DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, a National Center for Biomedical Computing

Genetic Variants and Ill Health

Scanning 500,000 SNPs Yields Gene-Disease Connections

PLUS

Computing the Ravages of Time

Using Algorithms To Tackle Alzheimer's Disease

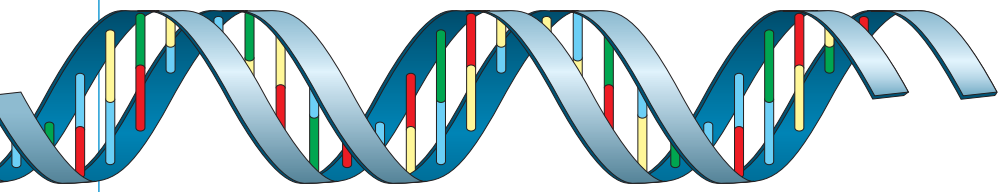
Fall 2007

contents

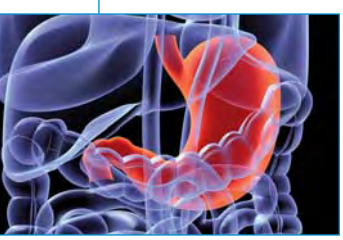
ContentsFall 2007

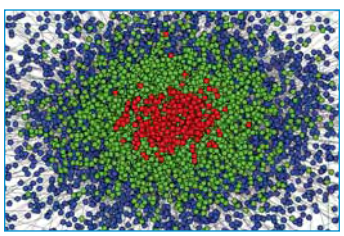
FEATURES

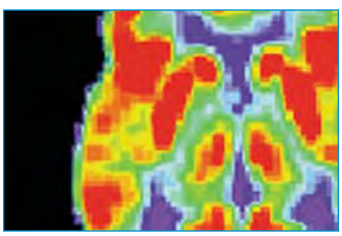
DEPARTMENTS



COVER ART BY
JAMES BENNETT OF AFFILIATED DESIGN







Fall 2007
Volume 3, Issue 4
ISSN 1557-3192

Executive Editor
David Paik, PhD

Managing Editor
Katharine Miller

Science Writers
Katharine Miller
Hannah Hickey
Regina Nuzzo, PhD
Louisa Dalton
Matthew Busse, PhD
Kristin Cobb, PhD

Community Contributors
Karin A. Remington, PhD
Ali Navid, PhD
Cheol-Min Ghim, PhD
Eivind Almaas, PhD

Layout and Design
Affiliated Design

Printing
Advanced Printing

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

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

Biomedical Computation Review is published quarterly by Simbios National Center for Biomedical Computing and supported by the National Institutes of Health 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)
Jennie Larkin, PhD (NHLBI)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Richard Morris, PhD (NIAID)
Grace Peng, PhD (NIBIB)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)
Jane Ye, PhD (NLM)
Yuan Liu, PhD (NINDS)

GuestEditorial

KARIN A. REMINGTON, PHD

It's not what you know, but what you don't know...

Bioinformatics and computational biology have told us a lot about biology—primarily that we know so little. Advances have led to many more unanswered questions, suggesting we know less and less all the time. For example, during the half dozen years since the human genome sequence was first published, the biomedical research community has been pouring through a variety of computational annotations on the human genome, including regulatory and protein coding sequences, evolutionary conserved regions, and a broad collection of potentially significant single-nucleotide polymorphisms (SNPs). Still, we have a very poor understanding of even these basic genetic elements, let alone the potential role of other new and yet to be discovered players. Similarly, modeling and computational simulations have advanced our understanding of the structural and chemical properties of the DNA molecule, revealing new significance for genomic regions long dismissed as "junk." What undiscovered control mechanisms might be hidden in the non-coding regions of the genome? Just this small sampling of recent findings aided by computational approaches reveals tantalizing new awareness of all that we don't know, and surprising new perspectives:

- A bioinformatics analysis, coupled with hypothesis driven lab work, led **Paul Pease, Oren Levy and Jerod Ptacin** at University of California, Berkeley to the discovery that specific short, asymmetric DNA sequences control the bi-directionality of the DNA translocase FtsK, a molecular motor involved in bacterial chromosome dynamics (*Science*, 2005).

- The ENCODE pilot project's recently published results (*Nature*, 2007), relying heavily on computational analysis and comparative sequence analysis, have demonstrated that even our most basic understanding of transcription must be revisited.

- **Eran Segal** and a team from the Weizman Institute and Northwestern University constructed a model of nucleosome-DNA interactions in yeast with a validation strategy that revealed a genomic encoding of the nucleosome organization (*Nature*, 2006).

- Computational simulations by **Jory Ruscio** and **Alexey Onufriev** at Virginia Tech have been targeted at the role of flexibility in nucleosomal DNA packaging, generating new experimental directions of inquiry that could elucidate nucleosome dynamics (*Biophysical Journal*, 2006).

- Work in **Carlos Bustamante's** lab at UC, Berkeley led to the development of new mathematical models of the DNA double helix from a mechanical perspective, inspired by their experimental observations of DNA winding and stretching properties that confounded earlier models and intuitions (*Nature*, 2006).

As Director of Bioinformatics and Computational Biology at the National Institute of General Medical Sciences, which has been privileged to fund many of these studies, I can sense a growing awareness within NIH of the extraordinary potential of asking new questions and revisiting long held assumptions of what we "know." The Institutes' recently announced, and aptly named, Exceptional Unconventional Research Enabling Knowledge Acceleration RFA (EUREKA, http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-08-002.html), reflects this sense of promise. In fact, the new world of science, where advances increasingly require large interdisciplinary teams, provides the perfect opportunity to thrive on what we don't know. Most of us realize, from first hand experience, the difficulty in establishing communication across disciplinary boundaries. This difficulty, though, ought to be its very strength. It gives us a rare opportunity to ask colleagues about what we don't know and what doesn't make sense to us. In turn, if we thoughtfully consider our responses to similar questions posed to us, we will allow ourselves to be enlightened by the questions that challenge our assumptions of what we know, and energized by the exciting paths to discovery that may result. □

DETAILS

Karin A. Remington, PhD
Director, Center for Bioinformatics and Computational Biology
National Institute of General Medical Sciences
National Institutes of Health

NewsBytes

Simulating Membrane Transport

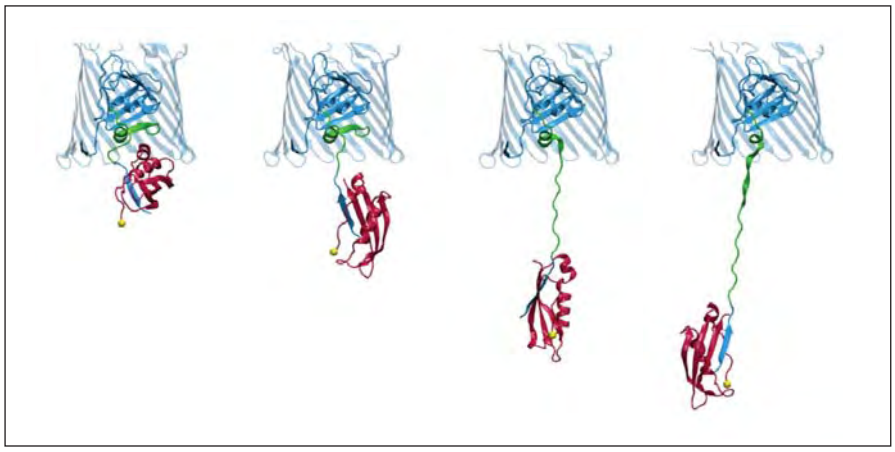
For a bacterium to admit certain large nutrients, a steady tug from inside might do the trick, according to computer simulations recently published in *Biophysical Journal*.

Bacterial membranes are loaded with a vast number of specialized transporters. For some of these to function, an energized inner membrane protein must transmit its energy to an outer membrane protein so that nutrients can enter the cell. The question is: How does that work? Does the inner protein shuttle across the membrane, unplug a pore, or simply yank open a gap?

"I don't know of any other model in which you energize a protein and then send it out to take energy to the outer membrane," says **Emad Tajkhorshid, PhD**, assistant professor of biochemistry, pharmacology and biophysics at the University of Illinois, Champaign, a co-author on the paper along with his student **James Gumbart**. "We are trying to improve the picture of this mechanism by doing this simulation," he says.

Bacteria can seem obsessed with finding and absorbing nutrients from the environment. Indeed, they dedicate more than 50 percent of their genetic material and 50 percent of their energy to membrane transport. For large nutrient molecules, such as vitamin B12, bacteria rely on TonB dependent transporters such as BtuB. These barrel-shaped molecules reside in the outer membrane with their tails (known as the plug domain) tucked in and plugging the barrel. TonB itself is anchored to the inner membrane with its tail end mating up with BtuB's plug domain, according to X-ray crystallographic results published last year in *Science*.

Based on experimental evidence, researchers know that when TonB is energized, nutrients can pass through the TonB dependent transporter. Scientists have proposed several possible mechanisms: TonB acting as a shuttle; or TonB forcibly pulling open the plugged barrel either by unfolding the luminal domain or by unplugging it entirely. Last year, with the new structural information about the TonB/BtuB complex, Tajkhorshid and his colleagues decided to simulate the "pulling" theory in order to determine where the force would be felt first.



In these four snapshots, as the end (yellow bead) of Ton B (red) is pulled inside the cell, the plug domain of BtuB (green) unfolds allowing nutrients to enter. Courtesy of James C. Gumbart and Emad Tajkhorshid.

"It was possible that if you pull on TonB it might just come off," says Tajkhorshid. But that didn't happen. The connection between TonB and BtuB held, while the luminal domain of BtuB unfolded. After pulling for 100 angstroms, this produced an opening wide enough for vitamin B12 to pass through. The team also tested the "unplugging" theory, but found that it took ten times as much force to remove the entire luminal domain. "Unfolding is much easier to induce than coming off as one piece," says Tajkhorshid.

But, he says, there are still plenty of unknowns. They didn't try to simulate the shuttle theory because it would require too much computing power to do so. And, "in our simulation, we had to pull for 100 angstroms to observe enough opening to let the substrate permeate the transporter. But it's obvious to me that there is no way you can have 100 angstroms of pulling from something on the inner membrane. There must be other things going on."

Susan Buchanan, PhD, an investigator in the Laboratory of Molecular Biology at National Institute of Diabetes and Digestive and Kidney Disease, agrees that a linear movement of 100 angstroms is unlikely. But, she says, "The simplest thing to do in simulations is to apply a linear force as he did. *In vivo*, it could be a combination of some sort of rigid body movement, conformational changes, and rotation. But those things are hard to simulate." What's important, she says, is that this work provides a model that people can look at further. "With the recently solved crystal structure," she says, "it's important to do simulations at this point because no one's been able to do this in vivo yet."

—By Katharine Miller

"I don't know of any other model in which you energize a protein and then send it out to take energy to the outer membrane," says Emad Tajkhorshid, PhD.

Flexible Molecular Computer Functions Inside a Cell

A newly created molecular computer works in human cells and offers the flexibility of a general-purpose circuit. The advance, described in *Nature Biotechnology* in May, brings closer the eventual possibility of placing bio-based computers inside cells to diagnose and treat disease on a cellular level.

"In theory, there is no limit to the decision-making complexity" that this system can handle, says bioengineer **Yaakov Benenson, PhD**, a Bauer Fellow at Harvard University's Center for Systems Biology. Until now, molecular computers have mostly been test-tube prototypes that tackled just one specific task, such as tic-tac-toe. Benenson, along with **Ron Weiss, PhD**, assistant professor of electrical engineering at Princeton University, devised a way to engineer a general purpose circuit by taking advantage of the cell's cell-regulation pathways.

The scientists first got their machinery to work inside a human kidney cell by mimicking a virus. They transfected the cell with genes that code for the circuit. The cell then took up the genes and created the computer network for them.

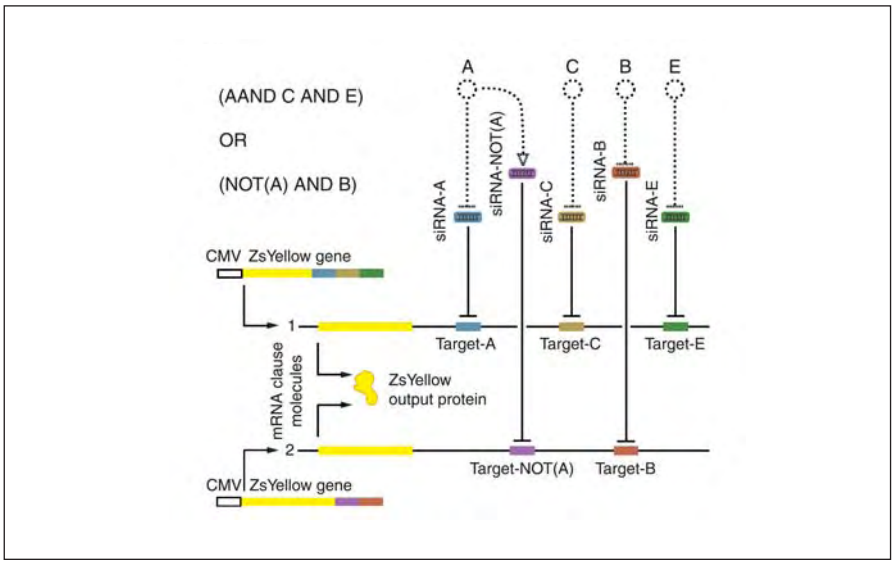
The network itself is made up of engineered mRNA strands that encode a chosen protein and smaller RNA strands that interfere with the translation of the mRNAs. Scientists can engineer these small, interfering RNAs (siRNAs) to bind any number of possible disease markers in a cell. (Weiss and Benenson did not experimentally validate sensing disease markers in this work.) In the simplest scenario, once an siRNA binds a disease marker, that siRNA can't interfere with translation of the mRNA, and the protein is made. The protein can be whatever the designer likes: a fluorescent tag or therapy for the diseased cell, for example.

By adding interacting pairs of mRNAs and siRNAs, the researchers can individualize the network to handle any problem that can be expressed as a Boolean logic formula—equivalent to Boolean operations run on traditional silicon-based computers. The formula could be simple: "If marker A or marker B is present, then make the protein." Or it could be much more complex: "If marker A is present and marker B is absent, or if marker A is present and marker C is present, or if marker D and E are both absent, then make the protein."

Weiss and Benenson tested their system using a network of five siRNAs and two mRNAs. Complex functions, Benenson says, are limited by the scalability of the molecular components.

According to **Darko Stefanovic, PhD**, associate professor of computer science at the University of New Mexico, many functions "will require unacceptably complex forms." Yet, Stefanovic comments, "the paper presents an innovative way of accomplishing logic computation using transcriptional networks." It's a promising direction, he says, for synthetic biology.

