



TOP7: FROM COMPUTER-AIDED DESIGN, A NEW PROTEIN

By Pam Frost Gorder

ONE OF THE BIGGEST NEWS STORIES AT THE CLOSE OF 2003 TOUCHED ON THE BIGGEST MYSTERY FACING BIOCHEMISTRY TODAY. BOVINE SPONGIFORM ENCEPHALOPATHY (BSE), or mad cow disease, is one of many incurable diseases caused by malformed proteins in the body. The discovery of a BSE-infected dairy cow in Washington state just before Christmas caused widespread concern in the United States, and with good reason: humans can contract a version of the disease by consuming infected meat, and nine out of 10 Americans eat beef.

How protein molecules form into useful shapes—and what causes proteins to go wrong—as with BSE—is unknown. It's a puzzle called the protein-folding problem, and it's key to developing treatments for diseases as diverse as Alzheimer's, Parkinson's, cataracts, cystic fibrosis, and diabetes' most common form.

One month before this latest BSE incident, scientists took one small but intriguing step toward solving the protein-folding problem by synthesizing a protein called Top7. Yet the news barely registered with the mainstream media—perhaps because this new protein, while showing that scientists might be on the right track to a solution, defies easy description.

The Problem

A good way to explain Top7 is to explain what it is not. It is not the first artificial protein—scientists have been creating custom protein molecules in the lab for decades. And though it is not found in nature, it's not the first unnatural protein structure ever produced. It's neither a new drug, nor a useful enzyme—it performs no known function. Top7 is, however, an innovative computing method product that mimics protein evolution in nature, and its mere existence suggests that the protein-folding problem is not intractable.

Scientists have been working to understand how proteins

fold since the early '60s, when chemists at the National Institutes of Health discovered that these molecules start out as long, skinny amino acid chains that twist and loop into specific shapes to carry out chemical functions. Proteins drive the most basic cellular processes in both plants and animals, so understanding what makes them fold up one way or another could revolutionize medicine.

One way to better decipher the hows and whys of folding is to try to imitate nature by designing a protein. The task is well suited for parallel-processor computing. In fact, David Baker, head of the Howard Hughes Medical Institute laboratory that created Top7, calls his software methodology “embarrassingly parallel.”

Landscape Designer

Baker and his colleagues named their program Rosetta, after the engraved stone that helped scholars translate Egyptian hieroglyphics. Just as the Rosetta stone helped turn pictures into words, the Rosetta software shapes an amino acid chain into a protein landscape, or topology.

Scientists can spell out a chain in letters (for example, “A” for the amino acid alanine, “C” for cysteine), but the sequence itself doesn't have much meaning without knowing how those amino acids connect in three dimensions. For Top7, the topology carries more meaning than the 93-letter sequence that spells out its amino acid chain (see Figure 1).

“Our original goal was to predict protein structure, to ‘read out’ the structure from the sequence,” Baker says.

What sets Rosetta apart from other protein-design software is that users can watch a sequence evolve through successive iterations. Special computer algorithms alternate between optimizing sequence and structure to create a protein with the lowest free energy possible, because, in nature, lower-energy molecules are more stable.

In the first step, Rosetta takes a given protein structure and alters the amino acid sequence to configure that structure with the lowest possible energy. In the next step, it does the opposite—it considers the amino acid sequence to be fixed, and optimizes the structure—again, to produce a sequence with the lowest possible energy.

Rosetta Software

Laboratories can obtain a license agreement for the software through the Baker Web site: <http://depts.washington.edu/bakerpg/>. The program is available free for research purposes, as long as results are shared with the larger community.

Solving the Puzzle

Building the protein is like creating a jigsaw puzzle, Baker explains. There are 20 amino acids commonly found in proteins, and each one can rotate to form some 10 different shapes. That means that 200 options exist for each puzzle piece's shape and orientation—and a protein can contain hundreds of pieces. But while the solution must account for several alternatives, the problem easily partitions into small sections. Running Rosetta on a mainframe supercomputer would be overkill, Baker says, because the individual nodes do not need to communicate with each other at high speed. So he and his colleagues design their proteins on a standard cluster of Linux PCs with 650 CPUs.

