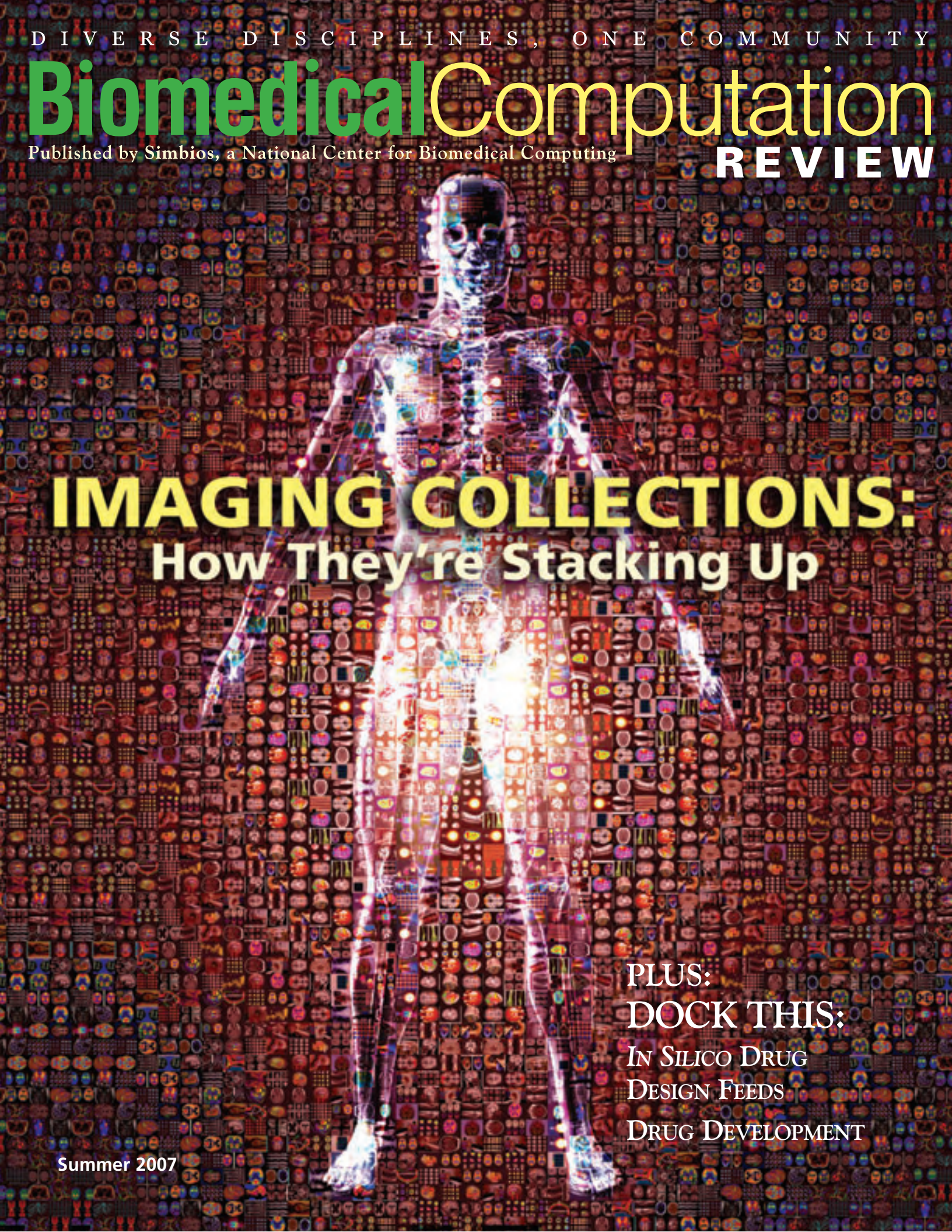
DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, a National Center for Biomedical Computing



IMAGING COLLECTIONS: How They're Stacking Up

PLUS:

DOCK THIS:
In Silico Drug Design Feeds Drug Development

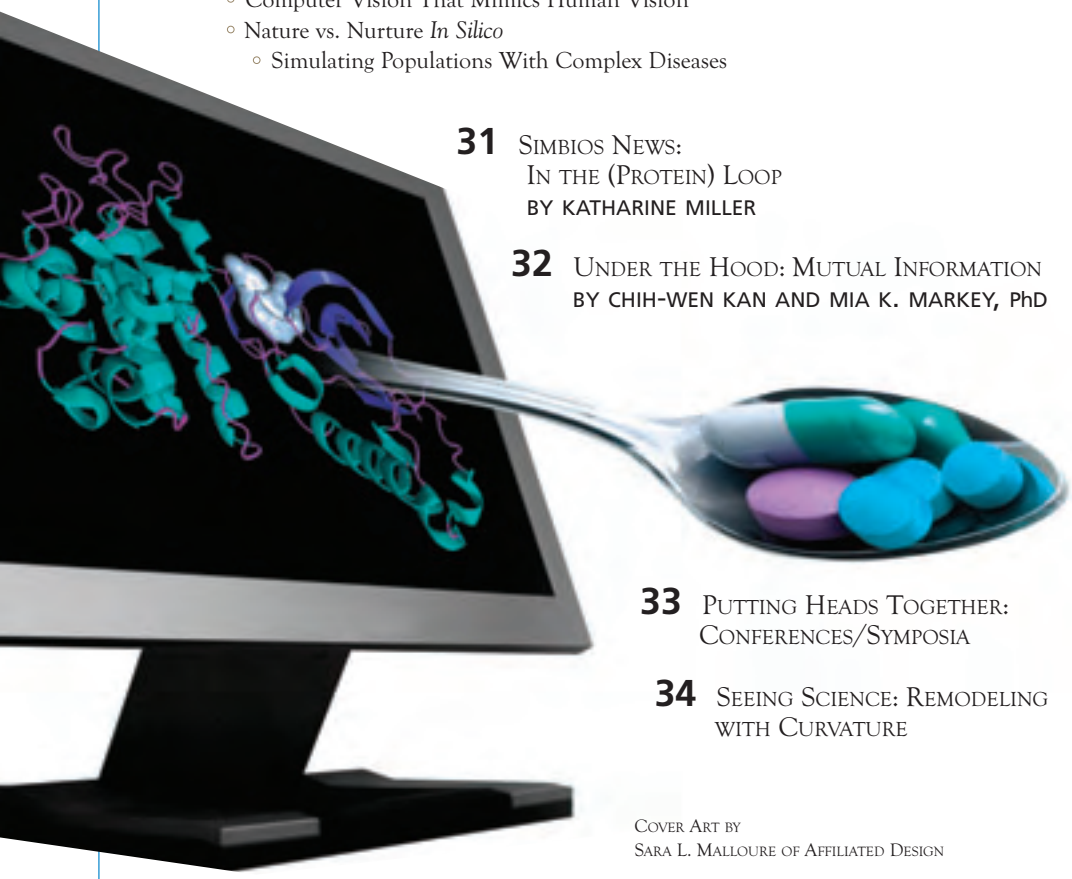
Summer 2007

contents

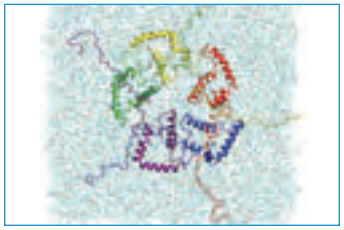
ContentsSummer 2007

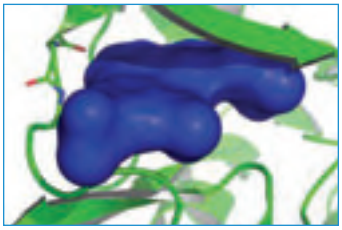
FEATURES

DEPARTMENTS



Cover Art by
Sara L. Malloure of Affiliated Design





Summer 2007
Volume 3, Issue 3
ISSN 1557-3192

Executive Editor
David Paik, PhD

Managing Editor
Katharine Miller

Science Writers
Katharine Miller
Louisa Dalton
Matthew Busse, PhD
Meredith Alexander Kunz
Kristin Cobb, PhD

Community Contributors
David Paik, PhD
Mia Markey, 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)

From theEditor

BY DAVID PAIK, PhD

The Active Transport of Ideas

How ideas spread gets at the very fabric of scholarly research and has been studied from many different angles.

Many studies examine person-to-person connectivity in social networks. Within a social network, the average path length between any two people is a key concept. By asking participants in Omaha or Wichita to mail chain letters that would get closer to selected recipients in Boston, Milgram's classic 1967 small world experiment demonstrated the six degrees of separation concept. Movie buffs have created a board game using this concept called the Six Degrees of Kevin Bacon and those interested in mathematical genealogy have adopted Erdös Numbers linking researchers by co-authorship to the prolific mathematician **Paul Erdös**.

However, a small world is not necessarily a robust world. In addition to path lengths, the connectedness between different parts of the social network is an important measure. A recent *Journal of the American Medical Informatics Association* paper by **Bradley Malin, PhD**, and **Kathleen Carley, PhD**, examines the connection between editorial boards of medical informatics and bioinformatics journals to describe the fragility of links between these two sister fields.

There are also many ways to examine the spread of ideas more broadly. The Rogers theory of diffusion of innovation states that depending on when they adopt new ideas, people form a bell curve as either innovators, early adopters, early majority, late majority or laggards and that the innovation penetration forms an S curve over time. The five stages are awareness of the innovation, persuasion of the value of the innovation, decision to adopt the innovation, implementation of the innovation and confirmation of the value of the innovation. Although broadly meant to describe the cultural spread of ideas and technology, it applies well in the narrower context of academic research. While the last four stages are well covered by traditional research activities, it is the initial stage of becoming aware of new ideas from far afield that is often the rate limiting factor and the least formalized in research.

As a great believer in the power of cross fertilization, I think that diffusion is too passive a metaphor; I prefer instead to think in terms of the active transport of ideas and places where I can search out sources that facilitate long range transport.

A foray into some areas that may seem off topic can provide a little dose of hybrid vigor to one's work.

I've recently found inspiration for orthogonal thinking from several unconventional sources. The TED (Technology, Entertainment, Design) Conference features a diverse set of inspiring speakers and is podcasted on the web. Edge Foundation is a web-based publication that includes the World Question Center annually featuring a grand yet simple question asked of numerous notable scientists. On the more focused topic of biomedical computation, the NIH Biomedical Computing Interest Group hosts webcast seminars, book clubs, tutorials and brainstorming events.

Although things are changing, academia is still hampered by the inertia of traditional boundaries between disciplines that form unintentional energy barriers against the diffusion of ideas. Just as a retreat or a sabbatical can provide a refreshing perspective, a foray into some areas that may seem off topic can also provide a little dose of hybrid vigor to one's work. □

DETAILS

Technology, Entertainment, Design (TED) Conferences: http://www.ted.com

Edge Foundation: http://www.edge.org

NIH Biomedical Computing Interest Group: http://www.nih-bcig.org

NewsBytes

Aquaporin Simulations De-Bunk Gas Exchange Assumptions

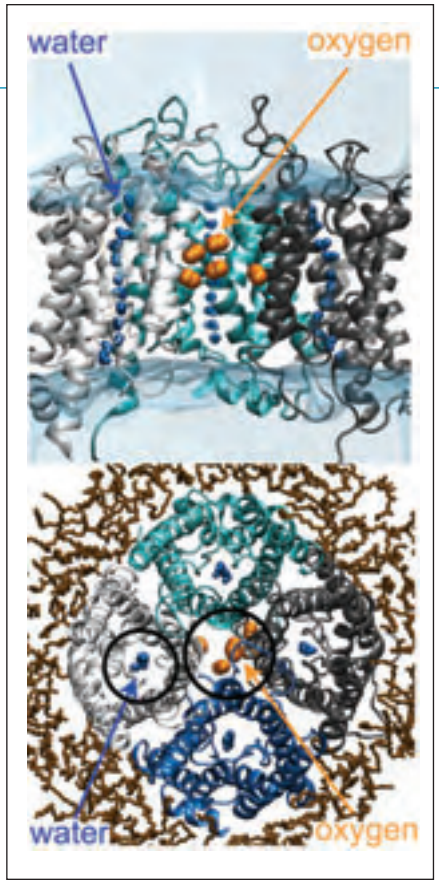
Biologists have long taken gas exchange for granted, assuming that gases simply seep through the cell's lipid membrane. Since 1998, however, evidence has been building that gases might also be exchanged through pores created by specialized proteins.

Now molecular dynamics simulations of aquaporins have weighed in on the question. The result: "It's now well established that these proteins can conduct gas molecules," says **Emad Tajkhorshid, PhD**, co-author of the work and assistant professor of biochemistry, pharmacology and biophysics at the University of Illinois at Urbana-Champaign. But, he says, some uncertainty remains: "Whether or not it's important in the human body, that's the controversial part." The work was published in the March 2007 issue of the *Journal of Structural Biology*.

Fifteen to twenty years ago, scientists believed that water permeation through lipid bilayers was enough for water transport into and out of cells. Gradually, though, researchers realized that some cells need to control water permeability, and other cells have lipid bilayers that aren't very permeable to water. Aquaporins, it turned out, carry water in and out in a controllable fashion. "I think the same might be true for gas permeability," says Tajkhorshid. "Gas permeability of a lipid bilayer is like an open free highway where everything can go through. With a protein, you can have a gating mechanism and some regulation."

One of Tajkhorshid's collaborators, **Walter Boron, MD, PhD**, professor of cellular and molecular physiology at Yale University, has been working on gas exchange experimentally for about ten years. To him, aquaporins are a likely suspect for gas conduction because they exist in places where oxygen must go in and carbon dioxide must come out. For example they are plentiful in cells that line the lung, in red blood cells, and in astrocytes—cells at the blood-brain barrier. But it's very hard to measure small changes in oxygen concentration at the surface of a membrane experimentally.

So Tajkhorshid's team pitched in with molecular dynamics simulations. Aquaporins occur in groups of four (tetramers), with four pores that conduct water (one through each aquaporin molecule) and one central pore where the molecules meet. The latter, until now, had no known function. When simulated using two complementary methods—explicit sampling with full gas permeation and implicit ligand sampling—the team found both oxygen and carbon dioxide were exchanged through that central pore. Carbon dioxide was also transmitted through the four water pores, while oxygen passed through those pores only rarely. The research also found, however, that a plain lipid bilayer conducts two and a half times as much gas as one embedded with aquaporin tetramers. "The question is whether this pathway is significant and makes any difference in terms of total permeability of the membrane," says Tajkhorshid.



Simulations of the aquaporin tetramer found that carbon dioxide and oxygen are exchanged through the central pore—a site of previously unknown function. Image courtesy of Emad Tajkhorshid, a faculty associate of the NIH Resource for Macromolecular Modeling and Bioinformatics, and his UIUC colleagues Klaus Schulten, Yi Wang, and Jordi Cohen.

"It's now well established that [aquaporins] can conduct gas molecules," says Emad Tajkhorshid. "Whether or not it's important in the human body, that's the controversial part."

The researchers hypothesize that, as with water permeability, aquaporins may be physiologically relevant to gas exchange when cells have dense, rigid lipid bilayers or when aquaporins occupy a major fraction of the membrane.

Tajkhorshid plans to introduce point mutations inside the central pore and manipulate the behavior of a gating loop to see how that changes the conducting properties of the central pore. Meanwhile, Boron's group is looking for a system in which gas conduction through aquaporins is a major pathway. Says Tajkhorshid: "Even if it's 30 percent of total gas permeability, it becomes physiologically relevant because then you can control it."

According to **Nazih Nakhoul, PhD**, research associate professor in biochemistry at Tulane University, "This idea of gas transport through membrane proteins is really gaining support. It's interesting to see molecular dynamics simulations confirm some of the earliest findings."

—By Katharine Miller

Parkinson's Culprit Modeled

Under a microscope, the curious protein clumps that dot the brains of Parkinson's patients stick out like the culprits they are. But no one has yet caught the protein—alpha-synuclein—in the act of causing disease. Now, investigators report in an April 2007 issue of *FEBS Journal* that they're getting closer: they've modeled alpha-synuclein's early aggregation and offered a detailed mechanism for its participation in neuron death.

"This is not just the first computational model of alpha-synuclein," says **Igor Tsigelny, PhD**, an author of the paper and a computational biologist at the San Diego Supercomputer Center. "Up to now, there was no molecular concept of the aggregation going on."

In the brain cells of Parkinson's patients, alpha-synuclein first starts to cluster as a proto-fibril. It then forms fibril chains, and finally ends up in the dense clumps of fibrils called Lewy bodies. Some researchers have suggested in the past few years that alpha-synuclein knocks off neurons right at the beginning of aggregation, long before it can be detected as a Lewy body. Biochemical and structural evidence hints that when a few alpha-synuclein molecules first self-assemble into proto-fibrils, they can form pore-like ring structures. These may interact with the cell membrane and allow ions to enter the cell. The entrance of ions such as $Ca2^{+}$ could lead to neuron death.

The computer model created by Tsigelny and his colleagues at the University of California, San Diego, supports this theory, providing detailed dynamics of alpha-synuclein hexamers and pentamers and their interaction with the cell membrane. What's more, the model shows that another synuclein in the cell—beta-synuclein—blocks alpha-synuclein's ring-making, suggesting at least one avenue for future inhibitory drug development.

