

NewsBytes

And the Winner Is... Computer Aided Protein Design

Each year, the American Association for the Advancement of Science (AAAS) gives an award to an outstanding paper that appeared in the pages of *Science*. This year the award—the Newcomb Cleveland Prize—went to researchers who computer-designed and then synthesized a protein that matched the design. The paper was published in the November 2003 issue of *Science*.

and people can make them with very high accuracy.”

Baker and his colleagues designed their novel protein using an iterative process. They started with a three-dimensional cartoon of a structure and used an existing design program to find the lowest energy sequence of amino acids to fit that structure. Like fitting puzzle pieces within an abstract shape, the sequence wasn't a perfect fit for the pre-designed structure. So they perturbed the structure to fit the sequence and then tried again to generate the low-

The work has helped researchers who struggle with the protein structure prediction problem, Baker says. “The prediction and design problems are closely related. The insights from the Top7 design have been helpful in developing methods for prediction, and the reverse is also true.”

Ultimately, Baker hopes to come up with novel protein machines and therapeutics. He's working on making enzymes that will catalyze reactions that aren't catalyzed in nature, and he's also trying to make better vaccines. It's

When produced in the lab, Top7 folded into a shape that very closely matched the computer design. And the shape is unlike anything found in nature.

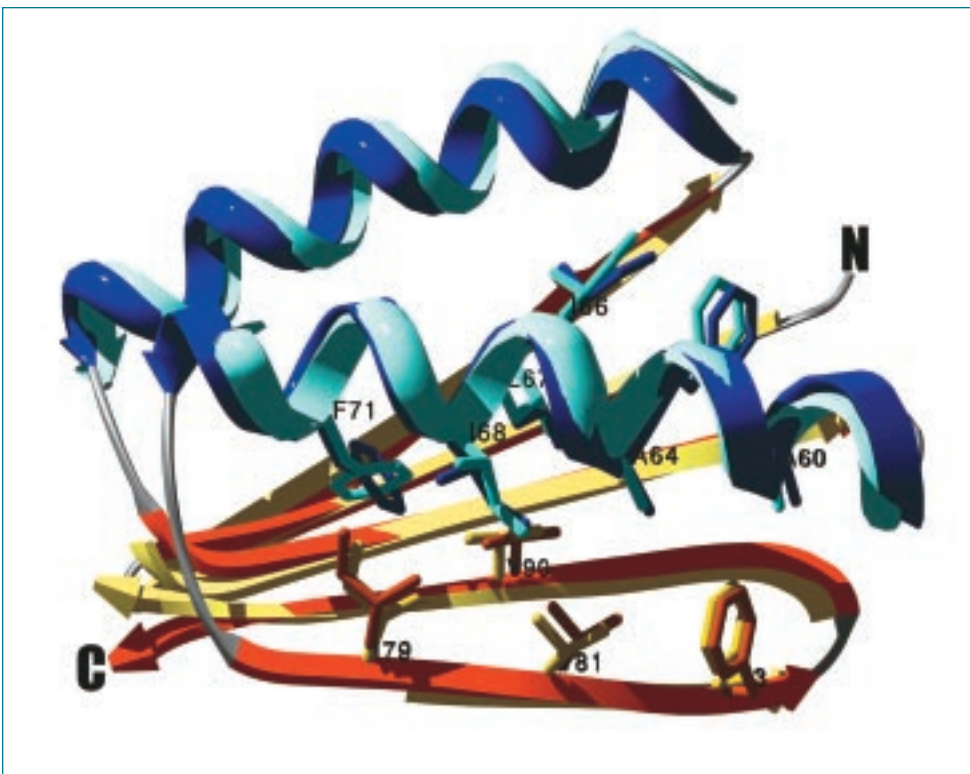
an open question whether vaccines might be best designed using a scaffold that doesn't occur in nature, Baker says. “Nature hasn't ever come up with a vaccine on its own.”

Baker foresees a time when computer-designed therapeutics will become a reality, so long as they aren't too immunogenic. The AAAS award suggests that Top7 marks an important step in that direction.

Spaced-Out Neurons

Do neurons need personal space like people in an elevator? Are they influenced by their neighbors or do they randomly find a home in the brain? If the arrangement is patterned, what is the cause of the pattern?