—By Louisa Dalton



Benenson's molecular networks utilize engineered messenger RNAs (mRNAs) and small interfering RNAs (siRNAs) to perform logic equations such as the one in this example diagram. If cellular marker A is present, it binds siRNA-A, keeping the siRNA from binding Target-A on mRNA-1. If siRNA-C and E are similarly occupied, mRNA-1 is freed for translation, and the output (ZsYellow output protein) is positive. The output is also positive if marker B is present, which occupies siRNA-B, and marker A is absent, because A activates siRNA-NOT(A) to bind Target-NOT(A). (This example uses human cytomegalovirus (CMV) immediate-early promoter.) Adapted and reprinted by permission from Macmillan Publishers Ltd: **Nature Biotechnology, *25(7):795-801, 2007.***

"In theory, there is no limit to the decision-making complexity" that this system can handle, says Yaakov Benenson.

The Spontaneous Brain

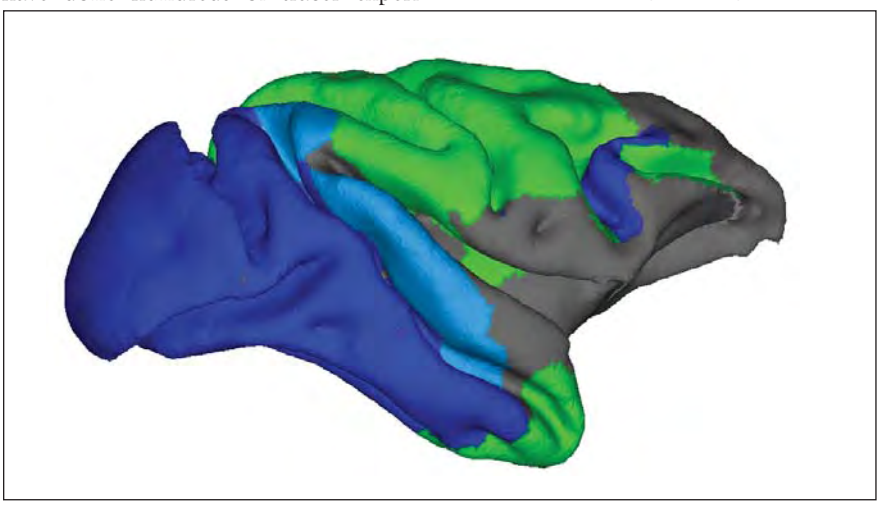
When people sit peacefully at rest, doing and thinking nothing in particular, their brains still buzz merrily along. In scans called functional MRIs, they light up in characteristic patterns. No one knows the purpose of this spontaneous chatter, but it accounts for up to 98 percent of the brain's activity and burns about three-quarters of the brain's energy. To help unravel its origins and significance, researchers at Indiana University built a new computational model of a macaque monkey brain, which they describe in the June 12 issue of the *Proceedings of the National Academy of Sciences*.

"With this work, we can shed some light on what is actually driving the pattern of activation and deactivation that is seen in the resting brain," says **Olaf Sporns, PhD**, associate professor of psychology, who worked on this project with his graduate student, **Christopher J. Honey**. The brain's activity at rest may ultimately influence how individuals think and behave and how the brain responds to injury and disease.

Sporns and Honey chose the macaque brain because its wiring diagram is well understood. Researchers have done hundreds of tracer experiments—where they inject dye into one area of the brain and trace its spread to other areas—to establish the connectivity patterns of the macaque brain. From these data, Sporns and Honey built a "connection matrix" that specifies which of 47 brain areas are connected and which are not. On top of this roadmap, they superimposed differential equations that describe the electrical activity of each brain area. Then they ran a simulation to see how their virtual brain lights up when it is just talking to itself, with no external inputs.

The resulting patterns of brain activity closely resembled those seen in imaging studies of the human brain at rest. Interestingly, though the model operates at a very fast time scale (sub-millisecond resolution) it generates the slower fluctuations seen on fMRI (seconds to tens-of-seconds resolution). "Despite the fact that we have fast dynamics, we get these very slow processes to unfold," Sporns says.

When they randomly scrambled the connection matrix in their model, they no longer saw the characteristic activity patterns of the resting brain. "So we have a good argument that what we see is actually because of the specific pattern of the connectivity," he says.



Monkey Brains. This virtual macaque monkey brain lights up in the characteristic patterns of a brain "at rest." Courtesy of Olaf Sporns; Reproduced from Figure 4c of Christopher J. Honey, Rolf Kötter, Michael Breakspear, and Olaf Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. PNAS 2007 104: 10240-10245. Copyright 2007 National Academy of Sciences, U.S.A.

"With this work, we can shed some light on what is actually driving the pattern of activation and deactivation that is seen in the resting brain," says Olaf Sporns.

"Their work makes this very important step of linking the anatomy—the connections between the brain areas—to the patterns of spontaneous activity. I think this is really the first study that makes this link explicitly," comments **Giulio Tononi, MD, PhD**, a professor of psychiatry at the University of Wisconsin. "They are able to explain many of the features that are observed in studies using fMRI."

The next step is to apply this modeling approach to the human brain, Sporns says. Though people cannot undergo invasive tracer studies, a new non-invasive technique—diffusion tensor imaging—is providing the connectivity data for human brains.

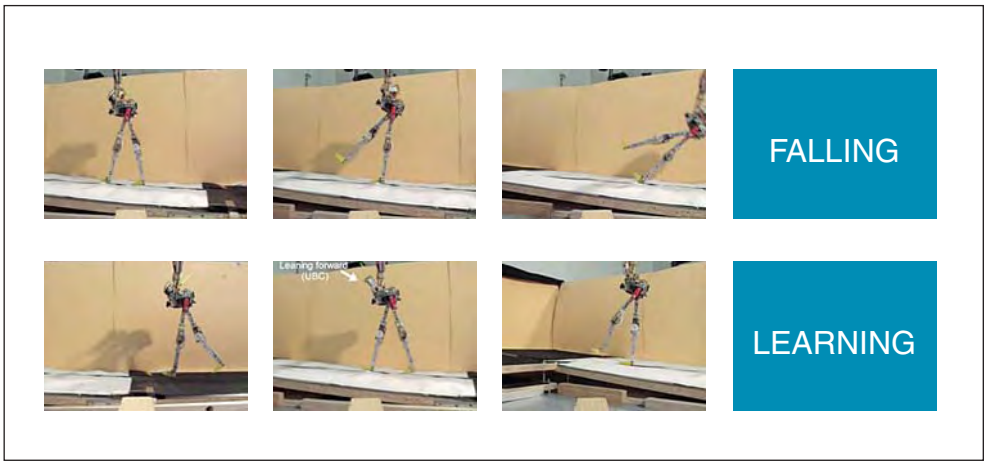
Using human models, Sporns plans to study how brain lesions interrupt the brain's network—its connectivity, spontaneous activity, and ultimately performance. "There is great potential here for understanding brain injury and recovery processes," he says.

He also plans to study how the resting brain's activity influences people's thoughts and behaviors. Every person has a unique pattern of spontaneous activity. "The open question is whether this spontaneous activity actually colors or somehow interacts with our ability to do a task," Sporns says. "If that were the case, that would be really interesting."

—By Kristin Sainani, PhD

An Uphill Challenge

RunBot, already the world's fastest bipedal robot, has now also learned to keep its balance when walking up ramps. "We have achieved a synthesis of different functionalities, between biomechanics, neuronal reflexive control, and adaptive control, which allows the machine to learn," says **Florentin Wörgötter, PhD**, head of the Computational Neurosciences Group at the University of Göttingen in Germany, and leader of the team that built RunBot. The work was published in *PLoS Computational Biology*.



On its first try at a slope, Runbot teeters backward and falls (top). But it learns from its mistakes: On subsequent efforts, Runbot makes it to the top of the hill (bottom). Courtesy of Florentin Wörgötter.

Creating a robot that walks as smoothly as a human is a long-standing challenge. Many walking robots are plodding and methodical, precisely calculating the trajectory of each step. The human gait is much more dynamic; our center of gravity is constantly shifting as we swing our legs forward. Last year, Wörgötter's group produced RunBot, a successful dynamic walker that could swing its legs almost as quickly as a human. However, RunBot was limited to walking on level surfaces, unable to adjust its balance to walk up an incline.

To address that shortcoming, Wörgötter's team added a learning mechanism that simulates synaptic plasticity, enabling RunBot to learn in a manner similar to humans. The learning mechanism allows RunBot to associate an infrared (IR) sensor, which detects changes in the angle of the floor, with an accelerometer, which detects the rapid acceleration of falling.

The first time RunBot's IR sensor detected a change in slope, the signal had no meaning, and RunBot continued to walk as normal until it fell, triggering the accelerometer. Over the next few trials, RunBot learned that the signal from the IR sensor requires a change in gait to avoid triggering the accelerometer. With guidance from the researchers, who predefined the direction in which RunBot could alter the parameters controlling its gait, RunBot experimented with different magnitudes of those parameters, resulting in different postures and stride lengths. After four or five trials, RunBot learned to lean forward and take shorter steps, similar to what humans do when walking up a slope.

RunBot is able to easily change its gait because of the hierarchical structure of its control systems. On the bottom level, each step is controlled by a reflexive neural network. Sensors in the feet, knee, and hip monitor the position of each joint relative to the other joints and the ground, and artificial motor neurons make minor adjustments to maintain stability. In this manner, the reflex control level autonomously generates a repetitive walking motion.

On top of the reflexive control lies an adaptive neural network, which controls RunBot's posture. By tweaking the activation parameters of the reflexive motor neurons, the adaptive control system causes RunBot to lean forward and take shorter steps when its IR sensor detects an upcoming slope.

In addition to creating robots with a more human looking stride, Wörgötter's work may be applicable to prosthetic legs. His lab recently started working with a major supplier of prosthetic devices, to apply similar neural networks in advanced prosthetics.

"RunBot is a successful demonstration of a small-scale 2-D biped that uses a controller that approximates a static neural network and a novel learning algorithm," says **Steven Collins**, president of Intelligent Prosthetic Systems and a doctoral candidate at the University of Michigan.

—By Matthew Busse, PhD

"We have achieved a synthesis of different functionalities, between biomechanics, neuronal reflexive control, and adaptive control, which allows the machine to learn," says Florentin Wörgötter.

Modeling Early Evolution

The fittest organisms survive and produce offspring, according to the Darwinian theory of natural selection. And the changes that make an organism fit happen at the molecular level: when genes mutate they produce different proteins generating traits that may or may not benefit the organism. Yet the relationship between proteins and organism fitness is not well understood.

Now, for the first time, a computer model has attempted to connect the dots between organism evolution and the evolution of proteins.

"People understand that somehow the properties of proteins determine the evolution of populations, but this is only words," says **Eugene L. Shakhnovich, PhD**, a professor of chemistry and chemical biology at Harvard University and lead author of the paper that appeared in *PLoS Computational Biology* in July 2007. "There's no detailed microscopic picture of how these two biologies talk to one another."

So Shakhnovich and his colleagues simulated an evolving set of proteins under selection pressure. "We developed a model, simpler than life, but still it's microscopic with these two levels intimately connected," he says. "So the properties of proteins in the model—like stability—are directly related to properties of model organisms that carry these proteins."

The simulation starts with 100 organisms, each with the same single primordial gene in their genomes. At each time step, the organism can replicate, die, undergo a gene mutation or duplication, or do nothing (with each event having a certain probability of occurring). The life expectancy of the organism is directly related to the stability of its proteins. The latter was determined using a "lattice" model that approximates a protein's actual structure. It's a useful approximation, however, because—for purposes of this model—a given amino acid sequence produces a specific measure of the stability of the native state of the protein, says Shakhnovich.

In about half of the 50 simulation runs, the organisms died off. But the successful organisms showed a characteristic pattern of protein evolution the researchers called "Big Bang" behavior. "At some point there is a discovery of a small number of advantageous protein structures and sequences that have evolvability properties," says Shakhnovich. "These serve as a nucleus for expansion of the protein universe."

Over time, the model reproduced other quantitative features of the existing protein universe, says Shakhnovich. "This makes us think that this model, while not the whole truth, captures essential aspects of early evolution."

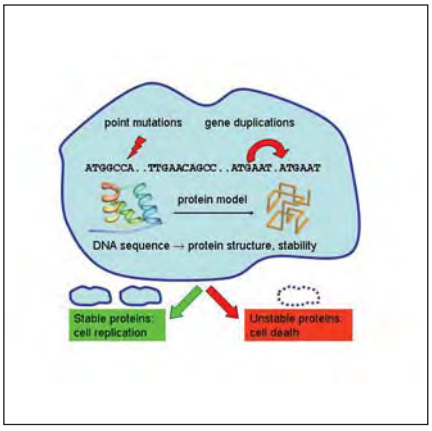
> "[We] think that this model, while not the whole truth, captures essential aspects of early evolution," says Eugene Shakhnovich.

For example, in nature, one finds large and small protein families co-existing. This is inconsistent with a random process. "It has been a mystery as to why this type of distribution pops up in protein science and genetics," says Shaknovich. "Our model suggests the source of it is in the evolutionary dynamics of proteins."

Next steps include adding more complexity: e.g., protein-protein interactions and immune responses. The researchers also hope to gain a better understanding of protein stability, possibly even using that information to develop more stable proteins, useful in drug discovery.

"This is the first paper where people have used a simple but realistic model of protein folding to simulate genomes containing multiple genes," says **Claus Wilke, PhD**, assistant professor of integrative biology at the University of Texas Center for Computational Biology and Bioinformatics. "I think that's an interesting approach, and I think that over time those kinds of simulations will lead to all kinds of interesting insights."

—By Katharine Miller



This schematic depicts a first-principles simulation of early evolution. One hundred organisms, each with the same single gene, begin to evolve. At each time step, the organism can replicate, die, undergo a gene mutation or duplication, or do nothing. The organism's life expectancy depends on the stability of its proteins as determined by a protein lattice model. Courtesy of Eugene Shakhnovich.

Teaching Resource: Computing Life

Explaining biocomputation to non-scientists can leave a person tongue-tied. Technical jargon gets in the way, and the breadth of the field resists encapsulation.

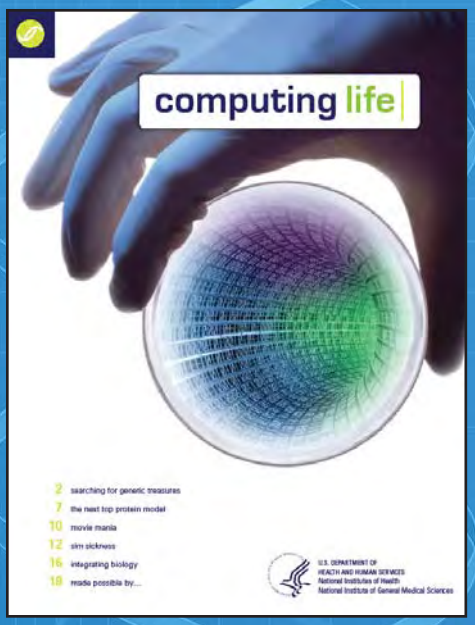
To help out, and to reach out to a new generation of future scientists, the National Institute of General Medical Sciences (NIGMS) has now published *Computing Life*. Due out in September 2007, the 24-page booklet presents snapshots of scientists' labs and brief overviews of what's happening across the field. The intended audience: high school and early college students.