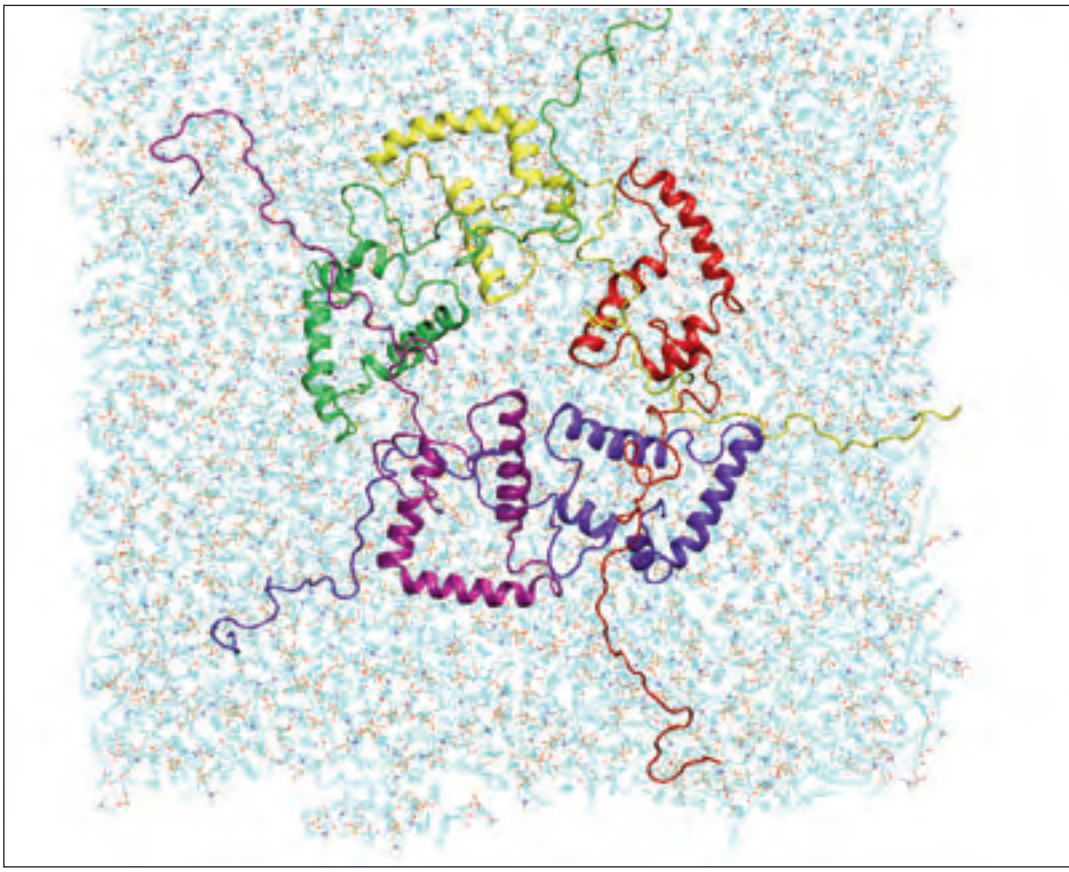
Modeling such a complex aggregation wasn't simple. Alpha-synuclein is a large protein (140 amino acids), and to model its hexamer interacting with the cell membrane required juggling around a million atoms, Tsigelny says.

Yet more than the size of alpha-synuclein, what made it difficult to model was its lack of structure. Alpha-synuclein is an intrinsically unstructured protein—one without a distinct three-dimensional shape. Most proteins consistently fold into a favored shape to do their jobs, a form that can be crystallized, imaged, and pored over. But unstructured proteins flop this way and that, even while performing their specific tasks, making them very difficult to pin down and study.

"We were not scared by an unstable protein," Tsigelny states. And he and his coworkers developed an unusual "all-dynamic" approach to modeling the protein. None of the conformations are final—they are all considered intermediate and each may last only as long as half of a nanosecond. Nevertheless, Tsigelny says, even such fleeting intermediates may aggregate. The pore-like aggregates, they found, are far more stable than single molecules of alpha-synuclein.

Having this model "is one step forward," says **Hilal Lashuel, PhD**, professor at the Swiss Federal Institute of Technology in Lausanne, Switzerland. The UCSD model provides a structural basis for testing the hypothesis that alpha-synuclein forms toxic pores, he adds. But Lashuel also cautions that only biochemical and in vivo studies can prove whether alpha-synuclein pokes holes in neurons. "Isolating the toxic species is really the most difficult question we are dealing with. You have to catch it in the act."

—By Louisa Dalton



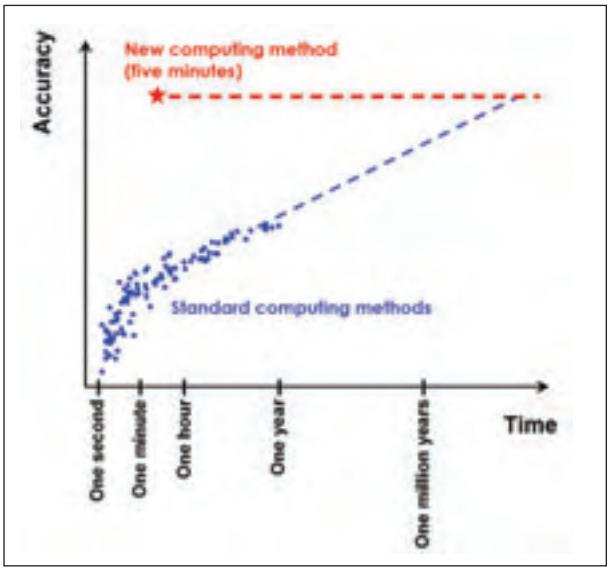
Alpha-synuclein poses as a pentamer, pore-like, on the surface of a cell membrane. Courtesy of Igor Tsigelny

Clustering Without Limits

Starting in preschool we all learn how to get organized. Typically, we start with pre-determined categories (dolls, trains, blocks); pre-set ideas about what belongs in each category (Barbie: doll; Thomas the Tank Engine: train) and a fixed number of bins to put things in.

But what if you started with none of those initial limitations? Could you still group the toys? It turns out that, in a computer, such sorting is not only possible, but extremely efficient. Using a novel algorithm called affinity propagation, researchers at the University of Toronto found that they can not only cluster lots of different kinds of data appropriately, but do it better and faster than other methods. The work was published in the February 16 issue of *Science*.

"Almost all existing techniques work on a hypothesis refinement basis: they start off with a set of assumed groups and iteratively refine them," says **Brendan Frey, PhD**, associate professor of electrical and computer engineering at the University of Toronto, co-author of the paper. "To our knowledge, ours is the first algorithm to consider all possible groupings at once."

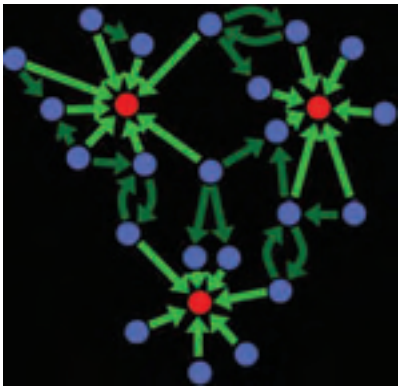


If asked to cluster facial images, a standard clustering method (k-means clustering) would take up to a million years on a single computer to achieve the accuracy achieved by affinity propagation after five minutes.

"Part of the attraction of the [affinity propagation] algorithm is that, although it was complicated to derive, it's quite simple to implement and to get an intuitive feel for it," says Brendan Frey.

The task sounds mind-boggling: There are a huge number of possible groupings. But affinity propagation handles that problem by sending messages between data points—pair-wise—so as to maximize the net similarity in each group. "Each message encapsulates or summarizes a whole distribution of possible groupings for one of the data points," says **Delbert Dueck, a PhD** candidate in Frey's lab. "No one has done that before."

Affinity propagation is based on an algorithm called belief propagation, which has been around in various incarnations for many years. But, say the authors, it's an approach that has never been applied to clustering. "Certainly not to generic clustering of any type of data," says Dueck. Indeed the algorithm is so generic that Frey and Dueck used it to analyze gene expression data, facial images, and airline routes, while other researchers have found applications in basketball statistics, the stock market and computer vision. And many tasks in computational biology require a computer to organize the data before using it to make predictions.



Frey and Dueck use affinity propagation to cluster data around "exemplars"—data points that best represent their compatriots. In this graphic, after starting with an equal chance of serving as an exemplar, candidates for that job have already emerged (red dots). Each data point sends messages to each candidate exemplar conveying how well it represents the blue point compared to other candidate exemplars. And candidate exemplars send messages conveying their availability to serve as an exemplar for particular data points.

"Part of the attraction of the algorithm is that, although it was complicated to derive, it's quite simple to implement and to get an intuitive feel for it," says Frey. There are basically only two equations to it. "Sometimes we'll give a talk and get emails from people who've implemented it the day after," he says.

When the researchers looked at how well the algorithm performed compared to other clustering methods they found it remarkably efficient. "A problem our algorithm could solve in about five minutes on one computer would take other methods up to one million years to solve on that same computer," says Frey.

Tim Hughes, PhD, of the Center for Cellular and Biomolecular Research at the University of Toronto, is considering using affinity propagation in his research. "It seems like it would do best when things really do form independent groups, and when the data are fairly sparse, so most of the correlation matrix can be dropped in early cycles," he says. "I think it will work well with exon-profiling data or genome-tiling data, where there is also a constraint that the groups have to correspond to regions near each other on the chromosome."

—By Katharine Miller

Computer Vision that Mimics Human Vision

Our brains can recognize most of the things we pass on an evening stroll: Cars, buildings, trees, and people all register even at a great distance or from an odd angle. Now, a new computer vision program can do the same thing. It successfully rivals the human ability to rapidly recognize objects in a complex picture because it mimics how information flows during the initial stages of visual perception.

"We've built a model to be as close as possible to what is known about the human visual system," explains **Thomas Serre, PhD**, a postdoctoral associate in the Center for Biological and Computational learning at MIT and lead author of two papers recently published out of the lab run by **Tomaso Poggio, PhD**, at MIT's McGovern Institute for Brain Research.

For decades, scientists have struggled to create computer programs that can recognize visual objects as well as humans can. Some computer systems excel at recognizing one particular object, but none are anywhere close to recognizing the wide range of objects observed by the human brain. Visual recognition is complicated by two conflicting goals: a program must be specific enough to discriminate between different objects, such as a person or a car, yet flexible enough to recognize the same type of object in different sizes, poses, and lighting.

"We've built a model to be as close as possible to what is known about the human visual system," says Thomas Serre.

To achieve these goals, Serre and colleagues used data recorded from real neurons in the visual system to program two fundamentally different kinds of virtual neurons called S (simple) and C (complex) units. S units recognize specific features of an image; C units monitor a range of S units in one area and allow for variation in position and size.

The researchers were surprised to find that a simple system, consisting of four alternating layers of S and C units, was able to classify pictures of a busy street scene as well as other leading mathematics-based computer vision systems, as described in the March 2007 issue of *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Serre's team then built a more complex system, consisting of many S and C layers designed to closely match the flow of information in a human brain during the first 100-200 milliseconds of perception. This enhanced system performed as well as humans on a rapid object recognition task: distinguishing animals from non-animals when images were flashed in front of humans and computers. The work appeared in the April 2007 issue of the *Proceedings of the National Academy of Sciences*. The computer system even made errors similar to the errors made by humans, suggesting that the model recapitulates the early processes of the human visual system.

The model will be used as a tool by neuroscientists to better understand the human visual system, and also has practical applications for surveillance, driving assistance, and autonomous robotics. According to Poggio, the team's next

When presented with a real-world street scene (left), Serre's computer vision system successfully recognized pedestrians, cars, buildings, trees, sky, and the street (right). Although not pictured, the model also successfully identified bicycles. Note the error in this example: the model mistakenly classified a street sign as a pedestrian. Graphic courtesy of Stanley Bileschi, PhD, McGovern Institute for Brain Research at MIT.

goal is to extend the model to include the "back projections" from other parts of the brain that allow feedback processing of visual information after 200 milliseconds.

"This is the first demonstration that a purely bottom up approach to visual object recognition, inspired by recordings from the neurons in the brain, is effective as a practical computer vision system," says **Terry Sejnowski, PhD**, head of the Computational Neurobiology Lab at the Salk Institute. "There is much more work to do, both to improve its performance, and also to use it to better understand how our own visual system works."

—By Matthew Busse, PhD

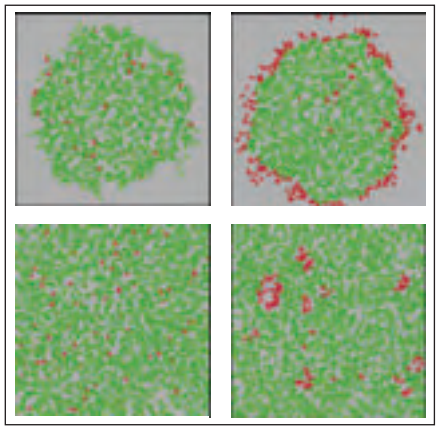
Nature Versus Nurture *In Silico*

Every generation, a few nonconformists crop up in tissue cultures of genetically identical cells. The question is: are the wayward simply born that way, or did something in the environment affect them? "You have these two possibilities—intrinsic or extrinsic, nature or nurture," says **Andras Paldi, PhD**, a biologist at Genethon in France.

Now, Paldi and his colleagues have modeled such cultured cells to determine whether extrinsic or intrinsic influences play a key role in the spontaneous emergence of phenotypic variation. It turns out that for spatial patterns beyond randomness to arise, there has to be some effect of sensing neighboring cells—i.e., extrinsic factors must play a role. And the extrinsic model resembles results seen in real cells. The work appears in April in *PLoS One*.

Paldi's work was motivated in part by the open question among stem cell biologists of what triggers a stem cell to differentiate. Why, in the same warm spot, getting the same rich media, do some cells differentiate and others stay stem cells? It is commonly assumed that this is because the decision to differentiate is intrinsic—that is, purely random.

To test that assumption, Paldi's group started by designing two simple, multi-agent based models of a tissue culture plate. In each model, all cells act independently and can switch between two cell types: A or B. In the "extrinsic" model, A cells turn into B cells when it gets crowded, and back to A cells when they have more space. In the "intrinsic" model, each cell has fixed probabilities of switching from A to B and back again.



Agent-based computer models predict the pattern (left) produced when genetically identical cells have an inherent probability of changing (from green to red and vice versa), and the pattern (right) produced when cells are triggered to change by an extrinsic factor, such as cell density. Top images represent exponential growth; bottom are at equilibrium. Courtesy of Andras Paldi.

When the scientists ran the models, they found each produces a stable, heterogeneous population, yet they differ in the cell patterns. The intrinsic model predicts lone A cells distributed evenly throughout a largely B population. Extrinsic predicts that the A cells will cluster. The result held even though the cells were allowed to migrate.

Why, in the same warm spot, getting the same rich media, do some cells differentiate and others stay stem cells?

This pattern difference allowed the researchers to compare their computational simulation with real cells. Using a muscle cell line that can switch between two distinct phenotypes, a stem-cell like progenitor state and a differentiated state, they found that the cell pattern mostly resembles that of the extrinsic model. Many of the rare, stem-cell like cells cluster; a few are solitary.

What's important here, Paldi says, is that they find environment playing a role—a significant one. In the case of stem (progenitor) cells, it means neighbor cells can affect the differentiation process. "The stem cell nature is not an intrinsic property of the cell," he says. "It is a property of the whole cell population." Paldi further believes the work supports the effort to find a way of converting adult, differentiated cells into stem cells (and avoid the need for harvesting embryonic stem cells)—a possibility that has not just scientific, but social and political implications as well.

Christa Muller-Sieburg, PhD, however, disputes that scientific conclusion. "The idea that mature cells can turn into stem cells is very attractive to many modelers but has little support through experimental data," says the professor at the Sidney Kimmel Cancer Center.

Sui Huang, MD, PhD, at Children's Hospital Boston, would have liked to see Paldi's group perturb the cell line or the culture to confirm their model. But both he and Muller-Sieburg believe the study addressed an important question, that of heterogeneity of a genetically identical population of cells. And, says Huang, it certainly "contributes to the discussion in the community."

—By Louisa Dalton

Simulating Populations with Complex Diseases

Diabetes, breast cancer, multiple sclerosis, Alzheimer's disease. All are associated with several genes' alleles interacting in complex ways with one another and the environment. Now, using a computationally intensive method known as forward-time simulation of human populations, researchers are hoping to gain a better understanding of how such complex diseases become established.

"In a real population you just see people with the disease," says **Marek Kimmel, PhD**, professor of statistics at Rice University and co-author of the work. "You don't see who in the population has the disease genes because people carrying these genes do not necessarily become diseased." But in the model population, he says, "you see both." And the researchers' approach allows them to simulate a very complicated scenario—including changes in types of selection pressure.

"This lets us evaluate how well statistical genetics tests determine what genes are responsible for the symptoms of a disease and how frequently those genes appear in the population." That's a non-trivial exercise, he says, because it has been impossible, until now, to compare the many existing gene-mapping methods head-to-head. The work was published in *PLoS Genetics* in March 2007.

