

D I V E R S E D I S C I P L I N E S , O N E C O M M U N I T Y

BiomedicalComputation

Published by Simbios, an NIH National Center for Biomedical Computing

REVIEW

BEHIND THE
Connectome
Commotion



PLUS:
Betting
on Genome
Interpretation

Summer 2013

10 Behind the Connectome Commotion

BY ALEXANDER GELFAND



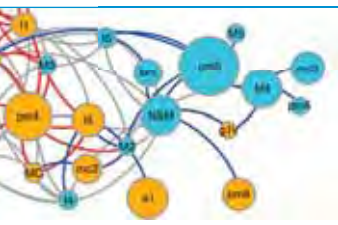
18 Betting on Genome Interpretation: Six Startups Jockey for a Place at the Table

BY KATHARINE MILLER

DEPARTMENTS

- 1** GUEST EDITORIAL | MISCONCEPTIONS OF TIME BY VIJAY S. PANDE, PhD
- 2** SIMBIOS NEWS | BENCHMARKS FOR MUSCULOTENDON MODELS BY KATHARINE MILLER
- 3** MODELING SPERM: THE FINER POINTS OF FERTILIZATION BY ROBERTA KWOK
- 5** DIGGING INTO PIXELS: RADIOGENOMICS EXTRACTS MEANING BY AMBER DANCE
- 7** MATTERS OF TIME: TICK TOCK GO THE SIMULATIONS BY KATHARINE MILLER
- 25** UNDER THE HOOD | VERTEX CLASSIFICATION IN GRAPHS BY JOSE LUGO-MARTINEZ AND PREDRAG RADIVOJAC, PhD
- 26** SEEING SCIENCE | TRAJECTORY OPTIMIZATION AND PHYSICAL REALISM BY KATHARINE MILLER

Cover Art: Created by Rachel Jones of Wink Design Studio using: Head with neurons image, © Kts | Dreamstime.com.
Page 7 Art: Created by Rachel Jones of Wink Design Studio using: Time image, © Agsandrew | Dreamstime.com.
Page 8 Art: Broken glass image, © Vladimir Yudin | Dreamstime.com.
Page 18 Art: Created by Rachel Jones of Wink Design Studio using: Casino image, © George Tsartsianidis | Dreamstime.com, and DNA image, © Anna Raspopova | Dreamstime.com.



Summer 2013

Volume 9, Issue 2

ISSN 1557-3192

Executive Editor Russ Altman, MD, PhD

Advisory Editor David Paik, PhD

Associate Editor Joy Ku, PhD

Managing Editor Katharine Miller

Science Writers

Alexander Gelfand, Katharine Miller, Roberta Kwok, Amber Dance

Community Contributors

Vijay S. Pande, PhD

Layout and Design

Wink Design Studio

Printing

Advanced Printing

Editorial Advisory Board

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

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

Office

Biomedical Computation Review
 Stanford University
 318 Campus Drive
 Clark Center Room S221
 Stanford, CA 94305-5444

Biomedical Computation Review is published by:



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

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

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

BY VIJAY S. PANDE, PhD

Misconceptions of Time



For those who are not practitioners of dynamical simulation methods, such as molecular dynamics (MD), one of the biggest misconceptions relates to time. Specifically, the mismatch between the timescales that the simulation can reach compared to what is experimentally relevant. Indeed, typical MD simulations are in the nanosecond to microsecond timescale regime. If the desired phenomenon of interest occurs on the second timescale, one would never see it.

This mismatch often leads people to say that these simulations don't work, whereas they're often doing exactly what they should be—reporting on the timescales that they purport to cover. In this sense, it's like saying that one's car doesn't work as a transportation device when you only let it run a few seconds and you never leave the garage.

a set of atoms, each modeled by its own sphere and interatomic interactions, one may set a single amino acid molecule (composed of several atoms) as the base unit, thus modeling at a much coarser scale. While coarse graining itself is quite old, what sets modern coarse-graining methods apart is the systematic derivation of models from more detailed models. For example, one might run an atomistic simulation for some time in order to derive the parameters for a simpler, coarse-grained model. This aspect of the work marks a shift away from intuition-based modeling toward data-driven, systematic methods, which has many appealing aspects.

Another approach has been to coarse grain not the interactions but the dynamics itself. If one cares about the millisecond (10^{-3} seconds) timescale, why would we want

As both spatial and temporal coarse-graining methods become even more systematic and statistically driven, one can imagine how they can start to merge to build the best physical model possible.

Once one recognizes this challenge, the natural next step is to devise means to defeat it—to get that car out of the garage, so to speak. In this issue of *BCR*, we're featuring a story that digs into this problem of time in dynamical simulations. In recent years we've seen a revolution in molecular dynamics simulations in this regard—one headed in multiple yet potentially complementary directions. Most of these approaches, at heart, depend on two key assumptions: 1) Typical atomistic simulations have detail that in some sense is not needed, and 2) One can build such models in a systematic, transparent and reproducible manner.

This duality is apparent in modern coarse-graining methods that systematically devise simpler models of molecular interactions by building up from atomistic simulations. For example, instead of representing a protein as

to run molecular dynamics with femtosecond (10^{-15} seconds) scale dynamical steps? One example of this approach is the use of Markov State Models which throw out the uninteresting, very fast timescales (femtoseconds to nanosecond) to gain a dramatic efficiency in calculations, especially in terms of parallelization, i.e., using many short trajectories to reproduce very long timescales. Generating 1000 trajectories each at the microsecond timescale and using them to predict the millisecond timescale is considerably more tractable than a single trajectory on the millisecond timescale.

It's interesting to consider the future of these approaches. As both spatial and temporal coarse-graining methods become even more systematic and statistically driven, one can imagine how they can start to merge to build the best physical model possible. This future gets particularly exciting when one considers how the "big data" approaches that are starting to revolutionize other fields might also impact dynamical simulation. With such a combination, reaching the millisecond timescale—now out of reach for all but a few researchers—could become routine, enabling molecular simulation to achieve goals that are barely conceivable today. □

DETAILS

Vijay S. Pande is professor of chemistry, structural biology and computer science at Stanford University.

BY KATHARINE MILLER

Benchmarks for Musculotendon Models

In simulations of human activities such as running, hundreds of individual musculotendon models turn on and off to swing the arms and legs. Naturally, these simulations can only be as accurate and efficient as the underlying musculotendon models.

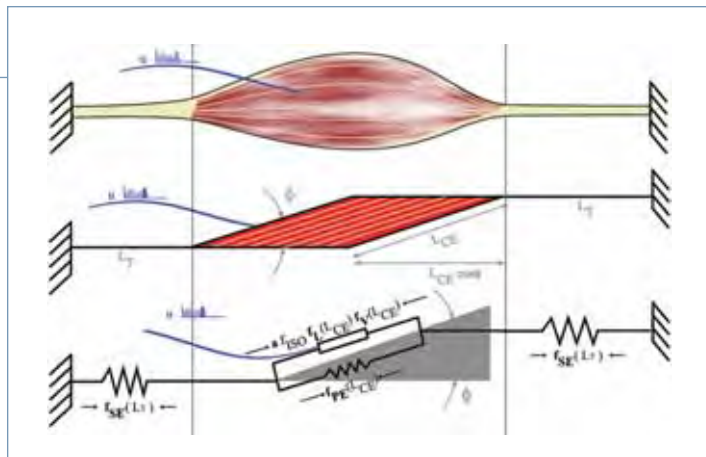
So how accurate and efficient are the most commonly used muscle models? It can be hard to say: Rarely do researchers determine how well their models match biological reality or how efficiently they use computational resources. To address that problem, Simbios postdoc **Matthew Millard, PhD**, collaborated with **Thomas Uchida, PhD**, **Ajay Seth, PhD**, and **Scott Delp, PhD**, at Stanford's Neuromuscular Biomechanics Laboratory to perform the most comprehensive evaluation yet. The work, which was published in the January 2013 issue of the *Journal of Biomechanical Engineering*, produced a new muscle model library that has been incorporated into OpenSim, a freely downloadable software for simulating human and animal movement.

These are “muscle models without compromises,” says Millard, now a postdoc at the University of Duisburg-Essen. That is, they are both biologically accurate and computationally efficient. And with the extensive suite of benchmarks the Simbios team developed and ran, they have the data to prove it. “With luck,” Millard says, “our benchmark tests will be extended and improved to serve as standard tests for the community.”

Muscle Models Without Compromises

Millard and his colleagues went to great lengths to ensure that the characteristic force-length and force-velocity curves that define their muscle models actually fit experimental data for real muscles. “The curves people are used to using have been made for convenience and in some cases are very different from the curves seen in experimental literature,” Millard says. He and his colleagues also made sure biologists would be able to easily and intuitively interact with and edit the curves in OpenSim.

The team also improved on the efficiency of simulating inactive muscles, which are computationally expen-



This schematic shows a simplified illustration of a muscle (top); a simplified geometric representation of muscle fibers and tendon for musculotendon modeling (middle); and a simplified musculotendon model (bottom).

sive to simulate for purely mathematical, nonintuitive reasons (a singularity in the state equations is approached as activation tends to zero). They added a damping effect to the commonly used equilibrium musculotendon model, which resulted in simulations that were up to ten times faster in tests using an explicit integrator (the most commonly used integrator for musculoskeletal simulation). This is an important improvement because many muscles are turned off during normal activities.

Proving Models' Mettle: Benchmarking

While long tendons must be simulated as elastic elements for accuracy, short stiff tendons can be approximated as inelastic to reduce simulation time. But how long can a tendon be before this approximation becomes inappropriate? Millard compared the forces generated by rigid muscle fibers attached to various lengths of rigid or elastic tendons and using different integrators. Turns out, if the length of the tendon is less than the length of the fiber, it doesn't stretch enough to make a big difference in the musculotendon's force profile.

Millard and his colleagues also benchmarked the biological accuracy of the equilibrium, damped, and rigid-tendon models by comparing them to biological muscle that is fully activated and partially activated, relying on data provided by the Sandercock laboratory at Northwestern University. Force profiles generated in simulations of maximally activated muscles using the damped muscle model were a close fit for experimental evidence, whereas simulations of submaximally activated muscles diverged slightly from experimental results, suggesting the need for further work to understand how muscles respond at less than full activation.

Now that the models are available online, along with the extensive benchmark tests and results, a researcher who wants to simulate a muscle with a specific architecture and specific type of integrator can choose an appropriately accurate and computationally efficient model. “It is our hope,” Millard says, “that our efforts will accelerate research to improve muscle models, and ultimately research of human and animal movement.” □

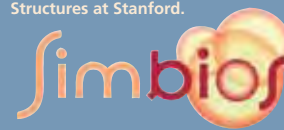
DETAILS (For more information)

Video: <http://www.youtube.com/user/SimbiosMovies>

Paper: “Flexing Computational Muscle: Modeling and Simulation of Musculotendon Dynamics,” *Journal of Biomechanical Engineering* 135(2):021005 (2013)

Benchmarking code and data: https://simtk.org/home/opensim_muscle

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



MODELING SPERM: The Finer Points of Fertilization

By Roberta Kwok

The essential elements of human fertilization are clear: sperm swim through the uterus, travel up the fallopian tube, and fertilize an egg. Not as well understood are the nitty-gritty details of how sperm navigate the curvaceous fallopian tube, boost their chances of reaching the egg, and pierce the egg's outer layer.

New computational models are helping researchers hone in on answers to these questions using such tools as agent-based simulations and classic mechanical engineering principles.

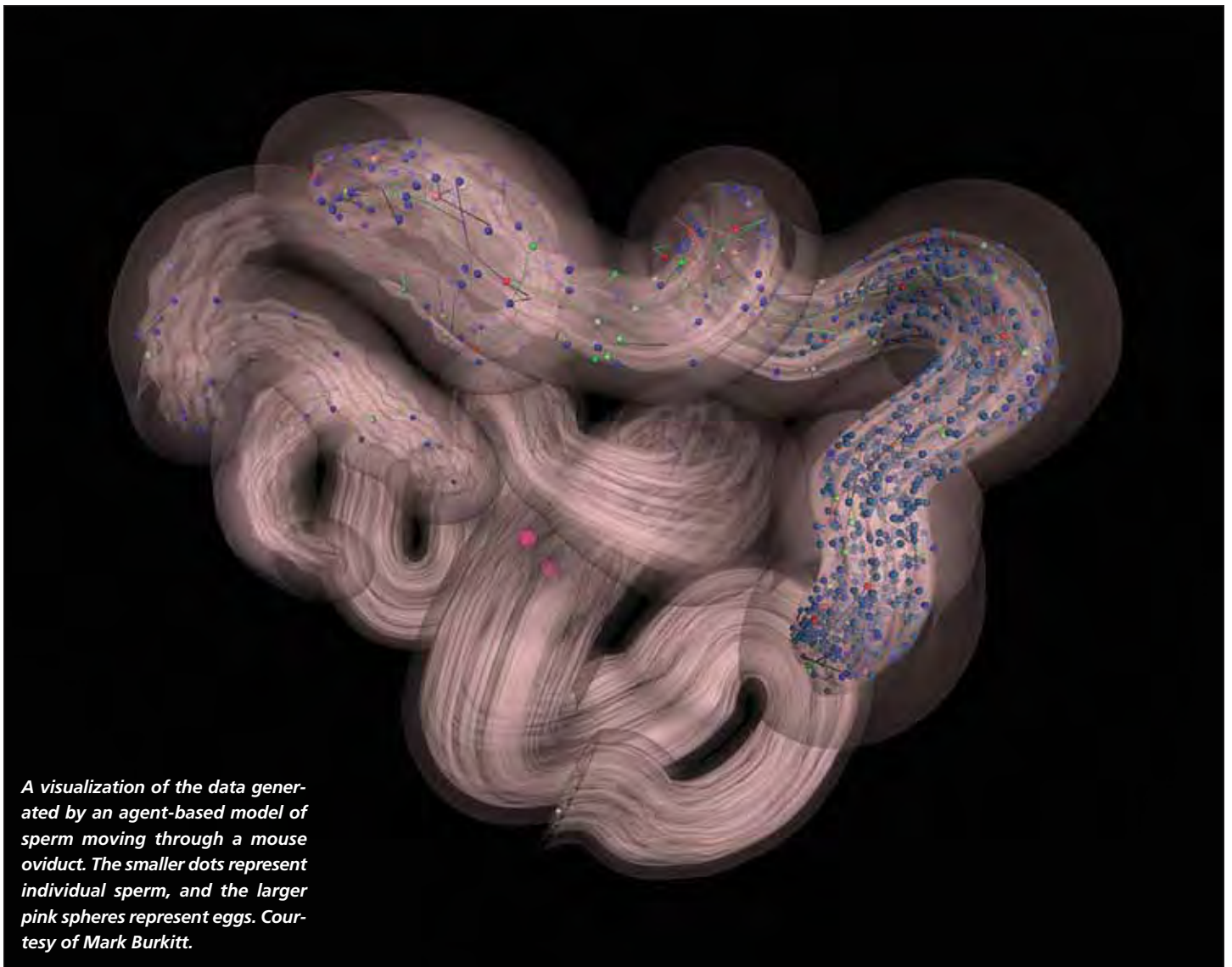
The studies could eventually improve di-

agnosis and treatment of infertility, a problem that's gaining more attention as couples increasingly wait until they're older to conceive. Clinics currently test basic properties such as the number and movement of sperm, but knowing which other characteristics are important for fertilization could help doctors pinpoint the problem, calculate a couple's chances of successful conception, and filter sperm for the best candidates. Scientists also could use this knowledge to design new male birth control treatments—for example, by knocking out functions essential for sperm motility. And sperm might

even inspire better “micro-swimming” devices that deliver drugs in the bloodstream.

