

*A biennial experiment  
helps scientists evaluate  
the best methods for  
predicting the structures  
of proteins.*

# The Art of Protein Structure Prediction

**F**ROM hemoglobin that carries oxygen, to enzymes and hormones that turn cells on and off, to antibodies that fight infection, proteins seem to do it all. There are many different types of proteins, each with a particular shape and function, and those shapes and functions are linked. For example, hemoglobin's shape allows it to carry oxygen; collagen's shape is ideal for connective tissue; and insulin fits in spaces like a key in a keyhole, enabling it to control sugar levels.

Disease can occur when a protein doesn't form, or fold, into its correct shape. Knowing that shape is critical for designing therapeutic drugs, for example, to treat human diseases that result from misfolding. However, predicting

protein shapes remains a daunting scientific challenge.

In 1994, Krzysztof Fidelis, a computational biologist at Lawrence Livermore, and John Moult, a professor at the University of Maryland Biotechnology Institute, received funding from the Laboratory, the Department of Energy (DOE), the National Institutes of Health (NIH), and the National Library of Medicine to organize the Critical Assessment of Techniques for Protein Structure Prediction (CASP). This biennial experiment, which is now funded by NIH with contributions from industry and international agencies, brings together groups of scientists from more than 20 countries with expertise in biology, physics, chemistry, and computer science to predict the structure of proteins.

CASP provides participants with the amino acid sequences for proteins whose structures are close to being determined experimentally by researchers. The participants then submit model structures generated by computer programs for these target proteins. Event assessors compare the prediction models with structures from experimental results.

“For decades, scientists would test modeling techniques using proteins whose structures were already known and think the problem was solved,” says Fidelis. “However, the methods would not necessarily work for other structures.” CASP allows organizers and participants to gauge which methods are most effective at predicting protein structure. At the sixth conference (CASP6), which is being held in Gaeta, Italy, this month (December 2004), participants will learn which models were most accurate for 76 target proteins.

### Solving Structures Experimentally

Predicting the shape into which a protein will fold is difficult because proteins are composed of 20 different amino acids that combine and can adopt one of several trillion shapes. Major steps in understanding the protein puzzle were taken by scientists working on the Human Genome Project,

which began as a DOE initiative in 1986 and culminated in 2000 when the DNA sequencing of the entire human genome was completed.

An organism’s genome is its full genetic instruction encoded and stored within each cell, providing all of the information the organism needs to maintain and reproduce itself. Each gene carries the instructions for making a particular protein. Once a

protein sequence has been determined, experimentalists perform the labor-intensive process of deducing its unique three-dimensional (3D) structure. To help experimentalists determine protein structure more quickly, CASP participants develop computational techniques for predicting structures.

The experimental methods most commonly used to determine a protein’s

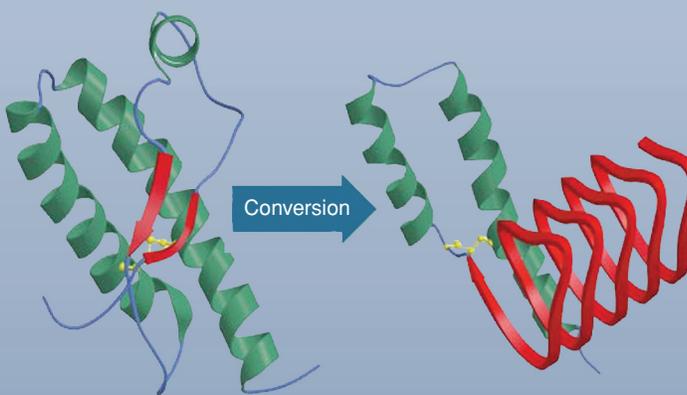
## Protein-Folding Diseases

Scientists have identified about 20 diseases caused by protein misfolding, which can be divided into two groups: diseases in which excessive quantities of wrongly folded proteins collect in certain bodily tissues and those in which a correctly folded protein is missing. The most familiar example of the first type is Alzheimer’s disease, which afflicts 10 percent of people over 65 years old and half of those over 85. Each year, Alzheimer’s kills 100,000 Americans, and about \$83 billion is spent to care for its victims.

