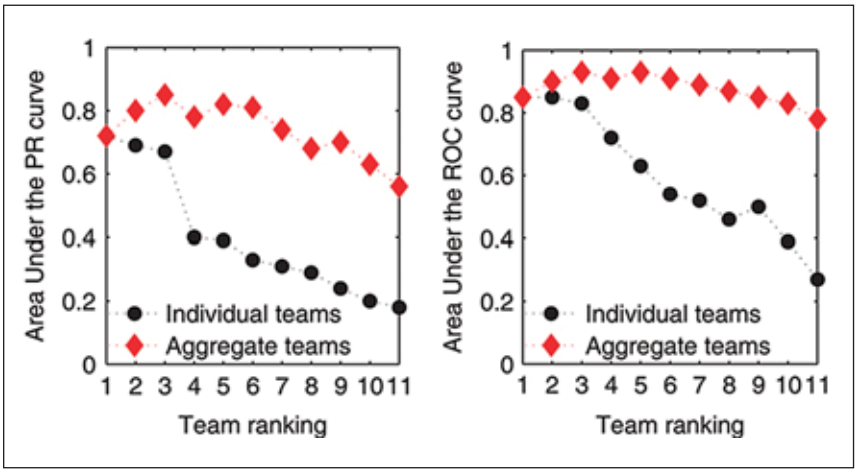
Safety in Numbers.** This picture shows 354 predictions (in gray) of the structure of a target protein (3dsm) from the eighth CASP competition. The actual structure is shown in colored ribbons. Reprinted from Keedy, DA, et al.,* **The other 90 percent of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models, Proteins: Structure, Function, and Bioinformatics*, 77:S9:29-49, 2009.

After each competition, results and data are made freely available to the community so that everyone can learn from the successes and failures. Competitions systematically reveal where people are "fooling themselves"; they also give a field insight as to which problems have been effectively solved. "As participants in a field, we've got much better feedback on what the real issues are and where we should focus our efforts," Moult says. DREAM organizers also aggregate

Competitions systematically reveal where people are "fooling themselves"; they also give a field insight as to which problems have been effectively solved. "As participants in a field, we've got much better feedback on what the real issues are and where we should focus our efforts," Moult says.

the best solutions—yielding a collaborative algorithm that often outperforms the best single method. "This is the wisdom of crowds," Stolovitzky says.

Competitions also increase confidence. "In the area of protein structure, before CASP got established, we were sort of a laughing stock with the experimental-



The Wisdom of Crowds. *Performance of models from the top 11 teams from one of the DREAM2 challenges. The challenge consisted of predicting transcriptional targets of the transcription factor BCL6; both "area under the curve" metrics measure how well the prediction matches reality (where 1 is perfect prediction and 0.5 is no better than chance). Even as the performance of the individual teams decreases (black line and circles), the integrated prediction of the best performer and runner-up teams (red line and diamonds) outperforms the best individual team. Reproduced with permission from: Figure 1 of Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average?* **Molecular Systems Biology** *October 2011; 7:537.*

ists. They all knew that we were exaggerating," Moult says. That's completely changed, he says. "In terms of people in the broader protein structure community having more confidence in the methods, it's had a huge impact."

Industry can benefit from competitions as well—as evidenced by the Netflix Prize for successfully predicting a person's taste in movies—but the rules need to be slightly different, Stolovitzky says. He and colleagues have pioneered a collaborative competition model for industry, called IMPROVER (Industrial Methodology for Process Verification of Research). IBM and Philips Morris codeveloped the first set of IMPROVER challenges in systems biology. They aimed to verify that computational approaches can use transcriptomic data to classify clinical samples into diseased and non-diseased (for specific illnesses, including multiple sclerosis, lung cancer, and chronic obstructive pulmonary disease). Entries were assessed using gene expression data from unpublished cohorts of cases and controls. Data, gold standards, and scores are available at sbvimprover.com

Predictive and "One-Click" Validation

In crowdsourced validation, the participants are blinded but the answers are known to the organizers. "Predictive validation" takes blinding one step further: predictions are made while an experiment is ongoing—in other words, when the answers are truly unknown. This type of prophetic validation has a certain "wow factor" that is particularly useful for convincing skeptics.

For example, in 2004, the American Diabetes Association asked David Eddy (CEO of the company Archimedes) if his healthcare model could predict the results of an ongoing clinical trial called the Collaborative Atorvastatin Diabetes Study (CARDS). The trial was testing whether atorvastatin could reduce the chance of heart attack or stroke in people at risk, especially diabetics. Months before completion of the study, Eddy's team simulated the trial, sealed the resulting predictions in an envelope, and FedExed them to the American Diabetes Association, the principal investigators of CARDS, and Pfizer (the drug's sponsor).

Their predictions were "right on the money" for three of four outcomes, Eddy says: they closely predicted the actual rates of heart attack and stroke in the control group and heart attack in the atorvastatin-treated group. They underestimated the drug's ability to prevent strokes, but even that "error" turned out to have value, Eddy says. In the absence of data, the modelers had assumed that atorvastatin's effects on stroke would be similar to that of other statins; but it turns out that atorvastatin may, in fact, be more effective. "The mismatch between our model and the real results was what alerted Pfizer to that fact. So that's opened up other research avenues," Eddy says.

The success of the predictive validation won over modeling skeptics at the American Diabetes Association, which went on to commission considerable work from Archimedes, Eddy says.

Modelers at the company have also pioneered a "one-click" validation tool that addresses the need to

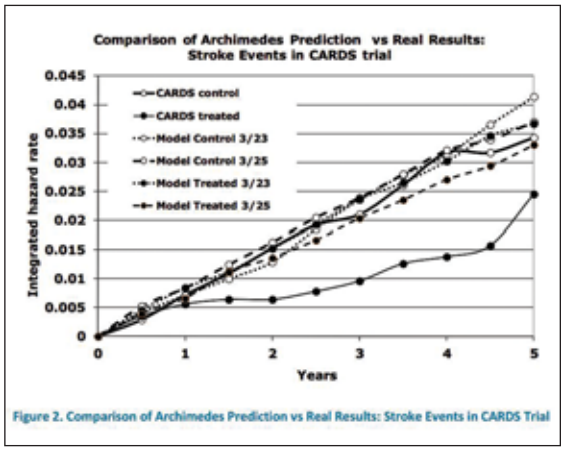


Figure 2. Comparison of Archimedes Prediction vs Real Results: Stroke Events in CARDS Trial

Right on the Money. *The Archimedes model accurately predicted three of four primary results of the CARDS trial months before the trial finished. Pictured here are the predicted rates of stroke in the control and treatment groups (from two separate simulations run on 3/23 and 3/25), as well as the actual rates of stroke observed in the trial. The model was right on the money for the control group, but overestimated the rate of stroke for the treated group. This "error" actually revealed novel insights into the drug's function. Courtesy of: Archimedes Inc.*

continually update and revalidate models. With validation, "there's no end point. It's not as though some hand comes down from the sky and says 'you've got it; you can rest, relax.' It's a constant process," Eddy says. One-click validation provides an automated way to revalidate models every time new medical evidence comes out. "With this one-click validation, we're getting much, much more efficient. We don't have to set up each new trial every single time," Eddy says.

With validation, "there's no end point. It's not as though some hand comes down from the sky and says 'you've got it; you can rest, relax.' It's a constant process," Eddy says.

Regulating Validation

Companies are increasingly using biomedical modeling in regulatory submissions for medical products, making validation a hot topic at the Food and Drug Administration (FDA).

Currently, decisions about validation are made on a case-by-case basis—and, correspondingly, what companies report to the FDA is highly variable. "What we get from manufacturers is just such a range in terms of defining their models, defining the limitations, defining what they're using the model for—the things that you think would be in any test report. We're not necessarily even getting those basics," says **Donna Lochner**, associate director for scientific outreach at the FDA's Center for Devices and Radiological Health.

The FDA is working to develop standards. "We want to promote greater use of computational models. One of the ways we can promote their use is to come out with clear expectations with respect to validation. That's where we are now," Lochner says.

So, the FDA is working to develop standards. "We want to promote greater use of computational models. One of the ways we can promote their use is to come out with clear expectations with respect to validation. That's where we are now," Lochner says.

Standardizing validation for biomedical models is a challenge. "The models in this space are very complex, particularly when talking about long-term interactions between a device and a patient," says **Tina M. Morrison, PhD**, a mechanical and biomedical engineer at the FDA's Center for Devices and Radiological Health. Though validation standards exist for hard-core engineering and physics-based fields, these don't necessarily transfer well to biomedical models—because data from living systems are harder to come by and highly variable, Morrison says.