Navigating the Oviduct

A team at the University of Sheffield has focused on interactions between sperm and the female oviduct, a tube connected to the uterus where fertilization occurs (known as the fallopian tube in humans). In the past, researchers had modeled individual sperm moving in a fluid. But little had been done to account for the shape of the oviduct which, contrary to popular belief, is not simply a cylinder but includes



A visualization of the data generated by an agent-based model of sperm moving through a mouse oviduct. The smaller dots represent individual sperm, and the larger pink spheres represent eggs. Courtesy of Mark Burkitt.

complicated bends and internal folds. “It’s the first time that anyone’s looked at how the sperm move about, using a conceptual model, within a representation of the oviduct environment,” says **Mark Burkitt, PhD**, who reported the results in his 2011 doctoral thesis and is now director of the consulting and software development company Scientific Online Systems Ltd.

Burkitt’s team chose to employ agent-based modeling, with each sperm represented as an individual entity. The sperm followed a specific set of rules—for example, they stuck to the oviduct wall and switched to a more mature, “capacitated” state with a certain probability; and died shortly after becoming capacitated. The researchers also analyzed histology images of mouse oviducts and developed algorithms to recreate the oviduct’s 3-D structure.

When Burkitt’s team removed bends and folds from the oviduct model, “we ended up with massive amounts of polyspermy,” he says. Polyspermy occurs when more than one sperm fertilizes the egg, resulting in a non-viable embryo. The results suggest that the oviduct’s complicated geometry prevents too many sperm from reaching the egg at once. “The purpose of the complexity of the internal system is to allow a slow progression of these sperm,” Burkitt says.

Asymmetrical Motion

To reach the egg, sperm have to swim in specific patterns. Each sperm’s tail, called a flagellum, beats in a sine-wavelike motion to propel the cell in a straight line. But the sperm also must enter a state called hyperactivation, in which the tail bends more in one direction than another and makes the sperm swim in circles. Hyperactivation might help the sperm free itself after getting stuck to the oviduct wall, and switching between linear and circular paths could improve its chances of finding the egg. “If you’re just going straight, you could potentially swim right by it,” says **Sarah Olson, PhD**, assistant professor of mathematical sciences at Worcester Polytechnic Institute.

In a study published in the *Journal of Theoretical Biology* in 2011, Olson’s team investigated how sperm switch to hyperactivated movement. Scientists know that calcium signals play an important role: To become hyperactivated, sperm need channels in the cell membrane that let calcium in. Olson and her colleagues hypothesized that the calcium influx makes motor proteins called dyneins on one side of the flagellum generate more force than normal, causing the tail to beat asymmetrically.

To test this idea, Olson’s team modeled

a simplified sperm moving through fluid and linked calcium levels in the tail to forces driving its movement. The team accounted for calcium flowing into the flagellum from its environment and calcium released from an internal store in the sperm’s “neck.” The model generated tail waveforms characteristic of hyperactivation, matching the patterns seen in mouse and bull sperm. And the virtual sperm swam in circles as expected.

Final Steps in Fertilization

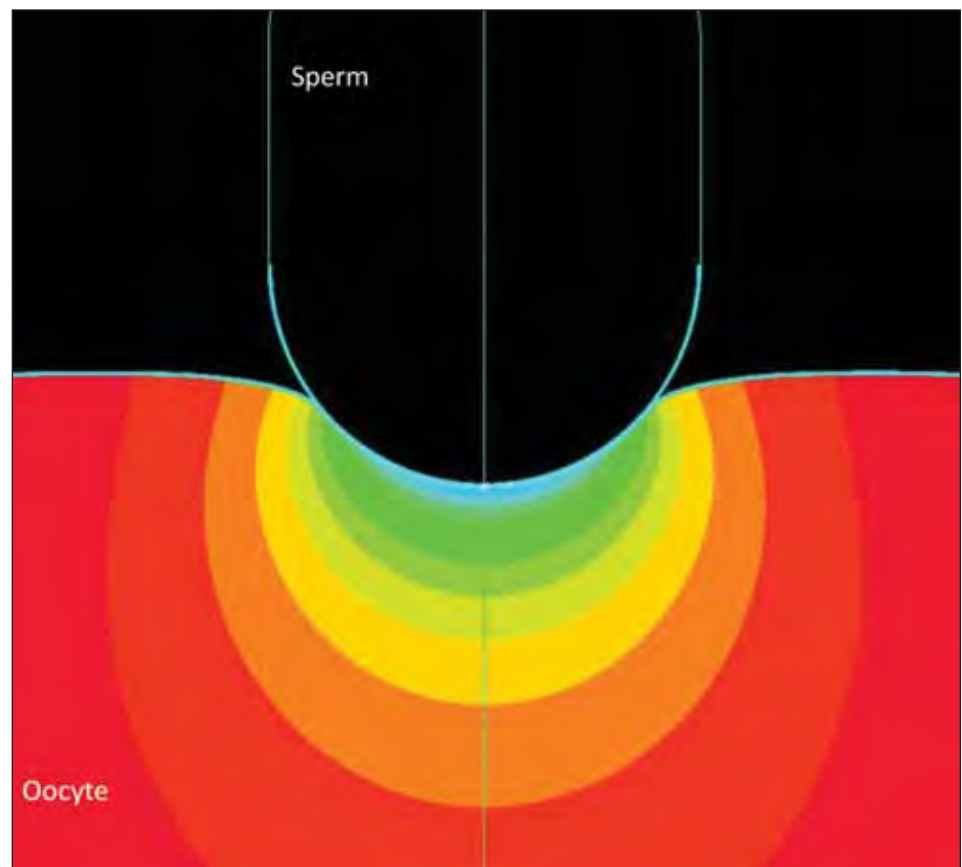
One of the last steps is penetrating the egg, which requires breaching an outer layer called the zona pellucida. Receptors on the sperm bind the egg’s surface, and the sperm releases enzymes to soften the barrier. Defective binding between sperm and the zona pellucida is a major cause of failure to fertilize an egg *in vitro*.

In a study published in the *Journal of Theoretical Biology* in 2012, **Amit Gefen, PhD**, associate professor of biomedical engineering at Tel Aviv University, and a colleague modeled this process using mechanical en-

the chemical forces were 4 to 17 times lower than the mechanical forces. But in a simulated sperm with an unusually sparse population of receptors (one-sixth that of the normal sperm), they were 63 times lower. With insufficient locking, Gefen speculates, a sperm’s powerful forward motion might make it slide across the egg’s surface instead of staying put.

This year, Gefen’s team published a study with a slightly different approach in *Computer Methods in Biomechanics and Biomedical Engineering*. The team used finite element modeling to represent the geometry of the sperm and egg wall more precisely. The researchers tested three shapes of sperm heads, ranging from sharp to blunt, and simulated softening of the egg wall. Not surprisingly, they found that sperm with sharper heads pierced farther into the egg. The model also suggested that the zona pellucida must soften to 10 percent of its original stiffness to allow penetration.

Gefen envisions that fertility clinics could one day test sperm’s ability to soften



A simulation of the deformations generated when a sperm head pushes on the surface of an egg. Courtesy of Amit Gefen, Department of Biomedical Engineering, Tel Aviv University.

gineering principles: They compared the mechanical forces generated by the sperm propelling itself forward to the chemical binding forces generated by receptors locking in place. In simulations of normal sperm,

an artificial, egg-mimicking material in the lab. But practical applications aside, he is fascinated by the science of how fertilization—the beginning of life—happens. “This is why we’re all here,” he says. □

DIGGING INTO PIXELS: Radiogenomics Extracts Meaning

By Amber Dance

In a radiological image, a tumor's edges might appear fuzzy or crisp; its shape could range from oval to many-lobed; and its density and texture might vary across the tumor. To determine whether and how those characteristics matter, researchers in the new field of radiogenomics are extracting as much information as possible from every pixel and correlating it with gene expression and cancer survival rates.

The work has the potential to increase our understanding of tumor biology and offer patients personalized medicine based on both imaging and genetics.

The field now seems poised to take off, says **Robert Gillies, PhD**, chair of cancer imaging and metabolism at the Moffitt Cancer Center in Tampa, Florida. "The whole idea of extracting large amounts of

tumor's "texture"—whether it's mostly uniform or mottled in density—is for computer software to compare the intensity of each pixel with those of its neighbors. The more alike the intensities are, the less textured and more homogeneous the tumor. By changing the parameters of the analysis, a computer can describe texture in numerous different ways, Gillies says.

He's using imaging features, such as density and contrast, to study the tumor microenvironment. Tumors can have a strong or weak blood supply from the vasculature, which shows up as a high or low perfusion of injected contrast agents on an MRI. And they can be more or less dense, which appears as high or low contrast with surrounding tissues. Gillies wants to classify tumors—in brain, breast, lung and

out how the cancer's molecular biology influences the pictures they see.

In a 2012 paper in the journal *Radiology*, **Sylvia Plevritis, PhD**, associate professor of radiology at Stanford, together with Napel and other colleagues, showed it could be done. The team examined CT scans and gene expression profiles from 26 people with lung cancer. They compared the gene expression data with 180 different image descriptors, both semantic and computational. Because the participants were newly diagnosed, with no data available on survival or relapse, the researchers looked for correlations between gene expression and prognosis in a public database. Combining those two analyses, they found that tumor size, shape and edge sharpness were most strongly linked to gene expression

"The whole idea of extracting large amounts of quantitative data from images has been around for a long time, but it's taken a while for the computational power to catch up with us," Gillies says.

quantitative data from images has been around for a long time, but it's taken a while for the computational power to catch up with us," he says.

Shades of Gray

In radiogenomics (also known as radiomics), researchers convert images to mineable data in high throughput, Gillies says. So the radiologist's "fuzzy edge" becomes a numerical descriptor, reproducible between physicians using the same rating scale or calculated by a computer based on the arrangement of pixels in shades of gray.

Radiomics data come in a few flavors, such as semantic and computational, says **Sandy Napel, PhD**, professor of radiology at Stanford University. Semantic features are word descriptors, such as round, oval or star-shaped, assigned by a human. Computational features are numerical values calculated from pixel patterns.

For example, one way to describe a

connective tissue—based on which microenvironments are present, and correlate that score with patient data such as treatment and survival.

Eventually, Gillies hopes radiologists will build databases of tens of thousands of patient images. A radiologist could log on with a picture from today's patient; compare it to images from patients past; and predict, based on those past patients' clinical histories, the best treatment options.

Images Plus

Beyond simply correlating images and clinical data, radiogenomicists can mix in molecular information, such as a tumor's genetic mutations or gene expression profile. Normally if physicians want to know about the biochemistry and gene expression in a tumor they need a physical piece of it. But that same information may be buried in the medical images obtained noninvasively—if researchers can figure

that correlated with prognosis.

For example, tumors that include air-filled structures, called internal air bronchograms, often upregulated a gene called KRAS. And KRAS overexpression, according to public databases, is an indicator that a tumor will likely recur. Other studies have offered conflicting results as to whether internal air bronchogram is a positive or negative sign, so more research is necessary. However, this and other correlations in the paper suggest the potential value of imaging in making prognoses.

Plevritis cautions that this was a small proof-of-principle study. "It just says that we should do more, and we are," she says. The team has recruited around 75 new lung cancer participants so far, and is also looking into similar studies with breast and liver cancer.