Another example in this group is the infectious diseases, mad cow and its human form, Creutzfeldt–Jakob disease. These conditions seem to occur when normal protein particles called prions misfold. The normal human prion is a component of the membrane of healthy nerve cells that fold and are disposed of without a problem. It can, however, misfold in a particular way that triggers a domino effect in healthy prions, forcing them to adopt its incorrectly folded form.

In the second group of protein-folding diseases, the lack of a correctly folded protein means that too little normally folded protein is available to do the job. This defect is thought to be involved in diseases such as cystic fibrosis, hereditary emphysema, and some cancers.

In the past two decades, scientists have discovered that most cancers result from mutations in the genes that regulate cell growth and cell division. Forty percent of all human cancers involve a gene whose sole function appears to prevent cells with damaged DNA from dividing before the damage is repaired or, if the damage can’t be fixed, to induce the cells to destroy themselves. Thus, the key to effective cancer treatment is to design drugs that can either stabilize the normally folded structure or disrupt the pathway that leads to a misfolded protein.



A normal prion resides on the membrane of a nerve cell, where it folds and is disposed of without a problem. In an infected person or animal, the prion misfolds in a manner that triggers healthy prions to adopt the abnormal form.

structure are x-ray crystallography and nuclear magnetic resonance (NMR). In x-ray crystallography, scientists determine protein structure by measuring the directions and intensities of x-ray beams diffracted from high-quality crystals of a purified protein molecule. NMR uses high magnetic fields and radio-frequency pulses to manipulate the spin states of nuclei. The positions and intensities of the peaks on the resulting spectrum reflect the chemical environment and nucleic positions within the molecule. Unfortunately, both methods are expensive and time consuming, and some proteins are not amenable to these techniques.

Scientists have been working to solve the protein-folding mystery for decades. In research that received the 1972 Nobel Prize in Chemistry, Christian Anfinsen showed that a completely unfolded protein could fold spontaneously to its biologically active state, indicating that a sequence of amino acids contains all of the information needed to specify its 3D structure.

Protein molecules—only 3 to 10 nanometers across—can self-assemble quickly, some as fast as a millionth of a second. But this brief period is long for computers to simulate. Two difficulties arise in mimicking the protein-folding process with a computer. “First, the number of possible conformations a protein chain can adopt is too vast to analyze even with today’s most powerful computer,” says Fidelis. “Second, the estimates of molecular interactions that we use in

simulations are simply not accurate enough to render a successful prediction.”

### Software to Assist Predictors

To help them predict a protein’s fold, scientists use computer programs that estimate the molecular forces between all of the protein’s atoms and the surrounding molecules. Thus, they try to determine if those forces cause a protein to fold in a certain way. Amino acids respond differently to the watery environment of a living cell. For example, some are drawn to water, while others are repelled by it. Researchers use such characteristics to develop algorithms that help predict structure.

Some prediction programs run molecular dynamics models to calculate the forces between atoms and determine whether those forces would cause the protein to fold a certain way. Other programs cut the protein into smaller sequences and then apply an algorithm that searches for similar protein fragments from the known structures stored in the Protein Data Bank (PDB). Originally developed by DOE’s Brookhaven National Laboratory, the PDB is the industry standard for protein structure and currently lists about 25,000 protein structures.