Morrison and colleagues at the Center for Devices and Radiological Health have begun drafting a guidance document specific to medical devices. Though in its early stages, some of the essentials are clear: "first and foremost, good documentation," Morrison says. We need companies to document "what they did, why they did it, what their results are, how confident they are in those inputs, and the use history of those models," she says. Secondly, validation will have to be more quantitative. It's not sufficient to say that the prediction and the data match by 20 percent and that's "close enough," Morrison says. Companies might need to perform formal uncertainty analyses (adding error bars) and sensitivity analyses, where they tweak the parameters and the assumptions in the model and see how much that affects their predictions. Finally, the FDA is creating an innovative scheme to risk-stratify validation requirements, so that the level of validation depends on how the model is being used in the regulatory submission.

For example, imagine a computational model that predicts which commercial hip implant would be best for a given patient, based specifically on his or her anatomy, bone density and activity level. If that model gets it wrong, the stakes are high: the patient could experience a bone fracture and require repeat surgery. So, validation will need to be rigorous. But if a company is just using a model to justify which sizes of its device it needs to evaluate with bench testing, the risks are lower and, thus, a less rigorous validation strategy might suffice.

"Right now, we're not making big decisions based solely on the computational models; therefore, the level of validation isn't high," Morrison says. "But if we start shifting where we make more important regulatory decisions based on the computational outputs, the amount of information that's going to be needed to support that model's credibility is going to change."

The FDA is also developing reference problems to attempt to benchmark model performance. For example, they challenged 28 labs to simulate flow out of a simple nozzle. The variability they see in the computations helps the FDA and the researchers understand how underlying assumptions affect the model's precision. Lochner says. "By validating the result for a reference problem, we can then gain confidence in the outcomes as more complexity is added to the model."

Comprehensive Validation

In the wake of the Columbia disaster, NASA developed a standard (7009) for assessing the credibility of models and simulations. The goal: to help decision-makers know if they can trust a model's prediction when it counts.

"When we talk verification and validation [V&V] at NASA these days, we're really pointing to something that's a little more globally inclusive, which is what we call the credibility score," says **DeVon Griffin, PhD**, project manager of the Digital Astronaut Project (DAP), which uses simulations to evaluate various risks to human health that arise during long-term space travel. "We always did V&V prior to 7009, but the standard provided a systematic way to do it. More importantly, the standard provided a vehicle to communicate with managers so they understand the requirement to do V&V."

"When we talk verification and validation [V&V] at NASA these days, we're really pointing to something that's a little more globally inclusive, which is what we call the credibility score," says Griffin.

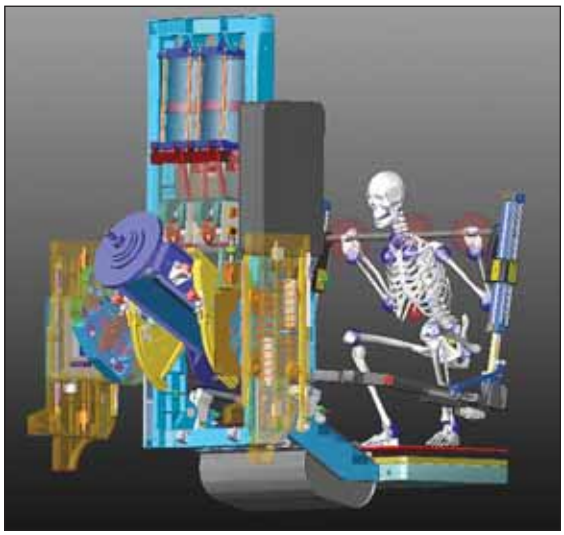
When Myers (who is DAP technical lead) first joined NASA's human research program the question of validation came up immediately, he says, "because we were getting very grandiose statements about what people's models could do." After Griffin identified and provided the standard to him, Myers says his reaction was: "This is like the greatest document I've ever read. It is a culmination of 50-plus years across NASA of understanding computational modeling." But the standard was designed for models in general and did not address the unique challenges of biomedical models. So Myers and his colleagues set about adapting the standard for their human research models. They are now writing up a formal guidebook on how to apply 7009 to biomedical models.

7009 is a synthesis of eight factors: verification, validation (comparison with experimental, simulation, or real world data not used to develop the model), input pedigree (how good are the input data), uncertainty quantification (error bars), model robustness (sensitivity analyses), use history, modeling and simulation management, and people qualifications. The eight categories are scored on a scale of 0 (lowest) to 4 (highest). The scores encompass both internal and external assessments. Though NASA currently uses the lowest of the eight scores as the overall score, NASA's Human Research Program is working to establish a process for calculating the overall score as a weighted average of the eight individual scores. The goal isn't necessarily to achieve a perfect 4.0, but rather to get as high as is reasonable for a particular modeling application. For example, a score of 2 or 3 might be the highest score that can be reasonably attained for certain biomedical modeling problems but may well be sufficiently high to meet customer requirements, Griffin says.

"I make it my business to look out into the field to see what's happening in V&V and credibility, and I still haven't found anything that's as comprehensive as 7009," says **Lealem Mulugeta**, DAP project scientist.

Most validation efforts only ask: how well does the model match the validation data? But 7009 additionally asks: how good are the data? "I think a lot of people just make the assumption that just because you have data, it's good data," Mulugeta says. "We go through the process of actually vetting our data to make sure that the data are credible and appropriate to use."

For example, the Digital Astronaut team created a model of an astronaut exercising on the Advanced Resistive Exercise Device (ARED)—the exercise device that astronauts use on the International Space Station to prevent muscle and bone loss. To verify and validate their model, they have to use data on joint torques and forces that were collected on earth or from other exercise models. So, when the models



***Digital Astronauts.** To maintain bone and muscle health while spending time on the International Space Station, astronauts exercise on the Advanced Resistive Exercise Device (ARED). Researchers at the Digital Astronaut Project create simulations, such as those pictured here, to predict the forces that the device places on muscles, bones, and joints in microgravity. These models and simulations are validated using the 7009 standard. Courtesy of: NASA Digital Astronaut Project.*

were extended to exercise simulations in microgravity, this resulted in a relative reduction in the overall credibility score of the models by about 25 percent. This lower score tells you that you will need to supplement the simulation results with other evidence to inform research or decision-making, Mulugeta says.

7009 also weighs people's qualifications and a model's use history—two features that can help increase confidence. "People I work with in the biomedical community will say, 'oh, this model doesn't take into account this parameter, so it's no good to me,'" Myers says. But then they look at the use history—what others have used it for—and that "tends to win people over pretty quickly," he says.

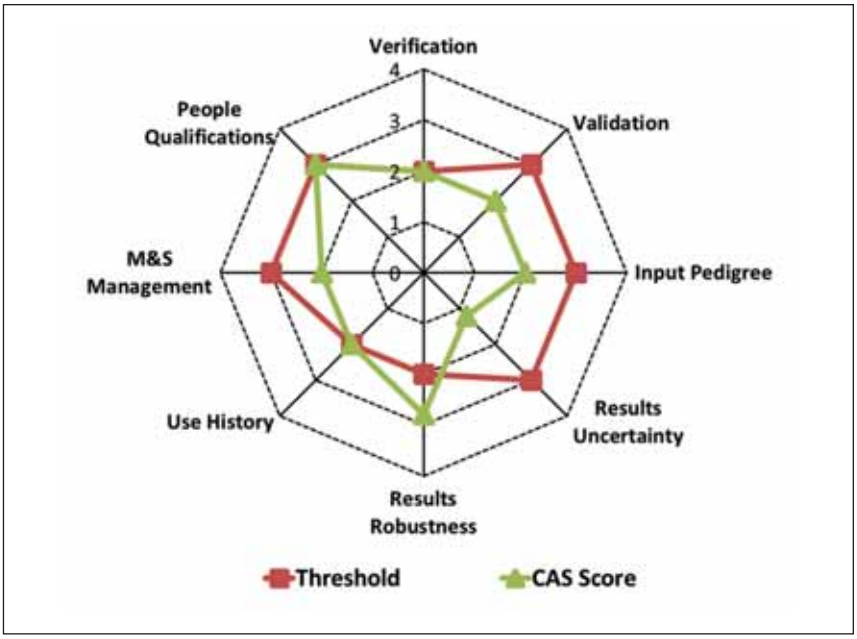
7009 has helped boost confidence in the Digital Astronaut Project. The models are now being used for completely unanticipated problems, sometimes without the modeling team's knowledge, Mulugeta says.