Before now, the most commonly used approach to simulating diseases in human populations—called the "coalescent" method—worked by coalescing backward in time to a most-recent common ancestor. But it's extremely difficult to take selection into account using the coalescent method, says co-author **Bo Peng, PhD**, a postdoctoral fellow at the University of Texas MD Anderson Cancer Center. Moreover, that approach gets too complicated if more than one disease gene is involved. So Peng and his colleagues turned to forward-time simulation, an approach that's been around for about one hundred years.

But that technique is not without its problems. When a population evolves forward in time, there are simply too many possible outcomes. Most notably, when you introduce a disease allele, it can rapidly be eliminated and replaced with new alleles. So Peng came up with a trick: He pre-sets desired disease allele frequencies in the current generation, extrapolates them backward, and starts the simulation from there. As Kimmel puts it, "We are restricting potential variability in one aspect of the present in order to produce a simulation that resembles something close to the actual variability that exists now."

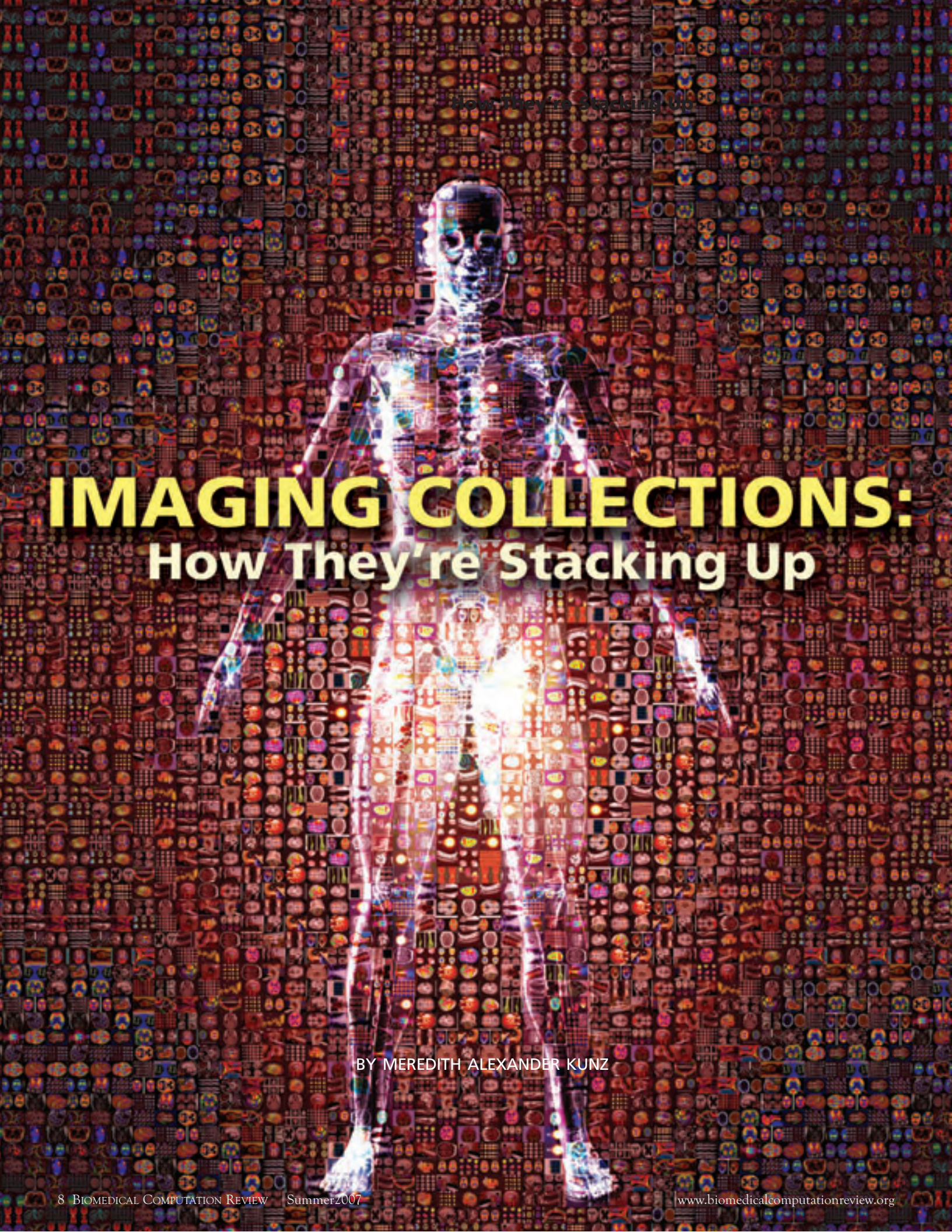
"In a real population, you just see people with the disease," says Marek Kimmel. "You don't see who in the population has the disease genes..."

The simulation uses a scripting language called simuPOP, a general-purpose forward-time simulation environment based on Python. The software is freely available at http://simupop.sourceforge.net, under a GPL license.

When Peng and his colleagues used their method to compare several gene mapping techniques they found that certain methods worked better for loci that were located distantly from one another; and other methods were more effective when loci were close together. Overall, though, says Kimmel, "We're mildly pessimistic" about current gene mapping approaches. "When the number of loci involved in complex disease is greater than two, the methods rapidly lose their power." Until recently, gene mapping for complex diseases has been disappointing, he says. Loci identified in such efforts have later turned out to be statistical artifacts. "Our modeling could figure out if this is inevitable," he says—and help guide people toward more effective approaches.

David Balding, PhD, a professor of statistical genetics at Imperial College in London, does similar work using forward-time simulations of large genomic regions. He has become pessimistic about the method's usefulness for understanding complex diseases because no one really knows what kind of selection is going on. Nevertheless, he says, this work can be useful for studying selection itself. "People tend to look at selection one allele at a time," he says, "But forward-time simulation lets us do it with complex interactions."

—By Katharine Miller □



IMAGING COLLECTIONS:
How They're Stacking Up

BY MEREDITH ALEXANDER KUNZ

In the beginning there was the Visible Human. It broke new ground by gathering some 2,000 serial images from a death row inmate's cadaver, and was the first time researchers had sectioned a single human being and gotten it right.

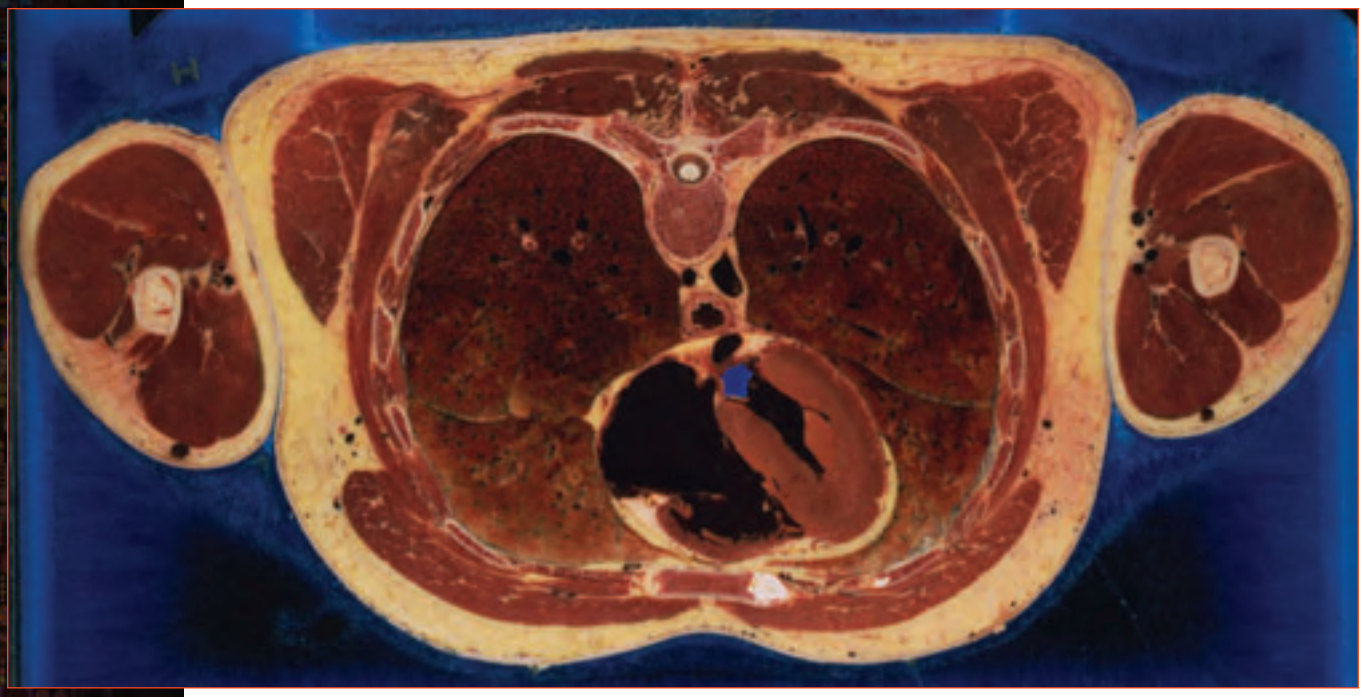
But the project broke new ground in another way as well. As the first large, publicly-available image collection, it proved that "If you build it, they will come," according to project director **Michael Ackerman, PhD,** of the National Library of Medicine (NLM).

The Visible Human was initially envisioned as a tool for teaching anatomy. But soon after the database launched in 1994, use agreements started pouring in from scientists who wanted to create 3-D images to test for radiation absorption or design artificial hips and knees, not to mention from artists illustrating anatomical injuries in court cases, to name just a few of the dozens of projects based on the Visible Human data.

Despite the suggestion that such large image collections could inspire new types of research, the Visible Human Project remained the only public imaging database available for many years. During that time, large public databases in other fields—most notably genomics and proteomics—created whole new realms of research.

Today, unlike genetic sequence data, which are centralized in GenBank, and protein structures, which reside in the Protein Data Bank (PDB), imaging

IMAGE COLLECTIONS: How They're Stacking Up



This section through the Visible Human Male's thorax shows his heart (with muscular left ventricle), lungs, spinal column, major vessels, and musculature. Image courtesy Michael Ackerman, Visible Human Project, National Library of Medicine.

Specialists carrying out imaging projects feel they should be the first to reap the benefits of the information the images contain, rather than having to share the data.

data still lacks a central repository. But an increasing number of people are hoping to create image collections from thousands of people, and not just one prisoner in Texas.

The question is whether the shift from examining images one at a time to looking at them in large groups will not only lead to better research of the type already done today, but will create something fundamentally different. Just as the field of genetics transformed into genomics when biologists moved from looking at individual genes and diseases to examining the whole genome, so too imaging could see a shift. A field that has traditionally studied narrowly defined problems using small collections gleaned from physician-collaborators could find itself faced with huge collections and the potential to reveal new correlations between diseases, genes, and anatomy. As in genomics, it will be possible to look at variation both within and between diseases like never before.

Before this transformation can happen, though, a leap of faith is required: Researchers must share their images now in hopes of greater rewards later. That's one of the current challenges researchers are tackling. There are others as well: Researchers must find ways to increase computer storage capacity; create a com-

"Neuroscientists who do complicated imaging studies are not that happy about having data out there before they can mine it," says Maryann Martone.

mon language for describing images; develop standards for "metadata" that will explain where an image comes from and what it shows; find ways to map images from different individuals onto an agreed upon "model;" and improve existing ways to analyze and interpret images consistently. They also must make images available remotely, so that physicians in rural areas will have access to large comparative collections.

As these barriers fall and imaging collections become more readily available, suddenly, imaging researchers will be able to do what genomics researchers do all the time: look at human systems in their entirety rather than in pieces.

But before we get ahead of ourselves, let's review the challenges.

Building and Sharing the Collection

Creating image data is easier than ever. Imaging capacity has increased by leaps and bounds. X-ray technology, developed in the 1890s, was followed by incrementally stronger imaging methods, from ultrasound (widely available in 1970s), to positron emission tomography or PET (1970s), to computerized axial tomography or CT scans (1970s), to magnetic resonance imaging or MRI (early 1980s) and functional MRI (early 1990s). New techniques are still appearing.

And with major improvements in data storage and networking, scientists do not worry as much about amassing bigger data sets. Big disks are relatively cheap—researchers might pay around $3,500 for a terabyte of storage—and the capacity of computer networks to transmit large images is ever improving. **Fred Prior, PhD,** of Washington University School of Medicine in St. Louis, recently purchased space to store new research images he expects will be generated during the next three years at the Electronic Radiology Laboratory which he directs. His team's new Network Attached Storage system from BlueArc can hold 102 terabytes, with an option to expand to 500 terabytes or, with an upgrade, to 4,000 terabytes (4 petabytes)—a number once unthinkable. And that does not even include clinical imaging, another huge figure.

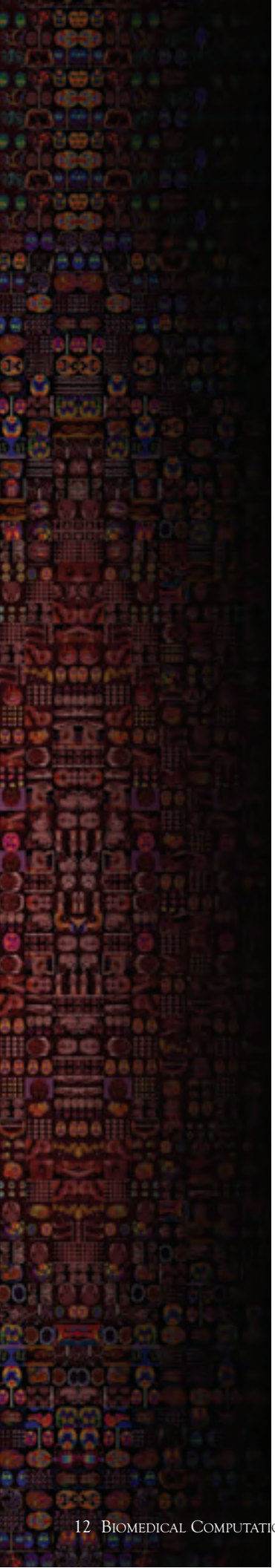
Even with such imaging, storage, and computing power in hand, a question remains: how to motivate other researchers to share their images? Scientists feel a sense of proprietary ownership over the images they have collected. While patients can perhaps stake the greatest claim to the images, most images are technically "owned" by the institution where they were made, and specialists carrying out imaging projects feel they should be the first to reap the benefits of the information the images contain, rather than having to share the data.

"Science is highly competitive. Scientists want to get the first publication, to gain funding, and get academic promotions," says **Arthur Toga, PhD,** head of the Laboratory of Neuro Imaging (LONI), at the University of California, Los Angeles.

Indeed, in 2000, a spat erupted in the brain imaging world when **Michael Gazzaniga, PhD,** director of the National fMRI Data Center, wrote to fMRI specialists who had contributed to the *Journal of Cognitive Neuroscience,* telling them they would be required to share their experimental data with the center if they wished to publish in journals including *Science* and the *Journal of Neuroscience*. Researchers immediately raised objections, sending a letter to the center's financial backers and 14 journals. Releasing their images, they argued, "impinges on the rights authors should have on the publication of findings stemming from their own work." The center decided to establish a "data hold" for a period of time, to allow authors to profit from their images first.

Maryann Martone, PhD, has run up against some of the same issues. As co-director of the National Center for Microscopy and Imaging Research

Researchers have shared abundant images in the Cell Centered Database. Here, a screenshot shows the types of images and movies available. Image courtesy Skip Cynar, National Center for Microscopy and Imaging Research, University of California, San Diego.



One of the most important parts of collecting large amounts of imaging data is also to capture each image's back story—the context in which it was made and the condition of the patient at the time.

(NCMIR) at the University of California, San Diego, she has led the creation of the Cell Centered Database (CCDB), one of the first Internet databases for cell-level structural data. She also coordinates a project supported by the Biomedical Informatics Research Network (BIRN) that investigates mouse models of human neurological disease.

"These resources were created with the idea that people were going to populate them from the community, but neuroscientists who do complicated imaging studies are not that happy about having data out there before they can mine it," she says. Because NCMIR is a "technology development center" funded by the NIH, she says, it has a mission "to serve a large collaborative community." So she decided to begin with her own center's data and hope that others would follow: "We do imaging that is unique. I figured, if we just took all the data around here and made it available, that would be helpful." It was: the project was one of the first web databases devoted to electron tomography when it launched in 2002. Since then, it has continued to give access to complex cellular and subcellular data from light and electron microscopy. Meanwhile, Martone and colleagues are still thinking about the best ways to encourage other research groups to share their data with the site.

As so often happens in the world of science, it is funders—in particular, big government-sponsored efforts—who are beginning to change the rules of the game. One project aiming to put its arms around as many images as possible is caBIG™. Launched in 2004 by the National Cancer Institute (NCI), it embraces 50 cancer centers and 30 other organizations. caBIG™ is an attempt to bring together the huge amounts of data gathered and tools created in NCI-funded cancer clinical trials. It aims to take an "open source" approach—creating an environment of sharing information in the work it funds. According to some, this is the wave of the future.