These are all unanswered questions

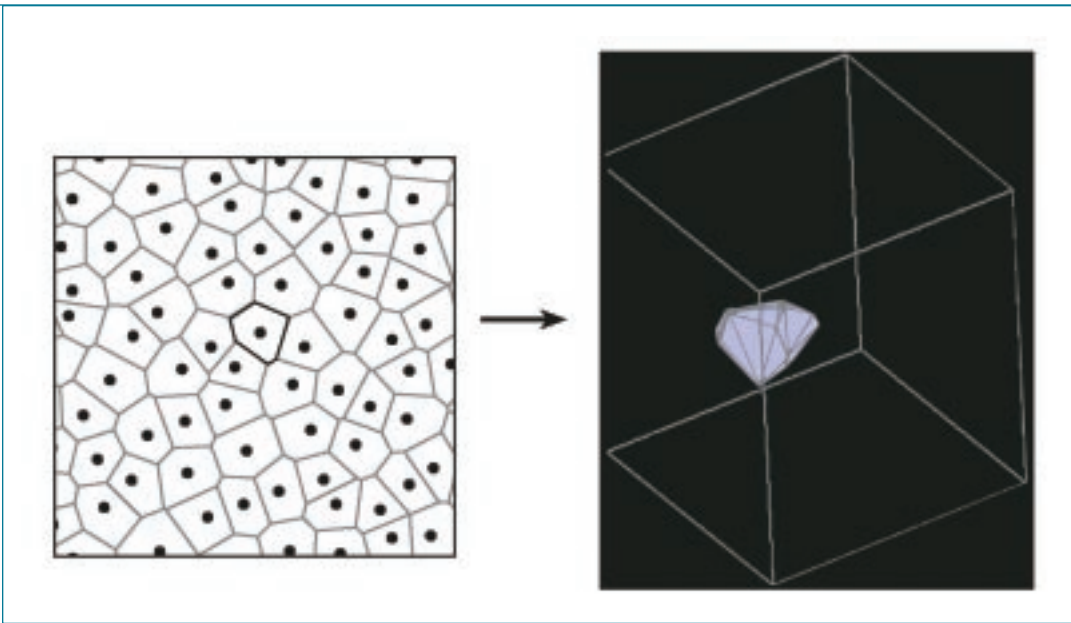


The Top7 computational model superimposed on the x-ray structure. The backbones are represented as ribbons (computational model: helices - dark blue, strands - red; x-ray structure: helices - light blue, strands - yellow), while selected amino-acid side chains in the protein core are represented as sticks. Courtesy: David Baker

“You see all these marvelous structures in nature and there has always been a question of whether there could be a larger set of proteins that don't exist in nature,” says David Baker, PhD, associate professor of biochemistry at the University of Washington and one of the paper's authors. “This paper showed such proteins do exist

est energy sequence of amino acids to fit the new structure. They went through this process ten times, and ended up with a protein they called Top7.

When they then produced that same protein sequence in the lab, Top7 folded into a shape that very closely matched the computer design. And the shape is unlike anything found in nature.



Measuring the geometrical relationships defined by the position of neurons in 3-D, as shown on the right, is far more computationally demanding than doing so for the 2-D version on the left.

a defined space with various constraints (e.g., a specified vicinity to similar, or other types of, cells) until the cells achieve the same density as is found within a region in the brain. The researchers will also generate experimental data using transgenic animals that express fluorescently marked populations of nerve cells. They will measure those neurons' x-y-z coordinates and feed them into the software

in developmental neurobiology, but that may soon change as a result of a National Institute of Mental Health grant to a group of multi-disciplinary researchers at the University of California, Santa Barbara and the University of Cambridge.

"We're creating software tools to analyze how neurons distribute themselves within the brain," says Benjamin Reese, PhD, principal investigator on the grant and a professor of psychology at UCSB. "We understand how neurons are born, the instructions governing their fate and how they then migrate, but virtually nothing about how they distribute themselves in three-dimensional space."

