

OCTOBER 17, 2007

New Approach Builds Better Proteins Inside a Computer

With the aid of more than 150,000 home computer users throughout the world, Howard Hughes Medical Institute (HHMI) researchers have, for the first time, accurately predicted the three-dimensional structure of a small, naturally occurring globular protein using only its amino acid sequence. The accomplishment was achieved with a newly refined computational method for predicting protein structure, which the researchers say can also improve the detail and accuracy of protein structures generated with experimental techniques.

A detailed understanding of a protein's structure can offer scientists a wealth of information - revealing intricacies about the protein's biological function and suggesting new ideas for drug design. Researchers often rely on x-ray crystallography to determine a protein's structure - bombarding the molecule with x-rays and analyzing the resulting diffraction pattern to piece together its structure. But not all proteins are amenable to this time-consuming technique, and those that are do not always yield the atomic-level data researchers would like to have.

"The overall lesson of this paper is that protein structure prediction, at least for smaller proteins, is now good enough to generate more accurate models from experimental data such as from NMR."

— David Baker

Computational techniques — such as the one described by HHMI investigator David Baker and colleagues in an October 14, 2007, advance online publication in the journal *Nature* - can complement this approach. Baker and his colleagues at the University of Washington and the University of Cambridge in England have shown that their technique can predict protein structure with remarkable accuracy. Their methods will help structural biologists overcome a challenge commonly known as the crystallographic phase problem.

The complex algorithms the researchers developed to carry out these analyses demand a tremendous amount of computing power. More than 150,000 home computer users around the world were an integral part of the project, volunteering their computers to participate in the quest for protein structures through Rosetta@home, a distributed computing project that is based on the Berkeley Open Infrastructure for Network Computing (BOINC) platform.

Over the past decade, Baker and his colleagues have made steady progress in developing computer algorithms to predict how a string of amino acids will fold into a given protein's characteristic shape. This intricate folding is molded by the complex molecular side chains that project from the backbone of the protein and can interact in myriad ways, making such predictions far from straightforward. Among the team's chief computational tools is a program called Rosetta that calculates which of a protein's potential shapes is most efficient, or lowest in energy.

One of the thorniest problems Baker and his colleagues have faced with their algorithm is that folding proteins can get stuck in partially folded structures. Predicting protein structure involves finding a structure that has lower energy than any other structures the protein could adopt. We might have developed a protein structure that is close to the right structure, but not quite there,' said Baker. You might think we could just wiggle the structure around and shake it computationally, but sometimes the energy barriers are so high that the protein just gets stuck in that shape. So, that's where we were stymied in our technique.

In the *Nature* article, Baker and colleagues reported a new strategy of targeted rebuilding and refining to overcome this hurdle. In this method, Rosetta identifies the regions most likely to give rise to misleading interim structures and isolates them for targeted rebuilding.

It's as if you have this complex coil of rope, and there is a section that you think just doesn't behave the way it should,' explained Baker. So you just cut it out, reconnect the ends, and computationally explore different conformations of just that section until you have a better model of its behavior.

If a single round of this rebuilding and refinement does not produce the lowest-energy structure of a folded protein, the researchers repeat the analysis, using a selection process inspired by natural evolution. Each iteration produces a set of structurally different models, from which those lowest in energy are chosen for the next round of computational rebuilding and refinement. Ultimately, the lowest energy model wins out.

It's as if you had many species of animals all competing with one another, said Baker. The idea is that you take the fittest from each population and let those compete, ultimately arriving at the fittest animal of all.

The paper represents a real breakthrough, wrote structural biologist Eleanor Dodson in a *News & Views* editorial also published online by *Nature*. Dodson writes, This approach demonstrates real progress in several respects: the use

of enormous computational power; the exploitation of known three-dimensional structures; the development of powerful search algorithms that relate those structures to new sequences; and the steadily improving tactics used to determine low-energy conformations of molecules.

The benefits will be seen in structure-based drug design and in improved models for crystallographic calculations. And in the future, this method might provide structural information about intractable molecules that are difficult to study experimentally, wrote Dodson, who is at the University of York in the United Kingdom.

Baker and his colleagues demonstrated the value of their technique by using it to improve data on protein structures derived using both x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. NMR spectroscopy analyzes the magnetic properties of atomic nuclei in molecules to gain insight into their structure. While both techniques are highly useful in analyzing protein structure, the data they yield have ambiguities that predictive protein structure modeling can resolve, said Baker. Specifically, they noted that the new computational method alleviates the crystallographic phase problem for small proteins by generating high accuracy atomic level models from which phases can be estimated.

The researchers also used their technique to successfully model numerous proteins whose structures were known. For many of these, they combined their computational analysis with data from experimental techniques. In the most dramatic test, however, they accurately predicted the three-dimensional shape of a protein based only on its string-like 112-unit amino acid sequence.

That is probably the most spectacular result in the paper, said Baker. In that case all we knew was the sequence of the protein; we had no NMR data and no related structures to base a model on. So given the sequence alone, we built models, and then chose the lowest-energy models, and they were very accurate. That was the first time it has been possible to take a globular protein structure and solve it without any additional experimental information.

The overall lesson of this paper is that protein structure prediction, at least for smaller proteins, is now good enough to generate more accurate models from experimental data such as from NMR, and for generating more accurate models based on other protein structures, said Baker. And in favorable cases you can get very accurate models starting from the sequence alone.

What's more, said Baker, the project proved the scientific value of using massive numbers of individual computers to contribute to such computational efforts. The Rosetta@home project was not only scientifically invaluable, but enabled us to build a science education activity around it, said Baker. People got very interested in the calculations their computers were doing and were prompted to learn more about proteins in particular and molecular biology in general, he said.