"With tremendous challenges ahead, we need to encourage young people to think about science with excitement and creativity," says **Karin Remington, PhD**, Director of the Center for Bioinformatics and Computational Biology at NIGMS. "Publications such as *Computing Life* help to light a spark, channeling what might've been untapped intellectual power toward the sciences, and building appreciation for the relevance and importance of scientific research in our lives."

To that end, *Computing Life* reads like a magazine. It's very visual, with plenty of colorful and intriguing graphics and tight, explanatory captions. "The booklet brings pop culture to the science," says editor **Emily Carlson** of the NIGMS Office of Communications and Public Liaison.

Topics covered include genomics, protein folding, infectious disease modeling, molecular dynamics simulation, and systems biology, among others. The booklet also provides links to online material including simulations and movies. "We plan to maintain a complementary web site," says Carlson. "We'll post new material there as our way of keeping the publication up to date."

—By Katharine Miller □



"With tremendous challenges ahead, we need to encourage young people to think about science with excitement and creativity," says Karin Remington.

Contents Include:
Searching for Genetic Treasures
The Next Top Protein Model
Movie Mania
Sim Sickness
Integrating Biology

To obtain free copies of Computing Life to use as a teaching tool visit http://publications.nigms.nih.gov/order/.

BY REGINA NUZZO, PHD

Computing the Ravages of Time

Using Algorithms To Tackle Alzheimer's Disease

In 1906, at a small medical meeting in Tübingen, Germany, physician **Alois Alzheimer** gave a now-famous presentation about a puzzling patient. At age 51, Auguste D.'s memory was failing rapidly. Confused and helpless, she was growing inarticulate and fearful of her family, Alzheimer reported. Auguste died four years later.

During the autopsy Alzheimer found dramatic shrinkage in Auguste's brain, with cells that were already dead and dying at the time of her death—plus two kinds of microscopic deposits that Alzheimer had never seen before. He summed it up in his presentation abstract: "All in all, we are faced obviously with a peculiar disease process."

Now, a century later, about 5 million people in the United States have Alzheimer's disease, at a cost of more than $100 billion annually. About one in every eight people 65 years and older has been diagnosed with the disease. With lifespans continuing to lengthen and waves of baby-boomers hitting prime-risk ages, the number of Alzheimer's patients could triple by the time today's college students enter retirement.

Thus far, no clinical treatment has been shown to stop Alzheimer's neurodegeneration. In addition to searching for new pharmaceutical targets, however, researchers are grappling with other disease fundamentals: how plaques and tangles form on the brain, how best to detect early onset of the disease before cognitive decline starts, and how to predict a person's genetic risk.

The stage is set for computational approaches to Alzheimer's, says **Arthur Toga, PhD**, a professor of neurology at the University of California, Los Angeles. The slippery, highly variable nature of the disease demands sensitive tools, an aging population creates the urgency, and new technology provides the power to meet those demands. "In some sense," he says, "we're now set for a perfect storm for Alzheimer's disease research."

"All in all, we are faced obviously with a peculiar disease process," said Alois Alzheimer in 1906.

Computational tools are extending researchers' reach at all scales. Molecular dynamics simulations visualize the protein clumps in the brain that experiments can't capture. With unprecedented ease, data-mining methods sift through the rapidly accumulating information about the proteome and genome. And sophisticated imaging analyses reveal changes in the structure and functioning of the entire brain.

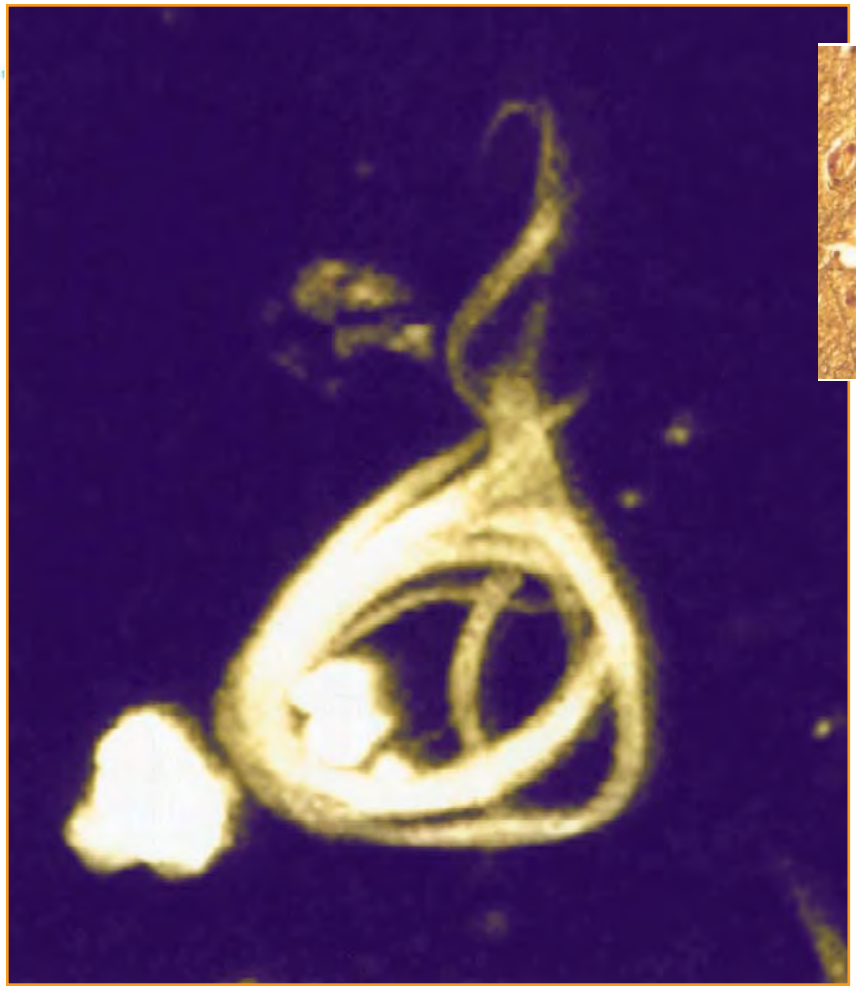
PROTEIN DYNAMICS: GETTING AT THE CAUSE OF ALZHEIMER'S

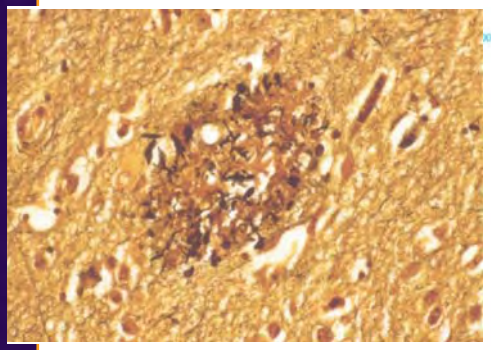
In the 1960s, researchers were finally able to use new electron microscope technology to see the molecular structure of the two types of mysterious lesions that Alzheimer first noticed in his patient's cerebral cortex—the so-called senile plaques and neurofibrillary tangles. The plaques, it turns out, consist mainly of amyloid beta peptides, while the tangles consist of abnormal forms of the tau protein. How these two proteins influence each other is not well known.

Some researchers postulate, however, that aggregates of amyloid beta—seen as senile plaques in their final form—are the proximal cause behind Alzheimer's disease, and the tangles and other neuropathological changes are a side effect of the gone-haywire amyloid beta assembly. Known as the amyloid cascade hypothesis, this suggests that understanding amyloid self-assembly could help crack open the puzzle of how Alzheimer's disease starts in the first place.

Researchers trying to study amyloid beta through experimental approaches run into problems, however, because many amyloid beta aggregates are unstable and short-lived. Computer simulations, on the other hand, provide the chance to study small amyloid beta aggregates in full atomic-resolution glory. Over the past two decades computational power has increased, allowing for better "all-atom" molecular dynamics simulations of short time-frames. And for longer simulation dynamics, coarse-grained protein models have been developed that can boil down a large number of degrees of freedom to a more manageable few, for instance by representing amino acids by a less complex structure of "beads."

H. Eugene Stanley, PhD, professor of physics and physiology at Boston University and director of the university's Center for Polymer Studies, models the folding and aggregation of amyloid beta peptides with a variety of approaches. In recent work, **Stanley, Brigita Urbanc, PhD**, senior research associate in physics at Boston University, and their students simulated these peptides using a coarse, four-bead protein model, in which amino acids are represented by three backbone beads and one side chain bead. Urbanc, Stanley and colleagues have been espe-





In Alzheimer's disease, beta amyloid, a protein fragment snipped from amyloid precursor protein (APP), clumps together and is mixed with other molecules, neurons, and non-nerve cells. Plaques develop, as seen here, in the hippocampus and in other areas of the cerebral cortex. Courtesy of the National Institute on Aging.

Healthy tau proteins stabilize microtubules, which themselves support neurons. In Alzheimer's disease, damaged tau begins to pair with other threads of tau and form tangles, as seen here. The microtubules disintegrate, and the neurons' support system collapses. Courtesy of the National Institute on Aging.

cially interested in investigating differences between the two most common protein forms seen in senile plaques: amyloid beta 40 and amyloid beta 42.

Their results, published in the *Proceedings of the National Academy of Sciences* in 2004, showed that the amyloid beta 40 and amyloid beta 42 peptides first folded into collapsed coil structures, then assembled into chains of different lengths. During the simulation, Stanley says, amyloid beta 42 tended to form longer chains, and the amyloid beta 40 shorter ones—in proportions consistent with laboratory results.

More recently, Stanley and his colleagues refined their simulation model to include the presence of electrostatic interactions between pairs of charged amino acids. Published in *Biophysical Journal* in June 2007, their results point to a specific spot on the amyloid beta 42 chains—the C-terminal region—that may be crucial for the molecule to aggregate. This suggests that inhibitors targeting this region could prevent chain formation or change the structure of the assemblies to reduce their toxicity, Stanley says.

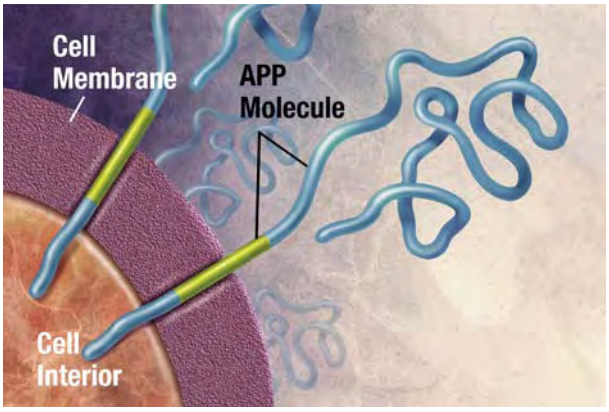
These *in silico* analyses are useful, Stanley points out, because they lead to predictions that lab researchers can test *in vitro*. And because computer simulations can reveal crucial, three-dimensional details of amyloid beta molecules, they can also aid in designing and testing drug molecules specifically for this target. "It's easier to design a key if you know the exact, three-dimensional contours of the lock," he says.

PROTEOMICS: SEEKING BIOMARKERS TO HELP DIAGNOSE ALZHEIMER'S

Before Alzheimer's disease can be treated, of course, it needs to be spotted—and the sooner the better. Evidence suggests that molecular mechanisms of the disease are at work early, perhaps even several years before neurons start dying and cognition starts to decline.

Yet tests that can accurately and reliably detect the disease at early stages have been hard to come by. As researchers understand more about the proteins involved in the disease process, they are also starting to investigate whether any of these molecules could serve as an Alzheimer's biomarker.

The answer isn't likely to be found in a single protein, however. The obvious candidates for biochemical markers—amyloid beta 40, amyloid beta 42, and the hyperphosphorylated tau protein—are indeed found at elevated levels in Alzheimer's



Amyloid Precursor Protein (APP) is associated with the cell membrane, the thin barrier that encloses the cell. After it is made, APP sticks through the neuron's membrane, partly inside and partly outside the cell.

patients, but they are also found in other neurological disease patients as well as in some normal controls.

Some researchers are therefore taking a big-picture, proteomic approach. They're looking for a combination of proteins whose expression levels in blood plasma serum or cerebrospinal fluid might yield a biochemical signature of Alzheimer's disease in early stages.

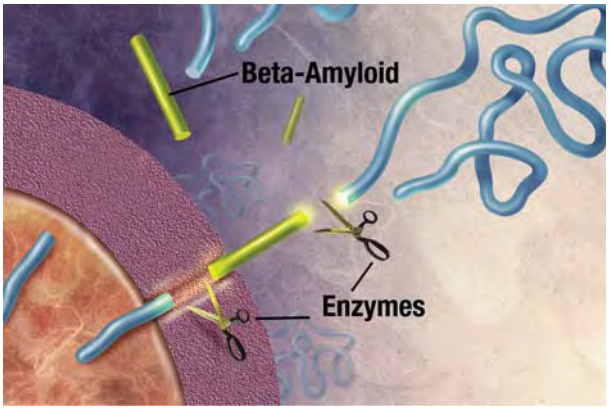
The lab of **Tony Wyss-Coray, PhD**, associate research professor of neurology at Stanford University, recently collaborated with Satoris, Inc., a biotechnology company Wyss-Coray co-founded, on such a project. Their focus: signaling proteins in plasma. **Sandip Ray**, chief scientific officer and cofounder of Satoris, came up with the idea. Using supervised learning software called Predictive Analysis of Microarrays (PAM), the researchers studied plasma expression levels for 120 immune response factors and other signaling proteins from an initial set of 43 Alzheimer's disease subjects and 40 age-matched unaffected controls. The algorithm honed in on a subset of 18 proteins that seemed to be characteristic and predictive for Alzheimer's disease.

Individually, each protein could not accurately classify the subjects as either a case or a control. But taken all together, the proteins' expression signature appeared to be good at predicting disease status, Ray says.

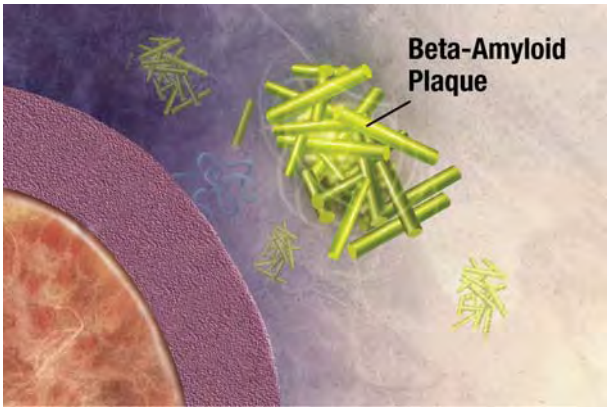
The researchers tested the 18-protein predictor on an independent test set of 92 subjects, which, like the training set, was drawn from seven different patient centers to minimize possible center biases, Ray says. The predictor reached a total accuracy of 89 percent.

The group went on to evaluate the expression signature's predictive abilities for a set of 47 patients with mild cognitive impairment, a condition which sometimes precedes Alzheimer's disease. The expression signature predicted that 27 of these patients would later develop Alzheimer's disease, and indeed, 20 of the 27 were diagnosed with the disease within six years. Overall, the predictor achieved an estimated 91 percent sensitivity and 72 percent specificity.