Reese and his colleagues have found that many types of neurons in the retina (essentially a two-dimensional space) respect one rule: they avoid being positioned near one another. This rule results in neurons being spread evenly across the retina, providing a uniform sampling of the visual scene—a characteristic required for good eyesight.

But neurons in other parts of the

brain might function under additional or completely different rules. Moreover, 3-D space is harder to model using current software. "The algorithms we've created for studying the distribution of cells in two dimensions are all Matlab-based scripts," Reese says. "Once we add the depth dimension, they become extremely cumbersome." So he and his colleagues, including co-principal investigator Steven Eglon, DPhil, a lecturer at the University of Cambridge,

"By July of 2006, we expect to have a website up and running with both two- and three-dimensional software available for others to download and use."

—Benjamin Reese.

are re-writing portions of the scripts in a lower-level language to improve computational efficiency.

The software will both simulate neuronal populations and compare the simulations to real biological data. The first simulation step: throw virtual cells into

program. The software can then determine the geometry of the simulations repeatedly, looking for the best fit to the real biological data.

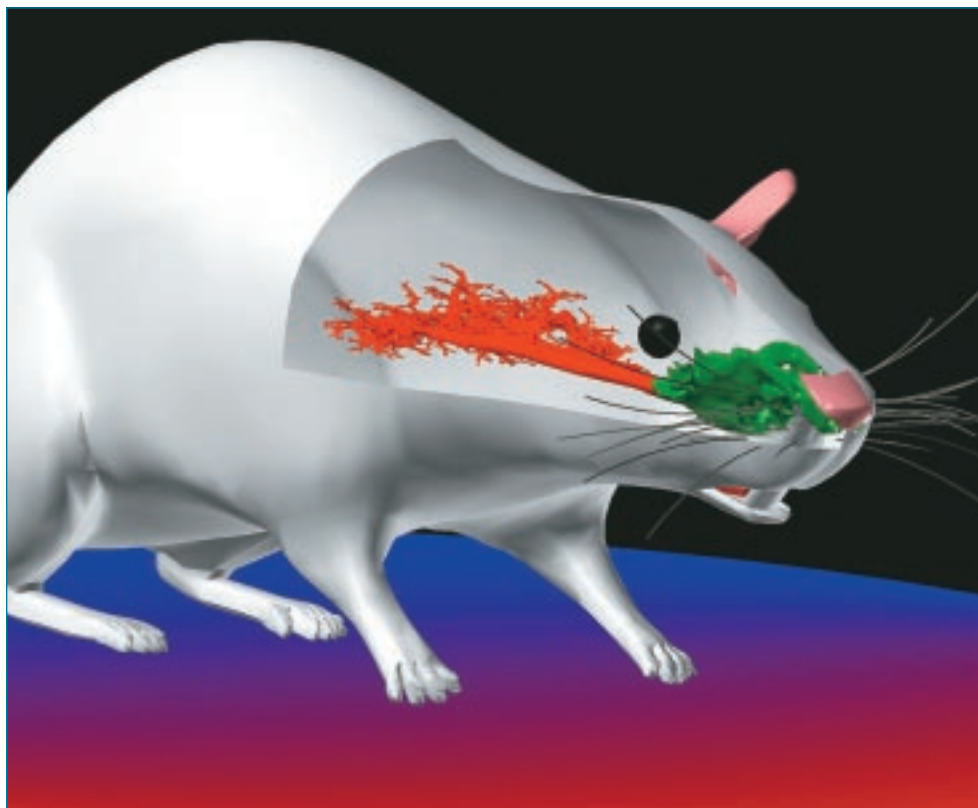
The group plans to make the software available to the public. "By July of 2006, we expect to have a website up and running with both two- and three-dimensional software available for others to download and use," says Reese.

Eventually, Reese would like to understand both cell spacing and its causes: "Is what spaces them apart a diffusible factor emitted by the cells, or is it contact-based, mediated by outgrowing dendrites?" Reese asks.

The understanding of neuron spacing may enlighten us about developmental disorders of the brain, Reese says. Mutations in genes that influence neuronal spacing may, in turn, alter the synaptic connectivity and circuit formation within the nervous system, altering brain function.