### Classifying Structures into Groups

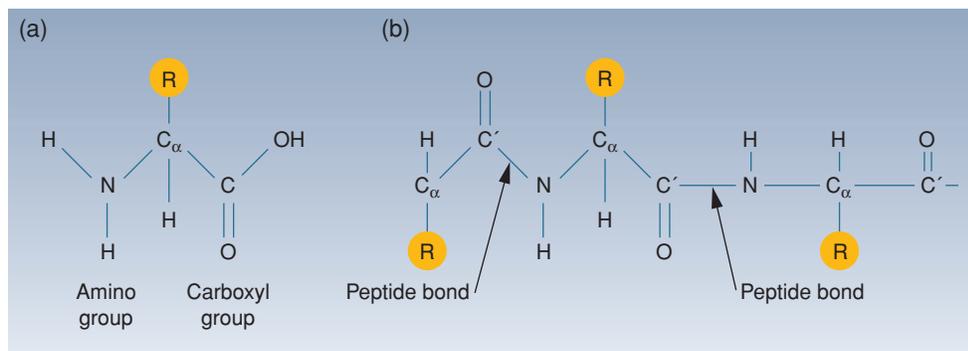
Fortunately for biomedical researchers, there are fewer classes of 3D folds than there are different sequences. In fact, many researchers believe that, because proteins are evolutionarily related, only several thousand

unique protein-structure families exist. Livermore computer scientist Adam Zemla explains, “Because there are 20 different amino acids, a medium-size protein with 300 amino acids would theoretically have  $20^{300}$  possibilities in sequence. In nature, not all combinations of amino acids can exist. Scientists estimate that the number of different protein sequences is close to a few million.”

All 20 amino acids have a central carbon atom, called carbon-alpha ( $C_\alpha$ ), to which are attached a hydrogen atom, an amino ( $NH_2$ ) group, a carboxyl ( $COOH$ ) group, and a side chain. The side chain distinguishes one amino acid from another. Amino acids join to form peptide bonds when the  $COOH$  group of one amino acid joins the amino acid group next to it to eliminate water. This process is repeated as the chain elongates. The repeating units, called residues, are divided into main-chain atoms and side-chain atoms.

The main chain is identical in all residues. It consists of a  $C_\alpha$ , to which is attached an  $NH$  group, a carbonyl ( $C=O$ ) group, and a hydrogen atom. The side-chain atoms are different for each residue and are bound to the  $C_\alpha$ . Each amino acid has a different side chain. Some sequences must be so precise that a change of even one amino acid can make a big difference, whereas in other sequences, any amino acid will work. For example, an amino acid change at one position in the protein beta-globin causes sickle cell anemia. (See the box on p 13.)

(a) Each of the 20 amino acids is composed of a central carbon atom,  $C_\alpha$ ; an amino group,  $NH_2$ ; a carboxyl group,  $COOH$ ; a hydrogen atom,  $H$ ; and a side chain,  $R$ , which is different for each amino acid. (b) Amino acids combine to form a polypeptide chain when the carboxyl group has formed a peptide bond,  $C-N$ , to the amino group next to it.



Polypeptide chains fold to form a 3D structure, which is composed of one or more regions, called domains. Domains can adopt any combination of three shapes, or secondary structures: alpha helices; beta strands, which combine to form beta sheets; and coils. Secondary structures can serve as modules for building up large assemblies of protein, such as muscle fibers, or they can form binding sites, such as those for enzymes.

### Evaluating the Difficulty

The level of difficulty in predicting a protein's structure is determined by the similarity of the protein sequence with that of a known protein structure. Scientists have classified protein-structure prediction methods into three categories. From least difficult to most difficult, they are comparative modeling (CM), fold recognition (FR), and new fold (NF). CM techniques are used when a protein's sequence closely resembles a known protein sequence in the PDB. With CM, the known protein then serves as a template. FR methods compare a specific sequence with all of the known folds in the PDB and estimate the probability of the unknown protein sequence having the same fold as that for a known sequence. NF methods are used when a protein has no detectable structural relative in the PDB. When working on proteins in the NF category, researchers use a combination of techniques to model the folds.

According to Livermore computational biologist Andriy Kryshchak, the response to the CASP experiments has been outstanding. "At CASP1, 35 groups submitted 100 predictions for 33 protein targets," says Kryshchak. "This year, at CASP6, we have 230 groups submitting more than 41,000 predictions for 76 targets."