For the Top7 experiment, the scientists wanted to see if they could create a stable protein not found in nature, so they selected a set of protein structures not found in the Protein Data Bank (www.rcsb.org/pdb/), a giant database maintained by the Research Collaboratory for Structural Bioinformatics. They searched the data bank using the Topology of Protein Structure server maintained by the University of Glasgow and the University of Leeds (www.tops.leeds.ac.uk/).

Then they ran the resulting proteins—named Top1, Top2, and so on, after the server—through Rosetta until the software couldn't optimize the structures any further.

"We stopped when the energy didn't decrease anymore from one iteration to the next, and when the sequence structure was what we'd expect for a naturally-occurring protein," Baker says.

When Top7 took its turn through Rosetta, the scientists repeated the op-

timization steps 10 times, and arrived at a structure that they predicted would have lower energy than any naturally occurring protein of the same size range.

Finally, the candidate structures became blueprints for real-life proteins synthesized in the lab. The amino acid chains for Top1 through Top6 did not crystallize into a solid structure, but the chain for Top7 did.

The scientists confirmed the structure with X-ray crystallography, in which X-rays bounce off a molecule's atoms, creating a diffraction pattern that indicates the shape. To their surprise, the Top7 molecule matched its blueprint shape. They had succeeded in designing a unique, but stable, protein that folded just the way they'd predicted.

Great Expectations

With precise control over protein structure and function, doctors could devise new therapies for diseases that arise from protein misfolds. The potential uses don't end there, though. Any industry that relies on chemical processing could benefit from the new enzymes and catalysts that could suddenly be made to order. Some experts have even speculated that protein folding could be harnessed to build electronic devices and machines for nanotechnology.

For Vijay Pande, chemistry professor at Stanford University, the creation of Top7 suggests that the protein-design community has the computational tools it needs to get the job done. He helped redefine the role that parallel processing plays in protein design with his distributed computing efforts Folding@Home (<http://folding.stanford.edu>) and Genome@Home (

Figure 1. Top7 protein. This is representative of an innovative computing method product that mimics protein evolution in nature

stanford.edu). Folding@Home had its first major success in 2002, when scientists accurately simulated the folding of a small protein called BBA5.

While Pande sees great potential for proteins in the future, he doesn't think that powerful electronics will necessarily result. "Moore's law ends when transistors are smaller than almost all proteins," he points out, so protein-based electronics couldn't be any smaller or more efficient than silicon-based electronics. "It's unclear to me whether proteins will really solve that problem, but it's fun to think about."

As to Baker's strategy of switching back and forth between optimizing protein sequence and protein structure, Pande and his team have been working to incorporate similar approaches into their calculations. "It seems like the right direction to go," he says.

David Jones, at University College London, feels that the creation of Top7 is a "really nice result," but the experiment must be repeated. Also, he says, Top7 closely resembles at least one previously known protein, so its topology lies "only just outside the 'envelope' of known structures."

The only way to tell whether the design methodology is truly robust, he says, is to use Rosetta to create different and more complicated folds.