Binary Breathing

In September 2004, researchers at Pacific Northwest National Laboratory (PNNL) in Richland, Washington, received a \$10 million grant to create a three-dimensional imaging and computer model of how the respiratory tract interacts with particles carried in the air. Ultimately, the researchers hope the effort will lead to a better understanding of what happens when



In 2001, Pacific Northwest National Laboratory scientists designed a virtual computer model of the nose, larynx, and lungs of a rat in hopes of better understanding how pollutants affect those systems. Now, they're taking that work further. Courtesy: Pacific Northwest National Laboratory.

people inhale either toxic substances or medications.

“We hope to develop a good predictive tool for modeling drug delivery or dosimetry,” says Richard Corley, PhD, principal investigator and PNNL environmental toxicologist.

Corley and his colleagues have been working in this area for some time. In 2001, they developed a virtual rat lung that breathes on a computer screen. Since then, his collaborators have also been working on virtual models of primate and human lungs—models that integrate movement, as well as cellular information.

At this point, says Corley, “We can go from animal, to image, to a mesh capable of doing air flow simulations within a day or two.”

The next step—generating a computational atlas of an animal’s respiratory tract—requires that the researchers first determine how variable the animals are. “There’s some fundamental biology we’re getting out

of this,” says Corley. “How many animals do we need in order to get an atlas? How variable are we? For the first time, we can get a statistical angle on that.”

Another important step is checking the accuracy of the model through lab experiments. “The computational capabilities predict where particles go,” Corley says, “but we need to measure it as well, to validate.”

While rapidly building up sets of data showing the geometry of the respiratory tract, Corley and his collaborators are also creating function and

“We can go from animal, to image, to a mesh capable of doing air flow simulations within a day or two.”

—Richard Corley

movement models. And they want to understand what’s happening on the cellular level as well—how each of the 40 different types of cells in the respiratory tract interact with particles that land on them.

Eventually, the project will produce a web-based program for interactive simulation modeling. Right now, Corley says, it’s important for people doing this work to solve a real medical problem early on. “What’s some low-hanging fruit out there for solving? We’re looking at drug delivery.”

Shining Light on Cells

When light hits an obstacle, its scattering pattern reveals information regarding the internal structure of the obstacle. If that obstacle is a cell, the scattering pattern might indicate whether the cell is healthy or cancerous. But studying and categorizing different cells’ light-scattering properties is no small task.

Now, with help from a National Institute of General Medical Sciences grant, Jun Qing Lu, PhD, assistant professor of physics at East Carolina University, and her colleagues are studying cellular light response using a promising mathematical approach called the finite-difference time-domain method (FDTD).

“We’re looking inside the cell without opening it. If there are any changes, we should be able to see them from the outside,” says Lu.

In the past, researchers used various approximation methods to study how light scatters from cells, but these simplified approaches can only provide

limited information about highly-symmetric homogeneous bodies. Since cells are irregular in both shape and contents, a different approach was needed. “FDTD can handle any kind of shape or structure,” Lu says. “But it’s very

computationally intensive.”

FDTD has been around for a while, but it has been applied to biology only within the last few years. It’s a numerical modeling technique that can be applied to interactions between electromagnetic waves and objects whose structural details are small compared to the wavelength of light. “Inside and in the vicinity of the target object, divide your space into a 3-D grid system and divide time into small steps,” Lu says. “When the light hits the object, the electric and magnetic field distributions at each point in the grid space are calculated for each time step. Then put

“We’re looking inside the cell without opening it.” —Jun Qing Lu

everything together to calculate the scattering pattern.”

With about a million grid points, about two thousand time steps, and six finite difference equations for each grid point, it’s clear why the process requires lots of computational power. If you also want to see how the light scattering changes with different cell types or the same cell in a different life stage, that requires even more power. “Parallel computing makes it faster,” Lu says.