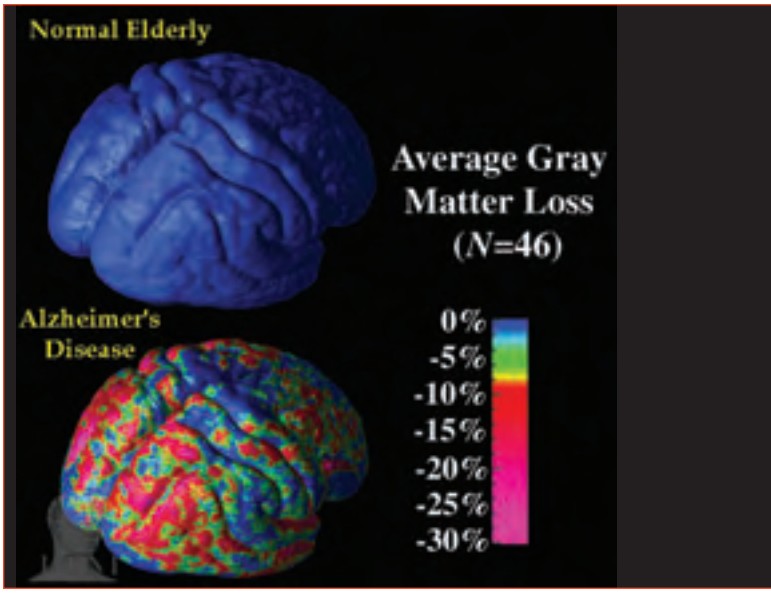
"Increasingly, the NIH is requiring that people share data," says **Daniel Rubin, MD, MS**, a clinical assistant professor and research scientist at Stanford University Medical Center. Clinical trial information, for instance, is becoming more readily available, Rubin says. He points to the American College of Radiology Imaging Network (ACRIN) as an example of this trend. This NCI-funded group hosts an imaging database that houses a large archive of clinical trial imaging data in cancer fields.

Toga thinks that it is ultimately in a scientist's self-interest to share. Lots of data is needed if scientists want to identify subtle differences between images, he says. "You can't possibly collect it on your own." What helps, he says, is when a couple of folks get together and say, "I'll share mine if you share yours," which is becoming more common.

Metadata: Capturing the Context

One cooperative project in which Toga has been involved is the NIH-sponsored Alzheimer's Disease Neuro-imaging Initiative (ADNI), which encompasses 60 different sites that are sharing image data on the disease. But if a researcher looks at an ADNI image without knowing whether the patient has a disease or not, or without access to the person's age or gender, or the drugs he or she has been taking, it becomes much less useful.

One of the most important parts of collecting large amounts of imaging data is also to capture each image's back story—the context in which it was made and the condition of the patient at the



Brain imaging studies are expanding into ever-larger populations. This enables digital atlases to be developed that synthesize brain data across vast numbers of subjects. Mathematical algorithms can exploit the data in these population-based atlases to detect pathology in an individual or patient group, to detect group features of anatomy not apparent in an individual, and to uncover powerful linkages between structure and demographic or genetic parameters. In this image, researchers from UCLA's Laboratory of Neuro Imaging (LONI) have used composite tensor mapping to show how Alzheimer's patients' brains exhibit loss of gray matter. Courtesy of Dr. Arthur W. Toga, Laboratory of Neuro Imaging, UCLA.

time. For images, efforts to create a framework for recording such information—known as metadata—currently lag behind efforts in other realms (e.g., the "MIAME" standards for microarray data). But work is now underway to improve the situation.

Some metadata—such as a patient's name, home address, and identifying features—must be removed before images enter a large database. The process of "de-identification of protected health information" follows federal privacy regulations.

But other useful information needs to be incorporated into image collections. Before image metadata can make sense, though, more standardization needs to be introduced into the field, many say. Radiologists have a long tradition of looking at images with expert eyes and dictating a free-flowing analysis, which becomes a text report that often uses terms in unique ways. That makes it difficult for other scientists or doctors to understand the image's context and content in a uniform way.

Attempts to collect and codify metadata are already well underway. One of caBIG™'s initiatives in its In-vivo Imaging Workspace is called "vocabularies and common data elements," an effort to standardize terminology in cancer analysis. Rubin, one of the group's co-leads, reports that they are trying to structure radiology imaging findings, to establish controlled terminologies for radiology, and to associate specific metadata about patients with each image gathered.

Indeed, such efforts do not end with cancer research, but could sweep across all aspects of radiology. Rubin is also involved with a project called RadLex, which is being created to offer a uniform lexicon for radiologists. RadLex plans to unify radiology term standards and to make the new terminology freely available on the Internet. Rubin sees these attempts to create a common vocabulary as the first steps in making metadata meaningful and useful for researchers and clinicians alike.

Comparing Images: Snapshots and Scales

The race to create useful imaging collections faces another hurdle: how can multiple images be compared in a way that makes sense? Each human's body parts are shaped differently, with variations that range from slight to immense. On top of that, describing shape is notoriously difficult. Though shape has been explored by the scientific community since the time of the Greeks, we still have no quantitative parameters for defining the shapes of "normal" human organs, let alone those suffering from disease. In addition, images are affected by the exact place and time they are taken, and the precise method used to take them. All this serves to undermine any straightforward database of imaging data. "Image data is a snapshot of one instance of a thing at one time under certain conditions. It's not a ground truth like a gene sequence," says Martone.

If all images can be standardized in the way they are conducted—that is, the types of equipment used, and the kinds of patients included, and the disease(s) being examined—comparison becomes easier. That is part of the success of ADNI, according to Toga: its research sites are required to follow strict protocols for their equipment and image acquisition.

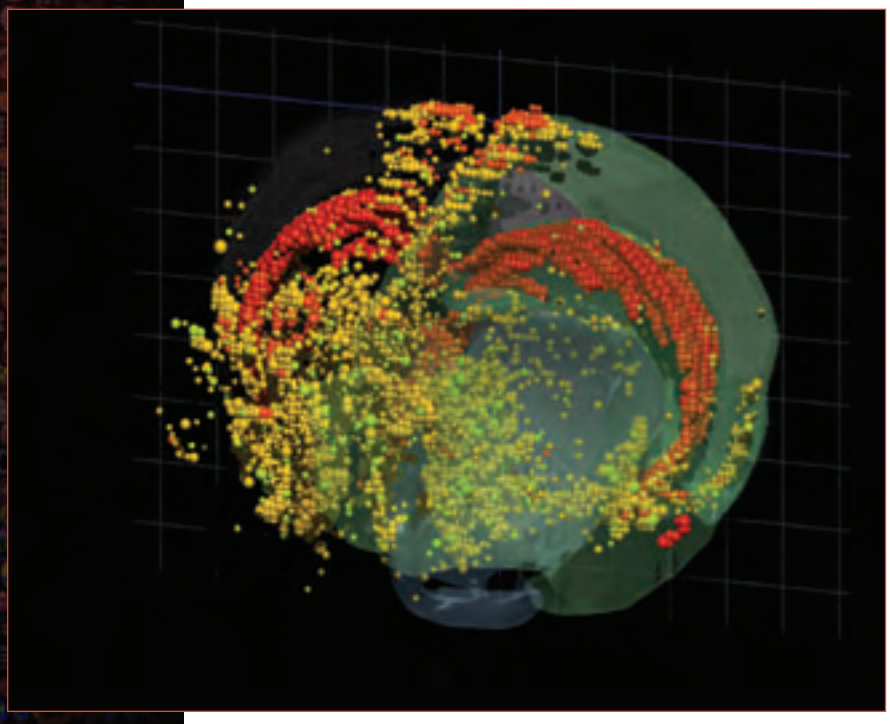
Imaging specialists have also come to rely on the best available scientific

means of shape comparison, and they try to incorporate this material into their collections. One example is in neuroimaging, where pictures of the brain are often linked to coordinate systems. Like a road map, these identify what parts are found where with reference to a grid or common starting point. For example, Talairach coordinates measure distances from a specific spot in the brain, the anterior commissure.

However, researchers find fault with existing coordinate systems because they fail to accommodate variation in large populations. While they may serve well for a single human or animal, they are not as helpful when scientists aim to "warp" many individuals onto a common model to illustrate the workings of a disease, for example. As a result, some recent brain atlases have developed their own, mathematically-complex methods for mapping variability in big groups onto a single framework.

In human brain mapping, researchers have found novel ways of dealing with natural variation between human brains. Toga reports that the 15-year-old International Consortium for Brain Mapping (ICBM) describes the brain in a probabilistic sense. For example, the atlas might tell viewers that there is an 80 percent likelihood that the basal ganglia is in a particular location that has been set out by coordinates.

Another means of handling variation is evident in the Allen Brain Atlas, an extensive mapping of the mouse brain's gene expression created by the Allen Institute for Brain Science in Seattle. The team behind this atlas created its own coordinate system to ensure extra accuracy. The ABA is a union of neuroscience, genetics, and informatics. To map gene expression onto the 3-D mouse brain model, a team of neuroanatomists drew all the regions of the brain, and then "we lofted those regions onto a 3-D model of the brain using informatics algorithms," says **Michael Hawrylycz, PhD,** director of informatics at the Allen Institute for Brain Science. Using high-level computations, an image of gene expression was then mapped onto the reference atlas's coordinates, creating pictures that form the database. ABA scientists chose one mouse to be the reference model, and the rest of the mouse data was warped to fit into the spatial framework of that single animal's brain. "We wanted a mouse that was held under exactly the same conditions that we were going to run the genes under," Hawrylycz says.



The Allen Brain Atlas produced this 3-D reconstruction showing normal expression of manosidase 1a in the adult mouse brain viewed from the front left. The translucent forms represent the left half of the brain and reflect the underlying standard anatomical reference framework to which the gene expression data was registered. Each colored sphere reflects expression of the Man1a gene in a 100 µm3 area. The size of each sphere corresponds to expression density, and the color reflects expression level. The large red arc indicates that this gene is turned on strongly in the hippocampus, a part of the brain known to be involved in learning and memory. The image was generated from the Allen Brain Atlas (www.brain-map.org) using the 3D visualization tool, Brain Explorer. Courtesy of the Allen Institute for Brain Science.

Another vexing challenge for image comparison is the issue of scale. Martone points to the problems confronted by brain researchers when they try to see the workings of a disease on multiple scales in a large set of images taken using different technologies. "We go from MRIs, to optical microscopy, to electron microscopy, then to X-ray crystallography," she says. "Every time you traverse scales, there are gaps. Every time you switch techniques, you lose continuity." Even the contrast mechanisms are different, so one scale may contain fluorescents while another is gray scale, disorienting researchers. It's like being confronted with a GPS tracking image of a moving vehicle one minute, and a Polaroid photo of the vehicle's front wheel the next.

To combat confusion, Martone's team is trying to create new coordinate and reference systems that ease the transition among scales when studying neurons in the brain. She cites a new software project that attempts to correlate microscopy with "feature-based matching systems" that describe the attributes of such cells in a uniform way.

Analyzing Images in Three Dimensions

Those who set out to compare images are also getting help from advances in image analysis software, a field that has advanced rapidly in the past few years. **Ron Kikinis, MD,** professor of radiology at Harvard Medical School, has helped lead the way. He and colleagues developed the "3D Slicer" image analysis software, initially a joint, open-source effort between the Surgical Planning Lab at Brigham and Women's Hospital, where Kikinis is founding director, and the Artificial Intelligence Lab at MIT. Created to help visualize medical image data in 3-D, it has been used with success in fields as far flung as astronomy and geology.

Although Slicer was conceived as an interactive tool for processing single images, it is also useful for researchers working with large sets of images, Kikinis says. "Now people are beginning to build informatics frameworks to hold and manage images; and soon people will shift focus to how to process those images," Kikinis explains. "With all the progress in image acquisition, you still need to turn data into medically-relevant information, and that requires image analysis," he says. The current version of Slicer is interoperable with BIRN's informatics frameworks and is also linked directly to the National Cancer Imaging Archive (NCIA)—a large repository of cancer trial images—as a recommended viewer for its images. Slicer can be used to review image sets for prototyping and results for quality assurance. For example, before processing hundreds of images, it's wise to test your algorithms and procedures with a handful first. That's where Slicer's interoperability with large databases can be used as a tool that offers essential functionality.

Another fundamental tool available to image users is the Insight Tool Kit (ITK), which Ackerman of NLM says took some three years to develop. Based on GE's Visualization Tool Kit (VTK), ITK's algorithm allows a user to identify a body part—for instance, the heart—and then ask the tool to draw a line around everything that looks like heart tissue. "Up until now, you'd have to do that by hand," says Ackerman. The tool saves users' time and is constantly being updated, making it ever more efficient.

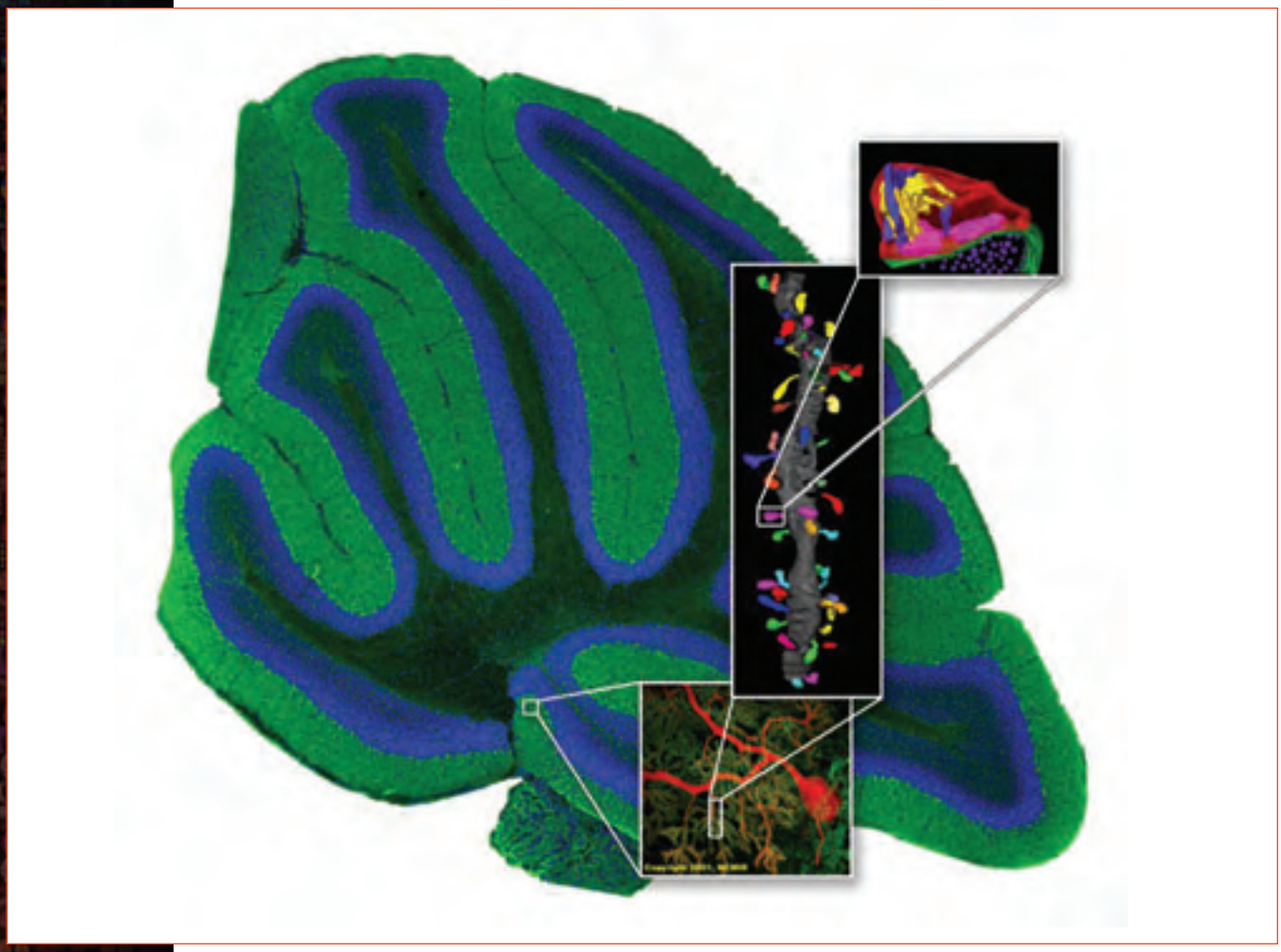
Other complementary efforts are working to ensure that researchers in distant labs can create their own image analysis applications on a lab workstation. Fred Prior has worked with other researchers to oversee creation of the Extensible Imaging Platform, or XIP. "The idea is that there are lots of commercial workstations that are optimized for clinical reading, lots of research packages like Slicer, and great toolkits like ITK that give you functionality, but what's missing is a way to build custom applications for these tools," says Prior. XIP will give users a "rapid development environment," he says, enabling researchers to do image processing more easily. XIP's initial targets are cancer researchers already working in the grid, but its potential is much greater.

"Image data is a snapshot of one instance of a thing at one time under certain conditions. It's not a ground truth like a gene sequence," says Martone.

"We're hoping we'll see a cottage industry building new applications in this XIP framework to do things like virtual colonoscopy and radiation therapy analysis," Prior says. The "slick part" in Prior's words is that such applications could be run through the grid and offered to other researchers remotely through the platform—creating a whole new level of sharing.