Three independent assessors with expertise in protein folding evaluate the CASP submissions. The assessors determine the category for each target, based on level of prediction difficulty. Then each assessor

evaluates submissions in only one category. Assessment is essentially blind—that is, the assessors are not informed of a group's identity or the method used until the submissions are evaluated and scored. An example comparison is shown below.

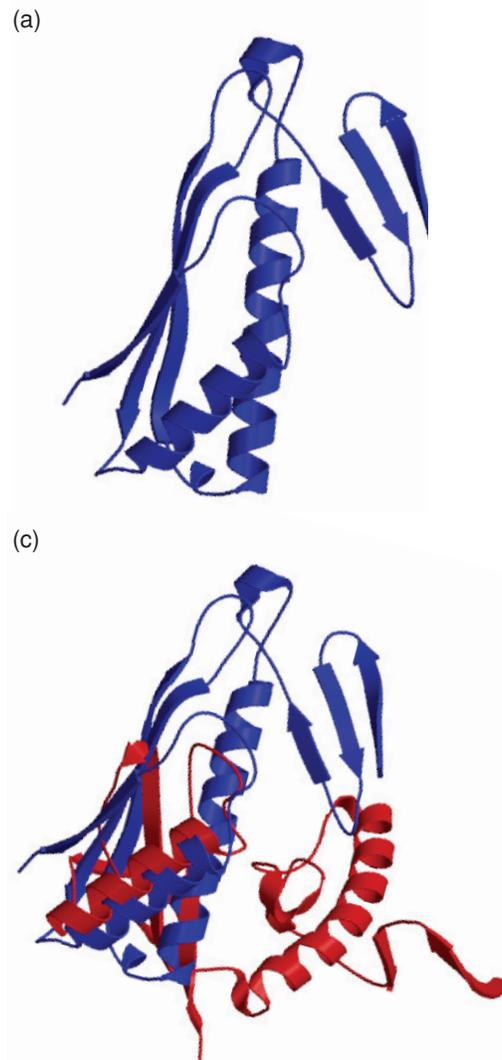
Groups can submit up to five partial or full predictions for each target. At CASP1, assessors manually evaluated the prediction models—an almost overwhelming task. In 1996, Livermore formed the Protein Structure Prediction Center to develop software tools for streamlining the process. Zemla then designed and developed computer systems that register predictor groups, collect targets, distribute target

information to the groups, verify format of submitted predictions, and provide numerical data on submissions to help assessors evaluate them.

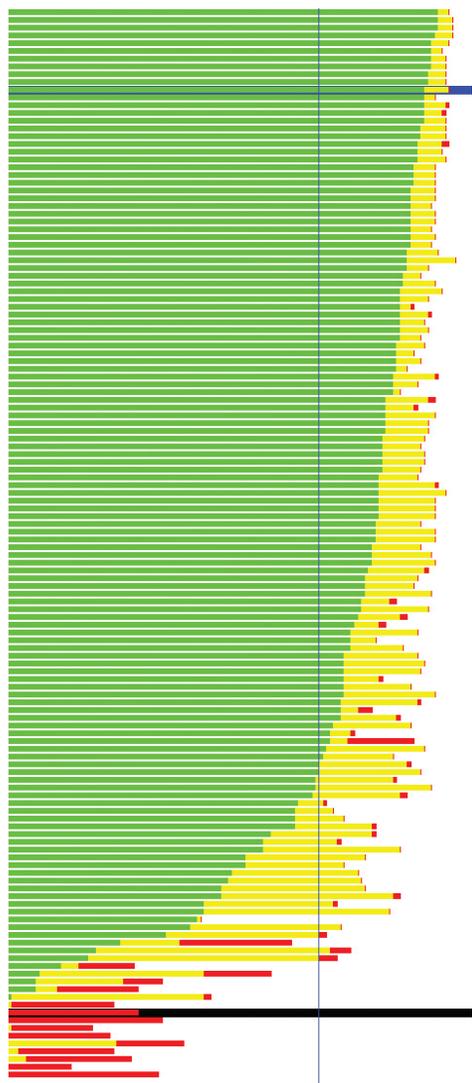
"Designing software to evaluate one model isn't difficult," says Zemla. "The challenge is determining which measures are most useful to assessors when evaluating prediction models against each other, especially when two or more partial model predictions for one target do not represent the same piece of the sequence."