"In some sense, we're now set for a perfect storm for Alzheimer's disease research," says Arthur Toga.



Enzymes act on the APP and cut it into fragments of protein, one of which is called beta amyloid.



The beta amyloid fragments begin coming together into clumps outside the cell, then join other molecules and non-nerve cells to form insoluble plaques.
Courtesy of the National Institute on Aging, www.nia.nih.gov.

Biologically, the 18 proteins seem to point to a systemic, not isolated, dysregulation in neuronal support, immune response, cell growth and cell death in Alzheimer's disease patients several years before clinical symptoms appear, says **Markus Britschgi, PhD**, a postdoctoral fellow in Wyss-Coray's lab and presenting author of a poster on the work in June at the Alzheimer's Biomarkers Meeting in Washington, D.C. The work has recently been accepted for publication in *Nature Medicine*. "But what we don't know at this time is whether these dysregulations are due to processes in the brain or processes only in the periphery," he says.

GENOME-WIDE ASSOCIATIONS: TYING GENES TO ALZHEIMER'S

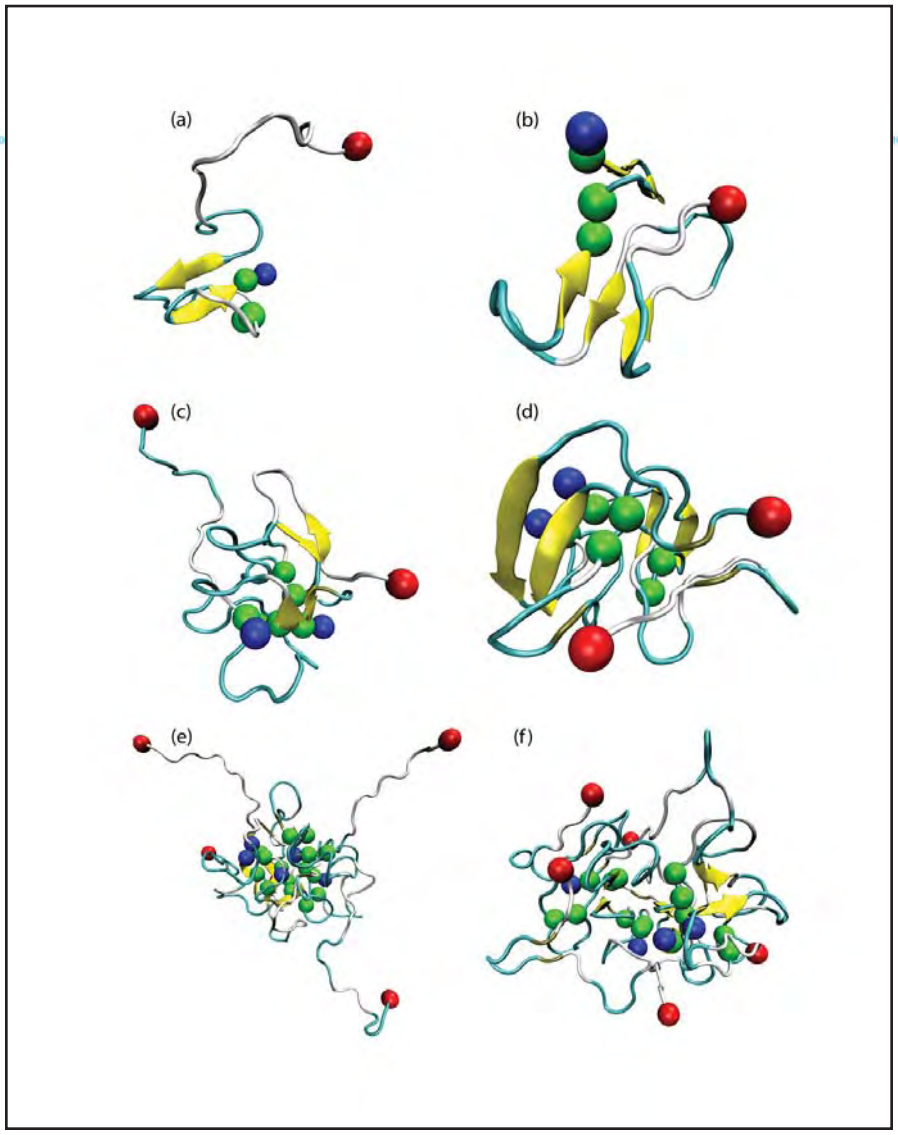
Studies of twins hint that up to 80 percent of Alzheimer's cases are due to genetic causes. Yet only three genes have been found on which mutations likely cause the disease through simple Mendelian inheritance: APP, which encodes the amyloid beta precursor protein, and PSEN1 and PSEN2, which encode presenilin 1 and 2. Mutations on these genes cause familial early-onset forms of the disease.

Most people, however, develop Alzheimer's disease after the age of 65 and do not have such a strong history in the immediate family. For this form, the most important known gene is ApoE, which encodes apolipoprotein E. Yet ApoE doesn't convey the whole picture: only about half of late-onset cases have a copy of the high-risk allele.

The search is on for other Alzheimer's susceptibility genes, but hard results have been elusive so far. It has been suggested that the survey of at least 300,000 single nucleotide polymorphisms (SNPs) from the whole human genome might be necessary for studies of genetically complex phenotypes. Until recently, however, published studies had not looked at more than 100,000 SNPs at a time.

Technology is changing that. "We're entering the era of high-density genome-wide association studies," says **Eric Reiman, MD**, executive director of Banner Alzheimer's Institute in Phoenix, Arizona. Thanks to advances over the past decade—in computing power, microarray technology and analysis tools, and human genome maps, for instance—genome-wide association studies are suddenly becoming feasible and successful (see the other feature story in this issue of *BCR*).

Their benefits extend beyond simple efficiency. The methods, which use high-throughput processes to examine about half a million genomic markers, can test many SNPs independent of any biases related to a researcher's favorite gene. "What's exciting about hypothesis-free genome-wide studies is that they can help uncover new mechanisms that people haven't thought about before," Reiman says.



Typical conformations of a folded monomer, dimer, and pentamer of amyloid beta 42 in the absence (a, c, and e) and presence (b, d, and f) of electrostatic interactions. Simulations by H.E. Stanley and colleagues suggest that the C-terminal region (marked here by a blue sphere) plays a key role in the formation of amyloid beta 42 oligomers—and the relative importance of this region increases in the presence of electrostatic interactions. Drugs targeting this area may be able to prevent the oligomers from forming or perhaps reduce their toxicity in the brain. Courtesy of Sijung Yun. Reprinted with permission from the Biophysical Society, **Biophysical Journal** ***92, 4064-4077 (2007).***

The first high-density genome-wide association study of Alzheimer's disease was published in *Neuron* in June 2007 by a 15-institution international team led by Reiman and **Dietrich Stephan, PhD**, associate director at the Translational Genomics Research Institute in Phoenix. That research was supported by 20 of the National Institute on Aging's Alzheimer's Disease Centers.

Using samples from 861 subjects with late-onset Alzheimer's disease and 550 elderly unaffected controls, they genotyped about 500,000 SNPs. These classifications were verified in more than 1,000 Alzheimer's cases and controls at autopsy. In three rounds of analyses, the researchers found six promising SNPs from a single gene that were significantly associated with the disease in subjects with the high-risk ApoE epsilon 4 allele. The SNPs all lay within the GRB-associated binding protein 2 (GAB2) gene.

In this particular study, the most significant SNP on GAB2 was associated with an overall four-fold increased risk for Alzheimer's disease, Reiman says. And people who carried both the epsilon 4 allele and the GAB2 high-risk allele had a 24-fold increase in risk for Alzheimer's disease.

The study's results need to be replicated with independent data, Reiman cautions. But for now they allow for possible mechanisms to be tested—investigating, for instance, whether the normal form of the GAB2 protein protects vulnerable neurons from tangles, he says.

"What's exciting about hypothesis-free genome-wide studies is that they can help uncover new mechanisms that people haven't thought about before," says Eric Reiman.

The researchers have deposited all of their data into the public domain. "We have just begun to have enough letters in the genetic book of life to understand the genetic story of Alzheimer's disease and other common phenotypes," Reiman says.

Letting Intermediate Phenotypes Stand In for Alzheimer's in Genomic Studies

One problem that genetic studies of complex diseases can run into is simply finding the right people to study. Clinical diagnoses of Alzheimer's disease in particular are not always accurate, and small errors in identifying the cases and controls in a study can mask or skew real genetic associations in the results.

One way to overcome this is to work with endophenotypes: intermediate quantitative traits that stand in for a more complex disease phenotype. Finding a good endophenotype for Alzheimer's isn't simple, however. It must be a trait that is heritable, associated with the causes and risks of Alzheimer's disease—which itself is still a mystery—and ideally be normally distributed within the population, says **Alison Goate, PhD**, professor of psychiatry, genetics and neurology at Washington University Medical School in St. Louis.

With the right endophenotype, however, the power of a genetic study can jump dramatically, Goate says. "If your quantitative trait represents something that is highly correlated with the disease but controlled by a small number of genes, then it should be easier to find those genes with the quantitative trait," she says.

Amyloid beta peptide levels are a natural endophenotype candidate, Goate says, because they are highly correlated with the presence of Alzheimer's disease and also with high-risk alleles in APP, ApoE, PSEN1 and PSEN2.

Goate and her colleagues are working with cerebrospinal fluid levels of two of the most common forms of the protein, amyloid beta 40, amyloid beta 42, plus the ratio of amyloid beta 42 to amyloid beta 40.

As part of an ongoing study, they recently looked at a set of 300 subjects in which two-thirds had a family history of Alzheimer's disease but were themselves unaffected and one-third had a diagnosis of mild Alzheimer's disease.

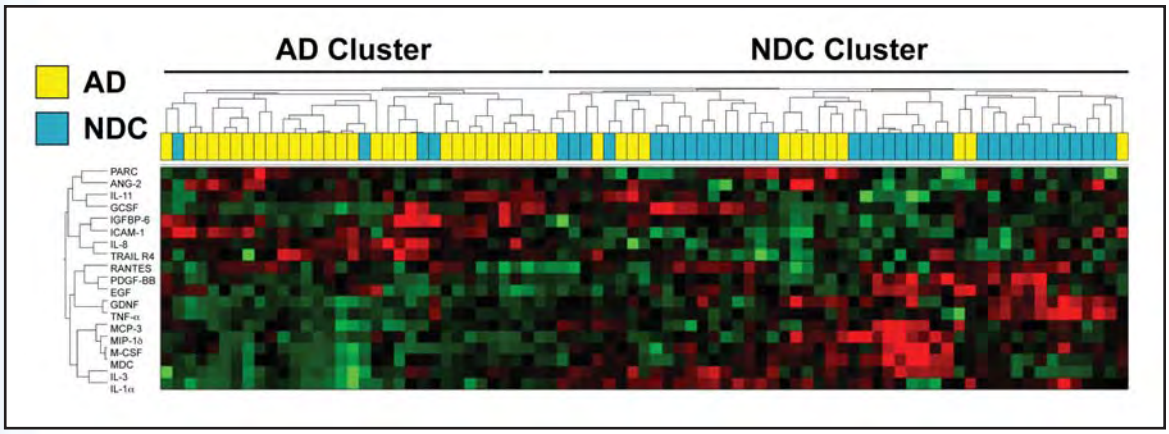
From a list of 19 candidate SNPs selected from the AlzGene database's meta-analysis, nine were significantly associated with the amyloid beta endophenotoype, with eight showing directions of association that were consistent with existing meta-analyses. "This is promising, because it suggests that these associations are likely to be real," Goate says.

The group is still collecting more samples, Goate says, and they hope to reach the point where they have a large enough sample to try out the amyloid beta endophenotypes in a broader set of SNPs across the entire genome.

IMAGING: Capturing the Brain on Screen to Diagnose and Track Alzheimer's

Brain imaging has long played a role in the diagnosis of Alzheimer's disease by helping physicians exclude the possibility of brain tumors or other ailments. More recently, however, researchers have become interested in using imaging tools for broader purposes: understanding the disease, detecting it at early stages, and tracking its progress over time.

As the tangles and plaques of Alzheimer's disease creep across a brain, its structure changes in subtle ways. With a skilled eye, radiologists examining brain magnetic resonance images one by one can quickly categorize the spread and degree of atrophy in the brain. But researchers would like to use assessments that rely less on subjective evaluations of skilled experts.



Normalized array measurements of 120 plasma signaling proteins from 43 Alzheimer's disease patients (yellow) and 40 non-demented controls (blue) were analyzed with the statistical program called significance analysis of microarray (SAM) to discover significant differences in protein concentrations. Samples are arranged in columns and proteins in rows. Increased expression in patients versus controls is shown in shades of red, reduced expression is shown in shades of green, and median expression is shown in black. Courtesy of Tony Wyss-Coray.

Voxel-based methods to catch early signs of disease

Some researchers are developing machine learning approaches that focus on individual voxels. **Clifford Jack, MD**, a professor of radiology at Mayo Clinic and postdoctoral fellow **Prashanthi Vemuri, PhD**, are investigating one such pattern classification method.

The technique uses a support vector machine algorithm, which aims to find a combination of brain image voxels that can best distinguish images of Alzheimer's patients from unaffected controls, Vemuri says. Their results, from a set of images of 380 Alzheimer's disease subjects and unaffected controls, were presented at the Human Brain Mapping meeting in June 2007.

The researchers first narrowed their attention to those brain regions that showed evidence of atrophy in Alzheimer's disease subjects. Within these regions, their tool found a subset of voxels that best classified the subjects into cases and controls. Altogether, the algorithm winnowed 10,000 voxels down to an essential set of 300, Vemuri says.

And these voxels, it turns out, form regional clusters that mirror the typical spread of neurofibrillary tangles. This provides an extra intuitive affirmation, Jack says. But the quantitative validation is what really counts: the method achieved 85 percent sensitivity and 85 percent specificity. Adding information about age, gender, and ApoE genotype further boosted both scores to 90 percent.

The process takes less than 15 minutes per case to run on a desktop computer. "Ten years ago, it might have required a supercomputer to do it," Jack says. "People in medical imaging are just now taking advantage of improved software available to the public."

Modeling Brain Contours to Find Alzheimer's

Another approach to analyzing structural brain images is to take a step back from the trees and look at the forest. Rather than analzying data on individual voxels, some methods model overall contours of brain regions, an approach that characterizes the shape of subcortical and cortical structures.

John Csernansky, MD, a professor of psychiatry and neurobiology, and **Lei Wang, PhD**, a research assistant professor, both at the Washington University School of Medicine in St. Louis, along with **Michael I. Miller, PhD**, a professor of biomedical engineering and electrical and computer engineering at Johns Hopkins

University, are working with surface-based methods that stem from classical mechanics.

When brain regions of Alzheimer's patients atrophy over time, they change shape in complicated ways. Miller has pioneered methods based on the principles of computational anatomy—which include tools such as large-deformation high-dimensional brain mapping—to model these variations.

