Teams Tackle Protein Prediction

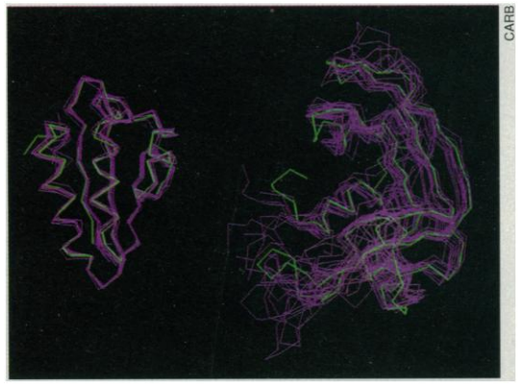
Only by pooling their efforts do many protein modelers think they can ever come up with methods to predict three-dimensional protein structures accurately

Over the next few months, some 70 research groups from around the world will crank up their computers and compete head-to-head with each other in a seemingly high-stakes competition. The "winners" will be the groups that come closest to predicting the three-dimensional (3D) structures of proteins from their amino acid sequences. There's no formal award, but the organizers of this novel event hope that the competition itself will help bring into reach a long-sought but elusive prize: the ability to determine the shape of a protein without having to coax it into forming a crystal and bombarding it with x-rays—all of which can take years of labor.

The contest, known as CASP2 (for second Critical Assessment of Techniques for Protein Structure Prediction), will formally end in December, when the contestants will get together to compare their results. At that point, the competition will turn into a giant collaboration, for the research groups will see what works and what doesn't, and each group will then build on these results to improve its own models. "Only as a community are we going to solve this problem," says modeler Manfred Sippl from the University of Salzburg in Austria.

Sippl speaks from bitter experience, for so far the models haven't worked very well. In an earlier round of competition, CASP1, which ended in October 1994, nobody really came close to predicting an accurate structure. "It had a sobering effect on the field," says physical chemist George Rose of Johns Hopkins University. But the lessons learned from CASP1—including the realization that collaboration will be essential for cracking the problem—have raised hopes that the improved models now under construction will fare better.

If they do, the achievement would have widespread implications. Structural information is crucial for determining how proteins interact, and it can be vital raw material for designing new therapeutic drugs, such as inhibitors that block the activity of potentially harmful proteins or small peptides that mimic the action of a large protein but are easier to make and administer. Knowledge of the crystal structure of a critical HIV enzyme called protease, for example, helped researchers construct inhibitors that are now revolutionizing the treatment of AIDS patients (*Science*, 28 June, p. 1882).



Match up. Predictions (magenta) based on sequence similarities matched the actual structure (green) when there were few gaps between the known and unknown protein *(left)*, but gave a poor match when there were.

The prediction problem

Until now, most protein structures have been worked out in much the same way Maurice Wilkins and Rosalind Franklin determined the crystal structure of DNA more than 4 decades ago. It is such a time-consuming process that scientists have figured out the 3D structures of only about 4000 of the more than 150,000 proteins now sequenced, according to an estimate by Tim Hubbard of the Medical Research Council's (MRC's) Centre for Protein Engineering in Cambridge, U.K.

It would clearly be a big step to go straight from sequence to structure, and researchers have been attempting to make that leap for some 30 years. By the early 1990s, many thought they were making headway. Several groups had published models that seemed capable of coming up with structures similar to those determined experimentally. A few even commercialized their products as user-friendly computer tools that anyone with a new protein sequence could use.

But even though the published models worked well for a few small, relatively uncomplicated proteins, when they were used to predict the structure of larger, more difficult proteins, they came up short. "They would beat that [test protein] to death," recalls Richard Judson, a computer scientist at Sandia National Laboratories in Livermore, California. "But their techniques were not robust enough to take the next step." That step: bona fide prediction of unknown structures.