In quite a different application of image analysis, some researchers are honing in on new ways to help scientists and doctors find the images they need using tools that analyze its image content rather than its metadata. Known as content-based image retrieval, these programs also strive to

IMAGE COLLECTIONS: How They're Stacking Up



The Cell-Centered Database, a project of the National Center for Microscopy and Imaging Research, brings together data from different experiments so that multi-scaled views can be created, helping scientists to study how higher order structures, such as cellular networks, are assembled out of finer building blocks, such as dendritic architectures. This montage shows seven orders of magnitude of scale from centimeters to nanometers. A slice through a centimeter-sized mouse brain was obtained by making a mosaic from thousands of multiphoton microscopic images. Then fluorescence microscopy was used to isolate a spiny neuron (first sub-panel). Correlating cell structures identified under the light microscope for subsequent examination under the electron microscope permitted biologists to visually reconstruct the three-dimensional structure of dendritic structures with nanometer resolution. The second and third sub-panels portray electron tomographic reconstructions of an unbranched spiny dendrite from cerebellum and its nanometer-sized synaptic complex (from hippocampus). Image courtesy Skip Cynar, National Center for Microscopy and Imaging Research, University of California, San Diego.

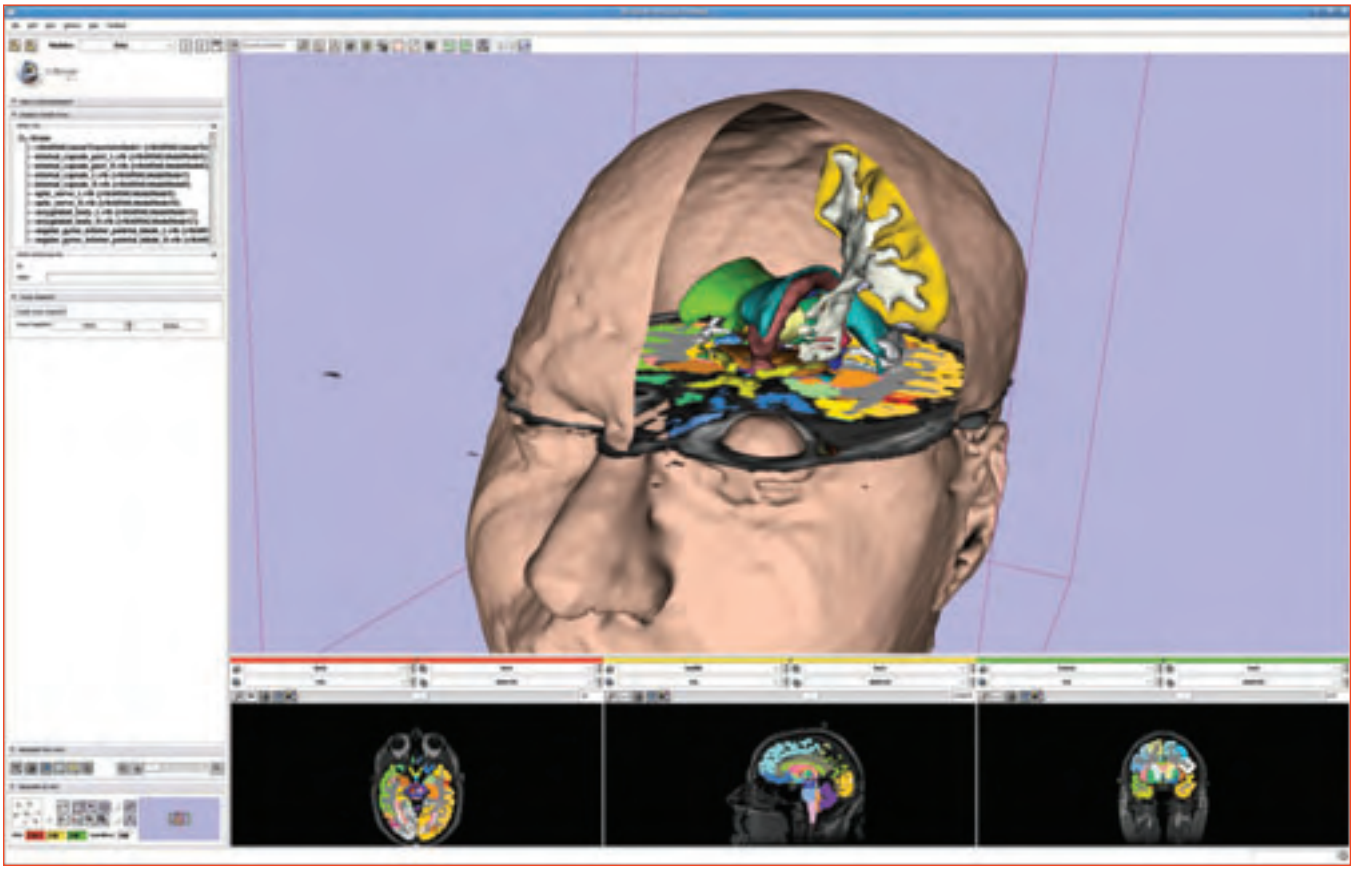
overcome errors caused when inaccurate text-based keywords lead to mismatches in retrieving images, write **Paul Miki Willy** and **Karl-Heinz Küfer, PhD,** of the German Fraunhofer Institut Techno- und Wirtschaftsmathematik in a 2004 paper. Content-based programs attempt to index images according to visual features such as color, texture, and shape. Ultimately, some hope that these systems might allow a physician to click on an image of a cancer in a particular patient and ask a database to show similar images for comparison. So far, this technology has not yet reached a wide audience; some believe more work is needed to ensure accuracy in such searches.

Accessing Image Databases: Connecting to the Grid

All these image collections will do little good if no one can access them remotely. Researchers at the crossroads of biomedicine and computational science are tackling that problem now.

One promising answer is to create "federated databases"—groups of unique imaging collections that are linked together by a sort of "grid," and that are accessible remotely via a seamless user interface that makes the data sets resemble one single virtual database. **Joel Saltz, MD, PhD,** professor and chair of the department of biomedical informatics at Ohio State University, leads a group that develops technologies that can enable "grid" access for large image collections to create such federated systems. His group has developed middleware to support complex distributed applications. It attempts to stitch together different bodies of images, making them available and searchable.

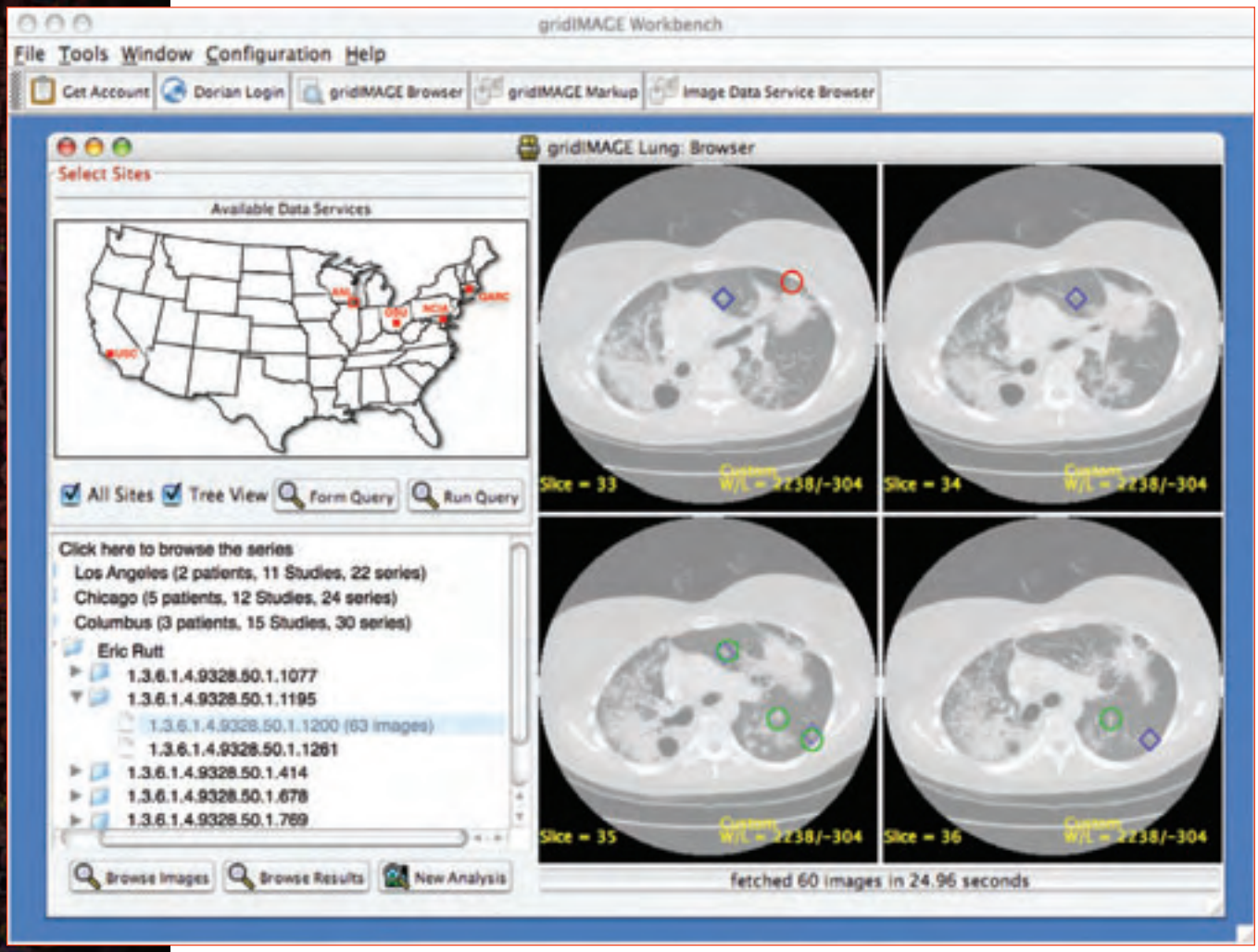
"The overall goal of the effort is to develop an infrastructure to connect multiple databases, to allow people to discover what images are out there, and to analyze both remote and local imagery and to integrate image data with information from molecular studies, clinical studies, and pathology specimens," Saltz says. The National Cancer Institute caBIG™ project has incorporated the Ohio State group's software in the caGrid software package. This was first distributed in December and, Saltz says, quite a number of funded efforts have begun to incorporate it. Furthest along in the process of opening up an image database to many users with



Slicer3 image analysis software is an integral part of the brain atlas created by the Surgical Planning Laboratory and the Psychiatry Neuroimaging Laboratory (PNL) at Brigham & Women's Hospital in Boston. This three-dimensional digitized atlas of the human brain is used for surgical planning, model-driven segmentation, research, and teaching. As this screenshot illustrates, Slicer3 enables users to outline and manipulate specific regions of the brain in three dimensions based on multi-modal volumetric input data including specialized MRI methods. An additional goal of this brain atlas is that it can be used as a template for automatically segmenting regions of interest in large new MR data sets. Image courtesy of Ron Kikinis, Surgical Planning Laboratory, Brigham & Women's Hospital.

Saltz's help is the National Cancer Imaging Archive.

These new systems may not be open to just any member of the public—at least some will require registration and credentials. But the incentive to participate is high. Researchers and physicians who gain access will be able to communicate with each other in new ways that could make a big difference to patients. A major benefit for those linking their images to a grid is the possibility of "central review," says Saltz. In central review, radiologists remotely read an image and provide their feedback via software that allows a user to capture mark-ups, pointers, and comments. For instance, a radiologist in Omaha might send out a CT scan of a patient's lung via Saltz' software to radiologists around the world as well as to computer-aided diagnosis algorithms available at supercomputers in research centers. She might hear back from radiologists in Mumbai, Tokyo, and Chicago, and from computers at a handful of universities, possibly discovering lung nodules she had missed.



This screenshot from the Saltz lab's gridIMAGE application shows how radiologists in remote locations can review and markup images from multiple collections. A radiologist accesses the interconnected or "federated" imaging databases through a single interface and can submit a review request to other participating physicians who use the same database. The reviewers can add marks and comments and then submit their marked-up results to a central result server, which transmits it to the radiologist who made the request. This application is based on the Saltz lab's In Vivo Imaging Middleware. Image courtesy Joel Saltz, Ohio State University.

The next generation of applications will reveal whether the rise of large imaging collections will create a new science, just as genetics spawned genomics.

Applications: Will They Come?

If researchers overcome the barriers described above, the question then will be whether it will prove worthwhile. Will innovative applications follow? In other words, if you build it, will they come?

Early indications are that they will. For some physicians, the near-term possibility of central review alone will make federated imaging databases worth the effort.

For neuroscientists, gaining insights into the brain's workings and connections requires large numbers of fine-grained images. In the past, scientists had done studies of specific parts of the brain, but few had tried to discover the overall structure of the brain. Large neuroimaging projects such as the ABA are attempting to change that. Indeed, some hope to one day map every single neuron in the human brain, creating a data set of upwards of 1 million petabytes. This "connectome," promises to be the image-based Human Genome Project of brain researchers. Its success will rely on computer-assisted image acquisition and analysis to map the structure of the nervous system, says **Jeff Lichtman, MD, PhD,** professor of molecular and cellular biology at Harvard.

In clinical trials for cancer treatments, image collections help in evaluating a drug's effectiveness, says **Carl Jaffe, MD,** diagnostic imaging branch chief for the cancer imaging program in the division of cancer treatment and diagnosis at NCI. The promise of using image collections to speed drug development is already beckoning. "The regulatory authorities are more willing to accept regression of a tumor as a sign of a drug's effectiveness...and imaging is the pivotal marker for this," he says. A large database of reference images helps to balance "reader artifacts"—that is, errors in radiologist's assessments—and to substantiate that a tumor has indeed changed size in an important way, he explains. Researchers could use a central review-style process to verify their reading of an image. "An image database allows you to go back to a larger community of observers and confirm whether or not something seems to be supportable."

For researchers studying rare diseases, the goal is to find others to compare against and to increase understanding remotely. For example, says Jaffe, in the old days, a researcher hoping to test a drug for a rare disease such as retinoblastoma—a cancer of the retina with an incidence of only 430 cases per year—would have to request MRI films from around the country to try to prove that his trial worked on a range of patients. But some films would come back too dark, some too light, and some without the right metadata. If all the data and images could be collected digitally in an online database, the researcher would more quickly understand the drug's impact. "What you want is an electronic, common pool of data and metadata," Jaffe says.

Surgeons and other physicians could also benefit from such systems as Rubin's efforts to use large groups of images to inform a doctor of how to diagnose and treat a patient. Using Rubin's decision support software, physicians can select from a series of structured annotations of an image and upload the image data. Then a computer program tells them the likelihood of disease. "We want to give radiologists a tool to help them decide when to biopsy based on what they see," he says. While it is partly based on the knowledge of expert radiologists, this type of technology will work even better when a large number of images are available to inform the program—hence the need for large databases filled with rich stores of metadata.

Increasing numbers of researchers on the biomolecular scale are also using imaging in their research, including scientists like Martone and the people who utilize the ABA and other such atlases. For example, labs are using the ABA to investigate risk factors for multiple sclerosis and to identify genetic hotspots associated with memory performance. And new databases at the cellular level are popping up, including the Open Microscopy Environment, a large public database focused on microscopy imaging data.

The New New Thing

Imaging is just one of many bioscience fields moving towards more and better information sharing and collecting. While the field faces its own hurdles—the difficulties of comparing images, for example—it falls within a larger trend of making data available and breaking down the silos of single organ or disease-focused work that for so long dominated the sciences. It's the same impulse that inspired the release of the genome and the dawn of genomics, and could cause a similarly radical shift in how people use image data.

The next generation of applications will reveal whether the rise of large imaging collections will create a new science, just as genetics spawned genomics. Ultimately, it might be possible to cross-compare between imaging and genomics. That's already happening in brain research projects such as the Allen Brain Atlas, but the trend could spread throughout the body. And as in genomics, the shift could generate an entire new field of research in which scientists could build an entire career.

If the Visible Human is any proof, simply building large, accessible collections of images will attract scientific curiosity and will launch a wealth of useful applications we cannot even imagine today. □

DOCK THIS:

In Silico Drug Design Feeds Drug Development

BY KRISTIN COBB, PHD

Once upon a time, not long ago, HIV/AIDS was a scourge, killing anyone who contracted the deadly virus. Now, many people are living with the disease, which they control with drugs initially developed in the 1980s and early 1990s using an approach called computer-aided drug design—the use of computer models to find, build, or optimize drug leads.

Armed with information about the 3-D structure of HIV protease, an enzyme essential to the HIV reproductive cycle, computational researchers designed molecules *in silico* to precisely fit the shape of the enzyme's active site—as though fitting a key to a lock. The resulting drugs, potent inhibitors of HIV protease and the HIV life cycle, were brought to market in record time and revolutionized the treatment of HIV/AIDS.