The techniques assume that differences in brain contours can be captured by "morphing" one brain anatomy into another through high-dimensional diffeomorphic transformations that smoothly change one shape into another. Essentially, Miller says, brain matter is modeled as if it had the physical properties of a viscous liquid. Since sets of differential equations describe the transformation, group differences can be efficiently characterized.

The group has applied their methods in a variety of settings. In a longitudinal study of 44 subjects published in 2003 in *Neuroimage*, the researchers used patterns of change in hippocampal shape over two years of follow-up to distinguish subjects with mild Alzheimer's disease from unaffected elderly controls. And in a study of 49 subjects published in 2005 in *Neuroimage*, variation in the shape of a particular part of the hippocampus surface could predict whether a subject would go on to develop mild Alzheimer's during five years of follow-up—and if so, how long it took for cognitive effects to show up.

With thousands of data points collected on each hippocampal surface and only a relatively small number of subjects, these methods demand some form of data reduction, Csernansky says. Early studies used principal components analysis to hone in on the most informative areas of the brain surface. More recently, however, the group has been working to make their results more interpretable to clinicians by using a simplified anatomical template of the hippocampus.

In a study of 135 subjects published in 2006 in *Neuroimage*, patterns of surface variation in particular hippocampal substructures could distinguish subjects with very mild Alzheimer's disease from elderly controls. In particular, changes in two specific areas of the hippocampus surface, one in the CA1 subfield and the other near the subiculum, significantly increased the odds that a subject had very mild Alzheimer's disease.

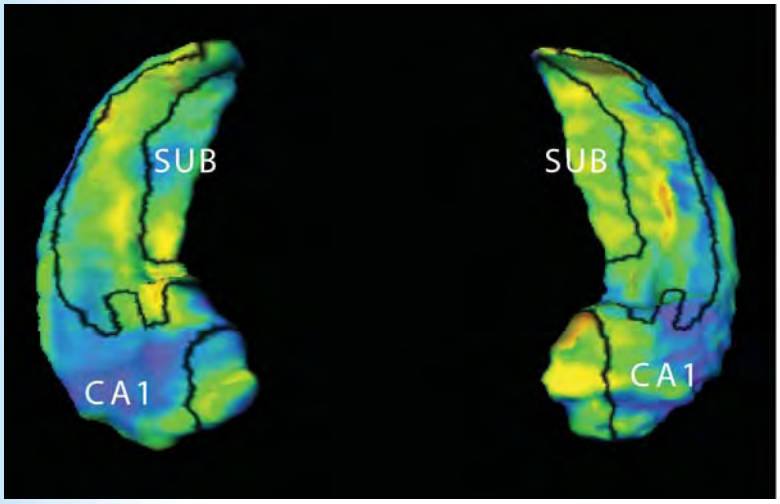
ADNI: Bringing it all together

Complex, degenerative diseases such as Alzheimer's can't easily be captured in small, cross-sectional studies. Researchers need large sample sizes to reach the statistical power necessary to tease apart subtle interactions, and ideally they would like to follow subjects over time in order to eliminate the large variability in how individuals age differently.

To that end, the National Institute on Aging in combination with the National Institute of Biomedical Imaging and Bioengineering, the pharmaceutical industry, and private foundations have been supporting a 5-year, $60-million longitudinal study that started in 2004. Known as the Alzheimer's Disease Neuroimaging Initiative (ADNI), the project's principle investigator is **Michael Weiner, MD**, of the University of California at San Francisco. The ambitious, highly collaborative undertaking aims to track the progress of Alzheimer's disease and its precursors, and develop validated biomarkers for Alzheimer's disease clinical trials.

The study is following 400 subjects with mild cognitive impairment, 200 subjects with Alzheimer's disease, and 200 elderly controls approximately every six months for two to three years, at about 50 sites across the nation. Researchers are collecting a variety of information from the subjects: MR images, clinical ratings, neuropsychological test results, and blood and urine samples from all the subjects, as well as [18F]-2-fluoro-deoxy-D-glucose (FDG) PET scans from half the subjects, cerebrospinal fluid from at least 20 percent of the subjects, and Pittsburgh Compound-B (PIB) PET scans from nearly 100 subjects. The data are immediately deposited into a repository that is freely available to the public.

As of June 2007, 804 subjects have been enrolled at 57 sites. Thousands of raw and processed images (scrubbed of the subjects' identities) have already been posted at UCLA's Laboratory of Neuro Imaging website (http://loni.ucla.edu/ADNI). Researchers expect that all studies and analyses—much of it computational—will be completed by the end of 2010.



With anatomical modeling tools, Csernansky, Miller, Wang and colleagues are able to capture regional changes in hippocampal shape from Alzheimer's disease. This image compares subjects with very mild Alzheimer's disease (as a group) to nondemented controls. Regions colored cool (purple and blue) are smaller in the Alzheimer's group compared to the controls; regions colored yellow and green are unchanged. The researchers found significant changes in the CA1 subfield and the subiculum (labeled). Courtesy of Lei Wang, PhD.

The group is looking now at how particular substructures change over time in Alzheimer's patients as compared to the normal aging population. And they hope to combine their own measures of surface deformations with other types of data, such as functional images or PIB-PET scans, Wang says. "With this type of metadata, we can understand how the disease progresses and also do a better job of prediction," he says.

Functional Imaging to See the Alzheimer's Brain in Action

With functional brain imaging, researchers can investigate the clinical aspects of Alzheimer's disease: how does the brain behave differently when it's affected by the disease?

Functional magnetic resonance imaging provides some of the most detailed clues to this question. Many fMRI studies have pointed out particular brain areas that show damaged functioning in Alzheimer's disease patients. But some researchers are now interested in how the entire brain might also change and adapt as the disease progresses.

Michael Greicius, MD, an assistant professor of neurology at Stanford University, is particularly interested in how brain regions connect and communicate among themselves. Recently, he and **Kaustubh Supekar**, a biomedical informatics graduate student also at Stanford, turned to analyzing a large network of brain regions for mathematical characteristics that were first used to describe social networks.

Their approach uses small-world measures, which have also been used to analyze a variety of other networks, including the Internet, global airline routes, and "six-degrees-of-separation" human networks. In social groups, a network node would be a person; in functional brain networks, a node represents a particular region of the brain.

Previous work has suggested that normal brains, like human social networks, exhibit small-world characteristics. This means that they encompass many tight clusters of nodes, and that information shared between any two nodes must likely pass through a large number of short-range connections. Greicius and his colleagues wanted to see if there were any small-world differences between Alzheimer's disease brains and unaffected brains.

The group recorded resting-state fMRI brain activity in 36 Alzheimer's disease patients and unaffected elderly controls every two seconds for six minutes. They then looked at activity in 90 separate regions of the brain—tens of thousands of voxels for each brain region—and created a time series of activity for each. They could then calculate the connectivity, or the amount of mutual information, between each of the 90 nodes.

The Alzheimer's disease patients, they discovered, had significantly less regional connectivity and displayed more impaired small-world functioning than did healthy controls. The results were presented at the Human Brain Mapping meeting in June, 2007.

It's not yet clear what the results mean biologically, Greicius says, but for now that's fine. "Intellectually it's less satisfying if there's not a clear biological interpretation, but from a practical, clinical standpoint we're agnostic as to what's driving the results, as long as they're reproducible and accurate," he says. A measure based on regional connectivity was able to distinguish Alzheimer's

disease patients from healthy controls with 73 percent sensitivity and 80 percent specificity.

The next step is to see if data reduction tools could construct a more simplified global network—based on 20 or 30 regions, say, rather than 90—that could better classify individual subjects as having Alzheimer's disease or not, Greicius says.

Challenges for the Future:

Future challenges for computational work in Alzheimer's disease research will likely center around the usual suspects, researchers say: data, people, and money.

"From the computational standpoint, researchers need more powerful ways to glean information from an increasing array of complex datasets," says Eric Reiman of the Banner Alzheimer's Institute, "and they need new ways to characterize the relationships among these potentially complementary datasets."

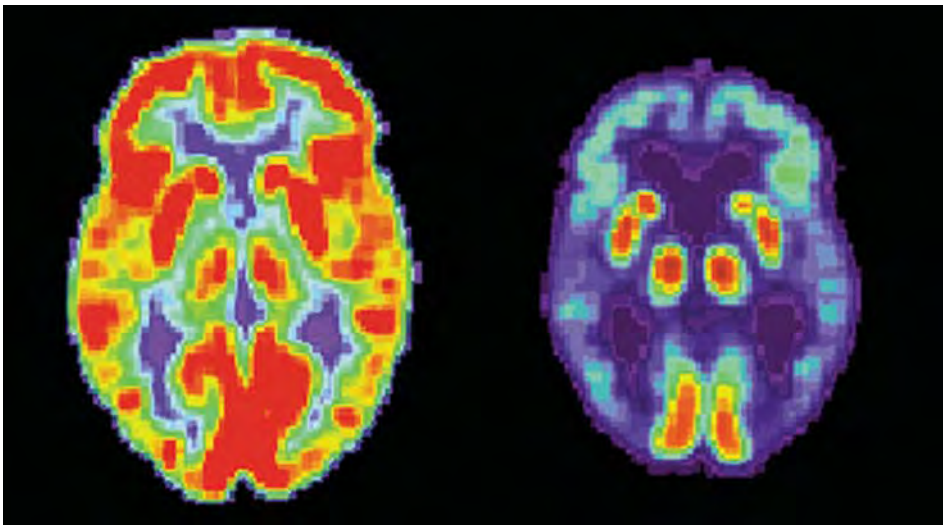
Indeed, simply using study subjects recruited in different ways from different clinical centers poses a real problem for the integrity of results, says Clifford Jack, MD of the Mayo Clinic. "The incompatibility of these patient groups is a huge confounder in our field that's not well recognized," he says.

What's more, the old, single-lab approaches to research likely won't survive in today's über-connected environment. "Increasingly, researchers from different scientific teams must work together to address their problems in a more fundamental way than any one team could do by iself," Reiman says. "That is both the challenge and opportunity now at hand."

Researchers might need to be jacks-of-all-trades, or at least forge connections with colleagues across campus. "We need computational researchers to enrich our information, but then that information needs to be transformed back into something that biologists and clinicians can comprehend," says John Csernansky of Washington University. "It takes time and a willingness to struggle together for a common understanding."

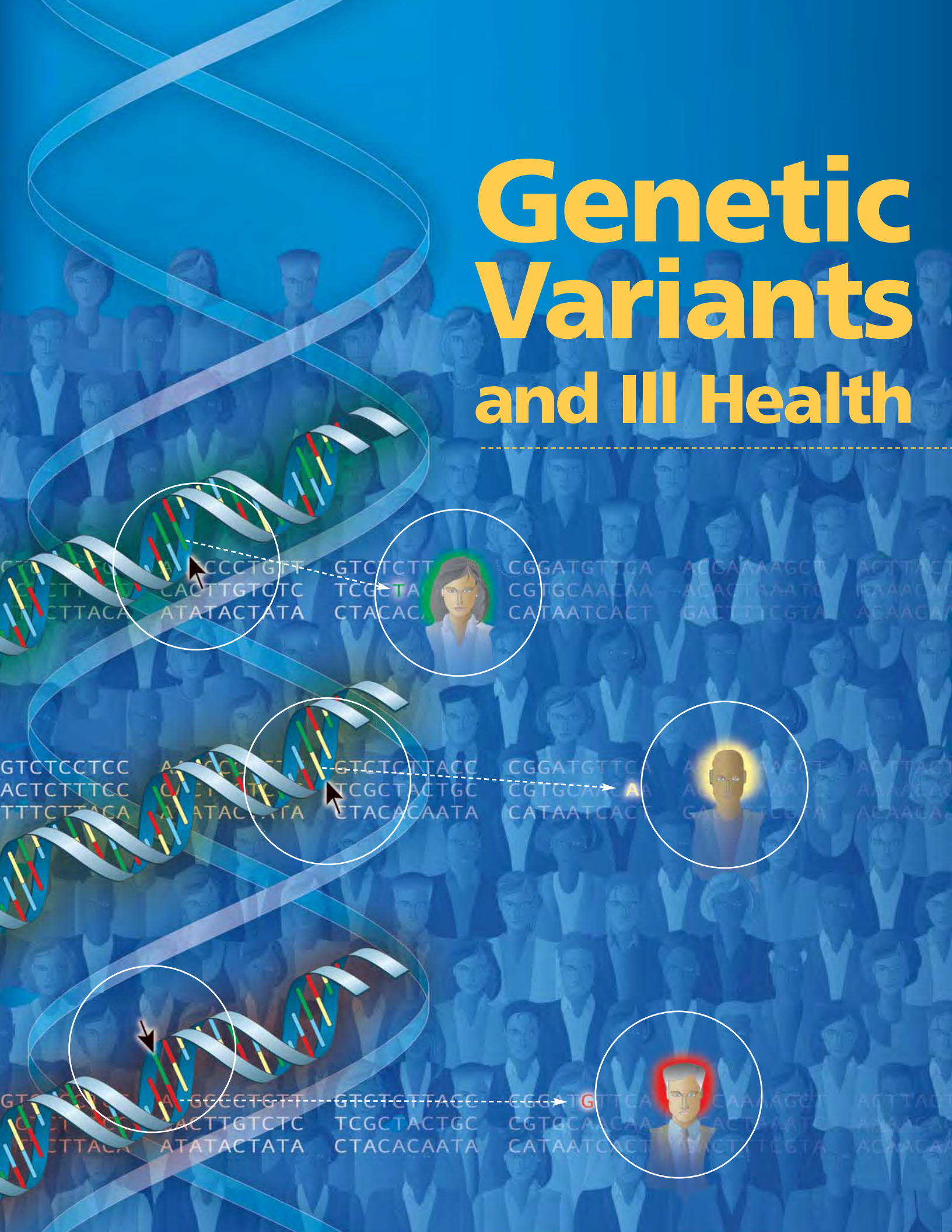
But real stumbling blocks to success in Alzheimer's disease research may lurk from sources beyond control. "It won't be from a lack of smart people, a lack of insights, a lack of new and useful things to do," says Jack. "The number one problem will be money."

Nevertheless, the field is a trendsetter of sorts, bringing together an unprecedented diversity of disciplines, data and people. "Alzheimer's disease is one of the gold standards of this research trend," says Arthur Toga of University of California Los Angeles. "It's motivated lots of people to try to do science in this way. That's very exciting." □



Since the early 1980s, researchers have used PET scans to help distinguish normal and Alzheimer's diseased brains, and work is ongoing to develop PET-based biomarkers for early-stage diagnosis. Here, two PET scans show the difference between a brain with advanced-stage Alzheimer's (right) and a normal brain (left). Courtesy of the National Institute on Aging, www.nia.nih.gov.