EDITOR IN CHIEF

Francis Sullivan, IDA Ctr. for Computing Sciences
fran@super.org

ASSOCIATE EDITORS IN CHIEF

Anthony C. Hearn, RAND
hearn@rand.org

Douglass E. Post, Los Alamos Nat'l Lab.
post@lanl.gov

John Rundle, Univ. of California at Davis
rundle@physics.ucdavis.edu

EDITORIAL BOARD MEMBERS

Klaus-Jürgen Bathe, Mass. Inst. of Technology, kjb@mit.edu
Antony Beris, Univ. of Delaware, beris@che.udel.edu
Michael W. Berry, Univ. of Tennessee, berry@cs.utk.edu
John Blondin, North Carolina State Univ., john_blondin@ncsu.edu
David M. Ceperley, Univ. of Illinois, ceperley@uiuc.edu
Michael J. Creutz, Brookhaven Nat'l Lab., creutz@bnl.gov
George Cybenko, Dartmouth College, gvc@dartmouth.edu
Jack Dongarra, Univ. of Tennessee, dongarra@cs.utk.edu
Rudolf Eigenmann, Purdue Univ., eigenman@ecn.purdue.edu
David Eisenbud, Mathematical Sciences Research Inst., de@msri.org
William J. Feiereisen, Los Alamos Nat'l Lab, bill@feiereisen.net
Sharon Glotzer, Univ. of Michigan, sglotzer@umich.edu
Charles J. Holland, Office of the Defense Dept., charles.holland@osd.mil
M.Y. Hussaini, Florida State Univ., myh@cse.fsu.edu
David Kuck, KAI Software, Intel, david.kuck@intel.com
David P. Landau, Univ. of Georgia, dlandau@hal.physast.uga.edu
B. Vincent McKoy, California Inst. of Technology, mckoy@its.caltech.edu
Jill P. Mesirov, Whitehead/MIT Ctr. for Genome Research,
mesirov@genome.wi.mit.edu
Cleve Moler, The MathWorks Inc., moler@mathworks.com
Yoichi Muraoka, Waseda Univ., muraoka@muraoka.info.waseda.ac.jp
Kevin J. Northover, Open Text, k.northover@computer.org
Andrew M. Odlyzko, Univ. of Minnesota, odlyzko@umn.edu
Charles Peskin, Courant Inst. of Mathematical Sciences,
peskin@cims.nyu.edu
Constantine Polychronopoulos, Univ. of Illinois, cdp@csrd.uiuc.edu
William H. Press, Los Alamos Nat'l Lab., wpress@lanl.gov
John Rice, Purdue Univ., jrr@cs.purdue.edu
Ahmed Sameh, Purdue Univ., sameh@cs.purdue.edu
Henrik Schmidt, MIT, henrik@keel.mit.edu
Donald G. Truhlar, Univ. of Minnesota, truhlar@chem.umn.edu
Margaret H. Wright, Bell Lab., mhw@bell-labs.com

Healthy Competition

Folding@Home and Genome@Home have succeeded in part because people around the world have joined to see who can donate the most CPU time to the projects. Researchers in the field are just as competitive.

To capitalize on that competitive spirit, John Moulton, a scientist at the National Institute of Standards and Technology, organizes a biennial event called the Critical Assessment of Techniques for Protein Structure Prediction (CASP). Given a set of proteins that experimentalists will synthesize in the coming months, CASP participants race to see who can devise the most accurate structure predictions. The number of entries grows every year—in 2002, more than 250 research teams came up with 30,000 prediction sets for proteins in different categories.

No formal “winners” are declared for the overall competition, but that doesn't stop participants from using the term.

Next Steps

At the University of Wisconsin, Sam Gellman is trying to build protein-like molecules from different building blocks. Called “foldamers,” his synthetic structures hold much the same potential for medicine and micromachines. To him, Top7 hints at the possibility that scientists could one day tune structures to perform new functions, with added benefits. For example, he says, “real proteins decay over time, but an unnatural protein might not.”

While Gellman's efforts focus on molecule synthesis rather than computational design, he says he would like to see user-friendly software tools like Rosetta that could mesh with his synthetic building blocks.

Jeffrey Gray, a former postdoctoral researcher in Baker's lab, is now an assistant professor of chemical and biomolecular engineering at Johns Hopkins University. To him, Top7 shows that computer models of proteins “must have some elements of truth in them. Keep in mind, though, that the computer models still do not capture reality—we still have trouble predicting the structure of native proteins accurately, so there are elements of physics that are not understood yet. But the elements we know now are on the right track.”

Gray's lab uses Rosetta to predict the structure formed when proteins “dock,” or bind to each other. He and his colleagues typically run the software on a Linux cluster with up to 60 processors. Because docking involves multiple proteins, the problem is larger—but still easily broken down as each processor calculates the protein's complex shape. One docking simulation requires a day of computing time.

What he learns could be used in the next step of the human genome project. “Now that we have our ‘parts list’ of

all the genes in a human, can we accurately predict how these parts fit together?" he asks.

As students and postdocs from Baker's lab have moved on to other institutions, Rosetta's development has spread as well. Gray and other former group members—including Brian Kuhlman, a coauthor on the *Science* paper that announced Top7 who is now at the University of North Carolina at Chapel Hill—are taking the work around the country. Another recently departed member is Carol Rohl, now at the University of California, Santa Cruz, who largely shaped Rosetta's overall architecture.

"We have a common, Internet-based source tree, and we are trying to modularize the code so that all the components work together more seamlessly," Gray says. "In the future, we hope we can tie together protein folding, protein-protein interactions, and protein design by picking and choosing components of each."