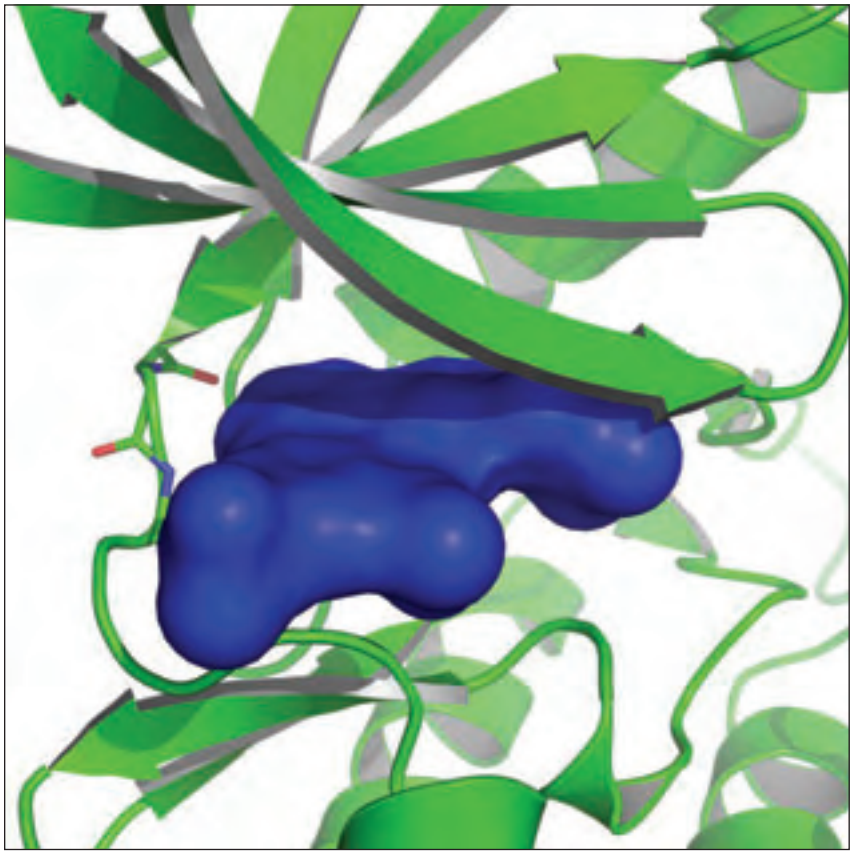
Around the same time, another anti-viral—Relenza, which treats influenza and was a forerunner to Tamiflu—was also designed using these methods. These HIV and flu drugs are among the best known success stories of computer-aided drug design (see page 23 for both stories).

Since those early successes, computer modeling has become an integral part of drug discovery. "Almost everything that has recently moved forward from big pharmaceutical companies to market has involved some sort of collaboration with computational chemistry. It's like asking, were there chemists involved? Of course there were. It is part of the process," says **Tara Mirzadegan, PhD,** head of the computer-aided drug design group at Johnson & Johnson.

"Almost everything that has recently moved forward from big pharmaceutical companies to market has involved some sort of collaboration with computational chemistry. It's like asking, were there chemists involved? Of course there were. It is part of the process," says Tara Mirzadegan.

Quite often, computers play a role without making the big splash they did with Relenza and the protease inhibitors. That's probably because no drug is created solely *in silico*; the computer is just one of many tools in this process. But as algorithms evolve, computing power explodes, and scientists solve a greater number of 3-D protein structures, computer-aided design has the potential to dramatically cut the cost and time of drug discovery. How? By narrowing down the field of compounds that might help treat a particular disease; by assembling novel drug molecules to disrupt specific disease pathways; and by providing new attack routes against traditionally difficult drug targets. Computers are also increasingly playing a role in optimizing drug leads for bioavailability and safety.

Despite the over-hype of computers as the saviors of drug development companies, many still expect this process to bear important fruit. Computer-aided drug design played a critical role in the design of several drugs that are now in late preclinical or early clinical development. Only time will tell which of these, if any, will emerge as drug success stories.



***Docked Drug.** This 3-dimensional computer graphic shows a candidate drug (a JAK2 inhibitor) docked in the active site of its target protein (JAK2). JAK2 protein is implicated in various myeloproliferative disorders (diseases that produce excess bone marrow cells, such as chronic myelogenous leukemia, or CML) estimated to affect 80,000-100,000 people in the U.S.. Courtesy of SGX Pharmaceuticals, Inc.*

Virtual Screening

How it works: In the ideal situation, the 3-D structure of the target molecule (usually an enzyme or receptor) is known, allowing scientists to directly visualize drug-target interactions *in silico*. Structure-based methods have evolved in two directions since Relenza and the HIV proteases—virtual screening and fragment-based design.

In virtual screening, the 3-D structure of a target is screened against libraries of potentially active small molecules. The computer "docks" each compound, or ligand, into the target's active site and scores its geometric and electrostatic fit.

Considerable progress has been made in docking programs in the last two decades, but scientists agree that the problem is complex and that they have yet to find a perfect solution. To start with, the ligand and protein target are often pictured as a rigid lock and key—but in fact they are dynamic, moving objects that continually change shape and adjust their shapes in response to each other.

"Imagine taking a fluffy ball and trying to mold it to optimally fit some kind of a binding site. There are just way too many configurations," says **Dimitris K. Agrafiotis, PhD,** vice president of

Continues on page 24

EARLY EXAMPLES: ANTI-VIRAL DRUGS

Relenza and the HIV protease inhibitors stand out as the two classic examples of computer-aided drug design.

Relenza was developed through a collaboration of Australian scientists, including **Jose N. Varghese, PhD**, head of structural biology at CSIRO Molecular and Health Technologies. In 1983, Varghese and his colleagues used X-ray crystallography to solve the 3-D structure of the enzyme neuraminidase, one of two potential protein targets on the surface of flu. Neuraminidase plays a critical role in the flu life cycle: after the virus replicates within a host cell, neuraminidase releases the newly formed viral progeny by cleaving a bond between the viral surface protein hemagglutinin and a sugar on the host cell surface, sialic acid.

A series of structural experiments revealed important insights. The active site of the enzyme was highly conserved in all strains of flu—both human and animal; the virus routinely escaped antibody recognition by mutating around the periphery of the active site but never changing the active site itself.

"Because it was so highly conserved, it seemed clear to us that it must have a very important function," Varghese says. "So, clearly if one made a molecule that went in there and blocked that site, it would be pretty effective."

A synthetic analog of sialic acid was known to inhibit neuraminidase, but without sufficient potency. Using the crystal structure of neuraminidase bound with this analog, the researchers set out to design a better inhibitor *in silico*. Computer predictions revealed that a particular guanidinium-for-oxygen substitution would give tight binding. Synthesis of this compound—Relenza—turned out to be tricky, but eventually succeeded.

"It bound in nanomolar binding, so it was very tight, and it certainly blocked the virus replication right down to its tracks," Varghese says.

Relenza was licensed to GlaxoSmithKline Inc. in 1990 and approved by the FDA in 1999. Following their lead—and capitilizing on a patent oversight, according to Varghese—Gilead Sciences developed the better-known neuraminidase inhibitor, Tamiflu (marketed by Roche). Both drugs may be important in the fight against bird flu, Varghese says.

Development of the HIV protease inhibitors lagged behind that of the neuraminidase inhibitors by several years, but the former won FDA approval sooner (in the mid-1990s) because of the pressing medical need.

Dale Kempf, PhD, who is now a distinguished research fellow in Global Pharmaceutical Research and Development at Abbott, was involved in Abbott's development of ritonavir (brand name Norvir), which started in late 1987.

"It's one of the first examples of the application of genomics for drug design," he says. When the HIV genome was sequenced and published in the mid-1980s, several groups recognized characteristic sequences suggestive of a protease enzyme.

Interestingly, the gene encoded only half a protein, which led Kempf and others to realize that the protease must be composed of a dimer—two identical halves that come together to form one active site. This provided a key structural insight even before X-ray crystal structures of the protease were available: the active site had to have a particular type of symmetry, known as C2 or two-fold symmetry (rotation 180 degrees around a central axis yields the identical structure).

Kempf's group used that insight to create a computer model of the protease active site and to design possible inhibitors *in silico* by starting with a known substrate, chopping off half of the substrate, and rotating the remaining half by 180 degrees.

"And when we went into the lab and made those compounds, they turned out to be very potent inhibitors," Kempf says.

Using a combination of the X-ray crystal structures of HIV protease (which had since become available) and computer graphics, they modified these compounds *in silico* to visualize how certain substitutions would improve characteristics like bioavailability. The first compound with sufficient oral bioavailability, ritonavir, was synthesized in 1991.

In 1996, the FDA approved ritonavir in record time (72 days). The total development time—about eight years—was roughly half that of a typical drug, due both to the structure-based approach and to the FDA's accelerated review. Several other HIV proteases emerged around the same time, including saquinavir (Roche) and nelfinavir (developed by Agouron, now a subsidiary of Pfizer). These drugs helped to revolutionize the treatment of HIV.

***Cancer Interrupted.** This three-dimensional computer graphic shows a drug candidate (MET tyrosine kinase inhibitor) bound to its target protein. MET receptor tyrosine kinase controls cell growth, division, and motility and is implicated in a range of cancers, including renal cell carcinoma, gastric cancer, lung cancer, glioblastoma and multiple myeloma. Courtesy of SGX Pharmaceuticals, Inc.*

Continued from page 22

informatics at Johnson & Johnson Pharmaceutical Research & Development. "Small molecules—unless they're very small—tend to be very flexible. They flop around a lot. They can assume a multitude of conformations in 3-D." If a molecule has five rotatable bonds, then each bond can rotate at many different angles, creating a lot of freedom to take on unique conformations.

Most docking programs now account for the flexibility of the ligand by sampling its many conformations and docking each one, but adequately accounting for the flexibility of the target protein is a much more challenging problem. Adding protein flexibility exponentially increases computing demands.

"The state of the art today is coming up with sensible simplifications that make the problem computationally tractable but still meaningful," Agrafiotis says.

Besides the flexibility of the protein, many docking programs do not adequately account for the influence of water—which surrounds all molecules in living systems. "The mathematical models for defining water and how it shapes itself around the receptor and the drug molecule are still pretty unclear," says **Kent Stewart, PhD,** a research fellow in structural biology at Abbott.

In addition, the algorithms estimate binding energies using classical Newtonian physics, rather than quantum physics—which also reduces accuracy. "You can calculate the binding energies from some sort of Newtonian point of view, treating atoms as sort of balls attached to springs. Or you can treat it from a quantum mechanical point of view. Now the quantum mechanical calculations, as you can imagine, are horrendous," says **Jose N. Varghese, PhD,** head of structural biology at CSIRO Molecular and Health Technologies. "At this stage, it is a computational challenge."

Methods of scoring how well a small molecule fits a protein's active site also must trade off between speed and accuracy. "The scoring function that we use has many shortcuts and approximations," says Mirzadegan. Her group will virtually dock the company's one million proprietary compounds (which it has purchased or developed over the years) against a given target, and pick the highest ranked 10,000 for biological testing. "We cannot afford docking one compound per day. That would be one

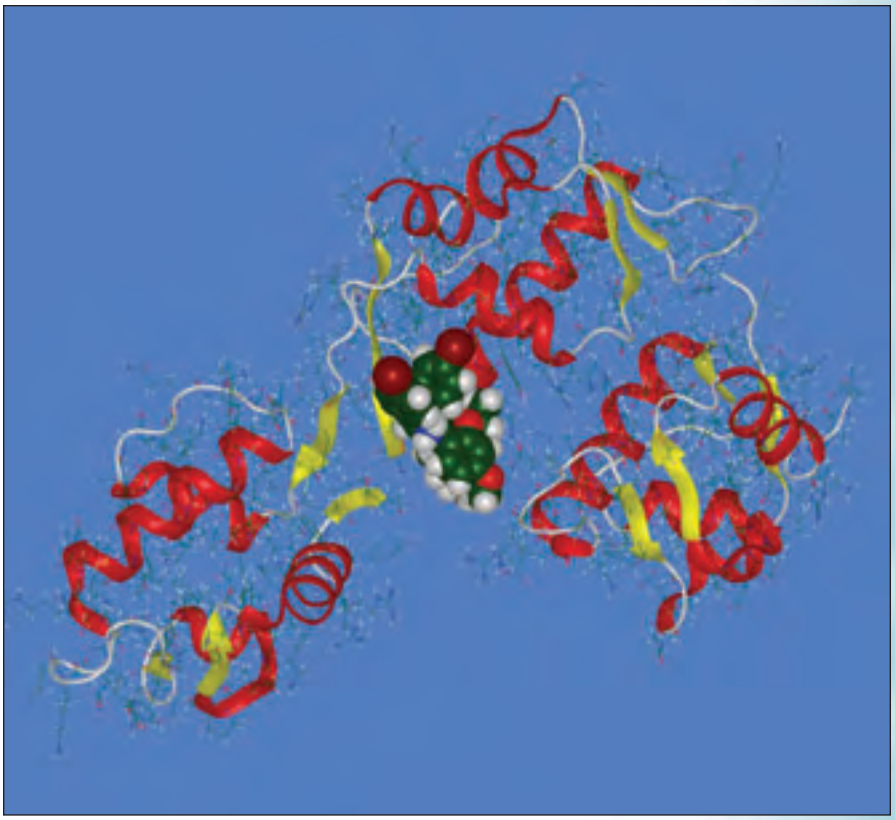
"The state of the art today is coming up with sensible simplifications that make the problem computationally tractable but still meaningful," says Dimitris K. Agrafiotis.

million days. So we have to do it in a matter of seconds or sub-seconds."

But increased computing power can help boost the speed of virtual screening without compromising accuracy. In 2000, for instance, **Arthur J. Olson, PhD,** professor of molecular biology and director of the Molecular Graphics Laboratory at The Scripps Research Institute, started the FightAids@Home project, which uses internet-based grid computing—as was popularized by the SETI@Home project—to do virtual screening for new anti-HIV drugs.

"If most people who have computers use only about five percent of the CPU cycles—and the rest of the cycles are just idle—how much wasted or available computing is there?" Olson asks. "It turns out to be an amazing number." His grid computing project makes use of that idle computer time and helps evaluate drugs for dealing with HIV proteins' habit of rapidly mutating to escape drug pressures. Fortunately, the 3-D structures have been solved for many of the mutant HIV proteins. With the help of about 500,000 volunteer computers, Olson used AutoDock (a popular docking program that was developed in his lab) to screen 2000 small molecules against several hundred different HIV protease mutants. The program took six months to run; he estimates that on the Scripps super computer, with 300 processors running, it would have taken 50 years.

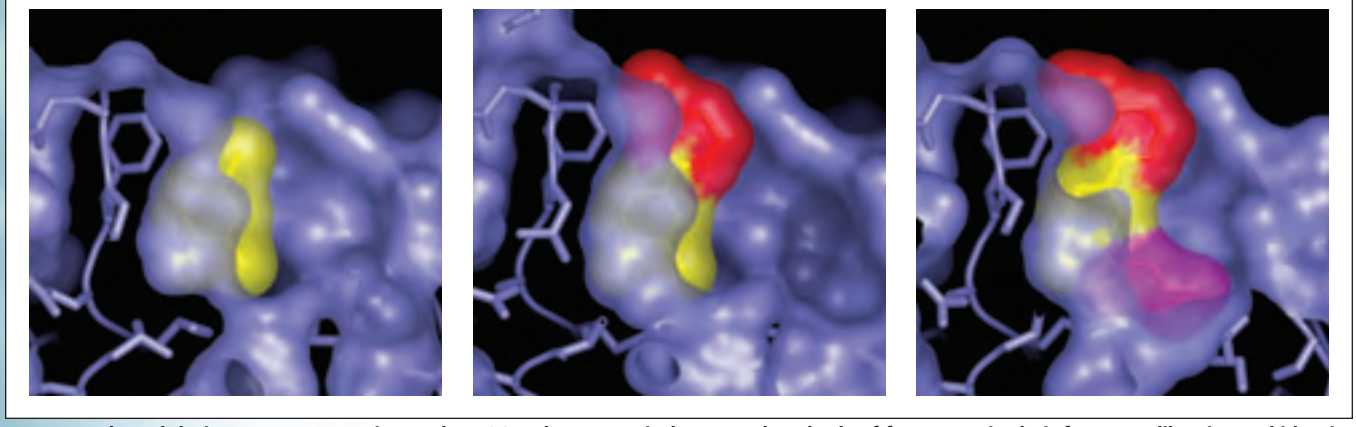
Besides identifying several drug leads, which are now in testing, Olson recognizes an even more important payoff: "When you do such massive dockings, you actually are collecting more than just an answer; you're collecting a lot of statistics." Such data could, for example, be used to identify a subset of mutants that represent a spanning set—one that captures all unique interactions with the ligands screened. "Doing docking on only this subset of mutants would free up computer time for screening larger libraries, using more dynamic representations of the protein targets, or using more accurate scoring functions," he says.



Anti-Cancer Key. An anti-cancer drug compound—nutlin—bound to the cancer-causing protein MDM2. Courtesy of RMC Biosciences, Inc.

The Folding@Home project at Stanford also uses grid computing for drug design. Led by **Vijay S. Pande, PhD,** associate professor of chemistry and of structural biology, Folding@Home focuses on simulating protein folding and misfolding, but "as our work matures, we have been looking into the next steps involved in computational drug design," Pande says. Using distributed computing, his group has devised new, more accurate algorithms for docking and for calculating ligand-protein binding energies. These algorithms are being used in the design of several new drugs, including new inhibitors of the cytokine-cytokine receptor interaction (involved in cancer); novel chaperone inhibitors (also involved in cancer); and novel antibiotics that target the bacterial ribosome.

"Distributed computing is a key



Fragment-based design. Drug companies, such as SGX pharmaceuticals, screen hundreds of fragments in their fragment libraries and identify hits that serve as the building blocks for novel drug candidates. Knowledge of the binding mode of each fragment to its target is combined with advanced computational tools to produce "engineered" drug leads. For example, in this series, a hit is first identified through crystallographic screening (yellow); then chemical groups (red and pink) are added to the bound fragment to increase its binding affinity. Courtesy of SGX Pharmaceuticals, Inc.