Genetic Variants and Ill Health



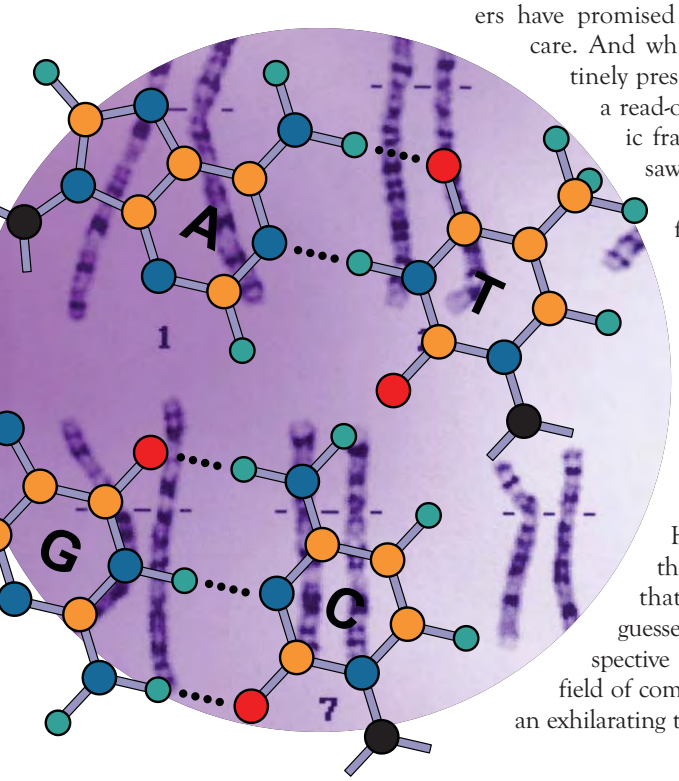
scanning 500,000 SNPs yields gene-disease connections

BY HANNAH HICKEY

For the past few months it seemed you couldn't open a journal without reading results of a new genome-wide association study. The results kept pouring in: four studies in April showing seven genetic links to Type 2 diabetes; in May, a paper in *Nature* showing six new links to breast cancer; and in June, a much anticipated study from the UK-based Wellcome Trust announcing genetic links to seven common diseases, from arthritis to schizophrenia. The list goes on and on.

This was unquestionably the year of the genome-wide association study—research that seeks connections between traits and common mutations in the human genome. The work relies on information gleaned from the Human Genome Project and its successors. Cheaper and more powerful sequencing technologies available in the last two years let researchers scan 500,000 genetic markers on a single chip. This technological muscle has ended a long wait in the use of genetics to study common, perplexing diseases.

"There's been a sea change in this type of work," said **Eric Topol, MD**, director of the Scripps Translational Science Institute in La Jolla and lead author of a commentary in *JAMA* on genome-wide association studies.[1] "[The research] is going at a breakneck pace. This week was HIV. Last week was Type 1 diabetes. It's an extraordinary chain of discoveries."



Is the Human Genome Project finally changing medicine? For years, scientists and policymakers have promised the dawn of personalized care. And while doctors do not yet routinely prescribe medications based on a read-out of an individual's genetic frailties, the summer of 2007 saw a great leap forward.

"After many years and a fair amount of frustration ...we have, in the last few months, about 50 discoveries of genetic risk factors for common diseases," said **Francis Collins, MD, PhD**, director of the National Human Genome Research Institute at the National Institutes of Health in late July. "Most of those point us toward targets that nobody would have guessed," he added. "From the perspective of people working in the field of common disease genetics, this is an exhilarating time."

Genes and Candidate Genes: A Slow Start

Medical genetics has spent a couple of decades in the doldrums. Early geneticists assembled family histories to study how specific traits were passed on. They then honed in on the genetic target. This method discovered genetic bases for such diseases as Huntington's and cystic fibrosis, as well as for about 2,000 other inherited traits. Though these were important discoveries at the time, today, some geneticists refer to them as the "low-hanging fruit"—easy targets involving a single gene producing a single, generally rare and deleterious, trait.

But many traits are more complex, involving multiple genes as well as the environment. And some aren't classically genetic—a person can have the risk allele and not develop the condition, while people lacking the risk allele do get sick.

The Human Genome Project seemed to offer a promising way to get at these more complex conditions through an approach known as candidate-gene studies. Researchers used their knowledge of a disease to study likely suspects in a particular pathway—insulin-production in diabetes patients, or cholesterol production in patients with heart disease. Despite a few successes—for example, a single-nucleotide variant that explains a large part of why people react differently to the blood-thinning drug, warfarin—progress was middling.

"The disadvantage of candidate-gene studies," said **Mary Relling, PharmD**, the chair of pharmaceutical sciences at St. Jude Children's Research Hospital in Memphis, Tenn., "is you will only find what you're looking for."

And for many common diseases, we don't yet know where in the genome to look.

The Genome-Wide Strategy

Researchers hoping for better results began taking a different approach. They look genome-wide in hopes of finding associations between genetic variations and a particular disease or drug response. Such studies are sometimes described as "agnostic" or "hypothesis-generating." One could describe them as "brute force" or even "shot in the dark" methods—applying purely com-

putational and fast-sequencing technology to the problem without any preconception about where in the genome relevant alleles will be found.

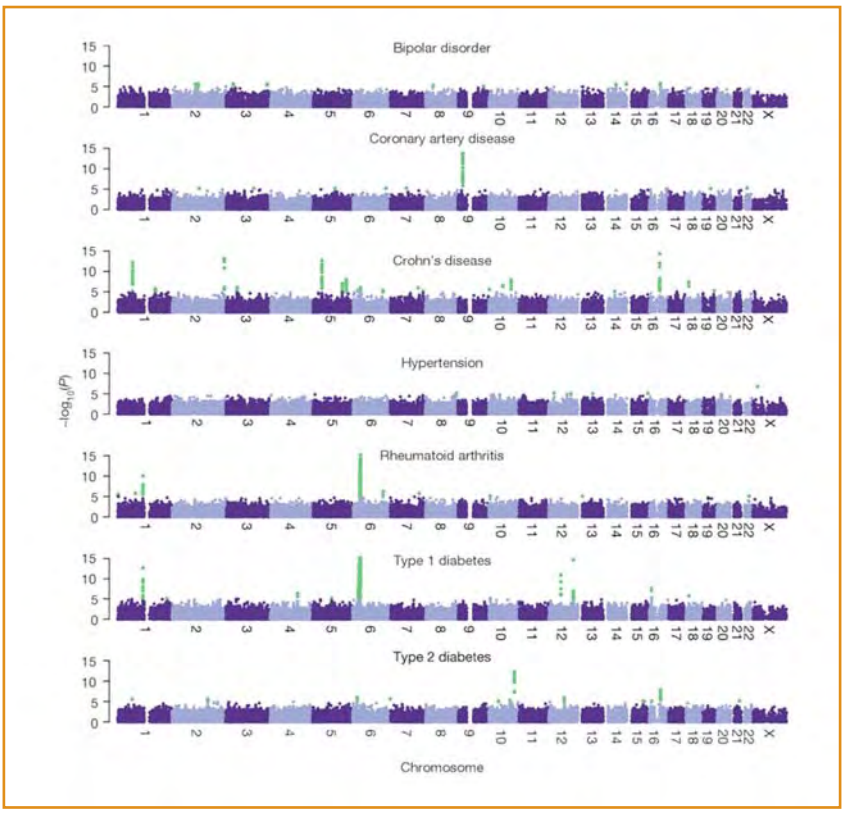
Biologists begin by assuming no knowledge of how the disease works. They assemble a group of people, some who have the disease and some who don't. And they scan about 500,000 of each person's single-nucleotide polymorphisms, or SNPs (said "snips")—single-letter blips in the genetic code. In this sense, the term "genome-wide" is misleading. Scientists don't yet sequence individuals' whole genomes. Instead, they sequence a few hundred thousand of the most changeable of the 3 billion letters in our genetic code. Those most changeable locations were identified by the HapMap project, which mapped locations where a rare allele crops up in at least 1 percent of people.

The straightforward SNP approach won't find duplications or changes in the structure of DNA, such as doubled chromosomes. And it won't detect addition or deletion of DNA units or flipping of large segments of DNA within a chromosome. It also won't find truly rare mutations that vary in less than 1 percent of the population, or the somatic errors that occur in cancer cells. What it will find are fairly common blips in the genetic code—blips that constitute about 90 percent of human genetic variation.

Early Successes: Macular Degeneration & Crohn's Disease

Before this summer, genome-wide studies had a trickle of early triumphs. Three independent groups reported in 2005 that a single point mutation increases a person's risk of developing age-related macular degeneration, the most common form of age-related vision impairment, by as much as seven times. The Yale University group used some 100,000 SNPs, a fairly small number by today's standards, in a tiny cohort of 96 cases and 50 controls.[2] That discovery has inspired ongoing development of drugs based on the complement factors H and B pinpointed by the genes.

Last fall, another group used a genome-wide scan to discover a novel genetic link to Crohn's disease, a common inflammatory bowel condition.[3]



In June, The Wellcome Trust Case Control Consortium in the UK published the biggest genome-wide study to date. It found genetic associations with seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). The study recruited 2,000 subjects for each of the phenotypes and 3,000 common controls, all of European ancestry. The graph shows correlations for each of the seven diseases. Chromosomes are shown in alternating shades of blue. Green indicates a P value of less than 10-5, meaning the association between the phenotype and this position is unlikely to have happened by chance. And although the scale is truncated at 10-15, a few markers – for example, the position on chromosome 6 that is linked to both Type 1 diabetes and Rheumatoid arthritis – exceed this significance threshold. Image courtesy of Peter Donnelly. Reprinted with permission from Macmillan Publishers Ltd: **Nature*, 447, 661-678, 2007.***

Strength in Numbers: Databases

Prompted by the need for ever larger studies, medical geneticists are learning to play together.

The Wellcome Trust Case Control Consortium (WTCCC) brought together over 50 research groups from the United Kingdom to carry out its recent genome-wide study, the largest yet. The group makes its data available on it's Web site. The study authors write that larger sample sizes—in their case, 2,000 for each disease and 3,000 shared controls, all of European ancestry—greatly increased the number of statistically significant associations they were able to find.

The **database of Genotype and Phenotype (dbGaP)** dbGaP was launched in 2006 by the National Center for Biotechnology Information and is housed at the NIH. It will receive data from NSF-sponsored studies, which strongly encourage researchers to deposit data to a public source. dbGaP, which also encompasses the Genetic Association Information Network (GAIN), is a public-private partnership between the Foundation for the National Institutes of Health and Pfizer's research branch. The national institute offers researchers a carrot: if they deposit phenotype data for clinical studies underway or already conducted, they will sequence the study participants' DNA.

For the genetics of drug response, **PharmGKB** was established in 2000 as part of an NIH-sponsored pharmacogenomics research network. The network offers data to investigators outside the network, and it actively recruits from any relevant studies. Persuading researchers to contribute data has become much easier in the past seven years, said **Teri Klein, PhD**, PharmGKB's director at Stanford University. PharmGKB curators also actively recruit data by scanning publications and contacting authors to submit.

Ideally, all these databases will support one another. PharmGKB will link to microarray data housed on the NCBI's Gene Expression Omnibus database, merely noting what data is available. The organizers have developed a similar relationship with dbGaP, Klein said, while PharmGKB will focus specifically on drug response.

Even more specialized databases are cropping up. The Bipolar Disorder Phenome Database, launched in July as part of Johns Hopkins Psychiatry's "BioinforMOODics" site, offers detailed symptom descriptions and complete SNP profiles for more than 5,000 people with bipolar disorder.

Hurdles remain. Scientists don't want yet another hoop to jump through when publishing results. And many are reluctant to share their data before results are published. From the study participants' side, issues of informed consent and access to the data are under discussion. But more data sharing may be inevitable: new evidence suggests that the bigger the study, the more possible associations will be found.

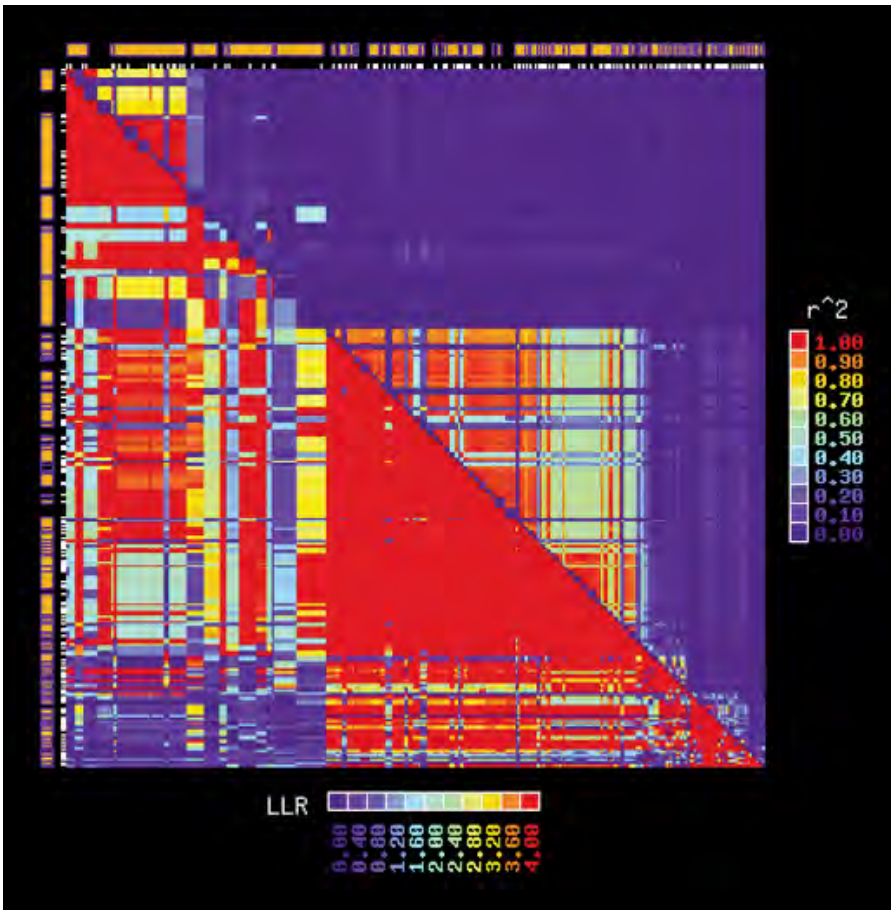
The study of 300,000 SNPs in about 550 cases and controls honed in on three statistically significant links. Two were known genetic markers for Crohn's disease. The third, IL-23R, was brand new. It has since been widely reproduced and prompted new therapeutic research.

These are the good-news tales in taking a statistical association to the realm of medicine.

Early Disappointments: The Replication Problem

Unfortunately, the last couple of years also saw many genetic associations that turned out to be embarrassing dead ends. Epidemiologists had faced a similar problem in the 1990s, when some worried that too many false discoveries threatened the field's credibility.[4] In 2005, a large genomic study for the time found 13 associations with Parkinson's, the degenerative muscle disease.[5] Follow-up studies the subsequent year didn't support a single one of the links.[6] A much-touted link to longevity for the gene MTP was, ironically, short-lived—the original study made headlines in 2003 but had been largely discredited within two years. Associations with obesity— GAD, ENPP1 and, most recently, INSIG2—failed to produce convincing follow-ups.

