Rosetta Tackles the

Extreme Origami of Protein Folding

David Baker's model is producing remarkably accurate predictions.

BY DENNIS MEREDITH

protein folding has been called one of the great unsolved mysteries of molecular biology, a process too complex and elusive to predict with accuracy. Recently, however, a team led by HHMI investigator David Baker at the University of Washington School of Medicine has begun making predictions that one admiring expert compares to a string of home runs.

Baker has developed a computational technique, called Rosetta, that predicts the ways in which proteins, which start out as the string-like amino acid sequences that emerge from the protein-synthesis machinery,

REX RYSTEDT

undergo a folding process that might be dubbed "extreme origami."

Unlike the intricate Japanese art of paper folding, the result of protein folding is not just an elegant shape; it is also a functional one, akin to folding sheets of metal to create a working gasoline engine. The strings of amino acids collapse into the globular three-dimensional structures of enzymes and other life-sustaining cellular components.

Late last year, Rosetta proved its worth during the fourth "Critical Assessment of Techniques for Protein Structure Prediction" (CASP4). In this biennial series of experiments begun in 1994, researchers are given the amino acid sequence of target proteins and then asked to develop three-dimensional models of the final folded versions. Their predictions are compared with the actual protein structures, which have been solved experimentally by x-ray crystallography or NMR spectroscopy, but not yet published.

In the CASP4 experiment, which began in April 2000, more than 100 research groups generated three-dimensional structures for 40 candidate proteins. They presented and discussed their results at a conference in Asilomar, California, in early December (*prediction center.llnl.gov/casp4*).

"The CASP experiments have been among the most important influences in advancing this field," says Baker. "One of the problems with structure prediction is that it is all too easy to produce a program that correctly predicts the structure of a protein if you know the correct structure in advance. By challenging researchers to produce models before knowing the right answer, the CASP experiments have provided an invaluable boost to the field."

RISING TO THE CHALLENGE

Protein-structure prediction methods fall into three basic classes, explains Baker. For proteins whose sequences closely resemble those of other proteins with known three-dimensional structures, the structure can be modeled using the known protein as a template—a method known as comparative modeling. A second class of methods, called fold recognition, attempts to identify a known protein structure that is a good match for an amino acid sequence. Researchers use these methods when there is relatively little "sequence similarity" to a protein of known structure. The methods can succeed when a protein under study has a structure that is similar to one already known, but will fail if its structure is very different from those that have been determined previously.

The third class of methods is *ab initio* structure prediction, which attempts to model proteins by starting from an extended chain and folding up the sequence on the computer. These methods have the advantage that they do not depend on the existence of an already determined structure to serve as a template. Until recently, however, success in *ab initio* prediction was considered highly unlikely, says Baker.

The most exciting progress at CASP4 was in this area of *ab initio* structure prediction. As participant Peter Kollman, an expert in computational molecular modeling at the University of California, San Francisco, explained shortly before his death in late May, "The evaluators of the structures for the *ab initio* predictions gave two points for a structure which was 'among the very best,' one point for a structure that was 'pretty good' and zero if the structure was reasonably far from the correct one."

Rosetta did quite well under these ground rules. "The amazing thing is that David Baker's group had 31 points and the next-best group had 8 points," said Kollman, who compared the results to a season when Babe Ruth hit four times more home runs than any other player.

HOW ROSETTA WORKS

In research that received the Nobel Prize in 1972, Christian Anfinsen showed that a completely unfolded protein could fold spontaneously to its biologically active state, which means that a sequence of amino acids contains all of the information needed to specify its three-dimensional protein structure. In the years that have followed, scientists have verified that a large number of proteins fold spontaneously to their biologically active states. They've accounted for these results with the hypothesis that a sequence of amino acids folds naturally into a protein structure that requires the lowest amount of energy, and the folding process is essentially a search for this structure.

Since most proteins do fold spontaneously to this correct native structure, why have researchers found it so difficult to mimic the process with a computer? There have been two main problems. The first is the sheer bulk of the calculations; the number of possible conformations that a polypeptide chain can adopt is too vast to analyze with anything less than a very powerful computer. Second, it is difficult to calculate with accuracy the energy of a protein chain in the watery environment of a living cell. Rosetta solves both problems by restricting the number of conformations that it considers for each short segment of the protein chain. It only considers conformations that the segment has actually adopted in proteins whose structures have been solved. It then searches through the possible local conformations to find the combinations that produce the most favorable "low-energy interactions" throughout the protein.