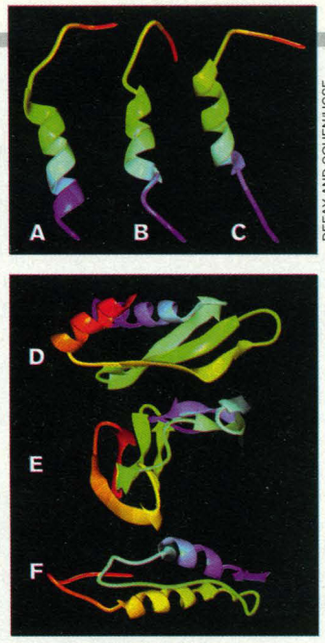
For a long time, however, far too few of those developing prediction models were willing to face up to this shortfall. "There was a lot of hype," Hubbard recalls. That's when John Moult, a protein modeler with the University of Maryland's Center for Advanced Research in Biotechnology (CARB) in Rockville, came up with the idea of asking his colleagues to try out their programs on a large number of protein sequences. "It was time to bring together researchers from all around the world and to make an evaluation of [their methods]," says Daniel Fischer, a computational biologist at the University of California, Los Angeles (UCLA). Moult and his CARB colleagues fleshed out this idea and got the Department of Energy, the National Institute of Standards and Technology, and his university to kick in funds for the contest. They then got on the phone and persuaded several friends in the prediction field to participate. With those participants signed up, Moult was then able to convince others they should take part.

Moult then canvassed crystallographers and nuclear magnetic resonance spectroscopists, who contributed 33 proteins and peptides, all of which had known amino acid sequences and the 3D structures of which they expected to solve by the end of the year. By the October deadline, 35 teams had submitted a total of 135 blind predictions of what they expected those structures to be.

The teams pitted their programs against one another in three different categories. In the first, called comparative modeling, the computer model compares a new amino acid sequence to existing ones whose 3D structures are known; if there is a close sequence match, the model bases the new structure on the old one. Computer programs in the second category, known as "threading," also rely on known protein structures, but the proteins do not have to have similar amino acid sequences. The programs simply "thread" the unknown protein's sequence—represented as a chain of spheres—along the backbone of a known protein and then determine whether the amino acid side chains protruding from the backbone fit comfortably in that arrangement. These programs—which were first developed several years ago by Sippl, Stephen Bryant of the National Library of Medicine in Bethesda, Maryland, and David Jones and Janet Thorton of London's University Col-

lege, among others—rely on the fact that proteins with very different amino acid sequences can still have very similar 3D structures.

The third, and most challenging, category encompasses ab initio approaches that try to discover structures without referring to existing sequence or x-ray crystallography databases. They rely instead on biochemical and biophysical data, such as estimates of the sizes and electrical charges of the amino acid side chains, characteristics that determine which ones are likely to tolerate being near one another. This approach would be particularly useful for proteins with no known relatives—if it can be made to work.



DEFAY AND COHEN/UCSF

From scratch. For a simple protein, two models (B, C) predicted the actual structure (A), but did not work for the more complex protein (D, actual; E, F, models) below.

The outcome of CASP1 proved humbling to contestants in all three categories. When the participants met in Asilomar at the end of that year to discuss their results, it became apparent that no group could consistently predict the proteins' structures. "What went wrong was the major part of the discussion," says Sippl.

And plenty did go wrong. The methods based on spotting and then aligning similar protein sequences stumbled when one of the pair lacked stretches of amino acids that the other had. The ab initio approaches also fell short of expectations. Some of these first fold the protein into protein substructures, such as helices and sheets, and then arrange those into folded regions called domains. Those predictions were no more accurate than ones made a decade ago with cruder methods. The next step proved even more troublesome: Within a complete 3D structure these substructures have to be linked in precise ways by loops of amino acid sequences, and the methods for constructing these loops "didn't work at all," Moult says. Other ab initio techniques, which simulate the whole folding process, worked only for very small protein fragments.

The most promising results came from the then relatively new threading programs, which were tried out on 10 of the competition proteins. Each protein was recognized by at least one of the threading programs, Moult reports, although "none [of the programs] came close to getting the whole set."

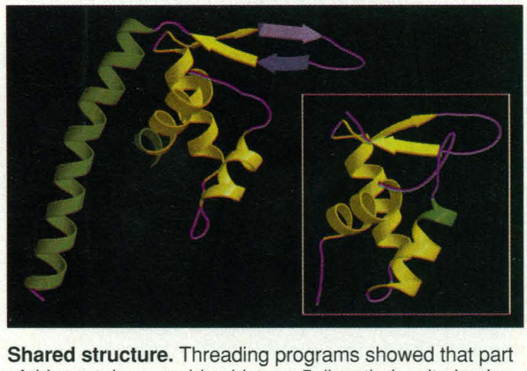
It was at this meeting that the collaborative nature of the venture kicked in. Indeed, Moult says, he encouraged all the participants to think of the competition more as a cooperative "experiment" rather than a contest to see whose method was better, and was pleased with the results. "There was a much franker exchange of views than at other scientific meetings," he says.

