Exploiting the Insights from Protein Structure

The search for patterns within the growing database of protein structures is leading to new approaches to rational drug design, understanding of the relation between structure and function, and ways of solving additional structures

THE three-dimensional structures of more than 300 proteins have been solved by x-ray crystallography. Each structure represents as much as several years work, and each one has, individually, yielded insights into the rules that underlie the translation of a protein's amino acid sequence into three-dimensional structure. Sometimes referred to as the second half of the genetic code, these rules remain frustratingly incomplete, although some significant advances have been made in recent years. Participants at a recent meeting on protein engineering described some of the current developments in structural analysis and exploitation.*

In some ways, analysis of proteins is now at the same point that molecular biology was in the 1970s, when it was struggling to appreciate the full potential of recombinant technology. The development of computer programs that can recognize features of interest in protein structures is still in an early stage. More surprising is the existence of only one facility, the Brookhaven Protein Data Bank, for the storage of protein structure data, as compared to the many DNA sequence archives.

As rules governing protein folding become clearer, the process of going from sequence to structure will rely less on the individual insights of investigators and become more of an established routine. Eventually, it will be possible to identify a particular structural conformation or function that is needed in medicine or industry—and then design a protein that will have it. Meanwhile, the trick is how to use combinations of data already generated to help pose and, in some cases, answer questions about protein design and structure.

For example, trimethoprim (TMP) is important clinically as an antimicrobial drug that inhibits the enzyme dihydrofolate reductase (dHFR) in bacteria and is frequently used to treat urinary tract infections. One drawback, however, is that at high concentrations it can attack the human dHFR as well as the microbial enzyme and thus be toxic to the patient.

Investigators at Burroughs Wellcome Co., North Carolina, have been comparing different dHFR structures in an attempt to design and synthesize TMP analogs with greater specificity for the bacterial enzyme. The TMP molecule has a certain flexibility—its three-dimensional conformation changes when it binds to dHFR. According to Lee Kuyper of the Wellcome Research Laboratories, he and his colleagues reasoned that if they could synthesize a rigid TMP analog that was already frozen into the conformation that it assumes when binding a bacterial dHFR, it would be less able to attack the human enzyme.

The Protein Data Bank is "not funded at a level commensurate with its importance to the scientific community."

Crystal structures for two complexes had been determined several years previously: one was a complex between *Escherichia coli* dHFR and TMP, and the other was a complex of chicken dHFR and TMP. By analyzing the differences between the conformations of TMP in the two structures, the Burroughs Wellcome group has designed and synthesized a molecule that shows promise as the parent for a new family of TMP derivatives.

Comparisons of crystal structures are also revealing unexpected relationships that may provide insight into ways that structural elements determine function. For example, different proteins can contain the same structural motifs. Triose phosphate isomerase has a highly characteristic domain, called the TIM barrel, which has the catalytic region (the active site) positioned at one end. This barrel is shared by 13 different proteins, which have no sequence similarities and, in many cases, catalyze very different reactions.

It has long been dogma that similarities in structure imply similarities of function, so the TIM barrel presents a challenge. According to Gregory Petsko of Massachusetts Institute of Technology (MIT), unless some functional similarities can be uncovered, the TIM barrel family would represent the "biggest uncoupling of structure and function" that has been reported. In this regard, it is curious that some of the enzymes in this family appear in pairs in their metabolic pathways: in other words, they catalyze consecutive transformations. There may therefore be some kind of association among these enzymes, with the structural similarity making it easier to link together different enzymatic steps in a particular "assembly line."

Triose phosphate isomerase is one of the most efficient chemical catalysts characterized, transferring protons at a rate of more than 5000 times per second. "Anything you try to do to it will make it worse!" Petsko states. By contrast, a commercially important member of the TIM barrel family, glucose isomerase, which is used to make the high fructose corn syrup vital in the soft-drink industry, is a much less efficient enzyme, catalyzing essentially the same reaction but only once every 2 seconds. This difference in catalytic power may reflect an underlying difference in catalytic mechanism. By comparing the structures of the two proteins, Dagmar Ringe of MIT, in collaboration with Petsko and Gerard Tiraby of the University of Toulouse, France, is trying to redesign glucose isomerase so that it will use the highly efficient mechanism of TIM.

In some cases it is very difficult to get enough pure protein to start growing a crystal for structural analysis. And, even if there is enough protein, it is impossible to predict how long it will take to grow a crystal. If the rules of folding were completely known, then three-dimensional structure could be "read" from the primary sequence. But, for now, what do you do until the crystal arrives?

Many investigators are attempting to combine the information about the 300 structures that have already been solved with the more than 8000 protein sequences that are available. The basic premise is that if two proteins are similar in their amino acid sequences then they will tend to fold into similar structures. The identical sequences are matched up and then analytical techniques, such as energy minimization, are used to determine the effects of the substitu-

*"Advances in Gene Transfer Technology: Protein Engineering and Production," Miami Bio/Technology Winter Symposium, 8 to 12 February 1988.

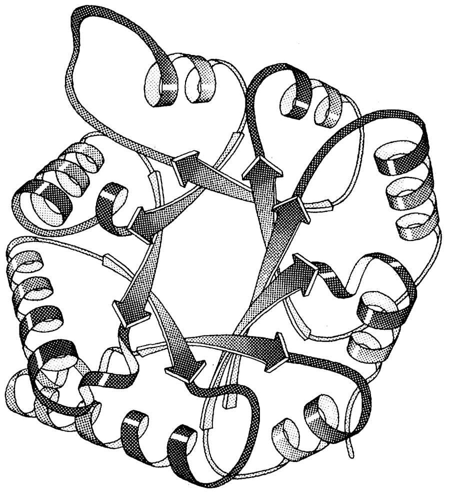
tions. The value of "sequence gazing," as Robert Fletterick of the University of California at San Francisco describes it, is that it enables predictions of which amino acids will be actually involved in catalysis or binding at the active site, or involved in governing the specificity or regulation of a reaction.