By greatly restricting the universe of options, Rosetta can search for low-energy conformations more quickly and eliminate numerous conformations that might be selected incorrectly with imperfect formulas. It's a procedure that mimics the refolding of real proteins, with segments of the protein chain "flickering" among different possible conformations until an overall conformation is found in which favorable interactions exist throughout the protein.

Recently, Baker's team improved the program by incorporating an insight from experimental data on protein-folding rates. "We noted a very strong correlation between how fast proteins folded and how close together along the amino sequence were the amino acid residues that touched in three dimensions in the structure," Baker explains. "It's actually very intuitive. In a protein in which most of the contacts in the three-dimensional structure are between residues that are close to each other along the sequence, the structure can assemble much more readily than if most of the contacts are between residues that are far apart. Proteins whose contacts are primarily local, or close along the sequence, fold much more rapidly than those whose contacts are primarily 'nonlocal.'

"Because the search for low-energy conformations in Rosetta is quite short," Baker continues, "we thought it might produce structures with primarily local contacts. This indeed turned out to be the case. One of our most important advances with the CASP4 predictions was to correct this bias towards proteins with all local contacts."

DRAWING INFERENCES

Where Rosetta can be improved, says Baker, is in predicting the folding of proteins with many nonlocal interactions among amino acids and by generally increasing the accuracy of the predicted structures. Still, he concludes, Rosetta's results in the CASP4 experiments demonstrate that enormous progress has been made in *ab initio* structure prediction. "Analysis of the predicted structures showed that for the majority of proteins with no sequence similarity to proteins of known structure, we had produced reasonable low-resolution models for large fragments of up to about 90 amino acids." By contrast, he notes, at the CASP2 meeting four years ago, there were few reasonable *ab initio* structure predictions.

"One of the exciting things about the results of [CASP4] is that it has become very clear that incorporating insights from experiments into our computational methods really helped a lot," says Baker. "For a long time there's been a hope that experimental study would contribute to structure-prediction methods, but it's been only very recently that such insights have actually contributed to making protein-structure prediction better." Other researchers share this excitement. "Nonetheless," Kollman cautioned, "there is still some way to go in predicting these structures to experimental accuracy."

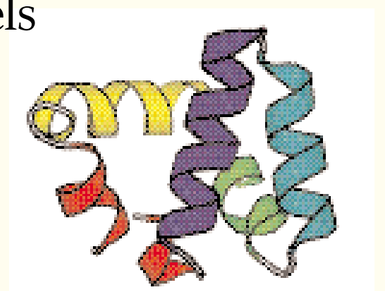
Baker acknowledges that "these three-dimensional structures are not detailed enough, for example, for structure-based drug design." Still, "they can yield invaluable insights into the function of unknown proteins," he says, "so our aim is to use our *ab initio* structure-prediction method to produce three-dimensional models for proteins of unknown function on the genome scale. A number of our CASP4 predicted structures provided insights into protein function that were not evident from the linear amino acid sequence, and we are optimistic that, using Rosetta, we can provide some insights into the functions of the significant fraction of proteins in recently sequenced genomes whose functions are not currently understood."

Baker's group is currently generating models for all large protein families whose three-dimensional structure is unknown. He and other computational structural biologists are also seeking to identify particularly interesting proteins whose structures are unknown. They're asking biologists to submit their top choices for a "10 Most Wanted" list of important proteins whose structures have not yet been solved (*predictioncenter.llnl.gov*).

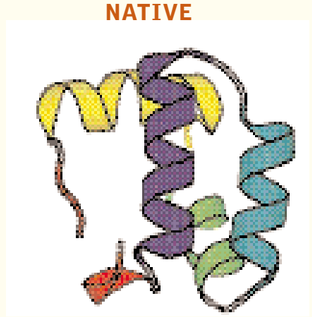
How the Models Compare with Reality

Two examples of how the protein structures predicted by Rosetta compare with the proteins' actual shapes. In both examples, the *ab initio* structure predicted by Rosetta is on the left, and the experimentally determined x-ray crystal structure is on the right. For clarity, the amino acid side chains are not shown and the protein backbone is colored to show the beginning and end of the chain. In both cases, the overall fold of the predicted structure is very similar to that of the native structure but has some details incorrect. The predictions provide valuable insights that are not evident from the proteins' amino acid sequences alone. When the protein predicted in the first example was compared with known protein structures in a database, for example, it closely resembled the structure of a protein that plays a role in killing bacteria. Sure enough, the predicted protein turned out to play a similar role, even though it has an unrelated amino acid sequence. The second example shows one domain of a large 811-residue protein that was found to resemble proteins with related functions but unrelated sequences.

MODEL



NATIVE



MODEL

NATIVE