With large enough datasets, computers might pull out tumor features that humans would never notice, Napel says. "The tech-

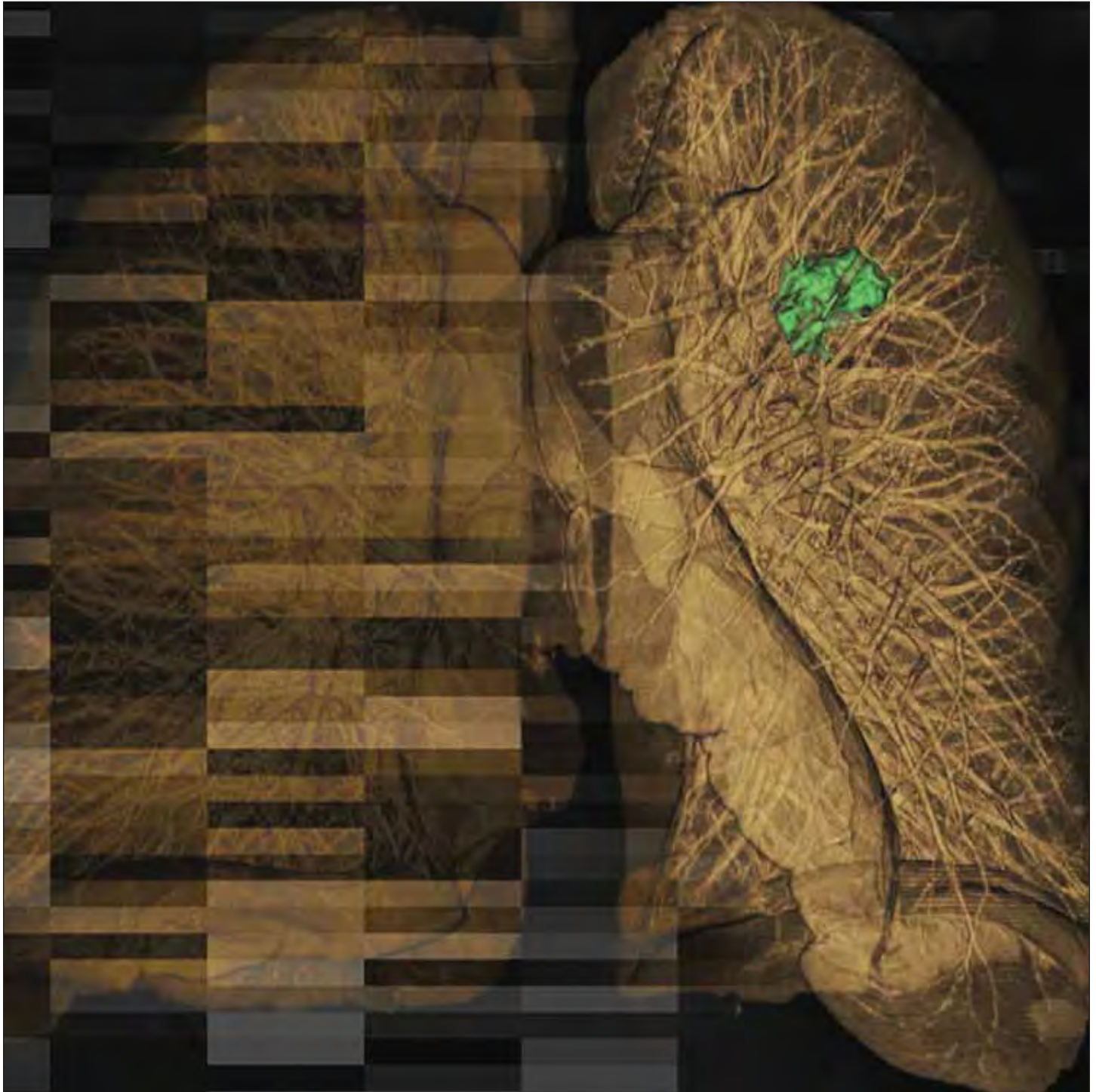
nology is here; it could be implemented today,” he adds. But there are logistical challenges. For one, data acquisition is not standardized across different institutions: Some use different slice thicknesses for 3-D images, or apply different filters. Similarly, no standardized methods exist for how people or computers convert those images into comparable, numerical data.

Other obstacles include time and money. Radiologists’ schedules are already full sim-

ply reading scans; they have no time to develop a new way of doing their business, Gillies notes. What’s needed, he says, is a “sandbox” where radiologists can experiment with information technology to build annotated image databases. He estimates the price tag for a single center like that at \$17 million—and there would need to be many such centers. Storing and moving the petabytes of data would cost a fortune, Napel adds.

But, Gillies points out, misdiagnosis due to incorrect scan interpretation is also costly. If computers could assist in diagnoses and prognoses, and catch human errors, he predicts the new methods would be “more than cost-effective.” And similar techniques could be useful in radiology and pathology image analysis beyond cancer, Napel says.

“I certainly don’t propose replacing radiologists with computers, but we need to incorporate them as allies,” Gillies says. □



A picture worth 1,000 genes. At Stanford, researchers are correlating radiological images—such as this CT scan of a lung tumor, rendered here in 3-D—with gene expression in the tumor and with patient survival. The work could lead to better understanding of tumor biology and personalized treatments based on imaging features. Courtesy of Amy Thomas and Shannon Walters, Stanford University.

MATTERS OF TIME: Tick Tock Go the Simulations

By Katharine Miller

Time flows like a continuous, steady river. And it moves forward—never back. These facts create inherent challenges for computer simulations of biological molecules in motion.

It would be lovely if time could be efficiently simulated as a flowing variable. But time has to march in discrete steps for computers to handle the complex movements of molecules. And that matters: The length of the step (be it a femtosecond, a millisecond, a minute, or a year) affects the stability and accuracy of a simulation; limits the amount of total time that a simulation can reasonably cover; and generates error terms that must be accounted for. In addition, researchers add inaccuracies of their own by coarse-graining models, simplifying the simulations in space as well as time to improve efficiency and cover the longer time spans of biological interest. And then there's the fact that time has a directional arrow—one that's hard to untangle from the energy landscape at microscopic scales.

Despite the challenges of simulating time, researchers remain committed to molecular dynamics (MD) simulations—including coarse-graining—because they provide insight, says **William Noid, PhD**, assistant professor of chemistry at Penn State University. Indeed, he predicts coarse-grained models will always be useful because, as he puts it, “the human imagination and computational demands will always progress at a rate far exceeding Moore's law.” But it's important to keep in mind that these simulations are models, not

Discretizing Time

Researchers simulate biological molecules to gain an understanding of how they function in living systems. These molecules move

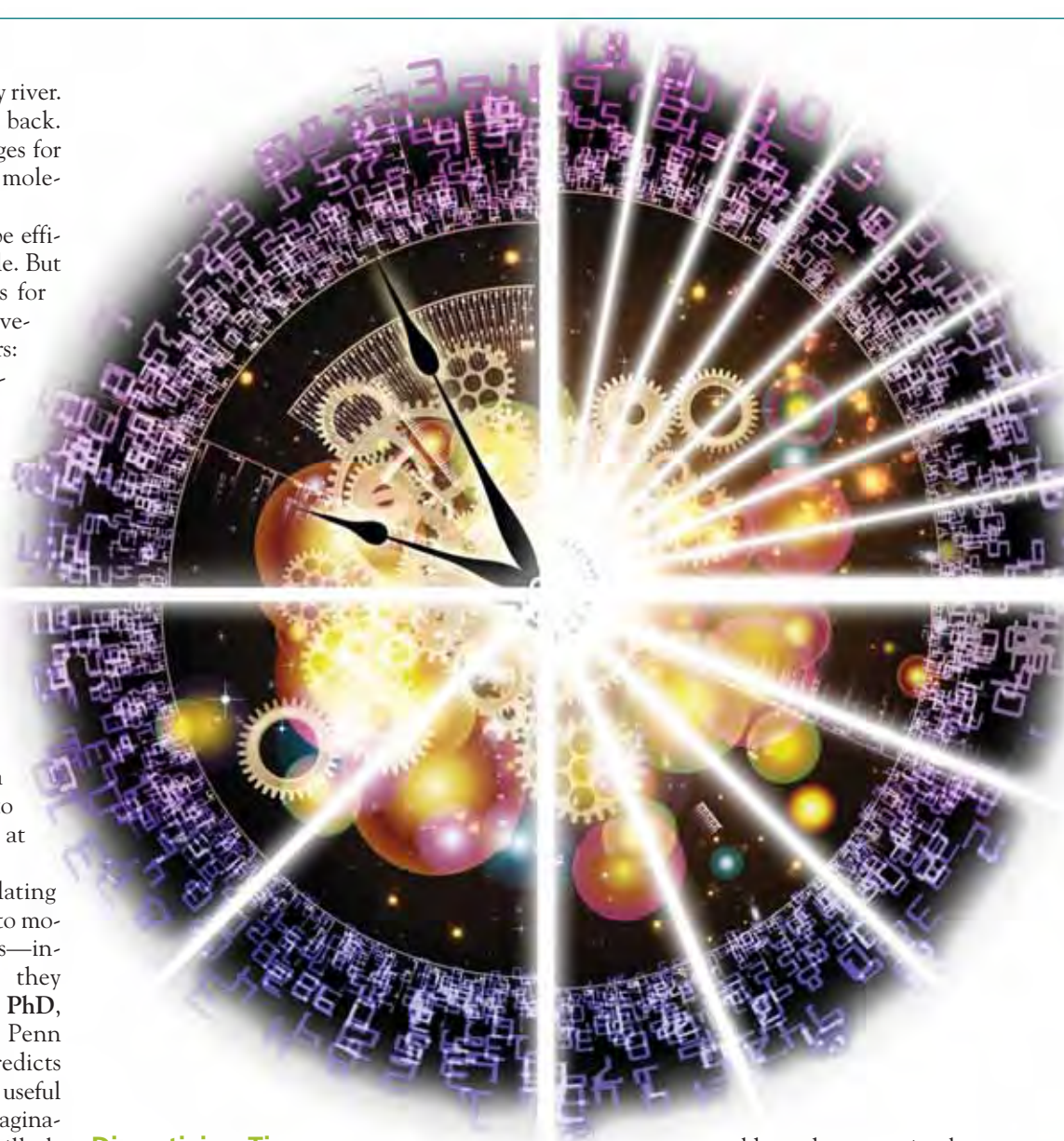
could say those are simulations with continuous time,” says **David Sivak, PhD**, a systems biology fellow at the University of California, San Francisco. But for more complicated molecular systems, the equations can't be solved exactly, even by a computer, he says. Breaking time into discrete steps becomes a way to make the calculations computationally tractable.

For example, a system of atoms or molecules can be described by a series of differential equations that evaluate how the particles' positions and velocities change over time in accordance with Newton's second law of motion ($F=ma$, or force equals

“The human imagination and computational demands will always progress at a rate far exceeding Moore's law,” William Noid says.

reality, Noid says. And the ravages of time are likely to always play a role in keeping it that way.

according to the laws of physics. For very simple systems, Newton's equations of motion can be solved exactly. “In that sense you



mass times acceleration). An MD simulation might then numerically integrate these equations of motion over a series of time steps. The computer calculates the forces on each particle based upon their current positions and then assumes that the forces on the particles are constant for a short increment of time, Noid says. During this short increment, the particle positions and velocities change due to the forces on them. The computer then updates the forces on each particle based on their new positions and velocities, and the process is repeated. Although researchers use many more sophisticated ways to integrate the equations of motion to achieve greater accuracy and efficiency, “most are only slight variations on this simple mechanism,” Noid says.

The basic assumption here—that forces are constant during the time step—inaccurately represents reality. “It’s like a movie; It’s really a series of discrete snapshots played fast enough that it looks continuous,” says **Greg Bowman, PhD**, a research fellow at the University of California, Berkeley.

Longer time steps reduce a simulation’s stability as well as its accuracy. Indeed, if the steps are too long, the molecules being simulated start to take on unfavorable conformations. “You get a cascading problem.

comes time for publication, Bowman notes, reviewers don’t necessarily notice the details unless the paper’s results don’t make sense. “Only then will they look back and question the parameters.”

A better practice, Bowman says, is to take the time to find the right time scale for the problem. One approach is to find the largest time step where things don’t blow up. Another way to think about it, Noid says, is to find the largest time step over which one can reasonably approximate the forces as being constant—where the particles haven’t moved enough to alter the forces appreciably. In the case of MD simulations, the appropriate time step is determined by the interaction that changes most rapidly, Noid says. For simulations of atoms, this ends up being on the order of one femtosecond—the rate of jiggling and wiggling of bonds or water molecules.

When a continuous process is simulated using discrete time, there are always errors—discrepancies between the calculated results and the true underlying behavior. Errors pose another consideration for choosing the duration of the timestep, Sivak says. So, for example, researchers might observe the error at the largest time step where things don’t blow up and then do the same

additional work on the system,” Sivak says. This realization allows researchers to quantify how far out of equilibrium a simulation is simply due to the discretization of time—even when the system otherwise would be in equilibrium. It becomes possible to characterize this “shadow work” and correct for it, separating the physically realistic aspects of the simulation from the artifacts of the computer method, Sivak says.

Temporal Coarse Graining and the Time/Space Connection

To overcome the limits of computer power, researchers often create simplified models that allow for more efficient MD simulations over longer time scales. The simplifications can be spatial—e.g., treating a group of molecules as a single ball; or temporal—e.g., using longer time steps. In reality, says **Thomas Miller, PhD**, professor of chemistry at the California Institute of Technology, the two go hand in hand. “You can’t coarsen spatially without coarsening in time,” he says. If atoms are clumped into a ball, the corresponding time scale for the movement of the ball is slower than it was for the atoms. “That’s two halves of the benefit of the process,” he says. “As you eliminate unnecessary spatial motions,

“You can’t coarsen spatially without coarsening in time,” Thomas Miller says.

It’s not a subtle thing: Your simulations just blow up,” says Sivak.

But many biological events of interest take too long to simulate using small time steps, given the limits of computational power. For example, proteins take milliseconds to fold—a process that would take more than a trillion femtosecond timesteps to simulate—beyond the capacity of typical computational resources. On the other hand, it takes only a thousand microsecond timesteps to simulate a millisecond. Researchers have to balance their desire to integrate the equations of motion as accurately as possible, against their need to make the problem computationally manageable.

