

SEPTEMBER 16, 2005

Computers Make Big Strides in Predicting Protein Structure

Computers can predict the detailed structure of small proteins nearly as well as experimental methods, at least some of the time, according to new studies by Howard Hughes Medical Institute researchers.

The findings, which were reported in the September 16, 2005, issue of the journal *Science*, provide a glimmer of hope that scientists eventually may be able to determine the structure of proteins from their genomic sequences, a problem that has seemed insurmountable.

"For more than 40 years, people have known the amino acid sequence of a protein specifies its three-dimensional structure, but no one has been able to translate the sequence into an accurate structure," said senior author David Baker, an HHMI researcher at the University of Washington. "The reason this research is exciting is that we're showing progress in predicting the structure from the sequence. It's not that the problem is solved, but that there is hope."

"The reason this research is exciting is that we're showing progress in predicting the structure from the sequence. It's not that the problem is solved, but that there is hope."

- David Baker

Proteins are biological machines, and scientists need to determine their structures to understand how the proteins work. Now, scientists determine structures exclusively by measuring the atomic characteristics of proteins in the lab. In contrast, "in this case, we never touched a test tube," Baker said. "We gave it to a computer and said, 'go.'"

In the study, a sophisticated computer program folded 17 short strings of amino acids into 100,000 possible variations. When the researchers compared the best predictions to the actual structures solved earlier by other scientists

using experimental techniques, they had the same success rate as the best hitters in major league baseball.

"For about one-third of our benchmark set of small proteins, we generated relatively high-resolution structure predictions, with parts of the structures predicted to near-atomic resolution," said first author Philip Bradley, a postdoctoral fellow in Baker's lab. "For us, it is a real step forward to achieve structures that are in some way comparable to what you can get by experiments."

The encouraging results come from a refinement of a sophisticated computer modeling program called Rosetta, first developed several years ago in Baker's lab. The program works on the premise that proteins collapse into their lowest energy state, like a ball that rolls down a hill until it comes to rest on level ground. The energies of hundreds of thousands of possible shapes generated by the computer are computed, and the lowest energy shape is selected as the prediction.

The prediction process happens in two steps, Bradley said. The first stage uses an approximate model which allows rapid calculation of the energy and so can be carried out rapidly, while the second uses a very detailed model for which the energy calculations take much longer but are much more accurate. A large scale search through possible structures is carried out in the first stage, and promising locations are then explored in detail in the second stage.

The first stage takes advantage of the fact that all amino acids have identical sections, which form the protein backbone. The computer adds a fuzzy picture of the protruding side chains that give each amino acid its unique identity. The sequence of side chains ultimately gives each protein its characteristic shape by the environment and neighbors they prefer.

Then the computer randomly twists, loops, and bends each amino acid sequence into 100,000 different shapes based on the preferred location of the amino acids. Some amino acids tend to dive toward the watery world of the protein surface while others take cover inside the protein. The computer also accounts for the social habits of the 20 amino acids; some want to be close to each other and others like their distance.

In stage two, Rosetta replaces the fuzzy picture of the side chains with detailed, physically realistic models with all the atoms represented. From the positions of the atoms in the sidechains and the protein backbone, the computer then uses a detailed physical chemistry based force field which favors close packing of atoms and hydrogen bonding to more accurately compute the energy of the structure.

"What seems to be critical is the packing of the molecule," Baker said. "The protein fits together perfectly with no holes in the middle, and no atoms on top of each other. It's about as densely packed as it could be. It's like a

three-dimensional jigsaw puzzle."

The researchers upped their odds of finding the right match by repeating the two-step process with 50 homologs of the proteins from other genomes, such as a mouse or fly. The protocol was first tested on a blind annual prediction test considered to be the highest standard for removing bias from protein structure prediction models.

"We can't compute the energies perfectly, but the biggest problem is the search through possible shapes," Baker said. "Where we were not getting the right answer on the computer, it was almost always the case that the actual structure had the lowest energy, so we would have succeeded if we had explored this part of the space."

In a related paper published in the August issue of the journal *Proteins*, Baker and his colleagues reported that similar approaches can be used to predict the structures of protein complexes. "For the first time, computational methods are able, for a subset of cases, to produce really accurate models," he said.

Baker compares the computer simulations of the proteins to the problem of trying to find the lowest point on the surface of the Earth for the first time. A simple way to find the lowest place on the planet is to send out as many explorers as possible. The more explorers there are the more likely one of them is to stumble onto the shoreline of the Dead Sea - the Earth's lowest point on land not covered by water. Each of the thousands of computer simulations is like one explorer.

Although the 33 percent success rate reported in the *Science* paper might be good enough to secure hall-of-fame status for a baseball player, Baker is quick to point out that it is not yet reliable enough for biology. Better models will depend on both smarter exploration strategies and more computer power. "If methods stayed where we are, we wouldn't solve the problem," Baker said. "On the other hand, we would do better with 10 times more computer time."

It takes less than one minute for a protein to fold into its correct shape in cells, but one oft-repeated estimate predicts it would take longer than the age of the universe for a computer to sample all the possible conformations of a folded protein. Baker's lab already receives help from supercomputing centers in San Diego and Illinois.

More help will soon be on its way from many of the 5,000 freshman entering University of Washington this fall. Using software developed to assist the Search for Extraterrestrial Intelligence (SETI) project, the students can put their computers to work at night while they are sleeping to search the atomic landscape for the lowest energy structure of proteins.

To improve protein structure prediction further, Baker's group has also started a distributed computing project that they are hoping will be aided by members of the public. The project, called Rosetta@home, is a scientific research project that uses internet-connected computers to predict and design protein structures, and protein-protein and protein-ligand interactions. The goal is to develop methods that accurately predict and design protein structures and complexes, an endeavor that may ultimately help researchers develop cures for human diseases such as cancer, HIV/AIDS, and malaria. More information is available online at <http://boinc.bakerlab.org/rosetta>.