In those frank exchanges, researchers could see how rival programs worked where their own programs failed, and they went back to their drawing boards with lots of clues about how to refine their techniques. More threaders, for example, now incorporate into their models algorithms that make predictions about substructures—such as helices—and check whether their threading alignments allow for those substructures. "You try to combine these methods and look for consistency," explains the MRC's Hubbard.

A new experiment

Two years later—and, they hope, somewhat wiser—those protein scientists are in the midst of their second competition, CASP2 (http://iris4.carb.nist.gov/casp2). The 70 participating groups will apply their computer models to as many as 50 proteins in the same three categories as before, with one addition: Modelers can also submit structures of just the proteins' docking sites. These are areas where a protein interacts with other molecules—the substrates acted on by an enzyme are one example—and are thus particularly critical to the protein's biological function. The catch is that the prediction must include both the docking site and the properly placed docked molecule.

This year's predictions should show some improvements over those made in 1994, even if the refinements are but "baby steps," as Sandia's Judson describes his field's progress. For one thing, hundreds of new protein structures have been determined since CASP1, expanding the set of known structures and sequences that new proteins can be compared with. And the programs themselves have been refined. More and more people are using neural network programs to predict the helices, sheets, strands, and coils, for example.


SHOSHONA WODAK/UNIVERSITY OF BRUSSELS

Shared structure. Threading programs showed that part of this protein resembles histone 5 *(inset)*, despite having a different sequence.

Many more groups are also tackling threading. Not only have they made improvements to older programs based on the CASP1 results, but some of the newer threading teams say they have come up with their own ways to align a protein with a known structure and to assess the fit of the alignment. For his part, Bryant decided that CASP1 showed he didn't know what parts of a protein's structure were most likely to remain the same and therefore align well with a different sequence. So he has been developing threading models in which he first compared the structures of several related proteins and identified what is common in all of them. Then those common "cores" are what he tries hardest to fit a new sequence into.

Others have been pounding out improved ab initio approaches. For example, Jean Garnier of INRA in Cedex, France, who does both sequence matching and ab initio work, reports progress in figuring out more specifically the placements of side chains in proteins with no known homologs, in part by switching to a more sophisticated computational method called simulated annealing, which helps him identify the promising arrangements more readily. In this method, the program adds virtual "energy" to the protein and then lets the protein settle into a low-energy state. The lowest energy structure that turns up after a series of these perturbations is assumed to be the correct one. Because of the computer time required with ever longer sequences, Garnier can use this approach on only a dozen amino acids at a time. But that's long enough to see what an active site looks like, even if it can't tell how the rest of the protein folds.

Garnier's basic strategy of looking for a structure with the lowest energy and thus the greatest stability is a common one among the modelers. But at Johns Hopkins, Rose and Rajgopal Srinivasan have taken a whole different tack, capitalizing on Rose's idea that proteins have evolved to fold into the easiest arrangements. Their modeling program, called LINUS (for Local Independently Nucleated Units of Structure), can thus ignore many of the detailed energy calculations that bog down prediction programs based on stability and simply allow the amino acids to reveal their preferred 3D arrangement. "It's a very simple idea, and a very simple simulation," Rose says.

In predicting a structure, LINUS shakes up the unfolded protein, pro-

gressing three amino acids at a time down the chain. It repeats this process 5000 times and freezes in place any arrangement that occurs 70% of the time until, lo and behold, substructures, sheets, for example, emerge. The hierarchical nature of proteins suggests that this approach will eventually yield an entire protein, although Rose has not yet demonstrated this. Even so, he is eager to match up the latest version of the program, Toddler LINUS, against other prediction tools. "We signed up [for CASP2] on the first day it was announced," Rose says.

He will know how LINUS fared by the end of the year: Participants in CASP2 will compare their results at another Asilomar meeting in December. But he and fellow protein predictors may not have to wait quite that long to get a sense of the competition. Moult has set up a Web site (http://iris5.carb.nist.gov:8000) with a set of examples that will enable those with ab initio programs to check how their programs compare with others in deciding which structures are the right ones.

In addition, a center for assessing predictions, based at Lawrence Livermore National Laboratory, is distributing the CASP2 test protein sequences and collecting predictions. Funded by the Department of Energy, this center will continue these activities even after December, to provide predictors with a continuous source of new test sequences.

Other steps are also being taken to encourage collective action by the community. Hubbard and his colleagues have written a computer program called Graphical Language for Assembly of Secondary Structure (GLASS) to visualize the results of all these computational efforts. "[GLASS] allows you to read in a lot of information from different types of predictions and it generates a 3D image," Hubbard explains. The scientist can look at the image and several variations on that image, superimpose related, known structures, and determine, for example, whether atoms that should be on the molecule's surface are buried. In this way, a researcher can use all the prediction tools and perhaps be better able to come up with the correct structure answer, Hubbard adds.