Picking a Timestep

In practice, many researchers don’t contemplate the size of the time step. They use the default settings or recommended time discretizations in readymade software packages, Sivak says. Or they copy the parameters used by others without necessarily evaluating where they came from or why they were chosen, Bowman adds. When it

at a smaller timestep to see how the error changes as the timestep shortens. If they know the level of error they are comfortable with, they can then pick a particular time step, he says.

Time At Work: An Intuitive Understanding of Timestep Error

Having selected a timestep and performed a simulation, researchers also have to correct for the errors the timestep creates. Recently, Sivak and his colleagues took a hard look at these errors and came up with an intuitive, physical way of thinking about them. The work was published in *Physical Review X* in January 2013.

Errors caused by time discretization turn out to be particularly important in so-called nonequilibrium simulations where the conditions are changing fast, such as where a protein is being stretched. Sivak and his colleagues found that just as you can mechanically put energy into a protein—by stretching it, for example—the discretization of time also puts energy into the protein. “The error arises because the simulation does

what’s left over moves more slowly so you can take bigger timesteps.”

In October 2012, Miller published in



the journal *Cell Reports* a coarse-grained simulation of the Sec translocon, a channel that allows proteins to pass through cell membranes. The feat required his team to coarse-grain out lots of faster molecular movements—from femtoseconds to hundreds of nanoseconds—in order to focus on the slower movements—from hundreds of nanoseconds to the full minutes it takes for a protein to pass through the channel. But before doing that, they had to determine the average effect of the faster motions. “We had millions of hours of underlying computer simulation time based on high-resolution models,” he says.

The Sec translocon paper demonstrates the degree to which complex biological machinery can be simplified while still capturing a wide array of experimentally observed phenomena in the system, Miller says.

Markov State Models: A Knob for Controlling Time and Space Resolution

Markov State Models (MSMs) offer another way to achieve longer time scales for MD simulations such as protein folding. An MSM can merge variations from thousands of successive protein-folding simulations and identify a set of relatively stable conformations along the protein’s many folding pathways. By choosing a timestep for the model as well as how many states to identify, whether 15 or 100,000, researchers can dial in the degree of complexity they seek.

The idea is that you’re removing the intermediate steps between these stable conformations, sort of like reducing the frame rate in a movie, Bowman explains. “We can use this time and space resolution basically as a knob to control how detailed our models are,” he says. The approach allows the simulation of

larger proteins for longer periods of time, permitting insight into how they function.

Tomorrow Differs from Today: Time’s Irreversibility and Biological Molecules

At the macroscopic scale, we have no doubt that time moves inexorably forward. A glass can fall off a table and smash to smithereens, but cannot jump back onto the

table in one piece. And we know instinctively when a movie of human-scale events is run in reverse.

But at the molecular scale, discerning forward from backward is much harder. That’s partially because everything is stochastic—tiny molecular machines fire randomly; they are not like steady car engines. Yet time’s forward arrow does exist at the molecular level thanks to the second law of thermodynamics which states that isolated systems spontaneously evolve toward maximum entropy. (All other laws of thermodynamics are equations that don’t care about time.)

It’s just that spotting entropy’s signature is tough at the molecular scale because the energy required to break time asymmetry—to move toward maximum entropy—is close to the entire local energy budget, says **Gavin Crooks, PhD**, senior scientist at Lawrence Berkeley National Lab. For example, an important molecule like ATP synthase—a tiny little molecular engine—functions at an energy level that is not much greater than the scale of energy fluctua-

tions in the environment.

Over the last ten years, Crooks and others have made progress toward spotting entropy’s signature against the fluctuating energy background in single-molecule experiments. It turns out that accounting for time asymmetry matters greatly in MD simulations of systems that are out of equilibrium—just the kinds of systems that interest Crooks. He has a grand vision of

[A]ccounting for time asymmetry matters greatly in molecular dynamics simulations of systems that are out of equilibrium—just the kinds of systems that interest Gavin Crooks.

thermodynamically realistic simulations of walking molecules, such as myosin stepping along an actin strand—a very non-equilibrium process. Such systems have their own intrinsic time asymmetry that needs to be untangled from the rest of the thermodynamics. “In the long run, I would like to do simulations of relevant biological systems that are active, that aren’t just at equilibrium. And I want to get the thermodynamics right,” he says.

Bridging Time Scales

The intrinsically molecular processes that govern our physiology include chemical reactions faster than a picosecond; bond rearrangements that take picoseconds to nanoseconds; changes in protein conformations that happen in microseconds; protein folding that occurs in milliseconds; and barrier-crossing events that take seconds to minutes, Miller says.

In biological systems, the separation of these time scales is not always clear. One process with higher time resolution may feed into a process with lower time resolution. “That complexity is potentially interesting but very challenging for the person doing the modeling,” says **Gerhard Hummer, PhD**, chief of the theoretical biophysics division at the National Institute of Diabetes and Digestive and Kidney Diseases at the National Institutes of Health. “To a large degree it’s an active area of research where there are no generally accepted and generally applicable solutions.”

Miller agrees. “Spanning these big ranges of time in biological systems is the big challenge of the field,” he says. “A whole lot of people with a whole lot of good ideas are trying to address that challenge.” □



BEHIND THE Connectome

Connectomics
is having
a moment.

Following on the heels of genomics, proteomics, transcriptomics, metabolomics, and microbiomics, the latest “omic” to seize the spotlight is generating the kind of buzz that makes other disciplines fluorescent green with envy. As the name suggests, connectomics maps connections—specifically, the ones between the neurons in an animal’s brain or nervous system.

The advent of high-throughput, computer-assisted techniques has led to an explosion of connectomic technologies and studies. The field is also amassing the sort of Big Science resources previously associated with efforts to land a man on the moon or decode the human genome. The Obama administration, for exam-

ple, recently decided to pump \$100 million into the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) project to develop methods for recording neuronal activity on a large scale; while the European Commission is investing €1 billion in Henry Markram’s Human Brain Project in Switzerland—a plan to build a computer simulation of the human brain, neuron by neuron.

At the same time, the connectomics frenzy has come under fire. Some scientists question the wisdom of devoting scarce resources to the pursuit of the human connectome when the task remains well beyond the scope of currently available technologies. As **Sebastian Seung, PhD**, author of *Connectome: How the Brain’s Wiring Makes Us Who We Are*, notes in a widely watched TED talk (more than half a million views and counting), the only complete connectome we have is for the tiny nematode worm *C. elegans*, an animal with roughly 300 neurons and 7000 neuronal connections; and the job took more than a decade using conventional electron microscopy, which was used to resolve individual synaptic connections between cells. By contrast, the human brain contains roughly 100 billion neurons and 100 trillion connections. This is why Seung describes the effort to map the human connectome as “one of the greatest technological challenges of all

time,” and asserts that it will “take generations to succeed.” Others question the wisdom of concentrating on the development of new technologies in the absence of clearly articulated scientific goals, and on focusing so intensely on wiring diagrams that cannot by themselves explain how our brains give rise to feelings, thoughts, and perceptions.

But researchers from a wide range of backgrounds—some trained in physics or engineering, others in neuroscience or medicine—are mounting a serious effort to demonstrate the viability of connectomics. Thus far, they have focused on two broad and equally important tasks: devising faster and better methods for building connectomes; and putting connectomic data to good use. Their collective spadework constitutes the true current state of connectomics—and their success, its true promise.

Building a Connectome by Taking the Middle Road

In 1993, Francis Crick coauthored a *Nature* article (“Backwardness of Human Neuroanatomy”) that lamented the lack of progress toward a “connectional map” of the human brain. Since then, not much has changed, argues **Partha Mitra, PhD**, Crick-Clay Professor of Biomathematics at Cold Spring Harbor Laboratory. In part, he says, that’s because neuroanatomists have tended to beaver away in their individual labs using different paradigms and techniques. As a result, their data often doesn’t integrate well, and the models of neuroanatomy and connectivity they develop stay locked inside their own brains.

In a 2009 *PLoS Computational Biology* paper, Mitra and a number of his colleagues proposed solving that problem by launching a coordinated effort to construct a mesoscale whole-brain wiring diagram for a vertebrate. That proposal led to the creation of the Mouse Brain Architecture (MBA) Project, an attempt to demonstrate the attainability of a large connectomic endeavor while also providing a testbed for developing practical neuroanatomical techniques.

As a theoretical physicist with an eye for the big

Commmotion

By Alexander Gelfand



picture, Mitra sought to develop a systematic method for constructing a complete connectome using currently available technology. Taking a page from the Human Genome Project, he developed a high-throughput, semi-automated pipeline for imaging multiple mouse brains using light microscopy, a technique whose resolution lies between that of electron microscopy, which is impractical for mammalian brains, and the non-invasive yet far coarser magnetic resonance imaging (MRI) techniques used on human subjects.

Last June, Mitra released the first round of gigapixel image data collected for the MBA Project. The images

can be viewed online and explored with a virtual microscope: Users can zoom in on individual neurons and their axons, the long, slender fibers that trail away

of a sphere-packing algorithm. The tracers are either absorbed into neuronal cell bodies and spread through their axons, or are taken up by axons at synapses and propagate up into the cell bodies. The brains are subsequently sectioned into 20 μ -thick slices and imaged using either brightfield (i.e., white-light) or fluorescence microscopy.

Then comes the tricky part. The resulting two-dimensional images must be assembled into three-dimensional stacks and registered to an anatomical reference atlas using a combination of off-the-shelf and custom software. Next, axons and cell bodies must all be identified, a step that presents a significant bottleneck. Mitra's lab is working on automating the process using machine-learning algorithms that can be trained to identify features of interest. For now, however, "somebody has to look through half a million sections," he says.

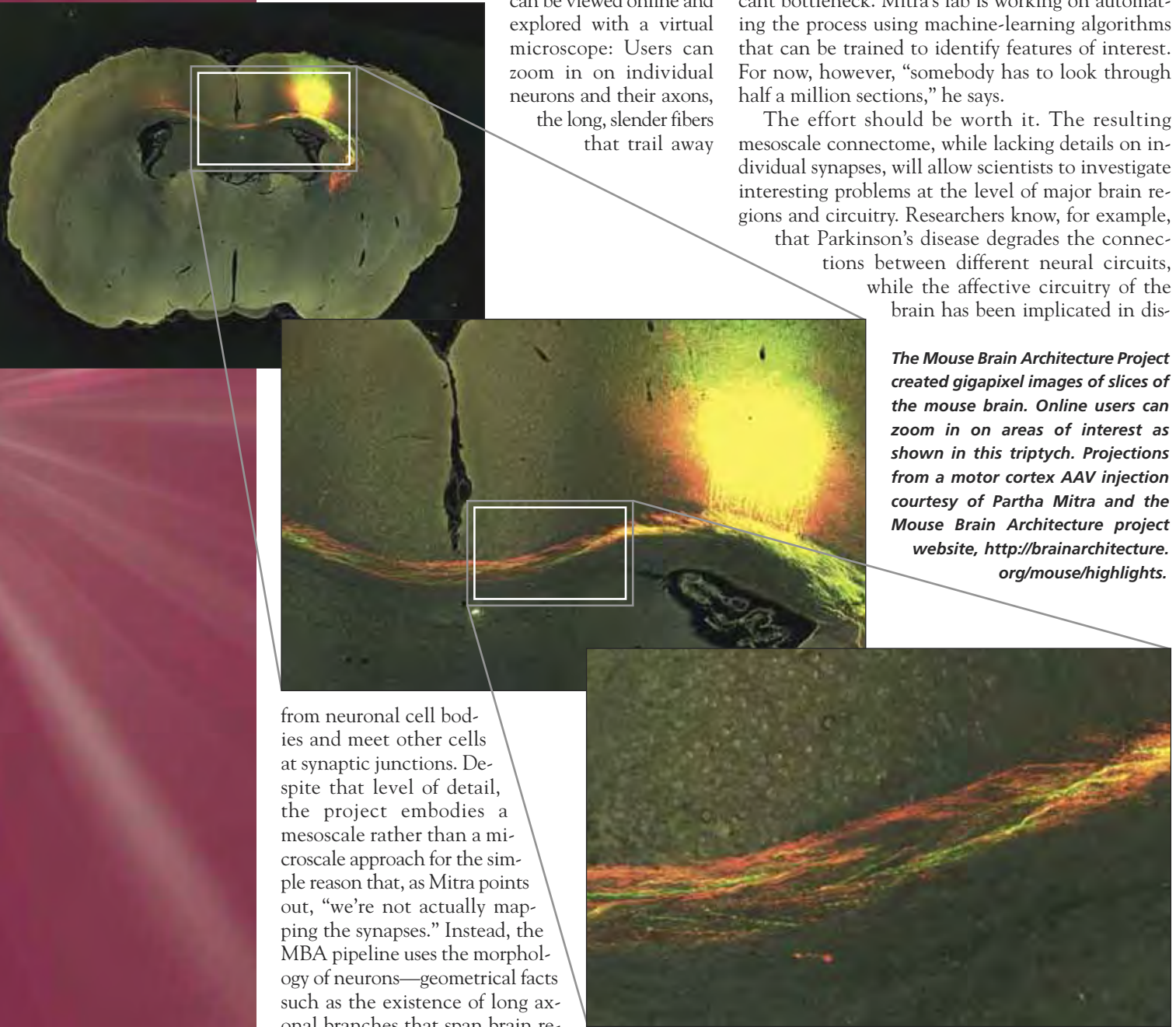
The effort should be worth it. The resulting mesoscale connectome, while lacking details on individual synapses, will allow scientists to investigate interesting problems at the level of major brain regions and circuitry. Researchers know, for example, that Parkinson's disease degrades the connections between different neural circuits, while the affective circuitry of the brain has been implicated in dis-

The Mouse Brain Architecture Project created gigapixel images of slices of the mouse brain. Online users can zoom in on areas of interest as shown in this triptych. Projections from a motor cortex AAV injection courtesy of Partha Mitra and the Mouse Brain Architecture project website, <http://brainarchitecture.org/mouse/highlights>.

from neuronal cell bodies and meet other cells at synaptic junctions. Despite that level of detail, the project embodies a mesoscale rather than a microscale approach for the simple reason that, as Mitra points out, "we're not actually mapping the synapses." Instead, the MBA pipeline uses the morphology of neurons—geometrical facts such as the existence of long axonal branches that span brain regions—to infer patterns of connectivity.