Another critical feature of 7009 is that it gives explicit weight to uncertainty and sensitivity analyses. "These two things are the keystone," Myers says. "Everyone worries about validation. But even if you get your validation close to perfect matching, you'll always have a cone of uncertainty that surrounds the data in your model." For decision-making, you have to understand uncertainty and sensitivity, because these are what indicate how far off the model's answer could be from the truth. "Uncertainty and sensitivity also imply how the model can be interpreted when used 'near' where it is validated, but not directly at the state in which it was validated," Myers says. This is particularly helpful when decision-makers have to make a decision involving multiple scenarios, factors, and mission goals, he says.

Uncertainty and sensitivity analyses can also clarify a model's weaknesses. For example, an allied team working on the Integrated Medical Model (IMM) modeled the risk of astronauts getting a hip fracture in space (a concern because astronauts experience accelerated bone loss) while wearing a cushioned spacesuit. Without data on how much a spacesuit actually reduces impact from a fall, Myers built his model using data on the cushioning effect of medical hip protectors (worn by the elderly to prevent fracture). But because the commercial systems they tested varied widely in terms of their abilities to dissipate and disperse an impact, the model produced large error bars around the risk estimates. So, NASA agreed to pay for a more appropriate dataset gathered using actual spacesuit material. This greatly reduced the uncertainty and improved the credibility score, Griffin says.

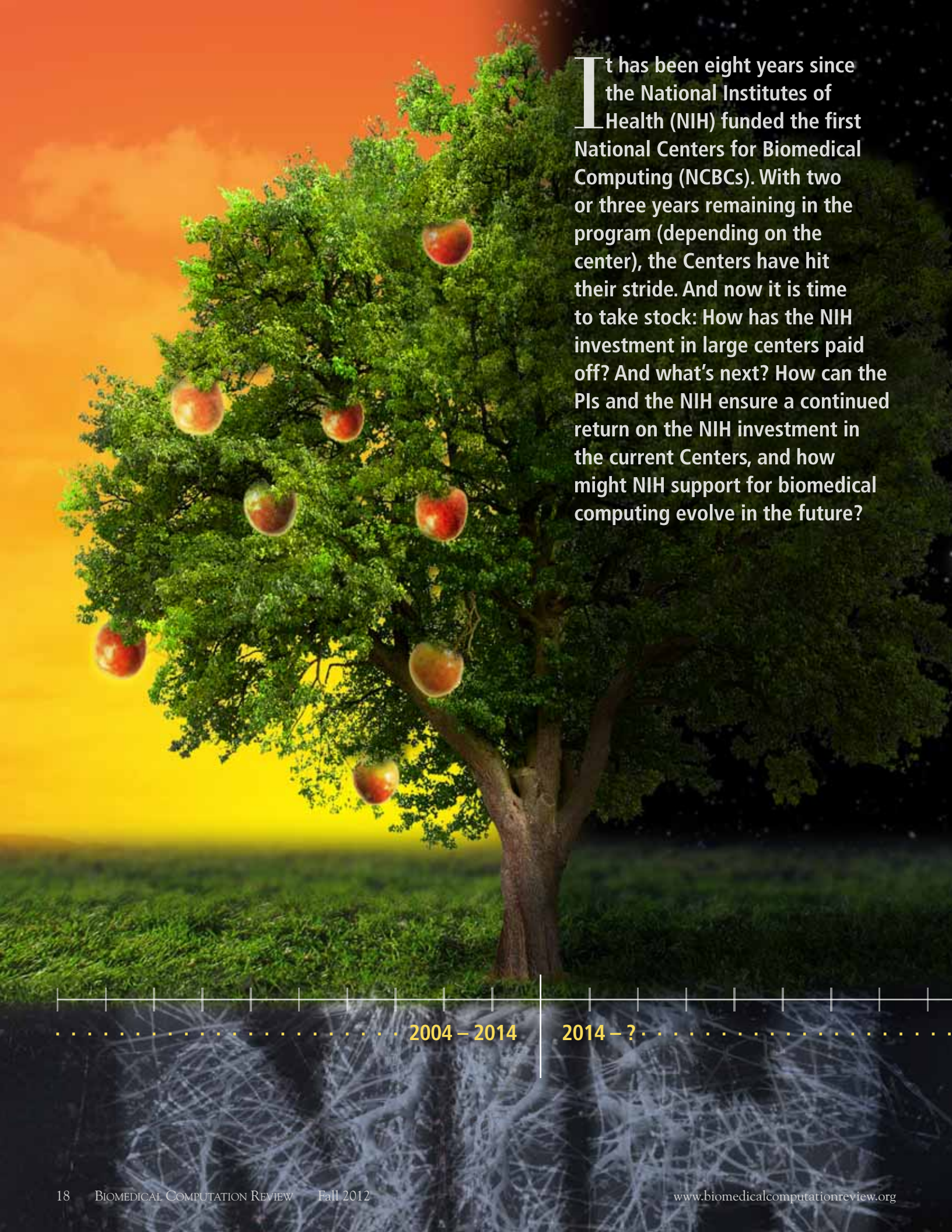
This is a good example of how, when done correctly, validation not only builds confidence in a model, but actually drives scientific research. "Validation tells you where to put your money and helps you make intelligent decisions about where to drive the science," Lyster says.

He adds: "I have been saying that validation is the organizing principle for scientific computing. A bold assertion, but I think it is that important. It's not about building a perfect model (there isn't one) but rather about seeking to quantify how imperfect your model is. In doing that you also understand more about the underlying science." □



Credibility Revealed. *The NASA 7009 standard is a synthesis of eight domains. This radar plot shows the detailed credibility breakdown for a hypothetical example. The threshold score for each domain represents the highest score that can reasonably be obtained for a given model and application—an acknowledgement that a 4.0 is often not possible for biomedical models. The Credibility Assessment Scale (CAS) score is the actual score obtained for the same model and application; some of the domain scores combine both internal and external assessments. The overall credibility is defined as the minimum of the eight domains (here 1.2—for Results Uncertainty), but NASA's Human Research Program is working to establish a process for calculating a weighted overall score based on the relative importance of each CAS factor for a given application. Courtesy of: Jerry Myers, Lealem Mulugeta, Marlei Walton, PhD, Integrated Medical Model Project Scientist, and Emily Nelson, PhD, senior research engineer at the Digital Astronaut Project.*

"I have been saying that validation is the organizing principle for scientific computing," Lyster says. "A bold assertion, but I think it is that important. It's not about building a perfect model (there isn't one) but rather about seeking to quantify how imperfect your model is. In doing that you also understand more about the underlying science."



It has been eight years since the National Institutes of Health (NIH) funded the first National Centers for Biomedical Computing (NCBCs). With two or three years remaining in the program (depending on the center), the Centers have hit their stride. And now it is time to take stock: How has the NIH investment in large centers paid off? And what's next? How can the PIs and the NIH ensure a continued return on the NIH investment in the current Centers, and how might NIH support for biomedical computing evolve in the future?

2004 – 2014 | 2014 – ?

NCBCs Take Stock AND Look Forward

Fruitful Centers Face Sunset

By Katharine Miller

The Payoff From National Centers

The NCBCs were established with an ambitious goal: to build a national infrastructure for biomedical computing. Such a mission requires investment and organization on a scale that goes beyond what a small entity can provide, says **Ron Kikinis, PhD**, Director of the Surgical Planning Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, and PI for the National Alliance for Medical Image Computing (NA-MIC). Kikinis compares the need for large, complex centers to the need for complex organizations to build superhighways and bridges. "A do-it-yourselfer might be able to build a room partition, but building a six-lane highway and a bridge across a river is not really something that's DIY," he says.

Moreover, the success of bigger, more complex projects must be measured differently from smaller ones: a partition need only stay put and look good; a six-lane highway and a bridge should carry traffic and take people somewhere—because it will be something large numbers of people rely on.

Thus, whereas some research projects are rightfully judged by whether they produce publications in high impact journals, the NCBC-built infrastructure must be evaluated by whether it is doing something bigger—creating computational environments and tools that could not have been created otherwise; and providing resources researchers and clinicians can—and do—rely on.

The NCBCs have done both. They have accomplished a number of things that would not be achievable with uncoordinated investigator-initiated R01-type research (the bread-and-butter of NIH research grants), says **Isaac Kohane, MD, PhD**, professor of pediatrics at Harvard Medical School and a principal investigator for the NCBC called i2b2—Informatics for Integrating Biology and the Bedside. And researchers everywhere are relying on NCBC resources.