### Improving Evaluation Software

Existing model evaluation software uses the root-mean-square deviation (RMSD), a



standard algorithm that compares distances between a model and a target. The RMSD is an average, so it looks for the best fit where all of the atoms for one set can be superimposed on all of the atoms for the



Each bar of this Local-Global Alignment graph represents a prediction for target TM0919. The colors represent the number of amino acid sequences that were correctly aligned (green), closely aligned (yellow), and poorly aligned (red). The predictions shown on p. 15 are highlighted: blue bar = successful model in (b); black bar = poor model in (c).

target. However, RMSD has one limitation: When two structures are similar in all but one area, this difference creates a large RMSD, which overstates the dissimilarity between the two structures.

To resolve this limitation, the Livermore team created a program that searches for local structural similarities between proteins. This method measures similarities between segments of residues rather than calculate the global (all-residues-based) RMSD. Because it allows for slight differences in residue position and focuses on matching segments, the program is better at detecting similar structures than RMSD software.

With funding from the Laboratory Directed Research and Development (LDRD) Program, Zemla also developed a software program called Local-Global Alignment (LGA). The LGA program compares distances between the protein structures for local segments and the global structure. The LGA scoring function has two components: longest continuous segment (LCS) and global distance test (GDT). The LCS algorithm identifies local regions in different proteins where the residues are similar within an RMSD cutoff. The GDT algorithm searches for the largest (but not necessarily continuous) set of equivalent residues from anywhere in the structure that fits with a distance cutoff. (See the figure on p. 17.)

LCS results are generated for a set of increasing RMSD cutoffs—0.1, 0.2, and 0.5 nanometer; for GDT results, the cutoffs range from 0.05 to 1.0 nanometer. These cutoffs are chosen because of the level of certainty in knowing a protein structure with complete accuracy: With x-ray crystallography, the level of certainty is about 0.05 nanometer, and with NMR, it is within 0.10 to 0.15 nanometer. For computer modeling, it may vary more than 0.4 nanometer.

The next challenge for the Livermore team was to convert the vast amount of numerical data generated by the evaluation

software into graphics formats that can be displayed on the Web site for the Protein Structure Prediction Center. In evaluating models, assessors use the center's Web site to define the parameters they need, such as RMSD cutoffs or side-chain residue sequences.

### Automated Servers Improve Effort

Because of its complexity, protein structure prediction has required researchers to be closely involved in the process. However, a growing number of CASP participants are using automated servers to calculate conformations of protein structure. At CASP4, assessors began evaluating the results from automated servers that were programmed with prediction algorithms. (A parallel organization, the Critical Assessment of Fully Automated Structure Prediction, developed the methods being used to assess the performance of automated servers that do not require human intervention.) At CASP5, FR-category results from the best metaservers were competitive with the best humans.

In the 10 years since the first CASP experiment, the Protein Structure Prediction Center has gathered an enormous amount of data. "For many proteins, especially for difficult targets, the best models are still not accurate enough to be useful for many applications," says Zemla. "The good news is that considerable advances have been made in the NF category and in the automated techniques used on servers."

### A Database of Models

Although the quality of predictions hasn't improved as quickly as desired, the results from each experiment are valuable. Prediction models also may prove to be useful in other research areas, such as evolutionary analyses. During the course of evolution, proteins in different organisms have diverged from a common ancestor protein. Changes have occurred in the amino acid sequence of the proteins, but their 3D fold and function have remained the same.