Distributed computing is key to developing better, more accurate algorithms for computer-aided drug design, says Vijay Pande. "It allows us to do calculations otherwise impossible."

aspect to this, as it allows us to do calculations otherwise impossible," Pande says.

Fragment-Based Design

Fragment-based methods take a "Lego" approach to drug design. In a lab, scientists create chemical libraries of small compounds, or fragments—perhaps one-third the size of a typical drug—that are easily linked together. They then screen the libraries for binding activity experimentally, using high-throughput X-ray crystallography (or NMR or mass spectrometry); when a fragment binds to the target, the crystallography provides an exact 3-D picture of the bound fragment in the active site. Next, with the help of computer modeling, fragments are turned into potent drug leads by adding new chemical groups to the initial core fragment or by stitching together several fragments that bind to different points in the active site.

"I think this approach is showing quite good promise," Varghese says. "In fact, with the advent of these modern synchrotrons, scientists can do this fairly quickly—and a lot of pharmaceutical companies are moving in this direction."

The approach offers a combinatorial advantage: "Instead of having a database of say four million compounds that a really large company would have, you take compounds that are say one-third of the size, and explore them combinatorically. If you explored ten fragments in three different positions, you'd actually explore 1000 combinations. So with a database of something like 400 compounds, you can explore a chemical space that is in the several millions," says **Sir Tom Blundell, FRS, FMedSci,** professor and chair of biochemistry at the University of Cambridge. In 1999, Blundell co-founded Astex Therapeutics to do fragment-based methods; the company is now testing a kinase inhibitor—a type of cancer drug—in clinical trials.

"The experiment is really one of using crystallography to do your screening. So you've pushed the crystallography technology to the point where you can do it so rapidly that it becomes effective to use as a screening tool," says **Siegfried Reich, PhD,** vice president of drug discovery at SGX Pharmaceuticals, another company that uses fragment-based methods. (Reich previously helped develop the HIV protease inhibitor nelfinavir at Agouron.) When it was founded in 1999, SGX was named Structural Genomix and its aim was to use high throughput X-ray crystallography to solve a record number of protein structures. But this was not sustainable as a business model. So, in 2000, the com-

"When you're talking about toxicity, it's much easier to give a compound to a rat than it is to dock against all possible proteins that are in the rat, even today," says Art Olson. "But someday, you might be able to do that. We're certainly creeping up on that."

pany changed its name to SGX Pharmaceuticals and put its crystallography power to use in drug discovery.

One of their lead candidates is a new inhibitor of BCR-ABL, a perpetually active kinase enzyme involved in chronic myelogenous leukemia, or CML. The BCR-ABL inhibitor Gleevec has had enormous success in treating CML patients, but 20 percent are resistant to Gleevec. So scientists at SGX cloned, expressed, purified, and crystallized the Gleevec-resistant protein. Then they screened their fragment library against the wild type and mutant versions of BCR-ABL to find compounds active against both. The fragment hit that eventually led to their lead candidate started with a low binding affinity of just 10 micromolars (i.e., a fairly high concentration of compound was required to bind at least half the protein).

This is where the medicinal chemists and structural biologists sit down with the computational chemists, Reich says. Computational chemists virtually build new compounds by adding chemical groups to the starting fragment. For example, they might try linking all the different simple alkyl amines to one of the fragment's "chemical handles" (sites on the fragment that easily bind to other chemical groups), Reich explains. The computer calculates the binding affinity for each iteration, until it finds one with tight binding. Specialized versions of docking programs are used to calculate the binding affinities. But because you already know exactly how the fragment binds, you start with more information than in virtual screening.

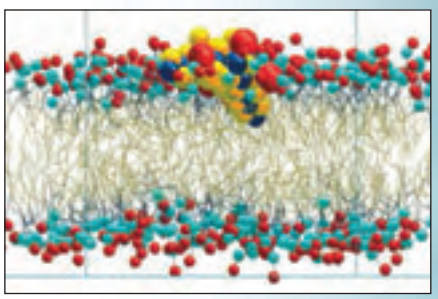
By elaborating their initial lead in this way, SGX got their first hit down to nanomolar potency—i.e. very little of the compound was required in order to bind the protein—in about three months. "That gives you a flavor for how fast this can go," Reich says.

Tricky Targets

Docking algorithms and fragment-based methods work well on soluble enzymes that are easily crystallized and contain well-defined pockets where ligands can bind—but many diseases instead involve membrane-bound receptors or protein-protein interactions.

Membrane-bound receptors transmit signals from outside to inside the cell. Because the proteins are embedded in the membrane, they cannot easily be crystallized and it is difficult to solve their structures. For example, 25 percent of the top 100 drugs on the market today target G-protein coupled receptors—including the dopamine and serotonin receptors in the brain—but the structure of only one mammalian G-protein coupled receptor is known.

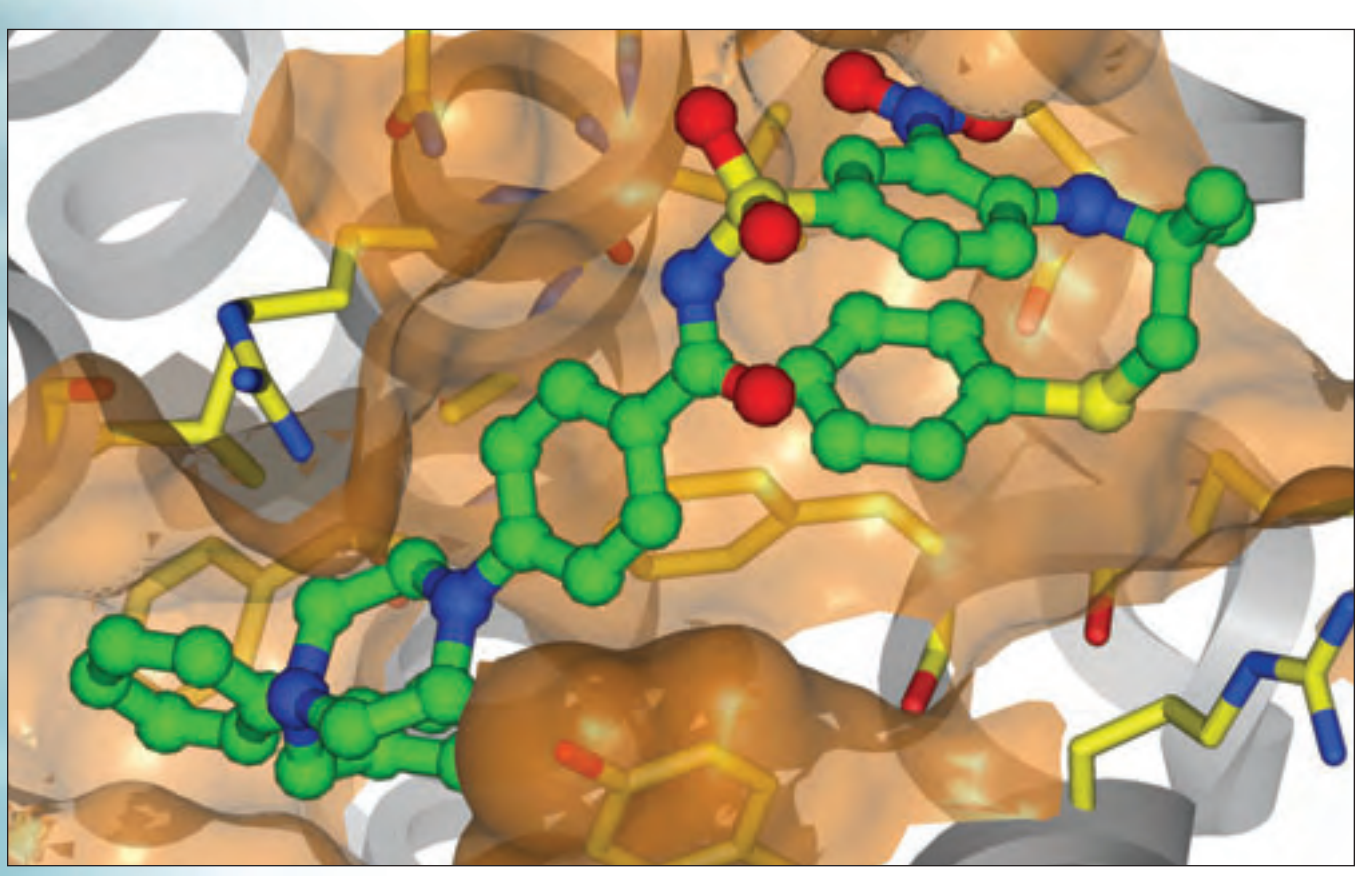
When structural information is unavailable, computational chemists use ligand-based methods to hunt for new drug leads. They superimpose a set of ligands with known activity against the target and compare their structural and chemical features. A common pattern, called a pharmacophore, emerges—key functional groups (such as hydrogen bond donors, electrostatic charges, and hydrophobic patches) must be in certain positions. This fingerprint is then used to virtually screen libraries for novel compounds with similar patterns. Ligand-based methods pre-date the structure-based methods and have helped develop many drugs, including drugs to treat high blood pressure, pain, and depression.



Tricky Target. This computer model of a bacterial cell membrane helped scientists at Polymedix design new antibiotics that mimic the action of the defensin proteins (natural proteins in the body that kill bacteria by puncturing their membranes). Courtesy of Polymedix.

Protein-protein interactions occur via surfaces that are often featureless and shallow, and binding affinities can be quite large—so it's hard for small molecules to disrupt these interactions, says Arthur Olson of Scripps Research Institute. You have to find or design drugs that can bind to multiple footholds, or hot spots, on the protein surface, which is challenging, he says. "I think that this is an area that is really still in its infancy."

But some progress is being made. Kent Stewart of Abbott Labs hopes to control BCL-2, a protein that is overexpressed in certain cancers. It blocks apoptosis (programmed cell death) and thus keeps cancer cells alive. Compared to HIV, Stewart says, which has an actual cave you can dock a molecule into, on



Cancer Interference. The oncogenic protein BCL-2 helps keep cancer cells alive via a protein-protein interaction. This Bcl-2 inhibitor—developed at Abbott using a fragment-based approach—binds to the BCL-2 protein surface and disrupts the protein-protein interaction. The compound is in late preclinical development. Courtesy of Abbott.

BCL-2, "there's no such thing as a cave; it's a very flat and open surface, so it's hard to get molecules that actually stick," So, using a fragment-based approach, scientists at Abbott linked together two fragments that bind to the BCL-2 protein surface, resulting in a potent compound that can disrupt the protein-protein interaction. The compound is now in late preclinical development.

Some companies have made these difficult targets their niche area. For example, Polymedix's mission is to develop drugs against membrane-bound targets, protein-protein interactions, and membrane-protein interactions, using a suite of computational tools specifically developed for these aims (by professors **William DeGrado, PhD,** and **Michael Klein, PhD** of the University of Pennsylvania).

Polymedix is working on a new line of antibiotics that mimic the action of defensins—natural proteins found in the body that kill bacteria.

"They work similarly to a needle or a corkscrew going into a balloon. They directly attack and perforate the bacterial cell membrane," says **Nicholas Landekic, MBA,** President, CEO, and co-founder of Polymedix. Because they do not target bacterial proteins—which can easily evolve to escape drug pressures—defensin-like drugs should not engender bacterial resistance, he says.

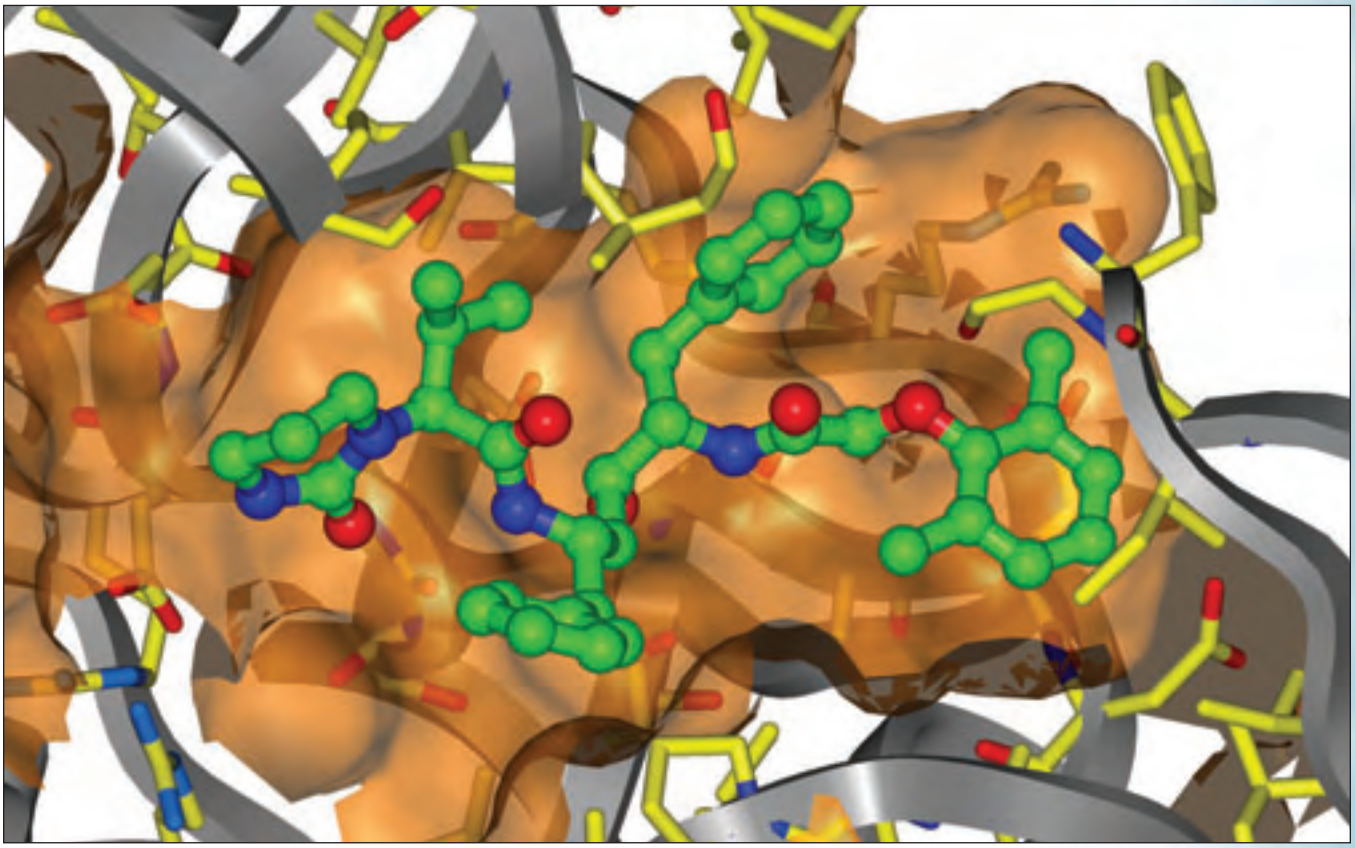
Scientists at Polymedix built a computational model of a defensin protein inserted into a bacterial cell membrane (a peptide-membrane interaction). Then they virtually transformed the defensin protein into a drug-sized compound. By swapping amino acid groups for chemically analogous small molecule groups, they shrunk the protein while preserving its chemical interactions (electrostatics, lipophilicity, etc.) within the membrane.

The result: drug leads one-tenth the size of the defensins, but about 100-fold more potent and 1000-fold more selective. "So we've been able to improve on nature," Landekic says. The compounds are now being tested in animal studies.

"We've spent less than 14 million dollars to date since starting Polymedix, so in terms of an efficiency and efficacy rate, I think that's pretty good," he adds.

Making Chemicals Into Drugs

Computer-aided methods can identify drug leads with potent activity against a target, but these compounds are far from being drugs. Drugs must also be bioavailable and safe. Safety problems derail many drugs late in development, so identifying potential safety snags early on could save considerable time and money.



HIV Protease Inhibitor. The second-generation HIV protease inhibitor, Kaletra, was developed at Abbott. Here Kaletra is shown bound to the active site of HIV protease. Courtesy of Abbott.

"How well can we evaluate bioavailability and toxicity *in silico*? It's pretty blunt and not a very popular answer: we don't do very well," Stewart says. "The biological mechanisms underlying bioavailability and toxicity are complex. So the mathematical models in those areas are still in their infancy,"

Olson agrees: We are a long way from being able to simulate a drug's effect on the entire human body. "When you're talking about toxicity, it's much easier to give a compound to a rat than it is to dock against all possible proteins that are in the rat, even today," he says. "But someday, you might be able to do that. We're certainly creeping up on that."

Computers do play a role today, however. Drugs must meet properties that fall under the ADME acronym: be Absorbed by the body, Distributed to the target tissues, and not Metabolized or Excreted too quickly. Software programs check molecules for key features (known as "Lipinski's Rule of Five") that are associated with favorable ADME profiles, such as having five or fewer hydrogen bond donors and a molecular weight below 500.

With enough computing power, scientists can also virtually screen a candidate compound against a large panel of proteins from the body, to make sure the compound will not cross react with other enzymes or receptors to cause side effects.

To ensure that molecules identified in the computer will have real-world

For the field to progress, says Anthony Nicholls, the current software needs to be more closely scrutinized—using prospective studies that directly compare the impact of computer-aided methods with more traditional drug design approaches.