Renin is one example of a protein that is being actively modeled by several research teams, including Michael James and his colleagues at the University of Alberta. The enzyme is important because it catalyzes the first in a series of reactions that lead to elevated blood pressure. Models of renin based on the solved structures of a similar aspartic proteinase, pepsin, have led to the design of nonpeptide inhibitors that mimic an intermediate of the reaction of renin with its substrate. It is possible that these inhibitors may eventually be useful in treating hypertension.

The validity of James' model, which is still being refined, has been at least partially confirmed by x-ray crystallographic studies done in collaboration with investigators at California Biotechnology Inc. in Mountainview, California. In the region of the active site, the difference between the location of an atom in the model and its actual location as determined by the crystal structure (called the root-mean-square deviation) is only about 0.4 angstrom which is considered to be very close indeed.

It is one thing to start out by comparing two very similar proteins, but some investigators are trying to go even further: specifically, by building proteins from fragments that do not necessarily have the same sequence but have the same main-chain carbon-backbone structure. T. Alwyn Jones of the University of Uppsala, Sweden, has been pursuing this approach—building proteins from spare parts—extensively. He derived algorithms for fragment fitting, which were used, with an electron-density map as a guide, to build retinol-binding protein from 22 fragments.

Shoshana Wodak of l'Université Libre de Bruxelles, in collaboration with Michel Claessens and others at Plant Genetic Systems in Brussels, Belgium, has built models of triose phosphate isomerase and lactate dehydrogenase from overlapping, unrelated fragments that were frequently less than six residues long. The ability to piece together fragments according to the distances between pairs of nuclei may become important as a way of determining three-dimensional protein structures from data generated by nuclear magnetic resonance (NMR) spectroscopy. Because NMR spectroscopy is done on proteins in solution, the technique bypasses the time-consuming process of crystallization, thus potentially speeding up the business of solving a structure.



Through the TIM barrel: *the active site of triose phosphate isomerase.* [*Courtesy Jane Richardson*]

Another approach is to look at structural elements derived from a whole set of proteins and determine whether a particular amino acid or kind of amino acid tends to be found in a particular location. For instance, Janet Thornton of Birkbeck College, London, has been systematically characterizing the amino acids in particular loops and turns that link such large-scale structural elements as alpha helices or beta sheets. These loops are the structures that vary the most among the proteins within a family and they may be important in binding and recognition. George Rose and his colleagues at Pennsylvania State University, as well as Jane and David Richardson at Duke University, have been trying to determine the sequence specificity that governs the formation of alpha helices.

Of course, the fly in this ointment is establishing how good these predictions really are, and there are investigators in the field who feel that local sequence information cannot incorporate global or long-distance interactions between amino acids. There are many existing models where the root-mean-square deviation is on the order of several angstroms. This kind of difference is not good enough to use in designing inhibitors, however, as deviations of 0.5 angstrom or less seem to be needed.

There is a consensus within the field that the greatest value of modeling is to help make predictions that can then be tested experimentally. "A model provides a three-dimensional environment in which the chemist can be very creative—a way to focus synthesis in more fruitful directions," says Jonathan Greer of Abbott Laboratories, Illinois.

None of this would be possible without easy access to the ever-increasing protein data being generated. The Brookhaven Protein Data Bank, located in the chemistry department of the Brookhaven National Laboratory, is currently the only structural library available, with distribution centers in Osaka, Japan, and East Melbourne, Australia. According to the director, Thomas Koetzle, the library now includes structural information on 330 proteins, with another 70 proteins about to be entered. Submission of information to the library is voluntary, although there is a vocal movement within the scientific community to make submission of atomic coordinates to the Data Bank mandatory before a scientific paper can be published.

It is hard to imagine that the small staff at the Data Bank will not become swamped by the flood of coordinates that will need to be entered in the next few years. As Carl Pabo of Johns Hopkins University observes, the Protein Data Bank is "not funded at a level commensurate with its importance to the scientific community."

While coordinates based on x-ray crystallographic studies represent the primary data entered, the library also includes a bibliography for each entry, amino acid sequences, comments from the authors, and locations of secondary structure elements such as loops and turns (although the identification of these elements depends on the subjective criteria of the contributors).

It is an enormous advantage to investigators to have all of the data in one place and in one format. However, the library is designed to be an archive, not an analytical tool, so that it is necessary for a user to have separate software to compare features among different proteins. This can be frustrating—almost like having a telephone directory in which all the names are associated with the right phone numbers but have been put in at random instead of alphabetically. In a few months, Thornton and her colleagues, under the sponsorship of the Protein Engineering Club (a group of British industrial firms who finance research at universities), will make available a database of protein structures that will have features tables for structural and sequence elements, which should make it easier to pull out the motif of interest from a set of proteins.

The advent of computer techniques for handling protein structural information, which includes not only storage but also the ability graphically to represent models and predict the effects of changes in sequence, has brought clear progress in elucidating the second half of the genetic code. ■

BARBARA JASNY

Barbara Jasny is an editor with Science.