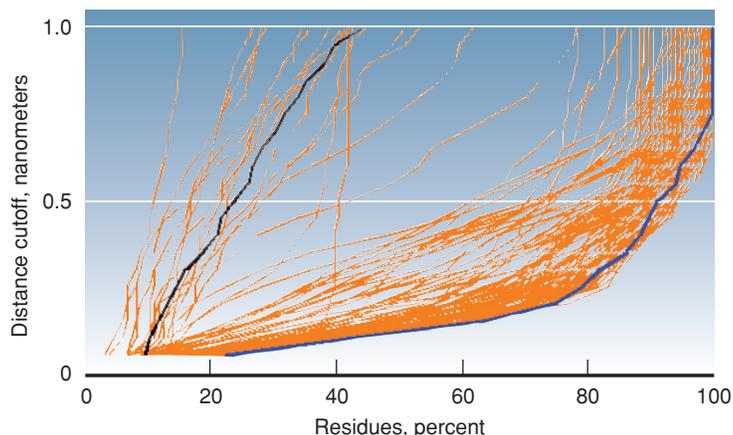
Thus, when two proteins have alignments in which the sequence identity is similar by more than 25 to 30 percent, scientists generally assume that the two sequences have diverged from the same ancestor.

More commonly, however, the sequence similarity has been lost along the evolutionary trail, so comparing structures may be the only way to identify their relationship. “Comparative structural genomics may become a powerful tool to identify the function of proteins and protein systems, helping scientists to better understand the corresponding mechanisms,” says Fidelis. “This improved understanding may, in turn, result in better control over engineered modifications that benefit such areas as environmental cleanup or producing therapeutics for human diseases.”

In 2003, the Livermore team received LDRD funding to create a database of protein models that will function as both a search tool and a model evaluator. When a protein structure is not found in the PDB, scientists can enter search criteria in the protein model database to look for models that may exist for the structure in question. The team’s prototype database currently stores more than 2,000 protein models. Team members are working with their colleagues who maintain the PDB to link the two databases.

### Software for Model Evaluation

Developing tools to evaluate models is more challenging than developing the database search function. Fidelis explains, “In CASP, we compare the differences between models and a target whose structure is known. It is much more difficult



Results from a global distance test show the percentage of amino acid sequences that each group predicted for one target within the designated distance cutoffs. The blue line represents the successful model shown in (b) on p. 15, and the black line represents the poor model in (c).

to compare two models that predict a protein whose structure is unknown and determine which one is more accurate.” The data collected from CASP may help the team evaluate a method’s performance. “We can look at which models came closest to predicting a target’s structure in a CASP experiment,” says Fidelis, “and then see which method the group used to achieve those results. Some methods work better for CM, some for FR, and others for NF.”

Many protein structures are unknown. Groups around the world are all attempting to determine the structures of proteins that are important for current research. The protein model database could help these researchers by combining all of the models to produce a single structural representation that is better than any one model alone.

The Livermore team also plans to use the database in Laboratory projects that study the function of proteins involved in the body’s response to infectious disease agents. This application could be particularly useful in support of the Laboratory’s

national and homeland security missions, for example, helping scientists develop methods to counter a bioterrorist threat.

Just as mapping the human genome led to the rapid discovery of thousands of protein sequences, researchers believe greatly improved protein structure predictions will lead to many discoveries that will benefit virtually every area of life. From designing therapeutics to developing pollution-busting bugs, the possibilities are endless.

—Gabriele Rennie

**Key Words:** Critical Assessment of Techniques for Protein Structure Prediction (CASP), global distance test (GDT), Laboratory Directed Research and Development (LDRD) Program, Local-Global Alignment (LGA), Protein Data Bank (PDB), protein folding, Protein Structure Prediction Center.

**For further information contact Krzysztof Fidelis (925) 423-4752 (fidelis1@llnl.gov). On the Web, see predictioncenter.llnl.gov.**