Although this kind of cobbling together of techniques may not be what the predictors had once envisioned—many had hoped for the fame and recognition of having solved the protein-prediction problem by themselves—many think it will ultimately be the most successful approach. "Everyone has these daydreams of being the Einstein of protein folding," says UCLA's Fischer. "But everyone realizes that we're just nibbling at the edges. Collectively, the *group* is ultimately going to solve the problem."

–Elizabeth Pennisi

GEOPHYSICS

Earth's Core Spins at Its Own Rate

Geophysicists with ever sharper seismological tools are constantly probing beneath the surface of the Earth, but the planet's deep interior remains its most remote frontier. So when a team of seismologists announced last week that Earth's solid-iron inner core spins faster than the rest of the planet, gaining almost a tenth of a turn during the past 3 decades, it sent a tremor of excitement through the community of deep Earth researchers. A week later, scientists were still digesting the news, but one thing was certain: Geophysicists at last had direct measurements to help guide their explorations of the frontier.

"It's an exciting result," says planetary physicist David Stevenson of the California Institute of Technology. Theoreticians aren't startled that the inner core rotates faster than the rest of the planet, he says, "but the surprising thing is that its rotation rate [perhaps as fast as once in 400 years] is as big as it is." The most important implication of this result, says Stevenson, is the additional constraint it will provide on models of Earth's magnetic field, which is generated in the molten-iron outer core. The new rotation rate offers modelers their first direct measure of what's going on in the core. It could, for example, provide a measure of the strength of the field in the core, a property only guessed at until now.

ENGINEERING ANIMATION INC.

Spin control. Earth's solid-iron inner core rotates around its axis faster than the rest of the planet.

Indeed, it was a model prediction last year that prompted Xiaodong Song and Paul Richards of Columbia University's Lamont-Doherty Earth Observatory to search for some sign of inner-core rotation; they reported on the results of that search in the 18 July issue of *Nature*. If the rotation rate were anywhere near as high as predicted, the team reasoned, they would see some change in recent decades in the speed at which seismic waves pass through the inner core. They based this assumption on two previously known properties of the inner core. First, for several years seismologists have been showing how the crystalline iron of the inner core has a "grain" much like a piece of wood (*Science*, 31 March 1995, p. 1910). This grain, presumably caused by an alignment of iron crystals, is revealed by the fact that seismic waves traveling along the grain, roughly north-south, move a little faster than those traveling across the grain parallel to Earth's equatorial plane.

Second, researchers have recently shown that this grain, or anisotropy, is not quite lined up with the north-south rotation axis of Earth and the inner core. This means that if the inner core rotates at a different speed from that of the rocky mantle, the orientation of the anisotropy—and of the "fast track" for seismic waves—would change over time, circling around the high latitudes like a searchlight. Anyone monitoring seismic wave velocities along a particular route through the inner core from one polar region to the other should see those velocities change over time, as the fast track becomes more aligned or less aligned with that particular route.

That is just what Song and Richards saw when they compared the travel times of seismic waves that passed through the inner core with those that didn't. For example, waves passing just outside the inner core from earthquakes in the South Sandwich Islands off the southern tip of South America arrived in College, Alaska, just as fast in 1967 as they did in 1995. But waves passing through the inner core made the trip 0.3 seconds faster in 1995 than in 1967. Song and Richards concluded that the inner core's fast path had been swinging into alignment with the South Sandwich–Alaska route at a rate of 1.1 degrees per year.

"I think they're right about the inner core rotating," says Kenneth Creager of the University of Washington, who has worked on inner-core anisotropy. And to everyone's pleasant surprise, the rotation found by Song and Richards is "roughly the same" as seen in one model last year, says Gary Glatzmaier of Los Alamos National Laboratory, who developed the model with Paul Roberts of the University of California, Los Angeles. Their model's inner-core rotation rate is within a factor of three or so of the observed rate and in the same eastward direction, suggesting that the properties of Earth's interior assumed in building the model "may not be that bad," says Glatzmaier.

If Glatzmaier and Roberts's model works anything like the real Earth, the inner core is being dragged eastward ahead of the rest of the solid planet by the powerful magnetic drag of two intense jets in the outer core, jets analogous to the jet streams in the atmosphere. In the model, these jets are part of the