"Big projects have huge benefits per dollar added," says **Art Toga, PhD**, professor of neurology at the University of California, Los Angeles and PI of the Center for Computational Biology (CCB). "With a coordinated effort, people complement each other in terms of specialties and disciplines. And collectively they create a whole that is bigger than the sum of its parts."

Here, we describe some of the major payoffs of the NIH investment in the NCBC program.

• Efficient Production of Hardened, Professional-Grade Software

"The NCBCs have produced a bunch of hardened, high quality software at professional or near-professional levels of quality that wouldn't exist without the NCBC program," says **Russ Altman, MD, PhD**, professor of bioengineering, genetics, and medicine at Stanford University and a PI for Simbios, the National Center for Physics-based Simulation of Biological Structures. "And people all over are downloading [NCBC software] and using it." As a fairly conventional measure of NCBC success, Altman says, this one is huge.

And it's novel: Academic centers, built around educating graduate students, aren't typically set up to create such professional products. "Software typically doesn't outlive an R01," Kohane says. That's because there's an 80/20 rule with software, he says: 80 percent of the success comes with a 20 percent effort, "but if you want anyone else to use it, you have to work on the hard side of the rule: use 80 percent effort to achieve the last 20 percent of the work."

To take on that hard side of the effort, the centers had to create a new kind of institution within academia—an institution with an executive director and professional programming staff. Once they were up and running, the professional products started to blossom and take hold. "Hardened software created over tens of man-years of effort has a much better chance of being taken up by others," Kohane says.

By building hardened software tools at large academic centers, the NCBC program also enabled economies of scale. For example, Altman says, a shared programming staff built Simbios' two

main products—OpenSim (biomechanical simulation software) and OpenMM (software for accelerated molecular mechanics simulations on high-performance computer architectures)—despite the fact that they operate at very different scales (musculoskeletal and molecular). It's an example of the NCBCs taking advantage of their size to produce infrastructure efficiently.

"Hardened software created over tens of man-years of effort has a much better chance of being taken up by others," Kohane says.

Moreover, says Lyster, "a lot of the software that has been created has been done in a mechanism where others could contribute to the code and algorithms." In this way, work that might have been developed in a specific biomedical context is being extended and enhanced to address similar or even fairly unrelated questions in entirely new contexts, he says.

For example, people who had nothing to do with the core of i2b2 are now building extensions, such as for natural language processing components. And a Canadian group has developed research software for adaptive radiation therapy based on the NA-MIC Kit. Similarly, some broadly used software packages have built NCBC products into their back end. OpenMM, for example, is now part of the widely used molecular dynamics packages CHARMM, TINKER and GROMACS.

• Well-Established Open-Source Software Repositories and Web Services

Almost every NCBC created a software repository using state-of-the-art software management systems, says **Peter Lyster, PhD**, program director in the Division of Biomedical Technology, Bioinformatics, and Computational Biology at the National Institute of General Medical Sciences. These allow developers around the world to contribute to the development process and include version control systems to track the provenance of software changes. "[This] gave developers the confidence to know that there's no mistake you can't undo." Lyster says. "And these centers brought that to fruition for biomedical computing."

Software repositories were also an area where the NCBCs fulfilled their charge of collaborating with and learning from one another, Lyster says. When they were first funded, he says, NA-MIC already had a highly advanced repository whereas Simbios was starting from scratch. "At working group meetings," Lyster says, "we would say: Take Kikinis' chief software engineer and have him tell Simbios how he set that up." And now Simbios has a highly professional repository called Simtk.org that looks and operates a lot like NA-MIC's original ITK/VTK repository. "That's an intangible advantage of these centers: It's very hard to quantify that we had working groups to make sure we all knew how to build professional-grade software repositories," Lyster says.

Open-source repositories are valuable for three different constituents, Lyster notes. They facilitate the work of developers who create projects in the repositories; they make software freely available to users; and they empower a large in-between group of user-developers—people who want to see what they are getting and then fiddle with it to create something new.

In addition, says **Lucila Ohno-Machado, MD, PhD**, associate dean for informatics and professor of medicine at the University of California, San Diego and PI for iDASH, the NCBC for integrating Data for Analysis, Anonymization, and Sharing, "The NCBC repositories benefit small and mid-sized institutions that often would not otherwise have access to biomedical data and computational infrastructure, such as high performance computing and processes to facilitate the execution of data use agreements."

The NCBCs also created extremely valuable web services such as the National Center for Biomedical Ontology (NCBO) BioPortal. This has become the go-to place for finding ontologies—sophisticated methods for annotating data to maintain deep connections that assist in revealing underlying knowledge. NCBO's Bioportal houses more than 350 biomedical ontologies and controlled terminologies, and its web services receive upwards of 3 million hits per month.

• A New Way to Locate Software Resources

To make it easier for people to reliably locate, publish, and access both software and data, the NCBCs created Biositemaps, a tool that enables contributors to annotate their software and data resources in a standardized way. The Biositemaps annotations are input into a web-based search engine called the Resource Discovery System (http://biositemaps.ncbcs.org/rds)—a joint creation of the NCBCs and diverse biomedical researchers—in what Lyster calls a "fascinating volunteer effort." Here, a user can, for example, search for "gene expression" and find 46 relevant tools including several from two different NCBCs and several more hosted at Simtk.org.

The NCBCs were uniquely positioned

"Surveying our users, many tell me that they'd have to abandon their research if our resource were to disappear," Musen says. "They'd have the rug pulled out from under them."

to create such a tool because they covered such a diverse set of biocomputational areas. Despite the system's breadth of coverage and ease of use, Lyster says, "Getting widespread adoption of any method for locating resources is still challenging." A number of other options exist and there's no community consensus on the best way to do it, he says. It's a problem that's ubiquitous, and not specific to the NCBCs or Biositemaps, he notes. "It's just one of the issues being tackled."

• Inspiring the Next Generation of Computational Biologists

In addition to directly training more than 400 computational biologists, the PIs say, the centers have inspired many others to consider the field as a career and have built a sense of professional identification with computational biology. Indeed, Kohane says, by funding the centers, the NIH sent a message to the quantitative community that, "yes, there are careers and support to be had in this area and so it's okay to invest your life in this field. At the time, that was not obvious to computational individuals."

As Altman puts it, "The centers have produced a cadre of ex-students and postdocs who now have a professional identification with computational biology." This includes many who are now in young faculty positions and have a research program in academia, he says. Because of their past ties to the center, many Simbios postdoc alums still use Simbios resources, get seed grants from Simbios, collaborate on Simbios workshops, use Simtk.org for dissemination, and work with the Center's software developers on enhancements to code they originally developed while at Simbios.

Moreover, the Centers' impact reaches beyond the funded trainees, Altman notes. "Simbios didn't fund that many graduate students, but students were affiliated with all of the Simbios projects, so there's this deflected involvement. It made grad students feel that computational biology was something they could do."

The centers also provide a rare opportunity for graduate and postdoctoral students to quickly turn new ideas into practical applications. By creating this setting and allowing trainees to be active participants, the NCBCs promote interest in industry careers for those who do not necessarily want to pursue academic positions, says Ohno-Machado. This expands the horizons for trainees and fills an important gap in building capacity in biomedical computing.

Leslie Derr, PhD, program director for the NIH Common Fund, agrees that the training component is one of the NCBC program's strengths. She also says that the training influences not only computational biologists but experimental biologists as well, the latter particularly at the NCBC for Multiscale Analysis of Genomic and Cellular Networks (MAGNet) which promotes a close integration of the two fields.

The NCBC program also helps MAGNet attract top-level students, says **Andrea Califano, PhD**, professor of chemical systems biology at Columbia University and PI for MAGNet. "Before we had an NCBC, most of the students we accepted went elsewhere (Harvard, MIT, Stanford)," he says. "Now they come here because of the effort to create an integrative program."

• Scientific Productivity

Great publications can happen with money in the absence of centers, Altman says. But the substantial sums provided for the NCBCs certainly enabled significant scientific productivity. All told, more than 1750 papers mention the eight NCBC grant numbers, and 35,000 others cite those.

Califano notes that about one-third of his center's publications are in journals with impact level above 15. He also gives MAGNet credit for specific developments in biomedicine: "We've come up with a new way of thinking of DNA as a molecule; we've combined structural and functional biology; we have a new ability to reprogram cells; the list goes on. It's a breadth of discovery that a center allows you to have rather than a single success story."

