

FEBRUARY 12, 2001

Rosetta May Hold Key to Predicting Protein Folding

A computational method developed by Howard Hughes Medical Institute investigator David A. Baker and his colleagues has proven quite successful in predicting the three-dimensional structure of a folded protein from its linear sequence of amino acids.

Rosetta, the name of the computational technique developed by Baker and his colleagues at the University of Washington, showed striking success in predicting the three-dimensional structure of proteins during the fourth Critical Assessment of Techniques for Protein Structure Prediction (CASP4).

"The power of these methods is that, since no information is needed other than the amino acid sequence, one can conceive of going through a genome and generating structures and possibly functional insights for every protein."

— David Baker

In the CASP4 experiment, which began in April 2000, more than 100 research groups generated three-dimensional structures for 40 candidate proteins. A candidate protein, or target, was considered to be eligible for CASP4 if its three-dimensional structure had been deduced through structural analysis but not yet published by researchers or made public in a protein structure database. Each research group was given the amino acid sequence of the target proteins, and they were asked to develop three-dimensional models of the folded proteins. Results of CASP4 were presented and discussed at a conference in Asilomar, California in early December.

Even a few years ago, says Baker, success in predicting how proteins assume their intricate three-dimensional forms was considered highly unlikely if there was no related protein of known structure. For those proteins whose sequence resembles a protein of known structure, the three-dimensional structure of the known protein can be used as a "template" to deduce the unknown protein structure. However, about 60 percent of protein sequences arising from the genome sequencing projects have no homologs of known structure.

Despite the lack of past success, researchers have pursued the problem of predicting three-dimensional protein structure only from the amino acid sequence—called *ab initio* prediction—because it is one of the central problems in computational molecular biology. Recently, the problem has taken on more importance as human gene sequencing efforts have provided researchers with massive amounts of raw gene sequence data

"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," Baker said. "By challenging researchers to produce models before knowing the right answer, the CASP experiments have provided an invaluable boost to the field."

The Rosetta computer algorithm for predicting protein folding draws on experimental studies of protein folding by Baker's laboratory and many others. "During folding, each local segment of the chain flickers between a different subset of local conformations," said Baker. "Folding to the native structure occurs when the conformations adopted by the local segments and their relative orientations allow burial of the hydrophobic residues, pairing of the beta strands, and other low energy features of native protein structures. In the Rosetta algorithm, the distribution of conformations observed for each short sequence segment in known protein structures is taken as an approximation of the set of local conformations that sequence segment would sample during folding. The program then searches for the combination of these local conformations that has the lowest overall energy."

The results reported using Rosetta at the CASP4 meeting revealed that enormous progress has been made in *ab initio* structure prediction, said Baker. For example, four years ago, at the CASP2 meeting, there were few reasonable *ab initio* structure predictions, he said. "In contrast, in the CASP4 experiment, analysis of the predicted structures showed that for the majority of proteins with no homology to proteins of known structure, we had produced reasonable low-resolution models for large fragments of up to about 90 amino acids.

"Interestingly, some of our predicted structures were quite similar to structures of proteins that had already been solved, and which turned out to have similar functions to the target protein, even though there was no significant sequence similarity. Thus, our predicted structures provided clues about function that could not be obtained by traditional sequence comparison methods," Baker said.

Peter Kollman, an expert in computational molecular modeling at the University of California, San Francisco, who participated in the CASP4 experiment, gives some additional perspective: "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.

"The amazing thing is that David Baker's group had 31 points and the next best group had 8 points. It is like baseball in 1927, when Babe Ruth hit 60

home runs and the runner up hit 14 [and] some *teams* didn't hit as many as he.

"Nonetheless, there is still some way to go in predicting these structures to experimental accuracy," said Kollman, "but all of us are hopeful this will advance also."

Baker concurs: "While these three-dimensional structures are not detailed enough, for example, for structure-based drug design, they can yield invaluable insights into the function of unknown proteins," said Baker. "So, our aim is to use our *ab initio* structure prediction method to produce three-dimensional models for proteins of unknown function. And using those models, we can search the database of protein structures to determine whether they are similar to proteins of known function. From this similarity, it might be possible to draw functional inferences about what those proteins do.

"We're very excited now about trying to do this on a large scale, to make functional inferences for the large fraction of proteins about which one cannot currently say anything at all," said Baker. "The power of these methods is that, since no information is needed other than the amino acid sequence, one can conceive of going through a genome and generating structures and possibly functional insights for every protein."