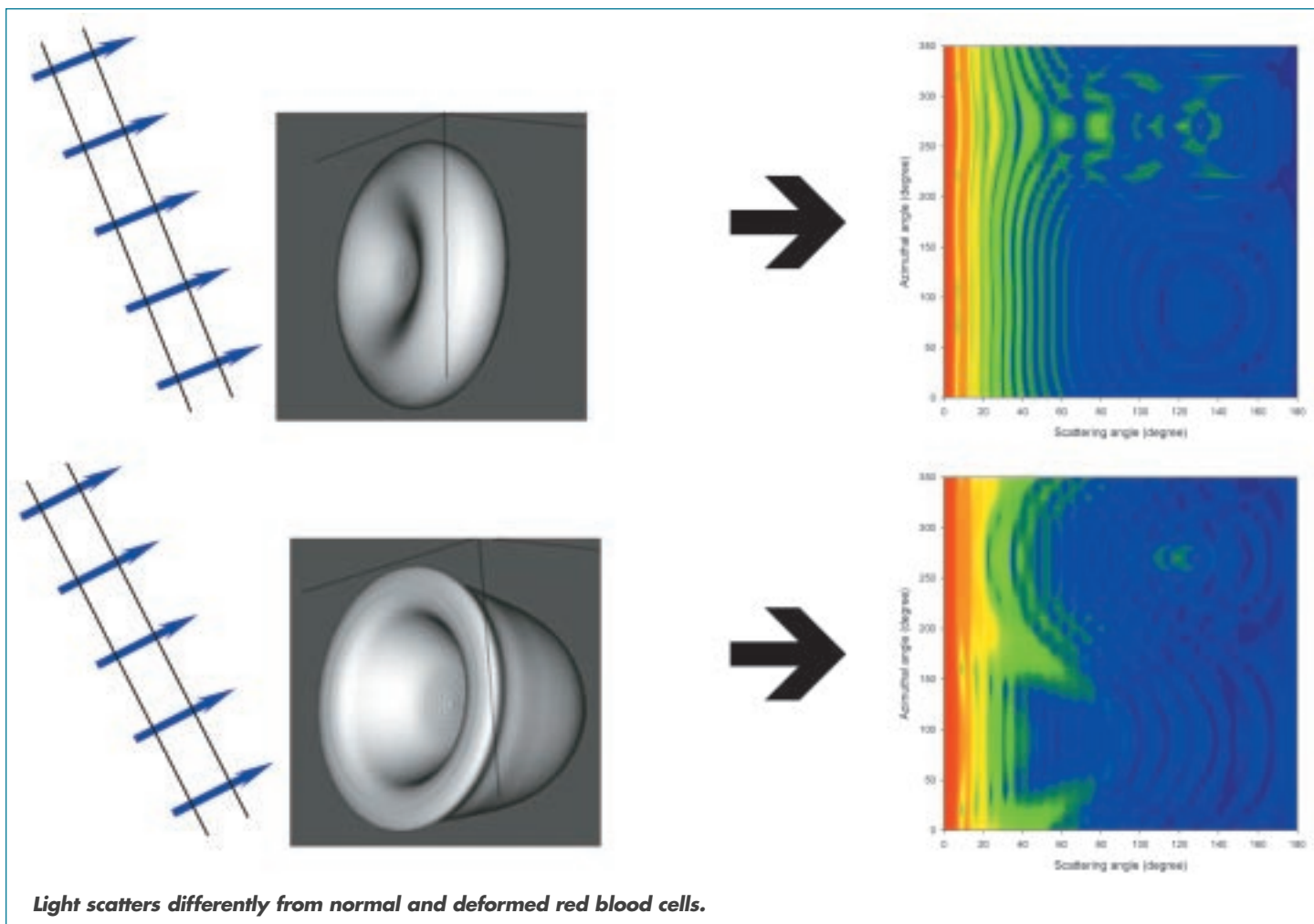
Lu and her colleagues work side by side doing computational modeling and experimental work. “I’m a theoretician,” Lu says. “But I have scientists

by my side doing experiments. So far, the models match reality pretty well.”

Thus far, Lu’s group has been studying light scattering by individual cells. Eventually, they will use the FDTD technique to do tissue studies—with hopes of distinguishing tumor from non-tumor. “People are showing lots of interest in this method,” says Lu. “It’s the right direction to pursue.”

Integrative Cancer Biology Program is Born

The National Cancer Institute launched the Integrative Cancer Biology Program (ICBP) in October 2004, providing a total of \$15 million to nine multidisciplinary centers. The goal: to use predictive cancer modeling to better understand how the disease



develops and progresses.