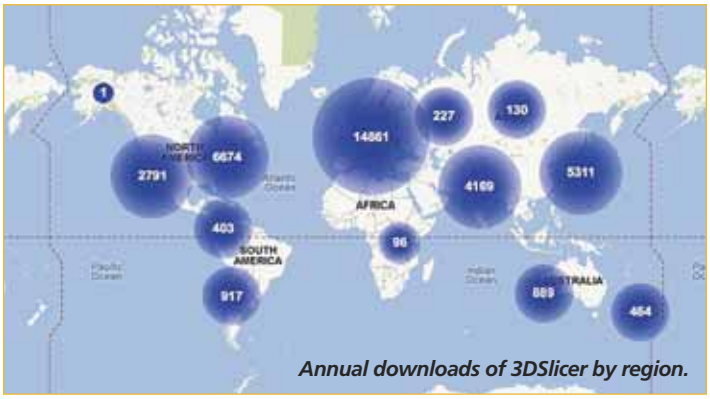
Kohane says the critical mass of the NCBC program also accelerated the arrival of solutions, some of which were fortuitous rather than planned. For example, i2b2 never had any ambition to do phar-

By funding the centers, the NIH sent a message to the quantitative community that, "yes, there are careers and support to be had in this area and so it's okay to invest your life in this field," Kohane says. "At the time, that was not obvious to computational individuals."

For Simbios, Altman says, "the big story is that GROMACS, TINKER and CHARMM [several of the most widely used molecular dynamics programs] now have OpenMM as their back end because it's faster and better." It means that OpenMM is now at virtually every drug company in the world, he says. "If that was the only thing Simbios did, it's huge."

macovigilance, but once they had liberated data from archived health records (with the goal of doing genomic research), they found that there were some amazingly low-hanging fruit in this area. "We could easily see appallingly obvious signals of drug-induced adverse events that had gone unnoticed." For example, they mined electronic health records to confirm the association between heart attack deaths and Vioxx and to identify a similar risk from the drug Avandia—information that is now on the drug's warning label. "The NCBCs experienced lots of such examples where in addition to the primary objectives, having a critical mass of people led to unanticipated progress," Kohane says.

As another example, the University of California organized a system to perform federated queries on data derived from electronic health records at its five medical centers, which collectively represent over 11 million patients. The hub of this system is provided by the iDASH NCBC coordinated at UCSD. "The system is designed to use software developed by three different NCBCs and lessons learned from all," says Ohno-Machado.



Annual downloads of 3DSlicer by region.

It's biomedical impacts like these that really made the NCBCs a scientific success story, the PIs say. And for that they credit the driving biological problems (DBPs) associated with each center. The DBPs keep the computational scientists focused on the science but also allow the centers to create robust software that can be extended and enhanced to address novel questions, Derr says.

• A Community and a Network of Leadership

Before the NCBCs came into existence, the field of biomedical computing had pockets of spontaneous collaboration in particular areas, but not the strong sense of community or common purpose that were enabled by a common fund and complementary expertise, Altman says. By empowering a disparate group of researchers to work together on a national infrastructure, he says, the NIH changed that.

"The NCBCs created a critical mass of computationally competent individuals working for a common biomedical purpose," Kohane says. "The critical mass raises the overall tenor and quality of the conversation. Otherwise everyone is an island."

"The NCBCs created a critical mass of computationally competent individuals working for a common biomedical purpose," Kohane says. "The critical mass raises the overall tenor and quality of the conversation. Otherwise everyone is an island."

And from that community a network of leadership at the NIH and at NCBC institutions emerged. "There is now a functional group that can think about and respond to issues of biomedical computing at a policy level," Altman says. "Before, there was nobody to point to and say, 'they can help us.' And now we have a group of centers and a variety of staff who researchers and administrators can come to for help, advice, or counsel."

INFRASTRUCTURE SUCCESS: Galaxies of Reliance

In addition to providing the payoff one can only get from large centers, the NCBCs made significant strides toward creating a national infrastructure for biomedical computing, says **Mark Musen, MD, PhD**, professor of medicine at Stanford School of Medicine and PI for the National Center for Biomedical Ontology (NCBO).

It's an infrastructure that can seem ephemeral, Kikinis says, because it's all "executed as electrons" and virtualized on computers. But he points to a map of NA-MIC's downloads as proof of a real infrastructure. "This shows for me the worldwide demand for what we are doing," he says. "In a nutshell: We are addressing somebody's needs."

And that's true for all of the Centers. After eight years, 84 hospitals rely on the i2b2 platform; 22,000 researchers use

"The NCBC infrastructure is always on the web and always virtualized, and so it's invisible," Kikinis says. "If NA-MIC were a supercomputing center I would show you the room with the big servers. The map of our many thousands of downloads is the equivalent of that, and it shows, for me, the worldwide demand for what we're doing. That's in a nutshell what we are accomplishing: We are addressing somebody's needs."

NCBCs by the NUMBERS

1 nonprofit created
by NCIBI to guide tranSMART code base and community development in collaboration with Pharma

84
Academic Health Centers adopted the i2b2 platform

62,000
downloads of OpenSim by 9600+ individuals including ~30 hospitals

35,000
annual downloads of 3D Slicer, a key part of the NA-MIC kit

1750+
papers in Google Scholar cite the NCBC grant numbers (and 35000+ other papers cite these)

2,000 *daily downloads of CMake, a NA-MIC software process tool*

22,000
members of Simtk.org

23,070
Downloads
of MAGNet software tools geWorkbench and Aracne

350+
biomedical ontologies on NCBO's Bioportal

3 million
calls per month for NCBO web services

20
end-to-end computational pipelines implemented by CCB

10,412,259
page views of wiki.na-mic.org, an open wiki for NA-MIC

3
widely-used molecular dynamics programs using Simbios' OpenMM as their back-end

11 million
Patients' data accessed via iDASH technology hub and shared across the University of California system using i2b2's SHRINE

400+
trained grad students and post-docs (and many others inspired)

10,000
downloads of CCB's Pipeline Processing Environment

83 Collaborative projects
use MAGNet methods and tools

15 billion
annotations in NCBO's Resource Index

1
Biositemaps Resource
a national standard for storing, representing and curating bio-medical data, tools, and resources

1 million annual web hits
for NCIBI's web services

Simtk.org; 3 million calls a month hit NCBO's web services; 83 collaborators count on MAGNet methods and tools; and daily, all around the world, upwards of 100 people grab the latest 3D Slicer from NA-MIC's web site and use it to analyze images of patients with a whole range of diseases. And then there are the thousands of other downloads that demonstrate widespread reliance on NCBC tools. A sampling of these are shown in the *NCBCs by the Numbers* chart on page 23.

A worldwide community depends on NCBC products. Perhaps this is the best way to think about the success of NIH infrastructure grants: While high impact journal publications count for something (and the NCBCs have produced more than their fair share as mentioned above), perhaps, Altman says, having a high impact in the world of clinical and biomedical research is a better indicator of whether the centers are fulfilling their mission.

The piece of infrastructure that can be enhanced as the centers mature is interoperability. At present, Lyster says, "The centers are like galaxies in that they are separate and non-overlapping. There are clumps of foci with some collaborations in interstellar space." As galaxies, he says, "they've been stellar," but the original centers were hamstrung in this regard because NIH funding didn't cover the waterfront: many areas of computational biology and medicine were not represented by the eight centers. Even with the NCBC Collaborations program, which created at least 33 spokes for the hubs, opportunities for interoperability only scratched the surface, Lyster says. "We don't know how much further we could have gotten with interoperability if we'd covered the waterfront better."

The June 2012 Draft Report of the Data and Informatics Working group of the Advisory Committee to the Director of NIH (ACD DIWG) noted that the problem was structural: "… due to the limited funding of the NCBC program and to the size of the overall research area, there is virtually no overlap of focus among the current centers. As a result, there has been less opportunity for synergy and complementary approaches of the type that have universally benefited the research community in the past."

Kikinis says that the centers naturally evolved into a hub and spoke model with ecosystems of collaborators. "It's a bit different from the way NIH envisioned it would be," he says, "but the way it evolved, the NIH got a lot of bang for the buck." The centers really provided a lot of enabling infrastructure for NIH grantees, he says. "So from my point of view, the program accomplished what the RFA [Request for Applications] intended it to accomplish."

Map of Centers that have adopted the i2b2 platform for clinical research.

- Clinical & Translational Science Awards centers (CTSAs) adopting i2b2 platform
- CTSAs evaluating i2b2 platform
- Academic medical centers adopting i2b2 platform
- Foreign medical centers adopting i2b2 platform

What Might Have Been Done Differently

• More Centers (and/or More Funds)?