To gather those geometrical facts, Mitra's team injects mouse brains with four different tracers, some of which express fluorescent proteins, in 262 uniformly spaced sites that were chosen with the help

orders such as anxiety and depression; and many believe that conditions like autism and schizophrenia are caused by pathological patterns of neural connectivity, or "connectopathies." Mesoscale connectomes could lead to better diagnostic tools for major disor-



ders, better drug therapies, and even to a better understanding of how genetic variation influences behavior by shaping the wiring of the brain.

Building a Connectome at the Microscale

Meanwhile, Mitra's Cold Spring colleague **Anthony Zador, MD, PhD**, co-founder of the Computational and Systems Neuroscience (Cosyne) conference, is working on a microscale approach to building connectomes that would dispense with microscopes altogether.

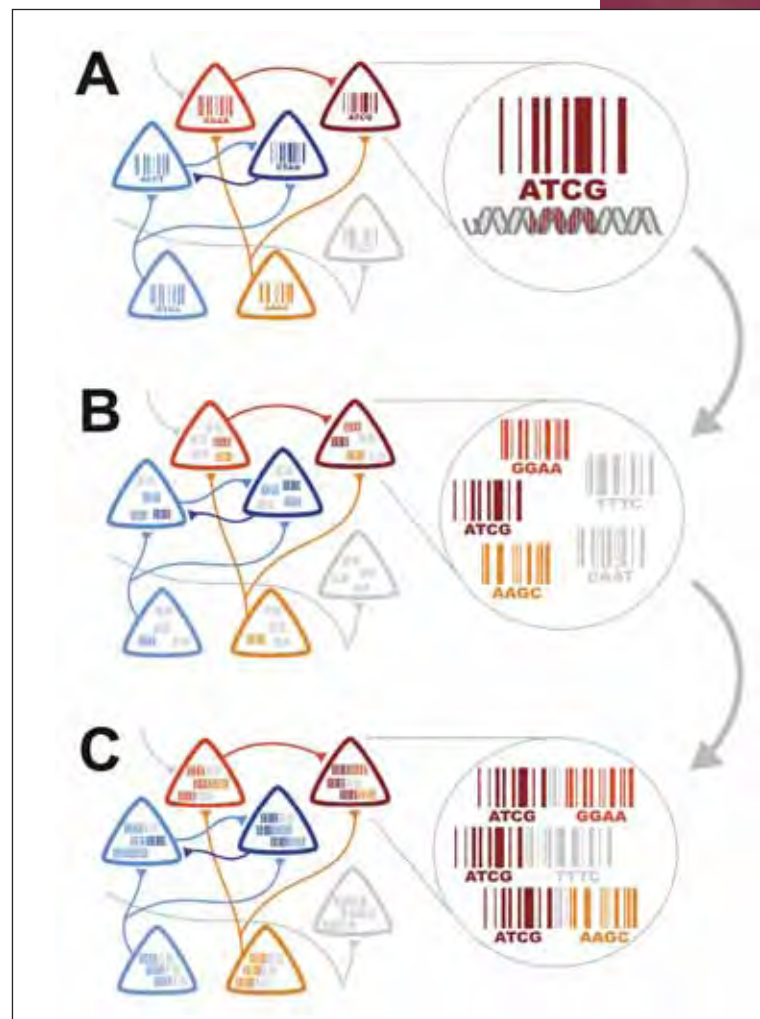
As someone who studies attention and auditory processing, Zador wants a way to model and interrogate neural circuits *in silico* in order to streamline the lengthy and laborious process of running experiments to determine how groups of neurons feed information to one another. Doing that, however, requires a quick and inexpensive method for tracing the synaptic connections between individual cells. If Mitra's mesoscale approach to identifying axonal pathways is akin to sketching the major highways in the United States, says Zador, he's interested in identifying "each and every street, road, and country lane."

Zador was initially inspired by the Brainbow technology invented by Harvard researchers **Jeff Lichtman, MD, PhD**, and **Joshua Sanes, PhD**. Brainbow uses genetically engineered neurons to express random combinations of up to four fluorescent proteins, producing brightly colored collections of cells that can be imaged using fluorescence microscopy. But the technique is limited to a palette of just a couple of hundred colors—not nearly enough to uniquely label all of the neurons in even a small sample of brain tissue—and offers limited resolution. So Zador began looking for optical alternatives to fluorescent proteins. Eventually, he realized that he could do away with them entirely and use DNA barcodes instead. "The readout for the fluorophores is a microscope. The readout for the barcodes is an Illumina," says Zador, referring to the ultrafast next-generation gene sequencing machines. Zador described his approach, called BOINC ("barcoding of individual neuronal connections"), in a *PLoS Biology* paper published last October.

With the help of a recombination enzyme that can scramble bits of genetic code, Zador instructs neurons to generate short random sequences of DNA. In theory, a sequence containing just 20 random nucleotides could uniquely label 2×10^{12} neurons, more than enough for the 100 million or so neurons in a mouse brain. Once the neurons are labeled with these randomly generated barcodes, Zador traces their connections by having a transsynaptic virus spread the barcodes from cell to cell. That has the effect of turning each neuron into a "bag of barcodes" that contains not only its own unique DNA label, but also the unique identifiers belonging to

each neuron that's connected to it. Some more genetic engineering technology is used *in vivo* to join each neuron's barcode with the barcodes from neurons to which it is connected by a synapse, creating sequences of fused barcodes that represent networks of neurons; harvesting and reading the fused barcodes with a high-throughput sequencer yields a connectivity matrix. Computation plays a role in several places: Zador and his colleagues had to develop novel algorithms to clean up the barcodes, correct for any sequencing errors, and determine the connectivity matrices. They have been running proof-of-principle experiments using neurons cultured in an incubator, and have successfully completed each step in the process in isolation. Despite some remaining technical hurdles, Zador expects to combine all of the steps within a matter of months.

Sectioning the brain before extracting the DNA for sequencing will allow Zador to identify the brain



Zador and his colleagues convert neuron connectivity into a sequencing problem that can be broken down conceptually into three components—labelling neurons with unique DNA barcodes; associating barcodes from synaptically connected neurons; and joining host and neighboring (invader) barcodes into pairs for sequencing. Reprinted from Zador AM, Dubnau J, Oyibo HK, Zhan H, Cao G, et al. (2012) Sequencing the Connectome. *PLoS Biol* 10(10): e1001411. doi:10.1371/journal.pbio.1001411, (2012).

region that each neuron comes from. And the use of genetic technology ought to allow Zador to determine the particular kinds of neurons (e.g., inhibitory, excitatory) involved, too. Inhibitory neurons, for example, express a particular enzyme that is encoded in mRNA; by tagging the appropriate mRNA in a given batch of neurons with barcodes, Zador should be able to identify the inhibitory ones. That would add a level of detail about the identity of individual neurons that would be hard to come by even using electron microscopy; and if his sequencing approach works, it would be cheaper and faster than anything currently out there. In his *PLoS* paper, Zador estimates that sequencing the connectome of a mouse cortex would cost \$40,000 at current rates, “and could easily drop several orders of magnitude in a few years.” Sequencing a fruit fly brain would cost one dollar, and doing *C. elegans* would be “essentially negligible.” That would put neuroscientists in a position to quickly and inexpensively map brain circuits, allowing them to develop testable hypotheses and design experiments far more efficiently.

Using Connectomes To Understand Behavior

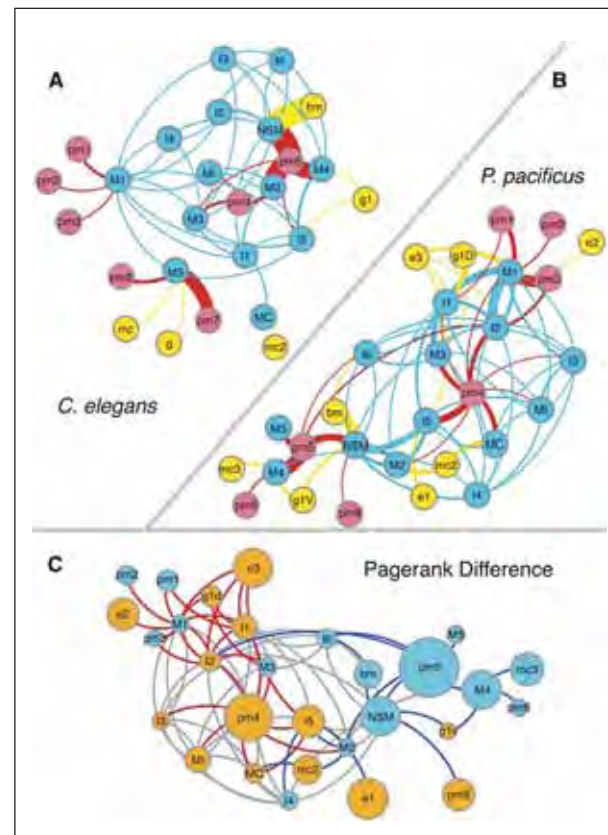
Some researchers are already using connectomics to understand behavior. In a December 2012 paper published in *Cell*, for example, **Daniel Bumbarger, PhD**, used differences in neural connectivity to help explain the divergent feeding behaviors of *C. elegans* and its nematode cousin, *P. pacificus*. Unlike *C. elegans*, which feeds exclusively on microbes, *P. pacificus* is capable of switching into predator mode and eating other nematodes. Typically, neuroscientists have explained such behavioral differences by looking at the physiology of the neurons involved, or the neurotransmitters that modulate them. But Bumbarger, who is a postdoctoral fellow at the Max Planck Institute for Developmental Biology in Tuebingen, Ger-

many, wanted to see if those differences in feeding styles could be related to differences in patterns of synaptic connectivity. So he compared the wiring diagrams for the pharyngeal nervous systems of the two worms—a task that first required imaging the 300 μ m-long pharynxes of several *P. pacificus* specimens using an electron microscope, something that in itself took nearly two years of work.

What he found was striking. Despite having basically the same number and types of neurons in their pharyngeal nervous systems—a remarkable conservation of cell identity—the two nematode species displayed a “massive rewiring of synaptic connectivity,” Bumbarger says, with *P. pacificus* demonstrating much higher and more complex connectivity than *C. elegans*.

To better understand how differences in connectivity might be driving differences in behavior, Bumbarger turned to graph theory, the branch of mathematics that gave rise to network analysis. Graphs are defined as sets of nodes connected by edges, or lines; consequently, graph theory can be used to analyze the characteristics of virtually any network, including neural ones in which the nodes are brain regions or neurons, and the edges are axons or synapses. To compare the relative importance of the various neurons shared by *C. elegans* and *P. pacificus*, Bumbarger computed a variety of measures that evaluate the centrality of nodes within their networks—measures like degree centrality, for example, which counts the connections associated with a node, and PageRank centrality, which gauges the probability of stopping at it. (PageRank helps Google rate the importance of webpages.) He also developed a new tool, called focused network cen-

Bumbarger and his colleagues compared the synaptic connectivity of the two nematode species C. elegans (A—based on previous work by others) and P. pacificus (B), both shown here in a two-dimensional representation. Nodes indicate neurons (blue), muscle cells (red), and other network outputs (yellow). Edges curve clockwise from the presynaptic to the postsynaptic node and are colored the same as their postsynaptic partners, with edge width indicating connection weight, or strength, according to multiplicity of synapses. Bumbarger and his colleagues also mapped differences in PageRank centrality onto the P. pacificus network (C). Node size is proportional to magnitude of the difference in PageRank between C. elegans and P. pacificus. Orange nodes have a higher centrality in P. pacificus, whereas blue nodes have a higher centrality in C. elegans. Nodes with connections to anterior pharynx output cells (red edges), including those nodes proposed to control predatory feeding, have a higher PageRank in P. pacificus than in C. elegans. Nodes with connections to posterior pharynx outputs (blue edges) have a higher PageRank in C. elegans than in P. pacificus. Reprinted with permission from Bumbarger, DJ et al., System-wide Rewiring Underlies Behavioral Differences in Predatory and Bacterial-Feeding Nematodes, Cell 152:109-119 (2013).



trality, to determine which parts of each network were most important to particular nodes.

Among other things, Bumbarger found that there was a general shift in network focus between the two species, with more going on in the anterior portion of *P. pacificus*' pharyngeal network than in the posterior portion—precisely the opposite of *C. elegans*. He also found that while information tended to follow the shortest path across *C. elegans*' pharyngeal network, information flow in *P. pacificus* was more indirect, suggesting more complex processing that could correlate with its more diverse feeding behaviors. And there were significant differences in connectivity and information flow associated with the two neurons that play the largest role in regulating feeding behavior in both worms.