"What we see in the media, or we have seen in the past ten years, is every now and then something will come up: 'Oh, there's a new gene for



A link to a single SNP is only the first step. That position might then be linked to the gene that's responsible. Many SNPs travel in packs—the 10 million SNPs in the human genome typically get passed on in clusters, and geneticists have whittled down the original number to a representative 500,000. After identifying an association with a SNP, biologists look at the related positions to see if a nearby mutation is the one that's actually responsible for disease. If two genetic positions are linked they get inherited together more often, and this is called "linkage disequilibrium." The figure shows linkage disequilibrium across an interval of the human genome for 173 SNPs in an interval of interest for autism. The correlation between markers is plotted as a heat map, with red marking the highest correlations and blue marking the lowest. On the upper right, r2, a measure of correlation, is plotted against the log likelihood ratio, a measure of the significance of the observed correlation. This plot was generated using the HaploBlockFinder software. Courtesy of J.L. Stone and S.F. Nelson at the University of California, Los Angeles.

"From the perspective of people working in the field of common disease genetics, this is an exhilarating time," says Francis Collins.

obesity,' or 'There's something new for cancer,'" said **Lon Cardon, PhD**, professor of bioinformatics at the University of Oxford and at the Fred Hutchinson Cancer Research Center in Seattle. "By and large, those things turned out not to be reliable."

Part of the problem was statistics. With hundreds of thousands of genetic suspects, the risks of finding one that looks guilty are high. Complex diseases are thought to involve tens or hundreds of associations, each exerting a small effect. Distinguishing a real effect against background noise, or against a genotyping error, was difficult. The p-values for these publications were a far cry from the typical values.

"People were optimistically taking exciting nominal p-values, 10^{-4}, 10^{-5}, and trying to construct interesting functional hypotheses," said **Mark Daly, PhD**, a population geneticist at the Broad Institute in Cambridge, MA. "What has changed is now people appreciate the statistics and that you will get some of those results by chance." To establish a biological link, he said, geneticists now replicate the association in another population.

A recent document by the National Institutes of Health, "Replicating genotype-phenotype associations," emphasizes replication and sets some standards: that the replication study should look at the same location on the chromosome, not a nearby position; that the replication study should look at the same or very similar phenotype; and that a similar population should be studied in the follow-up.[7] The report also includes guidelines for reviewing genome-wide association studies.

This being said, a negative result isn't always the last word. If the association was specific to one population, it might be that the second population didn't have that genetic risk, or that they were in an environment that didn't trigger the genetic expression.

This year's genetic results meet the higher standards for replication, Cardon and others insist. They have already been verified in another, usually bigger, population. For instance, a link between the FTO gene and obesity was first discovered by members of the Wellcome Trust in a study among 2,000 subjects and 3,000 controls. Then the researchers looked at that one position in relation to the physique of some 39,000 other people before announcing they'd found a reliable genetic link.

The Successes of 2007—And Their Limits

The phenotypes of most interest today—obesity, diabetes, heart disease, cancer—are complex traits that likely have many different causes. During the last few months, genome-wide association studies have produced results that offer hope of finding out why we inherit a risk for such traits.

For example, in May, four separate groups reported seven new associations with Type 2 diabetes, bringing the total number of associations to 10. But the new associations are hardly slam-dunks. Taken together, they explain between 2 and 20 percent of a person's risk of developing diabetes. The number of associations is expected to grow with subsequent studies. If the risks are additive, then as we discover more genes we will explain a larger percentage of susceptibility.

Drug response seems likely to be a similar story. Researchers studying warfarin had a big hit early on using the candidate-gene method. A single change in a vitamin K receptor gene, VKORC1, predicts almost a quarter of the patient's response to the common blood-thinning drug.[8] But follow-up studies, now using genome-wide associations, have found additional genes that explain fractions of the observed response: 10 percent, or 5 percent, or even smaller additional risk.

"If genetic factors can explain 5 percent of a phenotype, that's considered a big deal," said **Mark Rieder, PhD**, a geneticist at the University of Washington in Seattle and, lead author of the original warfarin study.

But from a patient's perspective, what does it mean to have a 5-percent increase in risk of getting a disease? And for a physician, what does it mean to have a 5-percent higher chance that a medication will cause side effects? Even a 50 percent increase in risk might not mean as much as it appears. If you initially had a risk of 3 in 1,000

Summer 2007 Results

DISEASE	STUDIES	ASSOCIATIONS	REFERENCES
Bipolar Disorder	1	1	Nature 447, 661-678, 2007.
Breast Cancer	3	8	See review article by Topol, E. J. et al. JAMA 2007;298:218-221
Celiac Disease	1	1	Nature Genetics 10 June 2007
Colorectal Cancer	1	1	Nature Genetics 39, 984-988 (01 Aug 2007) Letters; Nature Genetics 39, 989 - 994 (01 Aug 2007) Letters
Coronary Artery Disease	1	1	Nature 447, 661-678, 2007.
Crohn's Disease	1	9	Nature 447, 661-678, 2007.
Glaucoma	1	1	Science 31 January 1997 275: 668-670
HIV vulnerability	1	3	ScienceXpress 19 July 2007
Human Gallstone Disease	1	1	Nature Genetics 39, 995 - 999 (01 Aug 2007) Letters
Multiple Sclerosis	2	3	Eur J Hum Genet. 2007 Jun;15(6):703-10; New England Journal of Medicine 29 July 2007
Myocardial Infarction	1	1	See review article by Topol, E. J. et al. JAMA 2007;298:218-221
Obesity	1	1	See review article by Topol, E. J. et al. JAMA 2007;298:218-221
Prostate Cancer	2	2	Nature Genetics 39, 977 - 983 (01 Aug 2007) Letters;
Restless Leg Syndrome	2	3	Nature Genetics 39, 1000 - 1006 (01 Aug 2007) Letters; New England Journal Of Medicine online July 18, 2007
Rheumatoid arthritis	1	3	Nature 447, 661-678, 2007.
Type 1 Diabetes	1	7	Nature 447, 661-678, 2007.
Type 2 Diabetes	5	12	Nature 447, 661-678 (2007); See review article by Topol, E. J. et al. JAMA 2007;298:218-221

This summer saw an outpouring of results from genome-wide association studies. This table lists a partial summary of the results, including some that were previously listed in a commentary published in the July 11 issue of* JAMA. *Associations were found this summer for phenotypes that ran the gamut, from asthma to schizophrenia. Many of the results were announced simultaneously by more than one research group.

cases, then a person who carries the gene for a 50 percent increased risk has a 0.45 percent chance of developing the symptom, instead of 0.3 percent. Will a person stop eating burgers if he or she has the gene that increases risk of heart disease by 0.15 percent?

Because it's tough to answer these questions, screening tests for common diseases remain rare. But there are a few exceptions: Roche Diagnostics and deCODE Genetics, an Icelandic company, announced in July that they would offer screening for single mutations associated with increased risk of schizophrenia, Type 2 Diabetes and stroke. And in August, deCODE's chief executive, said the company will also use recent results for glaucoma to design a screen to improve diagnosis of the degenerative eye condition.

Still, many scientists caution against using the newly identified genes for screening whether a person is slightly more or less likely to get sick.

"It doesn't matter how much [the associations] increase risk," said Cardon. "It's the fact that they're a brand-new pathway about the biology of the disease."

This year's genetic results meet the higher standards for replication, Cardon and others insist. They have already been verified in another, usually bigger, population.

A better understanding of the disease may provide clues to new treatments.

A Mechanistic View: It's Not About Screening

Most researchers pursue the so-called "agnostic" studies to discover new genetic associations that could help to understand the disease mechanism and, ultimately, help to design drugs and genetic therapies.

A few of the discoveries, such as the August link to a major form of glaucoma, implicate a known protein and suggest a biological basis. But the biological basis of many new associations remains a mystery. Two groups reported in May that a single genetic variant on chromosome nine increases risk of heart disease by up to 60 percent. The variant is common, with a fifth of the European population having two copies. But it lies outside coding regions in an area of the genome with no known function. Intriguingly, the mutation lies close to the one of the Type 2 diabetes variants reported a week earlier.

In fact, most of the associations with Type 2 diabetes are located in "gene deserts," that have no known regulatory or coding function. None were in the locations for insulin resistance suspected of being linked to diabetes. Many of the new breast-cancer associations are similarly thousands of base pairs away from coding genes.

Collins says many geneticists are not surprised that the common diseases are influenced by variants outside coding regions. It makes sense that the contributions would be subtle. For example, rather than producing a different protein, the risk variants are being found in regulatory regions that might change the magnitude of a gene's activity.

"I think we all expected some of [the associations] to be non-coding," Collins said. "It's surprising just how many that applies to."

A recent study took a first step toward providing a mechanism. An international team of researchers led by **Bill Cookson, PhD**, at Imperial College London reported in July that genetic markers on chromosome 17 increase a child's risk of asthma by about two thirds. The researchers then recorded gene expression. The children with the variant had more of a different gene, called ORMDL3, in their blood. Their results imply that the genetic variant somehow causes more transcription of the ORMDL3 protein, and that may provide the pathway for the disease.

If genome-wide associations lead to advances in our understanding of biology, that's when they will really matter. As one researcher said: "If you have an association, you publish in *PLoS Medicine*. But if you have a mechanism, you publish in *Science* or *Nature*."

Environmental and Racial Issues

The mechanism for complex diseases may be especially difficult to tease out because many of these traits interact with the environment.

"We're in a stage where we have a lot of progress, but it's not yet clear with how well we've come to grips with the genetic basis of traits that are complex," said **Ken Weiss, PhD**, a geneticist at Penn State University. He points to complex phenotypes such as heart disease, obesity and breast cancer, where in a single generation the incidence of disease has skyrocketed—without any change in the population's genes. "Clearly, the population's gene pool hasn't changed in the last 30, 40, 50 years. But the disease risk has changed dramatically. Well, that has to be attributed to what we would call environmental factors, even if we don't know what the factors are."

Mysterious environmental triggers are yet another reason why scientists advise waiting before predicting risk. It may be that some populations live in an environment that sets off the genetic trigger. Other people, with identical genes, might react differently.

Just as touchy an issue as environment is race. Researchers in the U.K. have been careful to choose ethnically homogeneous samples. Otherwise, they risk finding differences between racial groups. And although the HapMap was created by comparing different ethnic groups, the resulting map describes European variation more thoroughly than the vast variety of genes that exist among peoples from sub-Saharan Africa.

Fuzzy Phenotypes: Taking The Measure of Disease

SNP discovery offers the computational biologist an obligingly quantitative trait. Each base-pair offers four possibilities that fit neatly into a ones-and-zeroes database. But researchers often struggle to fit phenotypes into an equally tidy box.

Sometimes it's easy. For example, pharmacogenomics associations are measured in a clinical setting. Computerized photographs of Petri dishes help assess lab cultures, and regular readouts of patient information measure drug response. One prime example is patients' reactions to warfarin. Because dosing is problematic, and mistakes can be fatal, delivery and patient monitoring happens in "a very tightly controlled environment with a narrow outcome, standardized across clinics. So that's a good phenotype," said **Mark Rieder, PhD**, a geneticist at the University of Washington.

But as genome-wide association studies set themselves more and more targets—autism, schizophrenia, obesity, asthma—the challenge becomes quantifying those traits in a meaningful way. Is anybody who's wheezed and self-identified as asthmatic an asthma patient? Does someone whose weight changes from one year to the next qualify as obese? Many of these traits may have a genetic component, but they also have decades of life experience.

Studies of irritable bowel syndrome carefully parsed out subjects and chose only those where symptoms were most similar. Genetic screens for bipolar disorder selected those patients for whom the illness began at an unusually young age, or those who also experienced panic attacks. Some researchers studying schizophrenia measure subjects' startle response as a quantitative proxy for the condition. Clever study design solved problems for Type 2 diabetes, which would seem to be rife with such problems. Many doctors would consider the condition to be a number of different diseases that manifest a similar set of symptoms.

"I think the most challenging part is still the fuzziness of the phenotypes," said **Michelle Carrillo, PhD**, a curator at the PharmGKB database, "because ambiguity makes it hard to aggregate [data from different studies] and that's what most investigators want to do." A group in Florida and a group in San Diego might both study hypertension and have genotyped their subjects. The genotyping occurs at particular positions and can be compared. But the phenotypes are harder because the tests might have been run differently in Florida and in San Diego.

The National Institutes of Health encourages researchers to describe methods and measurements in as much detail as possible so that subsequent studies can compare results. And PharmGKB is working on a phenotype ontology, so the vocabulary is standardized.

Pharmacogenomics researchers face an additional challenge: comparable studies must match up not only genotype and phenotype, but also drug dosage. To further research on warfarin, the blood-clotting drug, a new 21-institution consortium is working to develop standard dosing guidelines between members in the United States, United Kingdom, Israel, Japan, Korea and Brazil, said Klein. The Pharmacogenetics Research Network is now identifying other areas where a consortium would be beneficial, such as tamoxifen for breast cancer and statin drugs, Klein said.

"If genetic factors can explain 5 percent of a phenotype, that's considered a big deal," says Mark Rieder.

"Our tools are currently underpowered with respect to performing genome-wide association studies in populations of African ancestry," said **Malia Fullerton, PhD**, a bioethicist at the University of Washington School of Medicine in Seattle. "I think these differences are widely appreciated," she added. "But in the United States context, if we want to be true to our national demographics, and include a representative mixture of people in our studies, then we have to be paying closer attention to these issues."

The Future: Bigger and Better

The onslaught of data already raises new statistical questions, and those are only likely to increase. The newest chip from Affymetrix measures more than 1.8 million markers. And studies will also increase the number of people scanned for SNPs. Some complex diseases will likely involve rare mutations, generally classified as those that crop up in less than 5 percent of the population. Studying these rare mutations demands bigger and bigger sample sizes. If a variant exists in just 1 percent of a population, then a study of 100,000 people (far bigger than anything yet attempted) would count just 1,000 people with that variant.

Biologists already have more data than they know what to do with, asserts Weiss. And he foresees the day when genetic studies will expand to include hundreds of thousands of subjects and sequence more than a million markers for each person—or, likely soon, all 3 billion base pairs in the human genome.

This has already begun. The UK Biobank study this year began recruiting half a million volunteers of European ancestry aged between 40 and 69 for a long-range study that will look a links between genes, health and environment.[9] It will take blood samples and keep DNA for further testing, while carefully tracking subjects' health and environment. Similar projects have been discussed in other countries, including the United States. The current NIH budget does not permit such an undertaking here, Collins says. "I'm deeply concerned about that, because I think we're going to kick ourselves five or six years from now."

The recent successes will prompt more investigation into treatment, searches for new associations and a better understanding of what the existing associations mean. As nearly every aspect of our selves comes under study–dyslexia, autism, schizophrenia, obesity–the amount of data will grow.