When the NIH made the first announcement of NCBC awards in 2004, it was clear that some areas would be well served and others would not, Altman says. Eight years in, that problem has only grown worse as the importance of computation swelled, leaving unserved areas behind.

As examples, Altman cites genomics and natural language processing. With the advent of next-generation sequencing, genomics data has been confounding

"We never anticipated that there would be so much uptake nationally of i2b2 software across medical centers," Kohane says. "The fact that there was—despite the existence of alternative commercial offerings—told us in retrospect that there really was a huge need that required a concerted effort by a group of biomedically oriented computational researchers."

biologists. "People are now scrambling to handle that output," Altman says. "A center dedicated to handling and analyzing genomic data would have been a great idea." And as for natural language processing (NLP), "Guess what," Altman says. "Our entire understanding of biology and medicine is really contained in the published literature. And since people write in natural language, if you can't get computers to turn that information into databases and computable information, you're falling behind." Had there been an NCBC for NLP, he says, database managers wouldn't be hiring people to read the literature and distill it for others in computable format, which is what they're doing now.

The fact that the current NCBCs only covered a small portion of what is needed in terms of biomedical computing for the country is also described in the ACD DIWG Draft Report. It says: "the small number of active [NCBCs] has not covered effectively all relevant areas of need for biomedical computation or for all of the active contributing groups." The report specifically cites the lack of coverage for a number of grand challenges in biomedical computing such as multi-scale modeling; methods for "active" computational scientific inquiry; and comprehensive, integrated computational modeling/statistical/information systems.

• A Grand DBP

Kohane suggests that the NIH might have asked centers to participate in one grand driving biological project. "It's a sterile exercise to say 'let's share data,'" he says, "but to say 'let's solve this problem together,' that's much more tangible." He even has a problem in mind: obesity. "It would have involved new ways of imaging, new ways of doing genetics, new ways of integrating different modalities, possible simulations of the effect of weight on organs or the human body," he says. "It would have gotten interesting conversations going."

• Dedicated Training Funds

Knowing what they know now, the NCBC mission could have been well served by separate training grant funds or a companion center for training, say Musen and Altman. "Frankly, had we had that, we would have had a larger impact," Musen says. NCBO and other centers took advantage of existing training programs and leveraged those. "But if we really wanted to train the next generation of computational biologists who would inherit the work of our existing centers and carry that on, there could have been more dedicated training funds associated with the NCBC program to assure that," Musen says.

Lyster points out that, for the NIH, it's all about balancing competing needs. "I think that given that we started out with no centers and now have had eight with hundreds of students graduated, it's hard to think that's not a good thing," he says. "In fact, it's a very positive thing when a student is forced to confront both the biological and the computational question at once."

PREPARING THE CENTERS FOR SUSTAINABILITY

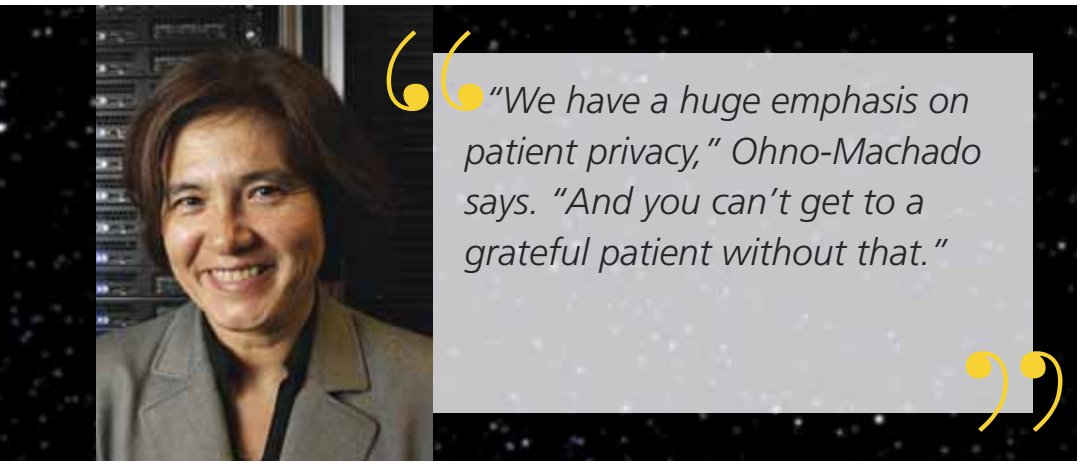
In their first eight years, the NCBCs made huge advances, but there is much more to do. And the needs that motivated funding for national centers haven't evaporated, Ohno-Machado notes. So what's next for the current centers when they hit their ten-year expiration date?

As Derr sees it, "These centers have always been aware that their funding ends within ten years, so they know that they need to think about sustainability." The center PIs need to think about what kernels they need to sustain into the future and how those might be funded, she says. "There are certainly other programs where they can compete for funds, and they would certainly be competitive in applying for them." Each center also has an external board helping them think about sustainability.

But the question remains: What can the PIs do to make sure the centers stay funded? Will there be a new program or will they cobble together a number of other approaches? As might be expected, the various PIs are each taking their own approaches to the problem. Califano, for example, plans to apply for one of the systems biology programs. And Kohane hopes that i2b2's open source repository will be taken over by the Harvard Medical School's Center for Biomedical Informatics.

Brian Athey, PhD, professor of computational medicine and bioinformatics at the University of Michigan and PI of the National Center for Integrative Biomedical Informatics (NCIBI) says his center, which was not renewed in 2010, has always taken sustainability seriously. They have already spun off two smaller efforts—a regional metabolomics center and a rare disease center—and are now deeply involved in tranSMART, an international effort to create a knowledge management platform that integrates, normalizes and aligns genetic and phenotypic data primarily for drug discovery. For financial support, Athey is talking to foundations, the FDA, the Veteran's Administration, and pharmaceutical companies. "Anybody who wants to sustain their efforts has to do that. And we are. We are deeply involved in that."

Kikinis and Altman point to the P41 Biomedical Technology Research Center (BTRC) program established by the former NCRR (National Center for Research Resources) as a good conceptual model for how the NIH might continue to support the NCBCs or portions of them. Like NCBCs, BTRCs support the development of technologies that are then made available to the research community, but unlike NCBCs, the biological problems driving BRTC technologies are funded separately. As a model for what's possible, the BTRCs have two additional features that Kikinis



"We have a huge emphasis on patient privacy," Ohno-Machado says. "And you can't get to a grateful patient without that."

says would work well for the NCBCs: They have no set end date and are subject to review by ad hoc study sections consisting of people with appropriate expertise. "To create innovative new infrastructure for biomedical research, the P41s are a good model," Kikinis says. "If I were in charge (which unfortunately, I am not!), I would treat the existing centers as resource centers and then do a limited RFA to capture a few additional centers—however many NIH is willing to fund."

Toga says that discussing whether P41s are a good model for the NCBCs is putting the cart before the horse. "The first decision is whether a national network of computational biology centers is a worthwhile endeavor." Answer that in the affirmative, he says, and shoehorning NCBCs into BTRCs is not the way to go.

Kohane is hopeful of a solution. "To the extent each NCBC has friends and supporters, I think there will be a lot of creativity both in the public and private domains to support continued efforts."

THE POSSIBILITY OF A NEW PROGRAM

Toga hopes that before the NIH starts designing a new program of national centers, it will conduct a meaningful programmatic evaluation of the current centers as well as of its own role in managing the centers. Just as Apple, Inc., reviews the performance of the iPhone 4S before designing iPhone 5, Toga says, the NIH needs to investigate how the research community benefited from the NCBCs and what it wants from future centers before designing the next iteration. And because the NCBCs were run as cooperative agreements, the NIH needs to turn that same critical eye on itself: Did the NIH manage, evaluate, and review the centers in the best way possible? Did their management impact the centers' success?

In the meantime, the ACD DIWG Draft Report is adamant that promoting further development of biomedical computing in a coordinated manner is critical to justify large investments in "big data" collection that will need computational analyses. Without such computational infrastructure, data will remain underutilized and stored in independent silos rather than made available as a national resource. The DIWG suggests possible next steps might include creating a larger number of national centers that are smaller in size, complexity, and scope.

"To the extent each NCBC has friends and supporters," Kohane says, "I think there will be a lot of creativity both in the public and private domains to support continued efforts."

BIG VS. SMALL

There is no definitive recipe for success moving forward. Altman, who is part of the DIWG, thinks smaller centers could work. He suggests 20 to 25 centers with smaller budgets—perhaps two million dollars per year rather than the current approximately four million. "With 25 centers covering biomedicine more broadly, you're talking about an infrastructure that would be really robust," he says. And there should be no official end date, he adds.