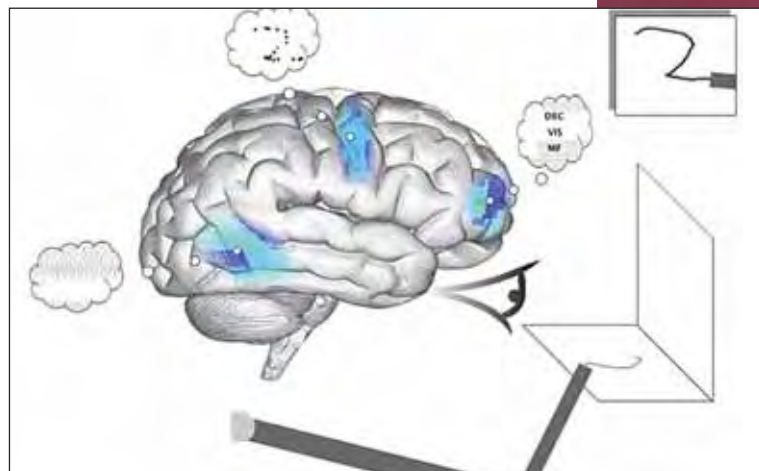
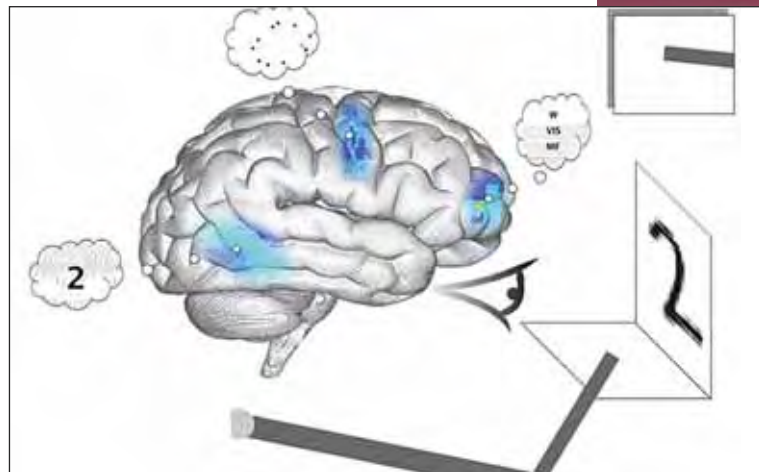
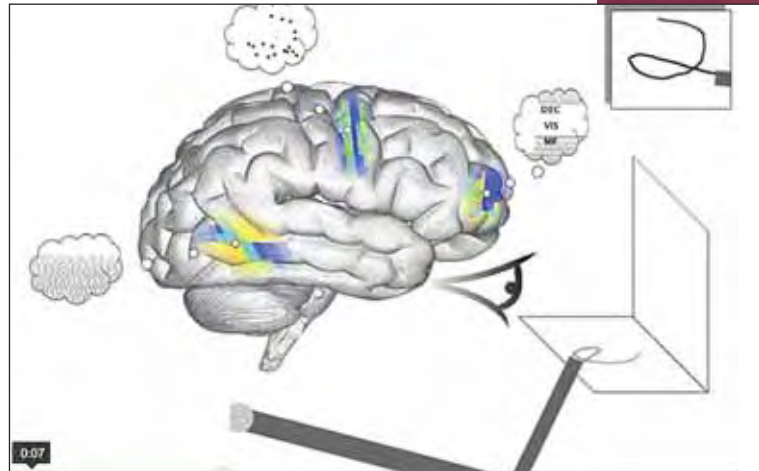
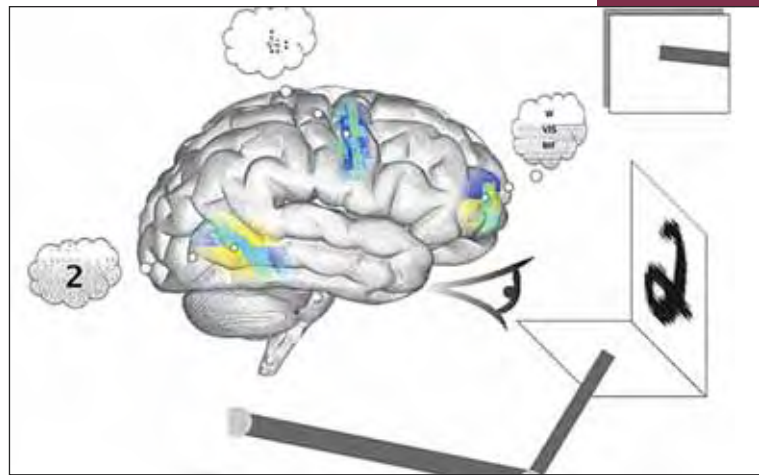
It's hard to know exactly what all this means in terms of function, but Bumbarger's findings point the way toward experiments that could help explain how differences in connectivity and network architecture affect behavior. He'd now like to do laser ablation experiments on both species—blasting away at those two neurons, for example—to see what, if any, changes in feeding behavior ensue.

Simulating a Human Connectome: Spaun

Worms are one thing, people another. And the amount of time it took Bumbarger to map just a tiny piece of *P. pacificus*' total connectome—let alone begin to understand its functional significance—gives some indication of why even the most ardent advocates of mapping the human connectome see it as a very long-term goal. But that hasn't stopped some researchers from taking the data that's already out there and using it to model human behavior.

Chris Eliasmith, PhD, a theoretical neuroscientist at the University of Waterloo in Ontario, Canada and author of the forthcoming book, *How to Build a Brain*, has developed a large-scale computational model of the human brain made up of two and a half million virtual neurons. Known as Spaun (for Semantic Pointer Architecture Unified Network), the model, which is described in a 2012 *Science* paper,

In these screenshots from movies of Spaun processing an input image, the spiking neural networks in the model are mapped to the corresponding anatomical areas. For example, the highest level of visual hierarchy lies at the back of the brain (in the inferotemporal cortex); the motor areas in the middle; and executive control in the front. When shown the number two, the visual area responds; and when prompted by a question mark to write the number, the motor area kicks in, recognizing not only the number but details such as the numeral two's loop (or lack thereof), demonstrating its ability to capture this subtle visual difference. Screenshots taken from <http://nengo.ca/build-a-brain/severaltasks>, courtesy of Chris Eliasmith.



can see with a simulated eye and write with a simulated arm. Eliasmith and his colleagues used as much neuroanatomical and biological data as they could to build Spaun; its simulated neurons are grouped into 20 anatomical structures (primary visual cortex, primary motor cortex, and so on) that are wired together in a realistic manner, mimicking functional brain areas that communicate with one another to reproduce a variety of cognitive behaviors.

Spaun was designed to perform eight basic tasks, including one that involves viewing a sequence of digits displayed on a screen, remembering them, and writing them down. Spaun can do it, but just like a real human being, it's better at remembering the items at the beginning and end of the list. It also performs about as well as most people would on a reasoning task that resembles the kinds of problems included on a common IQ test.

The fact that Spaun can do these things almost as well as people can, while also making the same kinds of mistakes they do, lends credence to the assumptions about brain wiring and function upon which it is based. For example, Spaun's ability to handle a diverse set of tasks is made possible by the way in which its virtual basal ganglia—a group of neurons that are associated with functions such as motor control and procedural learning—route information through simulated synaptic connections to different portions of its virtual cortex depending on the job at hand. To some degree, the system is even capable of changing its connection weights, or the strength of the connections between its neurons, a property that is believed to play a key role in memory formation, information processing, and behavior.

The model is far from complete. "It's large-scale in a way, but it's 40,000 times smaller than the

The practical benefits of a (relatively) large-scale model of a functioning brain are already apparent.

brain," Eliasmith says. And while it is capable of enacting very small variations on the routing schemes supplied by Eliasmith and his colleagues, Spaun will have to figure out how to rewire itself more substantially in order to learn new tasks. Still, the practical benefits of a (relatively) large-scale model of a functioning brain are already apparent.

On the one hand, Spaun should help neuroscientists figure out why specific connections matter, and how neural anatomy and physiology underwrite behavior. That, says Eliasmith, will help them understand how our brains relate to who we are and what we do. On the other hand, while Spaun can replicate normal cognitive behavior, it can also be used to

model the cognitive decline associated with aging, or the damage inflicted by diseases like Parkinson's and Alzheimer's.

Broken Connectomes: Understanding Brain Trauma

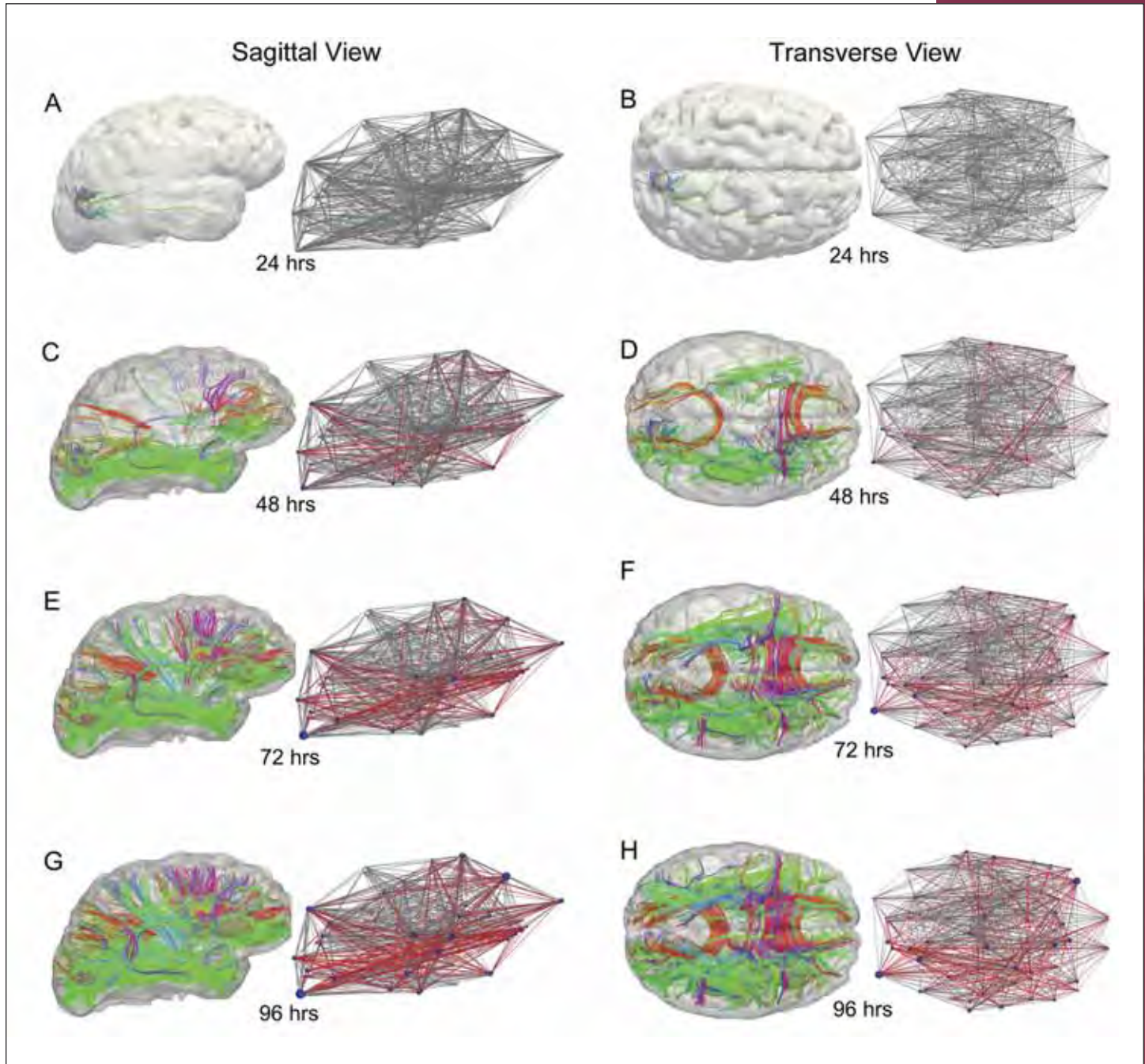
Damage is precisely what **Reuben Kraft, PhD**, assistant professor of mechanical engineering and member of the Institute for Cyberscience at Penn State University, wants to understand—and prevent. While working at the Army Research Laboratory, Kraft began trying to model traumatic brain injury (TBI), the signature injury suffered by American troops in Iraq and Afghanistan. Studies have linked TBI to chronic traumatic encephalopathy, a progressive neurodegenerative disease that affects both soldiers who are subjected to blast and concussion, and athletes such as boxers and football players who suffer repeated head injuries. Colleagues in the Translational Neuroscience Branch introduced him to connectomics, and Kraft, enticed by the combination of imaging and network analysis, was off and running. In a *PLoS Computational Biology* paper published last August, he and his collaborators used magnetic resonance imaging, biomechanical modeling, and graph theory to examine how a blow to the head might affect an individual's neural network.

Kraft used a technique known as finite element modeling to turn standard MRI scans taken from a graduate student at the University of California, Santa Barbara, into a three-dimensional model of a human head that included everything from skull and skin to brain tissue and cerebrospinal fluid. He then combined that model with a low-resolution connectome constructed via diffusion tensor imaging (DTI), a form of MRI that traces the approximate location of bundles of axons by analyzing the movement of water molecules along the fibers.

DTI provides far less detail than microscopy, but the technology is well suited to the kind of macroscale study that Kraft wanted to perform. By the time they were done, Kraft and his colleagues had a 3-D model kitted out with a connectome in which the nodes represented anatomical regions of the brain rather than individual neurons, and the edges represented the axonal highways that linked them together. By throwing in experimentally based models that predict how cells die off in the hours and days following an initial insult, Kraft was able to predict how the connectome would evolve over time after an impact to the head, with connections between brain regions degrading or disappearing completely as cells expired. He then analyzed the local and global effects on the network using graph theoretical measures of efficiency and connectivity. The system proved to be surprisingly robust: Even in the worst-

case scenarios, with many connections lost, none of the brain regions wound up being completely disconnected. Moreover, Kraft suspects that the extreme

protective portfolio” that might include recommendations about activities to avoid or precautionary measures to adopt—like not going out for the football



damage predicted at the outer limits of the model doesn't actually occur in nature, suggesting the existence of a protective or regenerative mechanism that has yet to be defined mathematically.