“Only high-level computation can handle the explosion of information that we’ve seen in the last ten years as a result of genomics, proteomics and molecular imaging,” says Daniel Gallahan, PhD, associate director of the Division of Cancer Biology at the NCI. “Cancer is such a complex problem that we really have to approach it with all the tools in our arsenal. By modeling how cancer develops from initiation to metastasis, we hope to predict and better understand the cancer process.”

Until now, cancer researchers have used computation only in a fragmented way. “Hard-core modeling hasn’t been addressed in the cancer community,” Gallahan says. “There has been some modeling of cell migration, some statistical analysis of microarrays, and some modeling of risk factors and predictors, but nothing at the level that we’re taking it to with the ICBP.”

Making the leap to more complex computation means that the cancer biologists who head up each of the nine centers had to enlist experts from other fields. “All of these grant applications had to include computation on an equal footing with biology,” Gallahan says.

Initially, the projects will be taking the steps necessary to integrate vast amounts of genomic, proteomic, imaging, and other data so that they are usable. Each center will then develop computational methods to make models that address a specific set of biological problems.

The nine centers cover the gamut of the cancer process—from initiation through signaling, DNA repair, tumor progression, invasion, angiogenesis and metastasis. One center, at Harvard, will be doing three-dimensional modeling of the tumor itself.

In principle, the ICBP should first lead to models at each step of the cancer process, but ultimately, Gallahan says, these should become modules that can be integrated. “Once these models are available in a modular way, we would then piece them together and look at how the cell transforms,” he says. “By increasing our understanding



Model of an enzyme, PanC, which is involved in the last step of vitamin B5 biosynthesis in *M. tuberculosis*. PanC is essential for the growth of *M. tuberculosis*, and is therefore a potential drug target. Credit: Mycobacterium Tuberculosis Center

of the cancer process, the models will help us identify and design better prevention and treatment strategies.”

A Crescendo of Protein Structures

A ten-year, \$600-million program known as the Protein Structure Initiative (PSI) has already, in its five-year pilot phase, greatly increased the speed at which protein structures can be determined, and added 1100 structures to the Protein Data Bank (PDB). Several thousand more may be added over the next five years. Completion of the project should lead to more rapid determination of protein function.

Medical Sciences (NIGMS), which funds the project. “Lots of interesting science will come from this large collection. It will allow people to think in structural ways when designing experiments or hypotheses. It will permit better attack on protein-folding problems. And it will lead to better and quicker work on target drug designs.”

A few thousand protein structures might not sound like a lot, given that the PDB—a federal repository for structural information about proteins—already contains about 30,000 structures. But the large majority of the banked structures are closely related to one another.

According to Jerry Li, MD, PhD,

The PSI is producing a catalog of structural information not only about a large number of proteins but about a larger variety of proteins than had previously been examined.

“The key is to make protein structures useful by getting them out there and in the hands of scientists all over,” says John Norvell, director of the PSI at the National Institute of General

program director at the Center for Bioinformatics & Computational Biology at the NIGMS, “We really have only a few thousand structures that are relatively unique,” says Li. “We

need a whole lot of structures that are not so homologous to each other.”

That’s why the PSI targets representatives of a wide range of protein families. As a result, the PSI is producing a catalog of structural information not only about a large number of proteins but about a larger variety of proteins than had previously been examined.

For 50 years, scientists have been determining the structure of proteins in order to better understand their function, but the PSI marks a shift in how structural biology is done. “The PSI is discovery-driven rather than hypothesis-driven.” Norvell says. “We’re systematically sampling the universe of protein structures.”

PSI’s efforts have also reduced the cost of determining protein structures, from \$420,000 per protein down to about \$125,000. Norvell hopes to reduce the cost even further to under \$100,000 or even as low as \$50,000.

The program is now moving into its second phase, with plans to identify more protein structures in two ways—in the lab and *in silico*. Under one set of grants, production centers will be established to elucidate 4000 or more additional protein structures over the next five years. Meanwhile, another set of grants will focus on improving methods for computational modeling of protein structures. The shapes of protein family representatives (PSI’s experimental targets) serve as rough templates for the other structures in the family, which will be determined using computer-based homology modeling.