Back at Howard Hughes Medical Institute, Baker is working to make Top7 dock with another protein—the next step toward building a functional enzyme. The process is similar to the one that created Top7. "We're doing the same sort of iterating," he says. "The concept doesn't change, but the details do."

Pam Frost Gorder is a freelance science writer living in Columbus, Ohio.

Award Winner

This year's National Academy of Science's Award for Scientific Reviewing went to Donald G. Truhlar, one of *CiSE's* editorial board members.

The US\$10,000 prize is awarded annually for excellence in scientific reviewing within the past 10 years. The 2004 field was chemical physics; Truhlar's current affiliations are Institute of Technology Distinguished Professor of Chemistry, Chemical Physics, and Scientific Computation, the Lloyd H. Reyerson Professor of Chemistry, and director of the Supercomputing Institute at the University of Minnesota, Minneapolis.

Truhlar was chosen to receive the award "for his incisive reviews on transition-state theory, potential energy surfaces, quantum scattering theory, and salvation models, which have informed and enlightened the chemical physics community for a generation."

EDITORIAL OFFICE

COMPUTING in SCIENCE & ENGINEERING

10662 Los Vaqueros Circle, PO Box 3014
Los Alamitos, CA 90720-1314
phone +1 714 821 8380; fax +1 714 821 4010;
www.computer.org/cise/

DEPARTMENT EDITORS

Book & Web Reviews: Bruce Boghosian, Tufts Univ., bruce.boghosian@tufts.edu

Computing Prescriptions: Isabel Beichl, Nat'l Inst. of Standards and Tech., isabel.beichl@nist.gov, and Julian Noble, Univ. of Virginia, jvn@virginia.edu

Computer Simulations: Dietrich Stauffer, Univ. of Köln, stauffer@thp.uni-koeln.de

Education: Denis Donnelly, Siena College, donnely@siena.edu

Scientific Programming: Paul Dubois, Lawrence Livermore Nat'l Labs, dubois1@llnl.gov, and George K. Thiruvathukal, gkt@nimkathana.com

Technology News & Reviews: Norman Chonacky, Columbia Univ., chonacky@chem.columbia.edu

Visualization Corner: Jim X. Chen, George Mason Univ., jchen@cs.gmu.edu, and R. Bowen Loftin, Old Dominion Univ., bloftin@odu.edu

Web Computing: Geoffrey Fox, Indiana State Univ., gcf@grids.ucs.indiana.edu

Your Homework Assignment: Dianne P. O'Leary, Univ. of Maryland, oleary@cs.umd.edu

STAFF

Senior Editor: Jenny Ferrero, jferrero@computer.org

Group Managing Editor: Gene Smarte

Staff Editors: Scott L. Andresen, Kathy Clark-Fisher, and Steve Woods

Production Editor: Monette Velasco

Magazine Assistant: Hazel Kosky, cise@computer.org

Design Director: Toni Van Buskirk

Technical Illustration: Alex Torres

Publisher: Angela Burgess

Assistant Publisher: Dick Price

Assistant Advertising Coordinator: Debbie Sims

Marketing Manager: Georgann Carter

Business Development Manager: Sandra Brown

AIP STAFF

Jeff Bebee, Circulation Director, jbebee@aip.org

Charles Day, Editorial Liaison, cday@aip.org

IEEE ANTENNAS AND PROPAGATION SOCIETY LIAISON

Don Wilton, Univ. of Houston, wilton@uh.edu

IEEE SIGNAL PROCESSING SOCIETY LIAISON

Elias S. Manolakos, Northeastern Univ., elias@neu.edu

CS MAGAZINE OPERATIONS COMMITTEE

Michael R. Williams (chair), Michael Blaha, Mark Christensen, Sorel Reisman, Jon Rokne, Bill Schilit, Linda Shafer, Steven L. Tanimoto, Anand Tripathi

CS PUBLICATIONS BOARD

Bill Schilit (chair), Jean Bacon, Pradip Bose, Doris L. Carver, George Cybenko, John C. Dill, Frank E. Ferrante, Robert E. Filman, Forouzan Golshani, David Alan Grier, Rajesh Gupta, Warren Harrison, Mahadev Satyanarayanan, Nigel Shadbolt, Francis Sullivan



IEEE Antennas &
Propagation Society

IEEE

Signal Processing Society