Kraft is already working on a follow-up paper that will use the model to examine network response and degradation in the face of blast injuries, which differ mechanically from physical impacts. Ultimately, he'd like to see this kind of biomechanical/connectomic model used to predict the likelihood that a particular person might be susceptible to long-term neurodegenerative disease following traumatic brain injury of any kind, whether sustained on the battlefield or the grid-iron. That information could form the basis of a “pro-

Finite element simulations coupled with cellular death predictions are used to specify injuries to white matter and subsequent damage over time. Damaged edges are shown in red, and node size increases as connections are lost. The predicted evolution of damage is shown for 24 (a and b), 48 (c and d), 72 (e and f), and 96 hours (g and h). Reprinted from Kraft RH, Mckee PJ, Dagro AM, Grafton ST, Combining the Finite Element Method with Structural Connectome-based Analysis for Modeling Neurotrauma: Connectome Neurotrauma Mechanics. PLoS Comput Biol 8(8): e1002619. doi:10.1371/journal.pcbi.1002619 (2012).

team, or wearing a particular kind of helmet.

It's the kind of idea—practical, useful, maybe even possible—that makes a proposal to map the human connectome seem like something that even the most skeptical critic could support. □

Betting on Genome Interpretation

By Katharine Miller

**Six Startups Jockey
FOR A PLACE AT THE Table**



A handful of startups are wagering that genome interpretation is the next big thing.

Why is this business space so hot?

“Once you can produce a better faster genome, thanks to Illumina and others, the bottleneck shifts downstream to processing, making sense of, and interpreting that data,” says **Jorge Conde, MBA**, co-founder and chief financial officer at Knome.

Each new startup seeks to turn whole genome sequences (or whole exome sequences—the portion of the genome that codes for proteins) into meaningful information. But each is also making a different bet about what approaches will succeed on the open market. Some companies plan to serve biotech or pharmaceutical researchers while others target clinical researchers or physicians and hospitals. Some focus on one step of the interpretive pipeline while others cover the whole shebang. Some provide cloud-based services, while others are betting on embedded platforms. And some rely on open-source algorithms and databases while others look to polish proprietary ones.

Which combination will ultimately prove successful is anyone’s guess. But right now, “there’s a lot of buzz,” says **Nicholas Schork, PhD**, founder of Cypher Genomics and professor of molecular and experimental medicine at the Scripps Research Institute. Researchers and medical institutions that are buying sequencing machines to do genomic profiling need someone to turn to who has a clue about the data they generate. “The time is definitely right to think about interpretation,” Schork says.

Sequence Crunching

Gaining knowledge from genomic data—the strings of C’s, G’s, T’s and A’s that constitute the standard output from a next-generation DNA sequencer—follows a series of fairly predictable steps, and some startup companies are putting their eggs in baskets defined by those steps.

Bina Technologies, for example, is focused on optimizing what’s called secondary analysis, the essential data-crunching step that happens immediately after the DNA sequence comes off the next-gen sequencer. That step requires software that first aligns an individual’s DNA sequence with a reference sequence and then picks out the differences between that individual’s sequence and the reference (a process known as variant-calling).

Bina is betting that the speed and accuracy of secondary analysis

will matter in a clinical context, says **Mahni Ghorashi, MBA**, director of marketing at Bina. Physicians don’t want to wait two weeks to determine appropriate cancer treatments based on genetic differences between tumor cells and normal cells, he says. They want that information now.

And for newborns whose lives are at risk, “they need the information in 48 hours or less,” Ghorashi says.

To meet that demand, Bina built a big-data plat-



THE BET:
Bina Technology

That the speed and accuracy of secondary analysis will matter in a clinical context.

form for genomics. Called the Bina Genomics Platform, it pairs specialized software with specialized hardware that’s designed to sit right next to the sequencer and analyze the data as it’s generated. “We’re able to take a process that used to take days or weeks and reduce it down to under four hours,” Ghorashi says. At Stanford, which is using the Bina platform as part of a pilot program for new-



Company launch date	June 2011
Funding	\$6.25 million in Series B funding announced March 2013
Staff size	14
Target customers	Researchers and clinicians
Products	Bina Genomic Analysis Platform—an embedded system of software and hardware for secondary analytics
Publicized successes	Pilots with Stanford University and the Palo Alto Veterans Administration

born screening, it used to take 10 days to process one genome on a shared computer cluster. “They are now processing 10 genomes a day, a 100x acceleration,” Ghorashi says.

Bina is only one of several companies that are focused on secondary analysis. DNANexus, Realtime Genomics and Appistry (not interviewed for this story) also focus on this area. In addition to its hardware version, Bina also offers its platform over the cloud, as does DNANexus. Realtime Genomics’ new genome analytics platform for the study of early childhood disease can be embedded locally or accessed on-demand in Amazon’s public cloud. And Appistry licenses GATK, the genome analysis tool kit developed by the Broad Institute. Each company makes claims of speed and accuracy akin to Bina’s, and other companies described below also incorporate secondary analytics as a part of their pipelines.


From Variants to Interpretation: Predictive Modeling

The output from secondary analytics software, such as Bina’s platform, is a variant file. To determine whether any of the variants are associated with disease, researchers or physicians must put those files through another computer pipeline (tertiary analysis) using a different product—one either developed in-house by the client or provided by another company—at least until Bina adds that step to its platform, which it plans to do. “We planted our flag upstream and our goal is to ultimately own the pipeline to guarantee accuracy,” Ghorashi says. But several other startups begin where the Bina platform lets off. Cypher Genomics and Knome, for example, are two startups that specialize in this space.

At the most basic level, tertiary analysis involves querying whether the genome contains variants that have already been discussed in the literature and are known to be associated with disease. But that’s just the beginning. Given that a

human diploid genome contains 3 billion base pairs and each genome has about 10 million variants in it, according to the National Library of Medicine, most variants found in an individual’s genome will not be described in the literature. Therefore, Conde says, at a bare minimum, interpretive pipelines need to include algorithms for predicting how a variant, though unknown, might be relevant for disease. Most companies, including Knome and Cypher, incorporate several open-source predictive algorithms to accomplish this goal. “If you don’t do that, you’ll only be looking for the keys under the streetlamp, as the story goes—because that’s where the light is,” Conde says.

In addition to open-source algorithms, Cypher li-



THE BET:
Cypher Genomics

That its analytical tools for genome interpretation, including some proprietary approaches licensed from Scripps Research Institute, will prove useful to pharmaceutical and other researchers.



Company launch date 2011

Funding Not publicly disclosed

Staff size 5

Target customers Pharmaceutical companies, research groups, and clinical partners

Products Cypher Analytics—a pipeline for ranking candidate gene variants in rare diseases; conducting family studies; and performing genetic association studies. The pipeline includes variant-impact prediction and gene-phenotype prediction.

Publicized successes Contracts with pharmaceutical companies and research groups

protein's structure or function—even if that variant has never been seen in the literature. Cypher also has tools for winnowing down from millions of variants to those that are likely responsible for a particular trait—be it a disease or drug response. And they have tools for leveraging annotations to make claims about groups of people, for example, individuals who do or don't respond to a drug.

Knome offers similar capabilities. “You need to be able to rapidly compare genomes to one another,” Conde says. For example, if a family member is sick and other family members are both sick and healthy, Knome's software can ask for mutations in genes that are predicted to affect protein function or structure in sick individuals where that variant is very rare and is not present in healthy individuals. “That very quickly filters you down to the needle in the haystack,” Conde says. In a study conducted on a family in British Columbia, Knome researchers used this strategy to find the sixth known genetic cause of Parkinson's disease. The same approach can work with unrelated individuals.


Knome's pipeline also includes “nifty algorithms,” Conde says, that look first for identical point mutations, then for mutations in the same gene, and then ultimately for mutations in genes that are part of the same pathway. In that final step, he says, “we tend to get very interesting hits.” This is important because unrelated people with the same disease or drug response are likely to have different mutations that fall within similar pathways.

For example, a pharmaceutical company asked Knome to look for gene variants that could explain why a group of unrelated people didn't respond to a particular drug. Knome's algorithms found that the nonresponders all had some level of mutation in different genes in the same network for metabolism of a particular starch. “To us it meant nothing,” Conde says. But the pharmaceutical company used that starch to stabilize the drug. “People with that metabolic deficiency were excreting the drug and the body was never really seeing it.”

Both Cypher and Knome provide genomics interpretation services for pharmaceutical and biotech

companies as well as clinical researchers.

In addition, Knome has created a product called KnoSYS™100, an end-to-end system for interpreting human genomes and exomes. The company is essentially betting that as the cost of sequencing goes down



THE BET:
Knome

That as the cost of sequencing goes down and the resolution of data goes up, clinics will shift from ordering an occasional test for a specific gene, to sequencing and storing whole exomes or genomes and querying them *in silico* whenever a test is needed.

and the resolution of data goes up, clinics will shift from ordering an occasional test for a specific gene, to sequencing and storing whole exomes or genomes and querying them *in silico* whenever a test is needed. “That's why our platform exists,” Conde says.

Diagnostic Odysseys

Some of the splashiest genomics news in recent years involved diagnostic odysseys—cases where whole genome sequencing was used to diagnose and treat patients with unique or very rare diseases. Both Knome and Cypher offer interpretation pipelines for diagnostic odyssey patients—ways to sift through genetic variants



Company launch date	2007
Funding	~\$12 million
Staff size	“Under 50”
Target customers	Pharmaceutical, medical, and academic researchers
Products	KnoSys™100—a fully integrated, locally installed, hardware and software system for the interpretation of human genome sequence data. KnomeDiscovery—an end-to-end solution for interpreting large numbers of human whole genomes and exomes—starting with sequencing and ending with a interpretation findings report.
Publicized successes	Discovered a new gene for Parkinson's disease

to find the likely culprit.

SVBio offers a combination of secondary and tertiary analytics to that same end. But SVBio differs from Knome and Cypher in its clinical rather than research focus. “Clinical companies come from a different mindset,” says **Dietrich Stephan, PhD**, SVBio’s CEO. “Rigor levels are much higher than for a research product.”

Because data that comes off next-gen sequencers is not in a form that can be used in the clinic, SVBio does a lot of massaging of the primary data in the alignment and variant calling step, Stephan says. And when assigning pathogenicity to a variant, they have to make sure they aren’t relying on a polluted public database.

In addition to accuracy, SVBio wants to be comprehensive. It’s not helpful to tell someone “you have a variant of unknown significance.” Instead, according to Stephan, SVBio can say with 99.5 percent precision, whether the variant is a mutation or polymorphism,

based on classifiers that are trained on all the historical data. Many labs do this, including Knome and Cypher, but according to Stephan, “Few have gone to the level we’ve gone to in terms of training complex classifiers on hundreds of attributes across 300,000 variants with publicly stated precision metrics around pathogenicity.”

In January 2013, the Mayo Clinic’s Center for Individualized Medicine teamed up with SVBio to build a robust software pipeline for interpreting a patient’s exome sequence—the portion of human DNA that codes for proteins—in a clinical setting. The system will go live in June.

What attracted the Mayo Clinic to SVBio? “We’d talk about sensitivity and specificity on a per patient basis,” Stephan says. “And we’d talk about the low probability of missing a diagnosis

across a specific number of patients. They liked that.”

One thing’s for sure: Getting the contract with Mayo didn’t hurt business. “This stuff is really complicated and nuanced and multifaceted,” Stephan says. “Being able to say Mayo Clinic is using it makes things a lot easier.”

Soup-to-Nuts Research and Diagnostic Services

Personalis¹ is one company that’s gone full bore into clinical research and diagnostics, with the goal of enabling accurate clinical grade insights into genomic data. They start with a DNA sample, sequence it in-house, do the alignment and variant calling, and analyze the variants to identify those with potential to cause disease. “It’s a kind of soup-to-nuts offering,” says **John West, MBA**, the company’s CEO. “We allow a customer to go from sample to insight.”²

By owning the whole process, West says, Personalis can innovate every step of the way. “We’re doing something novel in each area to achieve higher accuracy.”

So, for example, exome sequencers don’t actually catch all the genes. Coverage can be inadequate for many reasons, including sequencer bias against regions rich in guanines and cytosines (GCs), or because repeats are difficult to sequence, says **Richard Chen, MD**, chief science officer at Personalis. Because there could be something medically important in those gaps, Personalis has innovated to fill those holes—creating what they call ACE Technology™. “You don’t have to wonder if the variant you’re looking for isn’t listed because it wasn’t covered by the sequencing,” Chen says.

Similarly, Personalis has taken a close look at secondary analysis. “There are so many details there, and mastery of the process is not trivial,” Chen says. Many companies use standard tools and align against a standard reference that itself includes rare alleles. So Personalis has created its own reference genomes that contain the most common alleles for people of different ethnic backgrounds. “It gives us better alignment and variant calling,” he says. The company is also improving on public tools that are good for calling certain types of variants but do poorly on others, such as inserts/deletions and structural variants, Chen says. “For case-control analysis, accuracy in sequencing and alignment really matters so that the real biology can be dissected from the noise in the data.”

When it comes to the tertiary analytics—bringing biological meaning to genomic datasets—Personalis has also exclusively licensed and extended several large high-quality, manually curated databases in-

¹ Russ Altman is principal investigator for Simbios, which funds this magazine. He is also a founder and scientific advisor to Personalis as well as a personal friend of this author. He did not, however, play a role in the writing or editing of this story.

² Knome and SVBio also offer sequencing but they do not, so far, do it in-house.



The SVBio logo, consisting of the letters 'SVBio' in a blue, sans-serif font.