“In the end,” says Li, “the PSI will generate a few thousand experimental structures, but it will produce tens of thousands of modeled structures.”

Spit Diagnostics

If spit could talk, it might tell us whether we’re sick or healthy.


According to David Wong, DMD, DMSc—professor and associate dean of research at the School of Dentistry at the University of California, Los Angeles—the protein profile in our saliva might distinguish a person with oral cancer or breast cancer from one

who has neither disease. That’s why the National Institute of Dental and Craniofacial Research last fall funded grants to Wong’s group and two others who will identify all of the proteins in human saliva.

Because spit can be collected non-invasively, Wong says, it is ideal for diagnostic testing. But there’s a hitch: “Saliva contains all the information we know is in blood, but at much lower magnitudes,” Wong says. “So you need different tools to measure it.”

In recent years, those tools have been developed. Seven groups around the country, including at UCLA, have been building biosensors for saliva diagnostics.

Wong and his colleagues have already determined that the RNA transcriptome profile in the saliva of people with oral cancer is markedly different from that of healthy controls. So why bother with the proteome? “At the end of the day, we’ll have genomic, proteomic, and transcriptomic information,” Wong says. “The question will be



In the long run,
Wong expects that
people will spit into
a vial to be tested
for oral cancer,
breast cancer, or
other diseases.

© PHOTOGRAPHER: CATALIN STEFAN / AGENCY: DREAMSTIME.COM

which information by itself or in combination is most sensitive to predicting disease processes.”

Already, Wong’s group has identified 310 proteins in saliva. They expect to find 1500 to 2000 before they are through. Once they have the full list, Wong and his colleagues will be identifying the protein signatures of ten high-impact systemic diseases that might be detectable in saliva, starting with oral cancer, breast cancer, and adult-onset diabetes.

When the project is complete, a web-based Salivary Proteome Knowledge Base will contain the researchers’ proteomic data along with annotations of protein function and links to other databases.

In the long run, Wong expects that

people will spit into a vial to be tested for oral cancer, breast cancer, or other diseases. “It’s not that far away,” Wong says. “The proof of concept is there.” Biotech companies are now interested in the possibility of saliva diagnostics. “Our guess—in a year there will be a commercially available system for specific selected diseases and, eventually, many more,” Wong says. “And the test will be painless.”

Extra! Extra! Read All About It

The field of biomedical computation is increasingly seen as a hot topic worthy of coverage in publications other than *Biomedical Computation Review*.

In June 2005, *The Scientist* will publish a special issue on digital biology. The publication features a “vision” piece by a working group that includes Nina Fedoroff of the Huck Institutes of the Life Sciences at Penn State and Jeff Shrager from the Carnegie Institution of Washington’s Plant Biology Department, located at Stanford. They propose a hypothesis browser—HyBrow—that would “collect the hypotheses that survive experimental testing in a new kind of knowledge base comprising models with all their supporting and contradicting data and knowledge, indexed by the hypotheses themselves.”

That issue of *The Scientist* also includes a story about housing and maintaining data storage centers. The writer

Other publications
are recognizing that
digital biomedicine
is a hot topic.

visits the server farms at Sanger Labs and interviews individuals at NCBI, CERN, and even Google, which boasts a whopping 2 petabytes (about 2 million gigabytes) in their server farm. On May 23, 2005, *The Scientist* also ran a vision piece by Lincoln Stein of Cold Spring Harbor; and on June 20, 2005, they will publish a story about open source software.

The Scientist isn’t the only publication that’s taking a growing interest in the field. In February, *Communications of the ACM* ran a series of five features called “Medical Image Modeling Tools and Applications,” guest edited by Dimitris Metaxas. The stories explore the development of a surgical simulator for minimally invasive surgery; a computer-graphics alternative to optical colonoscopy; 3-D modeling and analysis of heart motion from MRI-tagged data; efforts to develop computer-based methods for teaching anatomy; and recent efforts to develop open source image-processing tools. □

© PHOTOGRAPHER: RODOLFO ARPIA / AGENCY: DREAMSTIME.COM