In a new model with smaller centers, Altman says, it's possible that several Simbios researchers would seek to create centers—perhaps a national biomechanical simulation center (with ongoing work on OpenSim) and a national center for molecular modeling (with ongoing work on OpenMM). "The big centers allowed us to have an umbrella over these two physical programs," he says. "In a new approach, these would be broken up, but they'd already have a bridge between them."

As for the details of funding smaller centers, several options exist. For example, Altman says, if the centers paid only for their part of the DBP projects (rather than for the biomedical research itself), the centers' budgets would drop by about 25 percent. "We can just say to application scientists, 'you already have money to do *x* from regular research grants, and we'll get the money to do the computational piece.'"

Kohane and Musen, however, think big centers are essential. "Big is good because of critical mass," Kohane says. Large centers attract good trainees, he says. And they empower computational researchers to produce durable software—rather than just answer biologists' question-of-the-day. If the centers were smaller, Kohane says, the power relationship between computational scientists and biologists would return to business-as-usual. "They'd just want their research done and not be particularly interested in us developing our software for others."

Musen has some other concerns about reduced resources for smaller centers. "This recommendation is probably the most pragmatic thing the advisory committee to the director could have recommended," he says, "but it will not in any way allow the kind of large scale development that the current NCBC program fostered." With fewer resources, he says, "We'd be doing more maintenance and less innovation. That's obviously not nearly as exciting for us."

"The value of the centers is not to have one success story," Califano says, "but to have a major impact in terms of the ability of other organizations to cooperate in areas that would otherwise be impossible to address."

COVERING THE WATERFRONT

If the NIH decides to fund more and smaller centers, the question remains: how to ensure that the centers cover the waterfront? Should the selection be top-down or bottom-up? Opinions differ. "I think I would target some areas of need instead of

being a free call," Ohno-Machado says. "The NIH could identify the needs that are most important and then highlight those for reviewers."

Toga agrees. "Researchers' computational needs might be better met by identifying where the biggest needs lie and developing clusters of opportunities around those, rather than having an open call," he says.

But Altman has a bottom-up view of the program. "I think you should put as much of the so-called decision-making into the hands of the scientists on the ground who write their best grant applications and let the chips fall." There's plenty of opportunity for coordination after the grants come in, he says. He has another big fear: "Two million dollars and a ton of mandates. To get the best people, you need to give them substantial freedom to do what they should do."

Meanwhile, Califano would ensure greater connectivity among the centers by creating two programs—one for clinical informatics and another for computational and integrative biology. "That would create a set of constituencies that speak the same language, where right now we have cats and dogs in the same room."

But Musen disagrees with this dichotomy. "Most of the centers are creating national infrastructure that's applicable in a variety of domains," he says. "I2b2 is very clinical and MAGnet is very molecular, but other NCBCs are in the middle. Imaging applies to cells and to people; physical simulation applies to molecules and muscles; ontology work involves data analysis from both clinical and life sciences domains."

Where the answer lies—and how NIH will respond to the DIWG—remains to be seen.

INCENTIVES FOR INTEGRATION

More centers might also enable interoperability by covering more of the waterfront. That hope is specifically stated in the DIWG Draft Report: "The NIH should also encourage and enable more overlap between centers, to facilitate collaboration."

Califano says that if the NIH launches a new program and wants more interoperability among centers, it needs to have more thematic overlap between them. Altman concurs: "One of the lessons learned is there needs to be a finer sampling of computational space. The gaps can still be significant but need to be bridgeable with a reasonable amount of effort." He would also encourage the NIH to provide incentives for integration, so that it happens spontaneously. "In the second generation, it might happen, but it can't come from top-down rules."

KEEP IT GOING

All of the PIs believe the current centers need to be sustained in some way. "If they were to terminate abruptly and no longer get any kind of funding, it would be a shame because they have been tremendously productive," Califano says.

Musen concurs: "Creating a national infrastructure is not something that can be done in 5 or 10 years. That kind of goal needs a longer time horizon for the development phase."

Kikinis points out that it takes five years or more for a newly funded NCBC to become fully functional. Now that the centers are operating properly and are past all the startup problems, he says, "In my opinion it would be a waste of NIH money to throw away these centers and start from scratch." Of course, he acknowledges, there's always a risk of complacency—but that's the role of appropriate and rigorous review and competitive renewals. Kikinis admits that he has a self-serving agenda, "but in all honesty, it's the right and proper thing; the centers haven't reached the point where they are non-productive." In the case of NAMIC, he says, "I can say without any hedging that there's a lot left to do."

Kohane says, "Whether it's us or others, we need to have centers." That's because, he says, "Biomedical research is underpowered with respect to its computational resources." Investment in biomedical computing, he says, "will lead to a big multiple in terms of yield because there is such a computational desert in terms of the availability of skillsets and people."

"Whether it's us or others, we need to have centers," Kohane says. That's because "biomedical research is underpowered with respect to its computational resources." Investment in biomedical computing, he says, "will lead to a big multiple in terms of yield because there is such a computational desert in terms of the availability of skillsets and people."

There's a need for centers because data acquisition is outstripping data analysis, Toga says. "I have hundreds of collaborators who have no idea what to do with their data." Funds for analytic strategies are needed, and large centers are the best way to produce them, he says.

Athey says there's a role for continued federal support of NCBC resources. "There's a growing dependency on these things and the government's role in continuing to sustain them is a legitimate conversation to have."

The 10-year sunset might have been necessary for a Common Fund program, Altman says, but it's not right for science. "Nobody thinks biomedical computing is important for these 10 years and then becomes less important. This program should become part of the ongoing fabric of the NIH with competitive renewals." Why? "Because the world needs help—the regular biologists and scientists with data are the ones who need the infrastructure we're building." □

"Nobody thinks biomedical computing is important for these 10 years and then becomes less important," Altman says. "This program should become part of the ongoing fabric of the NIH with competitive renewals." Why? "Because the world needs help—the regular biologists and scientists with data are the ones who need the infrastructure we're building."

Under TheHood

BY RONALD J. NOWLING, TREVOR M. CICKOVSKI

Prototype to Release: Software Engineering for Scientific Software

Having engineered several scientific software applications for public consumption, the authors know from experience that the process offers unique challenges. Typically, the algorithms being implemented are complex; the process involves numerous developers with various backgrounds and skill sets; and it all takes place in a fast-paced environment where new methods must be prototyped and tested regularly.

Set Clear Objectives

Our experience suggests that these challenges can be overcome by establishing a clearly defined set of objectives for the engineering process. First, there must be an intention to produce stable software releases for public usage on a regular or semiregular basis. Second, the process has to allow software designers to implement new experimental methods without disrupting the release schedule or introducing bugs or destabilizing the software. Third, the process must allow for regular, rigorous validation and testing of the software to prevent the accidental introduction of bugs and ensure ongoing confidence in the correctness of the software under continuous development. And finally, the process needs to be amenable to involvement by multiple developers with varying backgrounds and skill sets as well as different goals and priorities.

To fulfill these objectives, we recommend and describe here a development methodology that we've found to work: the iterative model. Here, we also survey some tools that we've found useful for implementing that process.

Other development methods and tools do exist, but many people don't use any process at all. We hope this article will give people some new ideas, and encourage them to both reflect on their process and research the options.

Iterate Using Split-stream Engineering

A key step is to split the engineering process into two concurrent "streams"—stable and unstable. The stable version of the code forms the basis of regular releases; it is the trusted code base known to be relatively bug-free. Each release of the stable version is guided by a development plan with well-documented features and goals. Goals should be narrow enough to be accomplished in the given timeframe and significant enough to warrant a release. And the plan should include sufficient time for testing to ensure that the release will be stable and trustworthy. If a desired feature or goal has unexpected large-scale consequences or might otherwise hold up release, it can be dropped or moved to the unstable stream. Otherwise, we follow the development plan to its completion.

Meanwhile, the unstable version serves as a basis for developing prototypes and experimental features, as well as for making large-scale changes. Occasionally, it may be necessary to "rebase" the unstable branch to the stable branch to keep them in sync. As various changes mature and are shown to be correct through rigorous validation and testing, developers can create a copy of the stable branch to work on as an extra unstable branch. Once the feature is solidified, tested, and stable, developers can merge it into the stable stream as part of a future release. If conflicts develop between the stable and unstable streams, developers resolve them in the unstable branch. This is a continual process.