Company launch date 2011

Funding Undisclosed funding by Sequoia Capital

Staff size 20

Target customers Hospitals and physicians

Products Cloud-based genome diagnostic services

Publicized successes Contract with Mayo Clinic

cluding PharmGKB™, a pharmacogenomics database; the Personalis Variant Database, a large database of disease-related variants; and Regulome software, from the ENCODE project. All three of these were originally licensed from Stanford for exclusive commercial use. They have also built an annotation engine that integrates over 30 different databases. “In doing so we’ve reached a level of accuracy and comprehensiveness that is beyond what others are doing,” Chen claims.

Personalis also runs sophisticated analytics (not unlike those run by Cypher and Knome) to identify differences between cases and controls at the variant

will nevertheless be a market for their services.

Ghorashi says that although the open-source movement has been “really good” at developing algorithms, Bina adds value by making sure the open-source tools interoperate optimally. “These algorithms are written by biologists who don’t necessarily take that extra step,” he says.

Schork agrees. “At the end of the day, the delivery of the information is just as important as the accuracy. And the open-source tools don’t do as good a job with delivery.” Companies are set up to make it easy to sift through the data and present results in an effective way, he says. “That is not typically in the domain of the academic or weekend scientist.”

But at least one company is making a different bet. “A lot of people starting companies are spinning them out of academic lab efforts and replicating what’s available in the open-source community,” says **Jonathan Hirsch**, CEO of Syapse. “The useful thing a company can do is make the use and delivery of those easier, not replace the algorithms.”

In secondary analytics, for example, Hirsch thinks the bioinformatics community wants to directly use open-source tools like GATK (from the Broad Institute) and Bowtie (out of Johns Hopkins University). “If there’s a choice between proprietary algorithms and the open-source algorithms, usually it’s the open-source algorithm that’s going to win,” he says. He points to Spiral Genetics and Seven Bridges Genomics as companies that are focusing on helping customers run the open-source algorithms

more efficiently by offering delivery mechanisms and a distributed computing platform. There’s also GnuBIO, he says, which integrates the secondary analytics on the sequencer. “In the future, you won’t need a separate secondary analytics process,” he says. “The machine will do the work, just as it performs the primary analytics today.” He says the same thing about knowledge bases: He predicts that the public ones will dominate, such as ClinVar.

These views have influenced



THE BET:
Personalis

That a soup-to-nuts interpretation pipeline with innovations around issues of accuracy at every step along the way will be the preferred product for clinical research.

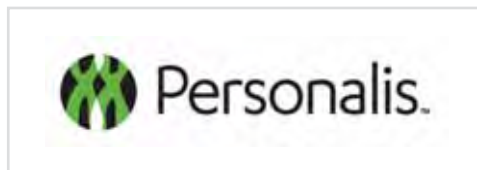
level as well as at the gene and pathway level. And their tools can analyze genomic data from an individual or family with the aim of discovering the genetic cause of a particular disease or characteristic. “We have developed a detailed process to apply what we know about the family and the biology to make the most likely variant stand out from the crowd,” Chen says.

Initially, Personalis is focusing on the clinical research market. In February, the Veteran’s Administration contracted with Personalis to analyze 1000 genomes this year, with an option to do the same for thousands more patients in the coming years. It’s a piece of the VA’s Million Veterans Program—an effort to build a vast DNA data repository and correlate it with the VA’s extensive electronic medical records.

Moving forward, the company would like to start working with hospitals and clinics. “It’s hard work to get to the level of accuracy required for clinical decision-making,” Chen says. “It’s a higher standard we’re holding ourselves to, but a necessary standard, whether you are doing research or using sequencing results to make life or death decisions for people.”

What about Open-Source?

All of the companies described here make use of some open-source tools. But they are betting there



Company launch date	2011
Funding	\$20 million
Staff size	25
Target customers	Clinical researchers but moving toward hospitals/physicians
Products	Soup-to-nuts diagnostics. From DNA sample to analytical report.
Publicized successes	Contract with VA to sequence and interpret 1000 genomes as part of the VA’s Million Veterans Program

Syapse's business model. "We don't do content," he says. Instead, Syapse is building what he calls "the software infrastructure for omics medicine." For content, Syapse hopes to partner with any or all of the companies described above. "We are not going to be the ones to choose the winner, so we want to partner with all of them."

Syapse is essentially a semantic computing company that builds a graph data structure that can leverage open-source ontologies for structuring biomedical terms and relationships. The goal is to make it easy to query the data to get a useful result. The company has two target audiences: data generators and clinics. For the data generators (hospital labs, diagnostic companies), Hirsch says Syapse will provide off-the-shelf software for managing, structuring, querying and reporting omics results. And for clinics, they will build an

omics medical record. It will allow the clinical site to connect omics data with the electronic medical record in a clinical decision support system that can recommend appropriate courses of action to physicians.

"Content is something that will eventually be free and open," Hirsch says. "So to me, the most important thing becomes the off-the-shelf software that enables users to make use of their data."

ask why the sequencing companies (such as Illumina and Life Technologies) aren't jumping in.

For secondary analytics, Ghorashi agrees that Illumina is interested. "Their customers don't want the raw data files. They want the alignment and call files," he says. But at the same time, he notes, "quite a bit of innovation needs to happen," and startups are better poised to move quickly.

For tertiary analytics, Schork says that the sequencing companies have plenty of activity just staying at the top of their own market without branching out into interpretation. "They see themselves as the iPad and these other companies are the apps," he says.

If the clinical space develops more, Conde says, "it's pretty clear that Illumina would want to dominate." But right now, "there's plenty of evidence to suggest that there's a role for new companies like ours."

How will it play out?

While it's anyone's guess which business model will prevail, one thing's clear: The appearance of multiple startups with an interest in genome interpretation foreshadows a potential sea-change toward personalized medicine. Torkamani says the time is ripe for making sense of the data in certain areas, such as pharmacogenomics, diagnosis of rare conditions, and cancer. "There's plenty of actionable information there," he says. And although there's still work to be done before genomics will make a dent in chronic common diseases, "even there, a few bits and pieces of information are starting to appear," Torkamani says. He points to ApoE for Alzheimer's disease and various risk markers for macular degeneration.

There remains a risk that the hype cycle for genome interpretation is only just starting. If patients go through testing and are told "you have these variants but we don't know what they mean" or, worse, are told predicted meanings that turn out to be false, companies could unintentionally cause harm to patients—and also to the entire industry.

Many of the things the current batch of startups hope to accomplish are entirely reasonable goals, says **Mark Gerstein, PhD**, professor of biomedical informatics, molecular biophysics and biochemistry, and computer science at Yale University, who is not personally involved in any genomics startups. But, he says, "There's a long way from an idea to having evidence that it's proven." For example, connecting variation to disease is still an area of intense research, he says. And being able to find different variants in the same pathway is not as straightforward as it sounds. Still, he says, "I think this business area is a good thing." The research community doesn't create production scale products that are ready for the clinic. "There's a lot of chaos in normal research," he says, "And extracting from that chaos hardened tested workflows that people can use is very valuable."

Torkamani hopes so. "We went into genetic research with the hope that it will impact peoples' lives," he says. "Now it's really possible to make that happen." □




THE BET:
Syapse

That off-the-shelf software that makes use of open-source databases and analytical tools will be the best way to help users make sense of their own data.

Where are the Big Dogs?

If genome interpretation is a hot niche, it seems reasonable to



Company launch date	2012
Funding	\$3 million
Staff size	3
Target customers	Hospitals
Products	Semantic infrastructure for omics generation and clinical reporting—agnostic as to knowledge base or analytical tools
Publicized successes	Provides infrastructure for several diagnostic companies such as InVita and Foundation Medicine, and for the Stanford Center for Genomics and Personalized Medicine

BY JOSE LUGO-MARTINEZ AND PREDRAG RADIVOJAC, PhD

Vertex Classification in Graphs

Graphs, or networks, have been widely adopted in computational biology, with examples including protein-protein interaction networks, gene regulatory networks, and residue interaction networks in proteins, to name a few. Graphs provide a single and methodologically well-understood way to describe high-throughput biological data as well as data from individual experiments.

Graphs are most useful when they are analyzed to draw inferences about the data. Such analyses fall roughly into two camps: unsupervised techniques for network motif finding (graphs that occur more frequently than expected) and clustering (grouping of data); and supervised techniques, which usually involve prediction tasks such as classification (prediction of discrete outputs) and regression (prediction of continuous outputs).

These supervised techniques can be applied to predict properties of a graph (graph classification) or of the vertices in a single graph (vertex classification). Below we describe how vertex classification techniques can be used

A structure of lymphocyte-specific protein tyrosine kinase (PDB id: 3lck) with a highlighted residue (Y394) that is known to be an autophosphorylation site.

to gain new insights into the residues that make up a protein.

When using graphs to analyze protein structures, the first step is to convert each protein structure of interest into a residue interaction network, where vertices represent amino acid residues and the links between pairs of vertices indicate that the two residues are in contact—often if the distance between them is within 3 to 6Å.

In the graph classification scenario, each protein can be seen as a different graph and the task may be to predict a structural or functional classification of such a protein, or graph—e.g., its fold class (e.g., barrel, globin) or its cellular role (e.g., catalytic activity, transcription factor activity). On the other hand, in the vertex classification scenario, all proteins are collectively considered as a single large disconnected graph, and the objective may be to predict some properties of interest regarding each residue. For example, the identification of functional residues (e.g., DNA-binding residues, post-translationally modified sites, etc.) falls under the vertex classification scenario, an example of

DETAILS

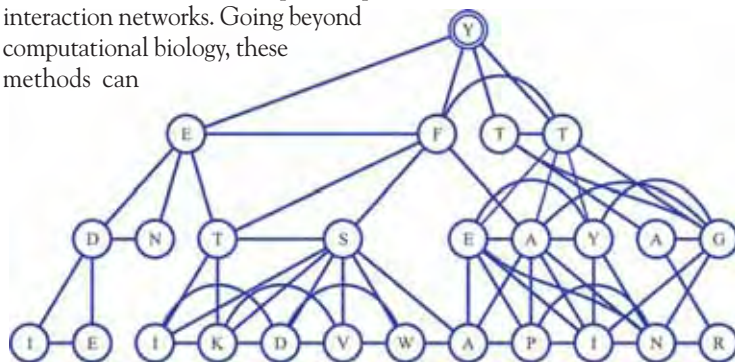
Jose Lugo-Martinez is a PhD candidate in computer science and Predrag Radivojac is associate professor of computer science and informatics at Indiana University, Bloomington.



which is shown here.

There are three principal approaches to vertex classification. First, for properties that tend to be localized, probabilistic graphical models (e.g., Markov Random Fields) can be used to propagate class labels across a graph, for example, from a group of DNA-binding residues to their neighboring vertices. Second, one can map each vertex together with its local neighborhood into a vector in the Euclidean space and then use standard machine-learning techniques for predictor development. Here, vertex properties such as degree, clustering coefficient, and others might be used to encode each vertex into a fixed-dimensional vector. Third and finally, if one has insight into how to effectively measure similarity between vertex neighborhoods, one might define a kernel (similarity) function over pairs of vertices based on their graph neighborhoods, for example, one based on simultaneous random walks starting at the two vertices of interest. Kernel functions can then be used by learning algorithms capable of working with similarities between objects rather than sets of object descriptors. In contrast to probabilistic graphical models, the latter two approaches assign class labels based on the similarity of vertex neighborhoods regardless of their location in the graph; however, they may be less effective in modeling dependencies between vertices. The final choice of a method thus depends on the problem at hand and domain knowledge.

In addition to identifying functional residues in protein structures, vertex classification is helpful for predicting protein function or disease associations from protein-protein interaction networks. Going beyond computational biology, these methods can



The local graph neighborhood for the Y394 residue (double circled). The graph was generated using a distance threshold of 6Å. Each residue is represented by a single letter amino acid code but the positional information is removed. The task of a classifier is to predict class labels (here, presence or absence of phosphorylation) for each vertex (local graph neighborhood) in the residue interaction network.

also help identify malicious web sites on the Internet or predict a person's voting preferences in a social network. As the volume and nature of data change with technology, development of vertex classification methods that can handle real-life (big and noisy) data, incorporate the wealth of auxiliary domain information in principled ways, and/or increase the efficiency of learning and inference will have wide implications not only for computational biology, but also for a number of scientific and industrial applications. □

Stanford University
 318 Campus Drive
 Clark Center Room S221
 Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

Trajectory Optimization And Physical Realism

An animated human figure seeking the optimal path from point A to point B typically relies on computationally expensive hard constraints that force the trajectories to be physically realistic. But contact-invariant optimization (CIO), as applied by Igor Mordatch, a graduate student in computer science at the University of Washington, can achieve physical realism more efficiently by changing the contact forces from binary (touching/not touching, which numerical optimizers can't handle in a smooth way) to a softer constraint that is more like a guideline. "It's like you have a jet-pack on your hands or feet," Mor-

datch says. As the optimization proceeds, it discovers for itself that the contact/no contact solution is optimal, while still preserving the physical realism of a smooth transition. "The gradual transition between contact and non-contact makes sure the numerical behavior is nice," he says. "That's kind of the primary trick."

Mordatch has used the approach to create animated figures that can stand from a prone position, do handstands, climb over walls, and pass objects. More recently, he has been adding physics-based muscle models in an effort to make the work useful for biomechanics researchers. He envisions a two-step process in which the simple models achieve the general motion that is then refined with a full physics-based model. "We haven't really tried that yet," he says. "It's exciting stuff for the future." □



For a trajectory-driven animation using CIO, the animator specifies a figure's initial position and target location (shown here as an "X") as well as the final stance pose (feet under the hips, feet shoulder-width apart, hands in a downward direction). In between, the optimization discovers when and where to place the hands and feet. Initially, the hands sort of slide across the ground as if flying with jetpacks. But after a while the hand contacts converge into single points that become the final solution. Screenshots courtesy of Igor Mordatch. Full movies viewable at <http://homes.cs.washington.edu/~mordatch/>.