"I think in the next seven to ten years, with the computational power that's coming on line here pretty soon and the steady development in algorithms, computer-aided design is going to make a huge difference," says Richard Casey.

value, computational scientists benefit from working closely with medicinal chemists during lead identification and optimization.

"Medicinal chemists would tell you that there's lots of intuition involved, so it's not all computational," says **Hans Wolters, PhD,** associate director of informatics at XDx, Inc. For example, he says that as computer scientists became more involved in making drugs, the molecular weight of candidate compounds began to creep up precipitously—to sizes that would not be easily absorbed by the human body. Medicinal chemists help recognize this type of problem early in the process.

Debating the Impact

In the past two decades, although computer-aided drug design has become an integral part of drug discovery, some remain skeptical as to whether these methods are delivering on their promise. The productivity of the pharmaceutical industry has actually declined in the past decade (The FDA approved 58 drugs from 2002 to 2004 compared with 110 from 1994 to 1996, according to the Tufts Center for Drug Development.) Though this is likely due to many factors—in particular, tightening safety standards and the enormous cost and time of clinical trials—the trend has left some wondering whether large investments in technology, including computer-aided drug design, are paying significant dividends.

Many modeling programs are unreliable, and they are not making a big difference in the real world, cautions **Anthony Nicholls**, President and CEO of OpenEye Scientific Software, which develops software for computer-aided drug design. "It's all done on faith. It's all done on the idea that 'oh, we're using computers, so it must be better,'" he says. "I think a lot of people are fooling themselves." He believes that, for the field to progress, the current software needs to be more closely scrutinized—using prospective studies that directly compare the impact of computer-aided methods with more traditional drug design approaches.

Other scientists agree that the algorithms are still being refined, but have a more optimistic outlook. They say that progress is steady and that computer-aided design is already having an impact. **Klaus Klumpp, PhD,** an associate director at Roche (who was involved in the development of the HIV protease inhibitor saquinavir), points to a suite of emerging drugs for hepatitis C virus (HCV) as a case in point.

HCV was discovered in 1989 and the virus was difficult to grow, so structural information for HCV polymerase and HCV protease became available relatively late—in the mid-to-late 1990s. By this time, computer-aided drug design was well integrated into big pharmaceutical companies. Several companies quickly identified binding sites and designed inhibitors, many of which are now in early clinical trials. "It is expected to completely change the treatment paradigm for HCV infected patients," Klumpp says.

Richard Casey, PhD, founder and chief scientific officer of RMC Biosciences, Inc., has also witnessed the dramatic effect that computers can have on drug design. His company provides computer-aided drug design services for small and mid-size pharmaceutical companies, which often lack in-house teams.

Recently, he made 3-D models and performed *in silico* docking studies for a mid-size pharmaceutical company that had identified active lead compounds but had no understanding of how they were binding the target, an RNA synthetase.

"When they saw this for the first time, it was the 'aha' effect: So that's why this compound has high activity and this compound does not. It was a real eye-opener for them," Casey says.

"I think in the next seven to ten years, with the computational power that's coming on line here pretty soon and the steady development in algorithms, computer-aided design is going to make a huge difference." □

SimbiosNews

BY KATHARINE MILLER

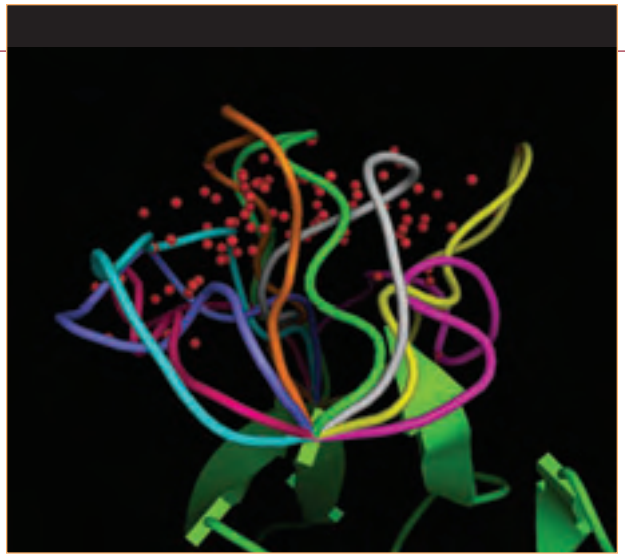
In the (Protein) Loop

In the gaps between the tight coils and flattened sheets that comprise most protein structures, flexible loops wave and bend. When crystallized, these loops can appear fuzzy in an electron density map—like moving objects captured in a still photograph. Often, loops may have an important role in a protein's function, but because they are so mobile, their structure and dynamics can be hard to study.

To better understand how protein loops move, Simbios researchers have created LoopTK, a toolkit that samples and visualizes many conformations of a loop, and provides various algorithms to manipulate and analyze loop structures. "We want to find answers that are distributed over all the motion space," says **Jean-Claude Latombe, PhD**, a roboticist and professor of computer science at Stanford University whose team developed the software. LoopTK is now available for download on the SimTK.org web site.

Latombe and his colleagues set out to place protein loops so that they correctly connect up with the protein's coils and sheets while avoiding atomic clashes in the loop and between the loop and the rest of the protein. "Solving both constraints simultaneously is the hard part," says Latombe. "That's what we do with LoopTK. And we can do it very fast. We can sample many conformations very quickly."

LoopTK relies on two techniques: seed sampling and deformation sampling. The seed sampling algorithm starts with nothing but the amino acid sequence of the protein. It then tries to place the loop in the full range of possible solutions. When several correct placements are found, the deformation sampling algorithm is used to deform the loop slightly without breaking the ends and without creating collisions among the atoms. "The two techniques are very complementary," says Latombe. "One gives you a global picture of the entire molecule in space, and the other allows you to explore specific regions of the motion space in more detail."



The Latombe group's seed sampling algorithm successfully defines the motion space for loops surrounded by empty space (as shown here) as well as for loops that are more constrained by the surrounding protein structure (not shown). In this picture, the red dots show the positions of the middle C atom of the loop in many sampled conformations, but for clarity only a small number of these conformations are displayed in their entirety. Courtesy, Jean-Claude Latombe and Peggy Yao.

Latombe's group is working with others on two applications of LoopTK. With the part of the Joint Center for Structural Genomics located at the Stanford Linear Accelerator Center, they are interpreting fuzzy electron density maps created from X-ray crystallography. "One would like to know the full range of loop conformations that could fit into this fuzziness," says Latombe. The resulting loop positions could then be submitted to the Protein Data Bank. "Biologists need to be aware of the flexibility of the loop and the uncertainty in the conformation," says Latombe. LoopTK can provide a sense of which conformations are more likely—a characterization of the distribution of possible conformations.

In a second project, LoopTK is being used for functional homology research. **Russ Altman, PhD**, chair of Stanford's bioengineering department, and his group are trying to extract structural knowledge based on partial knowledge about a protein's function. For example, if a protein X is known to bind to protein Y, LoopTK might help to infer possible conformations of the loop that are consistent with such binding.

"There might be dozens or more applications for this tool," says Latombe. "What we hope is that by putting it on the web site other people will explore those possibilities." □

DETAILS

LoopTK, a C++ based object-oriented toolkit, models the kinematics of a protein chain and provides methods to explore its motion space. In LoopTK, a protein chain is modeled as a robot manipulator with bonds acting as links and the dihedral degree of freedoms acting as joints.

LoopTK is now available for download at https://simtk.org/home/looptk. An application programming interface (API) lets users embed LoopTK in their application software.

LoopTK will be presented at the 7th Workshop on Algorithms in Bioinformatics in Philadelphia on September 8-9, 2007. (http://www.wabi07.org/)

Under TheHood

BY CHIH-WEN KAN AND MIA K. MARKEY, PhD

Mutual Information

Mutual information (MI) is defined in information theory as a measure of the dependencies between two random variables. There are many biomedical applications in which it is beneficial to quantify the information content using a measure such as MI. In classification problems, MI is used as a dependence measure to select features such that they are dissimilar from each other in order to reduce feature redundancy. MI can also be used in database retrieval. The MI is calculated between a query item and every entry in the database in order to identify the entry in the database that is most similar to the query item.

In image processing, it is also used extensively as a similarity measure for image registration and for combining multiple images to build 3D models. We will use the application domain of medical image registration to illustrate the utility of MI.

The mutual information of random variables A and B is defined as

$$I(A,B) = \sum_{a,b} p(a,b)\log\left(\frac{p(a,b)}{p(a)p(b)}\right)$$

where $p(a,b)$ is the joint probability distribution function of A and B, and $p(a)$ and $p(b)$ are the marginal probability distribution functions of A and B, respectively.

Thus, in the context of medical image registration, MI measures the distance between the joint distributions of the images' gray values $p(a,b)$ and the distribution when the two images are independent from each other. It is a measure of the dependence between the two images. Since the mutual information $I(A,B)$ is the reduction in the uncertainty of A due to the knowledge of B, when $p(a) = p(b)$, the uncertainty is minimal and the reduction of uncertainty is maximized.

In medical imaging, it is often necessary to compare images of a patient that are acquired at different times or by different modalities. For example, images may be taken pre- and post-operatively in order to assess the successfulness of a surgery. To facilitate the interpretation of such sets of images, registration—the process of aligning multiple images—is necessary. The goal of registration is to identify a transformation that maps each point in one image to the corresponding point in the other image.

One approach to image registration is based on defining landmarks or fiducial points in the images. By determining how to align those landmarks, one can determine how to transform one image to match the other. However, manual definition of landmarks is time consuming, may be difficult even for an experienced observer, and suffers from intra- and inter-reader variability.

Another approach to image registration is to determine a transformation based on a measure of the similarity of the images, such as MI. Since larger MI corresponds to more similarity of the two images, MI is maximized in registration algorithms.

In image registration, the goal is to determine a transformation of one image such that the MI between the transformed image and the reference image is maximized. Different types of transformations may be considered based on the application. The simplest class of transformations only permits rotations and translations. In medical imaging, a wider variety of scaling and shape changes are often needed, including non-linear transformations that allow for non-uniform changes across the image. An optimization algorithm is applied to dynamically search among transformations for the one with maximal MI.

In image registration, the goal is to determine a transformation of one image such that the mutual information between the transformed image and the reference image is maximized.

MI has been shown to be especially valuable for registering multi-modality images. For example, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) images of the same patient provide complementary information. Registration based on MI enables a healthcare provider to directly correlate the data from such different imaging techniques. MI has also shown promise for registering time series images. A series of images over time is often used to evaluate tissue function in addition to structure. □

DETAILS

Chih-Wen Kan is a graduate student in The University of Texas Department of Biomedical Engineering. She works on developing diagnostic decision support systems in Dr. Mia Markey's Biomedical Informatics Lab (http://bmil.bme.utexas.edu/).

PuttingHeadsTogether

The 6th Annual International Conference on Computational Systems Bioinformatics (CSB2007) coordinated by the Life Sciences Society.

WHAT: This conference is designed for any scientist interested in the interaction of biology and computing who wants to gain fast access to current research results; network with other life scientists; and listen to and meet scientific stars. CSB2007 will continue to be a five-day single track conference featuring 10 half-day tutorials, 30 referred papers plus keynote speakers, 150 posters and five full-day workshops. Special events for the evenings are being planned.

WHEN: August 13-17, 2007

WHERE: University of California, San Diego

MORE INFO:
http://lifesciencessociety.org/CSB2007/index07.html

Stanford's Bio-X Symposium: Life in Motion

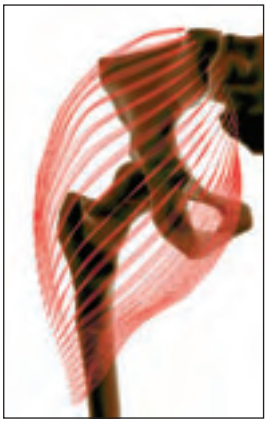
WHAT: Bio-X, Stanford's interdisciplinary life sciences initiative, hosts a major symposium each year. This year Bio-X has teamed up with Simbios—Stanford's National NIH Center for Physics-based Simulation of Biological Structures—to hold a symposium entitled, "Life in Motion". The goal of this symposium is to educate students and scientists from different disciplines about the exciting uses of simulations driven by the laws of physics and mechanics across a range of scales, from molecules to organisms. The talks will be presented by a series of experts and innovators from around the world. Confirmed speakers are: Sylvia Blemker; Joachim Frank; Robert Full; Jessica Hodgins; John Hutchinson; Roger Kamm; Mimi Koehl; Vijay Pande; Klaus Schulten; Demetri Terzoplulos.

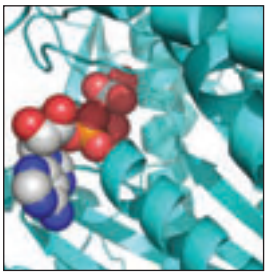
WHEN: October 25, 2007

WHERE: James Clark Center Auditorium, Stanford University

MORE INFO: *simtk.org/home/lifeinmotion*

The Pacific Symposium on Biocomputing (PSB) 2008





WHAT: The Pacific Symposium on Biocomputing (PSB) 2008 is an international, multidisciplinary conference for the presentation and discussion of current research in the theory and application of computational methods in problems of biological significance. PSB is a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, and other computational methods, as applied to biological problems, with emphasis on applications in data-rich areas of molecular biology. Papers and presentations are rigorously peer reviewed and are published in an archival proceedings volume.

WHEN: January 4-8, 2008

WHERE: The Fairmont Orchid on the Big Island of Hawaii

DEADLINES: Call for Papers—July 16, 2007;
Poster abstract submissions—Nov. 9, 2007.

MORE INFO: *http://psb.stanford.edu/*

OF NOTE: This year, Simbios will be holding a special session at PSB: *Multiscale Modeling and Simulation: from Molecules to Cells to Organisms*

WHY "PUTTING HEADS TOGETHER"?

This magazine strives to build connections among diverse researchers, all of whose work touches on biomedical computation. Because these highlighted conferences & symposia do the same thing, we are giving them a well-deserved spot in these pages. If you have a favorite conference you'd like to see appear in this magazine, let us know: editor @ biomedicalcomputationreview.org.

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

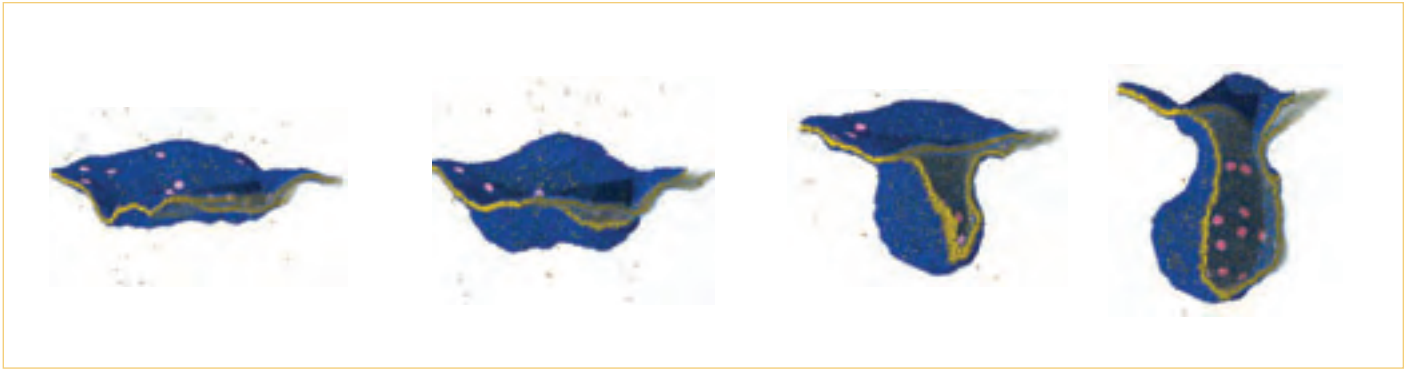
BY KATHARINE MILLER

Remodeling by Curvature

Whenever a cell needs to get rid of waste, transport materials, sort proteins, or build new organelles, membranes remodel themselves. Often that means forming small enclosed compartments called vesicles. Now researchers have gained a better understanding of that process using coarse-grained computer simulations. The work was published in the May 24, 2007 issue of *Nature*.

Researchers knew that specialized proteins are involved in triggering membranes to remodel themselves, but experimental and theoretical research could not explain how they do it. Because the energy required for major remodeling projects is greater than the energy used to bind the specialized proteins to the membrane (or to each other), some suspected that membrane curves themselves could carry the necessary energy.

Using coarse-grained simulations, **Kurt Kremer, PhD**, **Markus Deserno, PhD**, and their colleagues at the Max Planck Institute for Polymer Research in Mainz, Germany, showed that curvature-mediated attraction can indeed explain how membranes refashion themselves. Once a membrane starts to bend, proteins embedded in that membrane begin to cluster and draw the membrane into a curved shape—not unlike a vesicle.



The coarse-grained membrane simulation starts with a flat membrane containing 46,080 lipids and 36 large hemispherical "caps" (shown in pink) representing membrane proteins. Over the course of roughly one millisecond, the proteins begin to aggregate and form a large vesicle. The final image shows a cross-section of the vesicle in order to reveal the protein caps within. Courtesy of Kurt Kremer and Markus Deserno.