Goals should be narrow enough to be accomplished in the given timeframe and significant enough to warrant a release. And the plan should include sufficient time for testing to ensure that the release will be stable and trustworthy.

Use Appropriate Tools

Hosting and release platforms such as Simtk.org, SourceForge or GitHub can be used to facilitate development and to disseminate releases to the public. The appropriate platform should be chosen based on what team members are familiar with as well as available functionality.

Use of a versioning and revision control system (RCS, e.g., Subversion, Git or Mercurial) is essential as it tracks the history of a project's source code. As programmers reach various milestones, they can commit code to the repository, which takes a snapshot of the current state of the source code and determines line-by-line what changes from the previous version have occurred. Such systems also allow programmers to develop features independently from the work of other programmers, synchronize any changes, and resolve conflicts.

RCS systems tend to follow one of two models: (1) centralized, where all commits immediately appear in a cen-

tralized repository and (2) decentralized, where all commits are made to a local copy and are later copied to the centralized repository at the programmer's convenience. The decision of which RCS to use can be based on a balance among such factors as developer comfort, preference for centralized or decentralized repositories, and availability of support by the chosen hosting platform.

Test and Validate

In many scientific software domains, software testing and validation are also crucial. Testing refers to automatically identifying any changes in output. Validation is the process of determining if an output is correct, mathematically or scientifically speaking. Here, we do not further discuss approaches to validation because they typically depend on the specific scientific question being addressed. However, the details of testing merit some discussion.

Complex scientific algorithms are not only sensitive to small bugs but also susceptible to their introduction. Testing can serve as an "early warning" system for accidental bugs, help localize bugs to a specific function, and give developers a chance to think through the correctness of any changes.

Testing can be done at two levels: unit and system. Unit testing assesses an individual unit of code (such as a function or class) by providing a bit of input data and comparing the output against expectations. Unit tests are often written in conjunction with the code being tested and they are used early and often to check code as it's being written. Because they test parts independently, the developer should design the software as independent modules and reduce concrete dependencies using techniques such as abstract interfaces. In one approach, test-driven development (TDD), the programmer actually writes the unit tests before writing the implementation. Unit testing can often be automated and made easier through frameworks such as CppUnit (C++), JUnit (Java), unittest (Python), or FUnit or FRUIT (Fortran).

System-level tests check the integrity of the entire software system. These do not help localize bugs but can be used to determine proper integration of components, ensure dynamical run-time behavior has not changed, and enable comparisons with other software packages. Moreover, they are capable of testing higher-level results such as an entire simulation.

Document Assumptions

Developer documentation (which is distinct from user documentation) is another key part of the development process. It can help prevent mistakes by keeping everyone aware of assumptions and expectations. Basic forms of documentation include the use of descriptive variable names; clearly identified units; and other commentary. Document generators such as Doxygen (C++, others), Javadoc (Java), or Pydoc and Sphinx (Python), can automatically compile all API documentation assuming properly formatted comments. A more advanced documentation method, the design by contract methodology, provides a mechanism for specifying formal requirements for input (pre-conditions), guarantees for output (post-conditions), and maintained properties (class invariants) and can be enforced programmatically through language features such as assertions (which are manually specified by the programmer) and in more automated ways through preprocessors such as GNU Nana (C/C++) and Contract4J (Java).

Producing high quality scientific software starts with developing a culture or mindset that emphasizes quality. Here, we've discussed how requirements and a clear process can make a difference; described several types of helpful tools; and offered suggestions for good practices. We hope that these software engineering principles are helpful to others facing the challenges of preparing scientific software for public use. □

DETAILS

RJ Nowling is a PhD student at the University of Notre Dame in South Bend, Indiana, and has experience with several scientific software projects. He worked with the CONNJUR group to implement and release a file converter for NMR data and a visual programming integration environment for NMR data processing software; MimoSA, a database-driven system for annotating minimotifs from literature; DARWIN, a cataloging software for dolphins that uses semi-automated computer vision techniques; and ProtoMol, a molecular dynamics package designed for prototyping new algorithms. For the last two years, RJ has TA'd for Programming Paradigms, a required course for juniors that includes software development practices (unit testing, design by contract, design patterns). RJ recently visited Stanford as a Simbios OpenMM Visiting Scholar and was awarded a GAANN Fellowship.

Trevor Cickovski is an assistant professor at Eckerd College in St. Petersburg, Florida, where he teaches courses such as Data Structures, Compilers, and Operating Systems. Trevor has worked on ProtoMol; MDLab, a Python environment built on top of ProtoMol; CompuCell3D, a modeling environment and PDE solver for simulating cellular behavior; and GPU DePiCT, a GPU implementation of software for designing degenerate DNA primers.

Most recently, RJ and Trevor have worked on OpenMM FBM, a GPU implementation of the Flexible Block Method, a fast approximation method for the diagonalization of matrices in the context of normal mode analysis of proteins.

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

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

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

seeing science

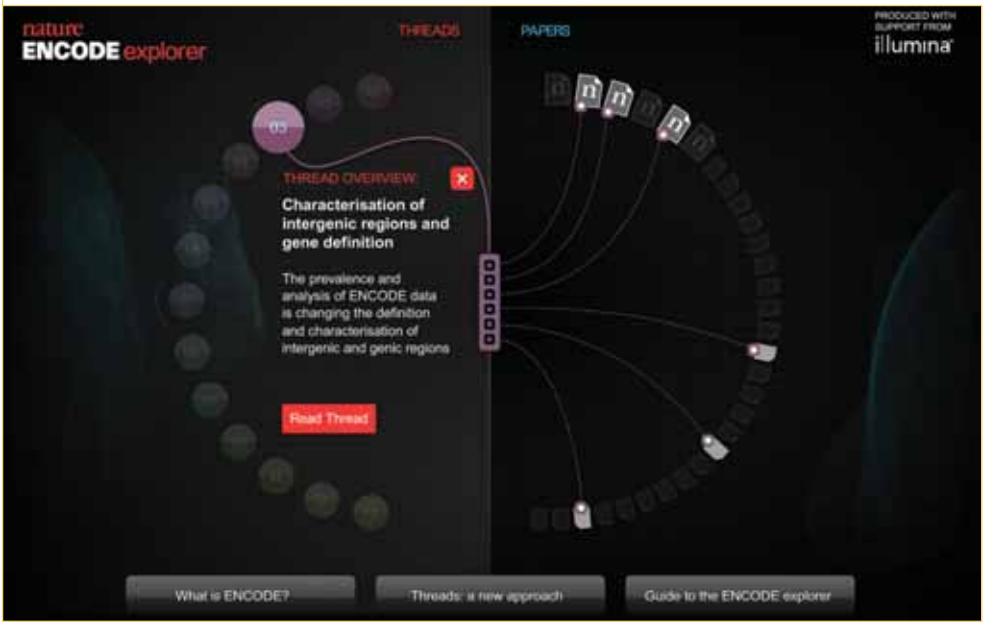
SeeingScience

BY KATHARINE MILLER

ENCODE's Threads

When a large research project generates lots of data over a long time, that data can tell many different stories. Such was the case when the ENCODE (Encyclopedia of DNA Elements) project geared up to publish its first wave of results. "They had to decide which stories were the most prominent and most complete to be told within the confines of traditional research papers," says **Magdalena Skipper**, the *Nature* editor who worked with the ENCODE project's authors.

Unfortunately, by choosing a set number of topics, she says, "other stories became fragmented and told across multiple papers." To address that problem, the researchers created a set of "threads" that pull together 13 of these otherwise fragmented stories. They then manually collected the relevant portions of each thread—a process akin to highlighting the portions (including figures and tables) of 30 papers that relate to a specific topic. So, for example, the tale of machine learning approaches to genomics became one thread; and three-dimensional connections across the genome became another.



The threads don't have a classic identity. "They aren't indexed in PubMed." Skipper says. But they provide a tool for exploring the published information through a different lens. "In an ideal world," Skipper says, "one would be able to generate these threads automatically on any topic." But current text-mining tools lag a bit—they can't, for example, adequately extract relevant figures or other display items.

To maximize the utility of a group of related papers, Skipper says she hopes *Nature* will do something like threads again. "Researchers appreciate it—it's visually appealing and the content is useful." □

To visualize the ENCODE Threads,* Nature *created this graphic. When readers click on one of the thirteen threads, the topic pops up with lines connecting the topic to the specific ENCODE papers that were extracted to form the thread. "The graphic is a guide to explain the relationship between the papers and the threads," Skipper says. Copyright © 2012 Nature Publishing Group, used with permission.