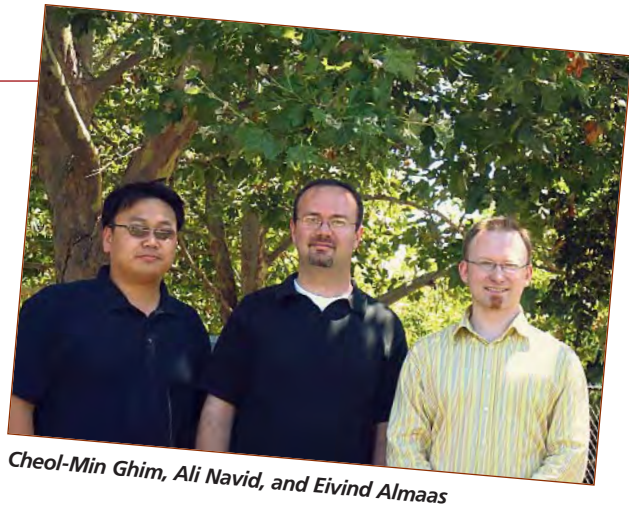
"It's not yet clear, I think, how much more information versus more confusion this huge amount of new data is going to cause," Weiss said. "We'll have to wait and see what people will attempt." □

1. The Genomics Gold Rush. Eric J. Topol, MD; Sarah S. Murray, PhD; Kelly A. Frazer, PhD JAMA. 2007;298:218-221.
2. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
3. Duerr, RH, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science (2006) 314: 1461-3.
4. Epidemiology faces its limits. Taubes, Gary; Mann, Charles C. Science; Jul 14, 1995; 269, 164-169.
5. Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant KPV, Frazer KA, Cox DR, Dennis G, Ballinger DG (2005) High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685–693.
6. D.R. Myers, Considerations for genomewide association studies in Parkinson disease, Am J Hum Genet 78 (2006), pp. 1081–1082.
7. Replicating genotype-phenotype associations. NCI-NHGRI Working Group on Replication in Association Studies et al. (2007)Nature 447(7145):655-60.
8. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, Blough DK, Thummel KE, Veenstra DL, and Rettie AE. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose New Engl J Med (2005) 352: 2285-93.
9. UK Biobank: bank on it. Palmer LJ (2007) Lancet 369(9578):1980-2.

Under TheHood

BY CHEOL-MIN GHIM, PhD, ALI NAVID, PhD, AND EIVIND ALMAAS, PhD

Scale-Free Networks in Contemporary Biology



Cheol-Min Ghim, Ali Navid, and Eivind Almaas

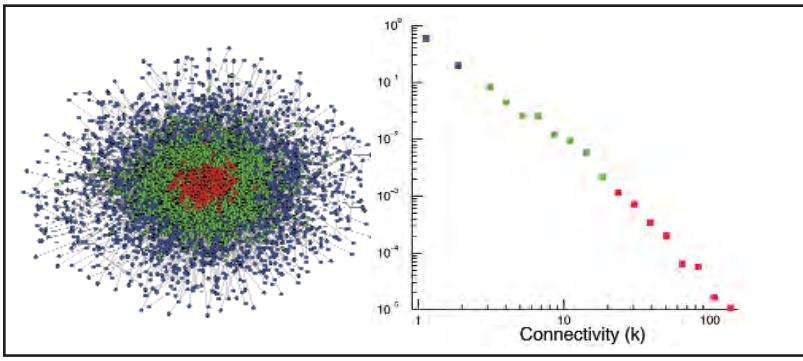
A standard dictionary definition of a network is "an interconnected or interrelated chain, group, or system." A cursory look at our surroundings shows that networks are ubiquitous. For instance, we can describe a single-celled organism as a highly complex web of multi-scale networks, ranging from the gene-regulatory system, to protein interactions, to metabolism. A chief challenge in modern biology is to develop a system-wide understanding of cellular function on the basis of genomic information. Network representations of complex biological systems have successfully served this purpose. Key characteristics of life, such as adaptation or robustness, can be translated into the interplay between network topology and dynamics. Recent discoveries of strong similarities in the architectural features of complex networks spanning the social, technological, and life sciences, have opened a new horizon in our understanding of the principles that govern and shape biological systems.[1]

Theoretical and experimental results have established that most biological systems are "scale-free," meaning that their connectivity distributions can be approximately described by a power-law function (see Figure). This "systems" feature has long been recognized in the context of economics (Pareto's law) and linguistics (Zipf's law). In biological networks such as signal transduction, transcriptional regulation and metabolism, the majority of the nodes (being either genes, transcription factors or metabolites) have only a few network connections, while a few nodes, the "hubs," have hundreds or even thousands of connections. We may interpret these hubs as ubiquitous metabolites, promiscuous enzymes, or versatile transcription factors.

A simple drawing of the wiring diagram demonstrates that biological networks are so compact that on average, any pair of molecules and chemical reaction events is connected in just a few steps. While scale-free networks are more vulnerable to intentional attacks targeting the hubs, they show a remarkable resilience to random failures, which are a common feature of evolutionary events. These discoveries may lead to an entirely new approach to pharmaceutical target identification, where the chance for development of drug resistance is significantly reduced.

Another significant feature of biological networks is their modularity: Similar to a computer program, it is possible to identify biological "subroutines" with a clearly defined function. However, biological functional units are highly interwoven on all levels, creating a hierarchical structure, suggesting that communication between the highly connected network regions is maintained by the hubs. This structural organization makes it possible for the systems to evolve using random mutations without disrupting the integrity of the whole system.

Trying to answer the old question of how living organisms achieve a robust state of being, the benefits of network approaches to biology become clear. Teasing apart the driving forces that have shaped these networks not only sheds light on evolutionary paths, but also provides guiding principles for current efforts in synthetic biology that aim to construct robustly operating biomimetic systems. □



The protein-interaction network (PIN) of Homo sapiens (from TheBiogrid 2.0.29), where proteins correspond to nodes, and a link indicates that two proteins physically bind. The low connectivity nodes (blue) coexist with nodes of intermediate (green) connectivity and hubs (red) with high connectivity. The connectivity distribution is scale-free.

REFERENCES:

1. A.-L. Barabási and Z. N. Oltvai, Network Biology: Understanding the Cells's Functional Organization. Nat. Rev. Genet. 5, 101 (2004).

DETAILS

Cheol-Min Ghim, PhD, Ali Navid, PhD and Eivind Almaas, PhD, are scientists in the Biosciences & Biotechnology Division at Lawrence Livermore National Laboratory. Their research is focused on understanding the function and robustness of microbial systems.

SimbiosNews

BY KATHARINE MILLER

A Giant Leap for Open Source Simulation

Researchers can now create musculoskeletal models and simulations on an open source platform. In August, Simbios researchers released OpenSim 1.0. This freely available software can, in about 20 minutes, create 3D simulations of human movement at a level of detail and accuracy that just a few years ago would have taken weeks or months.

The software should benefit the entire field of computational biomechanics, says **Scott Delp, PhD**, professor of bioengineering and mechanical engineering at Stanford and co-PI of Simbios. "Until now, it has been difficult to reproduce the results of simulation papers. With OpenSim, we hope to promote continuity across the field."

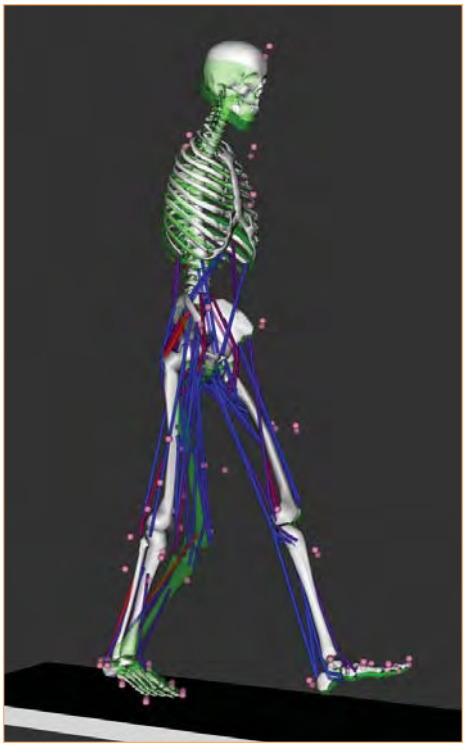
Musculoskeletal simulations allow researchers to visualize complex movements; estimate forces that are difficult to model; perform "what if" scenarios; and look for cause and effect relationships. Ultimately, they can guide doctors to plan appropriate surgeries and physical therapy regimens.

Until now, biomechanics researchers doing dynamic simulations have either purchased commercial software or invented their own. "With models being developed in the same platform, it will be possible to exchange them between labs," says **Kurt Manal, PhD**, director of the Center for Biomedical Engineering Research at the University of Delaware, who attended OpenSim training in August. "That's not possible when each lab is using its own application."

OpenSim has an intuitive, windows-based graphical-user interface. Users can modify a large variety of physical parameters at a click of the mouse; overlay models on one another; and easily animate and make movies of their simulations.

At a recent OpenSim training course, Delp and his colleague **Clay Anderson, PhD**, led fifty eager attendees in a twenty-minute simulation of a person walking with and without a weakened soleus muscle. When the two animations were overlaid on one another, the difference in gait was obvious: there was a distinct knee drop for the individual with the weakened soleus.

Manal, who attended the training, says the tutorial worked perfectly. He and his students will use OpenSim right away. And when graduate students move on to post-doc or academic positions, they will be able to continue their work without interruption.



This snapshot shows two simulations overlaid on one another. The model in green has a weakened soleus muscle that leads to an exaggerated dipping of the knee. Courtesy of Clay Anderson, PhD.

Rick Neptune, PhD, an associate professor of mechanical engineering at the University of Texas, says OpenSim is computationally efficient and does many things really well. "It will be a valuable tool for a lot of research labs."

Over the next few years, OpenSim will continue to evolve and grow. "It's our goal to seed the community with this," says Delp. "But we need the community's help to improve it, develop plug-ins, models and simulations and distribute them to the biomechanics community." In a few years, if people have produced open source models for individuals with stroke and Parkinson's disease, models of the wrist, the foot, the upper extremities, walk/run transitions, cockroaches, T. rex, etc., "Then we've really got something," he says. □

DETAILS

Simbios released OpenSim 1.0 in August 2007. It can be downloaded at https://simtk.org/home/opensim.

Simbios is a National Center for Biomedical Computing located at Stanford University.

PuttingHeadsTogether

MICCAI 2007, the 10th International Conference on Medical Image Computing and Computer Assisted Intervention.

WHAT: MICCAI typically attracts over 600 world leading scientists, engineers and clinicians from a wide range of disciplines associated with medical imaging and computer assisted surgery.

WHEN: October 29 to November 2, 2007

WHERE: Brisbane, Australia

MORE INFO: *http://www.miccai2007.org/*

BIBM 2007—The IEEE International Conference on Bioinformatics and Biomedicine

WHAT: IEEE BIBM 2007 will provide a general forum for disseminating the latest research in bioinformatics and biomedicine. This multidisciplinary conference brings together academic and industrial scientists from computer science, biology, chemistry, medicine, mathematics and statistics. It will exchange research results and address open issues in all aspects of bioinformatics and biomedicine and provide a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, simulation, ontology and other computational methods, as applied to life science problems, with emphasis on applications in high throughput data-rich areas in biology, biomedical engineering. IEEE BIBM 2007 intends to attract a balanced combination of computer scientists, biologists, biomedical engineers, chemists, data analyzers, and statisticians.

WHEN: November 2 to November 5, 2007

WHERE: Fremont, California

MORE INFO:
http://www.cis.drexel.edu/faculty/thu/bibm/index.php.htm

2nd International Biocuration Meeting

WHAT: This meeting will provide a forum for curators and developers of biological databases to discuss their work, promote collaboration, and foster a sense of community in this very active and growing area of research. Participants from academia, government, and industry interested in the methods and tools employed in curation of biological data are encouraged to attend. Talks will be selected from poster submissions.

WHEN: October 25 to 28, 2007

WHERE: San Jose, California

MORE INFO:
http://tesuque.stanford.edu/biocurator.org/Mtg2007

Search and Knowledge Building for Biological Datasets

WHAT: New biotechnologies and the accumulation of vast amounts of biological data have created a fertile ground for quantitative scientists. Issues of appropriate strategies for search, what to search for, and how to turn massive quantities of biological data into useful knowledge have moved to the forefront. Contributions from diverse areas such as combinatorics, graph and network theory, differential equations, machine learning, data mining, statistics and statistical physics have been used to create more powerful information search and knowledge management.

This workshop is intended as a convergence of quantitatively oriented researchers addressing these issues in their quest to answer important biological questions. It will provide an opportunity for researchers with quite different perspectives and interests to share their approaches with one another and cross-pollinate their ideas.

WHEN: November 26 to 30, 2007

WHERE: Institute for Pure and Applied Mathematics, University of California, Los Angeles

MORE INFO:
http://www.ipam.ucla.edu/programs/sews4/

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

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

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

seeing science

SeeingScience

BY KATHARINE MILLER

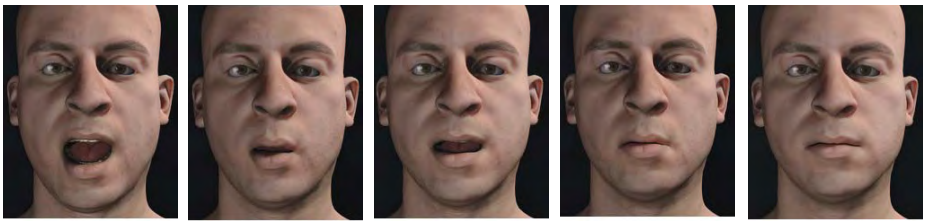
Talking Heads

He speaks: "Algorithm." And you can just about read his lips.

The movie was created using muscle-driven physics-based animation. Other techniques might produce images that look just as real, but they are much less versatile.

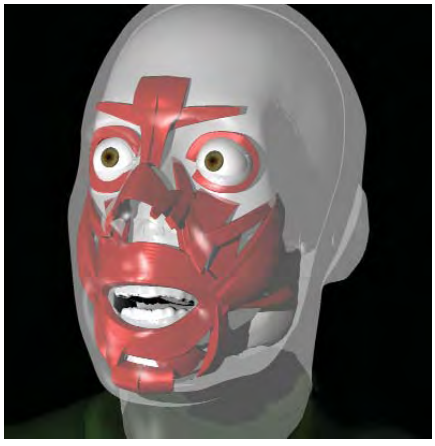
The animation starts with a highly detailed model of the head and neck that was created by **Eftychios Sifakis**, a PhD student, and his colleagues in **Ron Fedkiw's** lab at Stanford. They used data from the Visible Human project to create the model, and then morphed it to fit data obtained from both laser and MRI scans of a living subject.

To animate the model, the researchers estimated muscle activations, head position, and jaw articulation using motion captured performances of a living person—Sifakis himself. For ten minutes in front of eight cameras, Sifakis spoke a full range of phonemes with 250 markers attached to his face to get fully three-dimensional information. From this, the researchers constructed a phoneme database that described how muscles activated across a full range of possible facial movements for specific (and phonemically appropriate) periods of time. They then used these data to synthesize Sifakis' face speaking words that were not captured on film.



This sequence of images shows a synthesized utterance of the word "algorithm" using a physically-based facial muscle model. Courtesy of Ron Fedkiw's lab at Stanford.

The work was published in the *Proceedings of the Euro-graphics/ACM Siggraph Symposium* on Computer Animation in 2006. It could prove valuable not only for the entertainment industry but also for predicting the effect that facial surgery will have on expression. □



A cut-away image of the embedded muscle structure. Courtesy of Ron Fedkiw's lab at Stanford.