Protein Structure

Carl-Ivar Brändén, Uppsala

(Dept. of Molecular Biology, Swedish University of Agricultural Sciences)

A number of important biological phenomena can be rationalized in terms of the function of specific protein molecules. This function is, in turn, entirely dependent on the three dimensional structure of the proteins. Examples of such biological functions are enzyme catalysis, regulation and replication of genetic information, muscle action and the immune response. Knowledge of specific protein structures and of the principles that govern the formation of these structures is thus important for our understanding of biological processes. Furthermore, such knowledge has now become essential for applications in biotechnology with the recent discoveries of simple genetic methods to introduce mutations at a pre-determined site of a gene. From knowledge of the structure of one protein we can now engineer changes which give new functions to this protein better suited to our needs either in industrial processes or in living organisms.

Linear Structure of Proteins

Protein molecules are linear chains of small units, amino-acids. There are twenty different such units, all of which couple to each other in the same way. All protein molecules thus have a chemically similar backbone. The side chains that protrude from this backbone are, however, different for these twenty units. Viewed in this way one can make an analogy with strings of pearls with twenty different colours. The number of amino acids in a chain varies from 50 to 1000 with a mean value around 300.

The sequence of amino acids along the chain determines the three dimensional structure and thus the function of each protein molecule. These sequences are genetically determined in the form of nucleotide sequences along the genes. A translation machinery exists in each cell to convert this genetic code to protein sequences. Most of our genes are thus information storage units for the amino acid sequence of protein molecules. It has been estimated that we have around 100000 such genes in the human chromosomes, and a smaller number of genes in biological systems that we believe are less complex than human. About 1000 of these have been sequenced and we know the three dimensional structure of around 100 of the corresponding proteins.

The amount of variability of a sequence is very high, although in comparison with the number of sequences that 300 amino acids can produce in theory, namely 20^{300}, the number of different sequences actually used in present day biological systems is almost vanishingly small.

The Folding Process

Each amino acid sequence found in Nature folds into a unique three dimensional structure. This folding process can be reproduced in a test-tube but not simulated in a computer. The most important unsolved problem in molecular biology today is understanding this process. We cannot from first principles predict the three dimensional structure of a given amino acid sequence. We know that the final structure is at some energy minimum, either global or, more probably, at a local energy minimum determined by semi-stable states along specific folding pathways. The energy difference between a folded and an unfolded protein is however small, less than one kcal per amino acid residue. The degrees of freedom on the other hand are large, at least two per amino acid in the backbone and from zero to five in each side chain. These degrees of freedom are rotations around covalent chemical bonds. Present theoretical models for energy calculations that can be used with present computer capacities are not sufficiently accurate to cope with this problem. A fair number of attempts have been made and the results are on the whole frustratingly disappointing.

Efforts are now being concentrated on a different and more pragmatic approach, namely to treat the folding problem as a pattern recognition problem. In order to understand the rationale behind this approach we should, however, first have a look at the results from the experimental side of protein structure determination.

Experimental Protein Structure Determination

X-ray crystallography is the only method which so far has given accurate models of the three dimensional structure of protein molecules. This method has in the past been limited by four essential bottlenecks; crystallization, data collection, phase determination and structure refinement. Crystallization of protein molecules is still essentially an art and not a science although recent experience has shown that crystallization is facilitated when a protein is produced in large quantities by genetic cloning methods. Data collection and structure refinement have benefited from recent technological advances. The rate of data collection has been improved by one or two orders of magnitude by using commercially available area detector systems and synchrotron radiation sources for large molecules. Structure refinement has also been improved by the use of advanced graphic display systems with efficient software packages and super-computers for the least squares refinement of the several thousand parameters.

The phase problem remains a bottleneck. The diffraction pattern of a crystal gives direct information on the amplitudes of the diffracted waves whereas direct information on the phase relationships is obtained via two experimental

methods: isomorphous replacement with heavy metals and measurement of the anomalous scattering of intrinsic or extrinsic atoms in protein crystals. There has been no significant breakthrough leading to easier or faster applications of these methods.

Nevertheless, indirect information on the phase relations is contained in the diffraction pattern, and statistical methods for extracting this information have been very successfully developed for small molecules by the winners of the 1985 Nobel Prize in Chemistry, J. Karle and H. Hauptman. For protein molecules more powerful methods are now being explored mainly based on the maximum entropy method. In addition one can use the fact that protein crystals frequently contain several identical molecules in the asymmetric unit. The diffraction pattern from such crystals contains information on the transforms of several identical molecules in different orientations. This redundant information can be used to improve existing limited phase information.

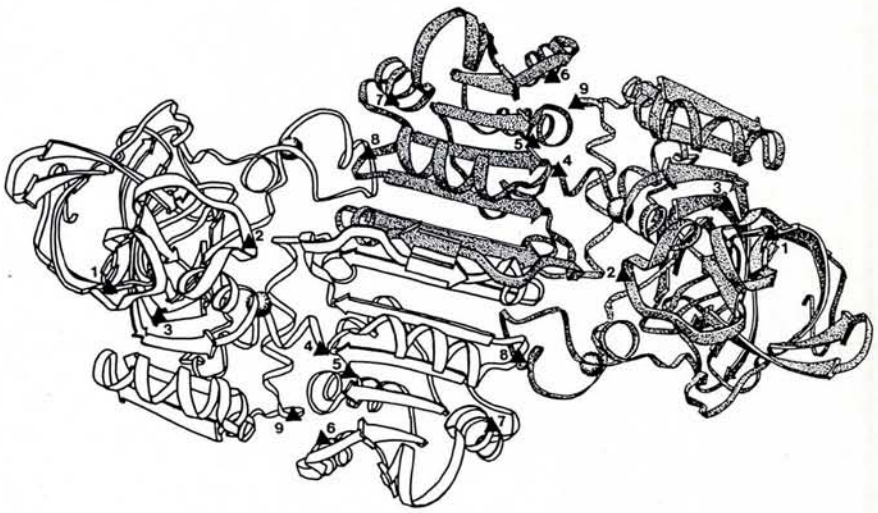

Fig. 1 — A schematic representation of the structure of an enzyme, alcohol dehydrogenase. The dark part is one protein chain of the dimeric molecule. Each of the identical protein chains is divided into two domains: an α/β domain in the middle region and a β domain at the two ends. The active site regions of the enzyme are in the cleft regions between the domains.

Architectural Principles of Protein Structures

The most important result that has emerged from the structures which have been experimentally determined is that certain general architectural principles of protein structure have been discovered. These principles are as follows: A large protein chain is usually folded into several globular units, domains, of a size around 150 residues. Within each such domain the protein chain is divided into a number of segments where each residue has the same backbone conformation. This repeat arrangement is of two different types, one giving rise to a helical arrangement of residues called α-structure, the second giving linear extended strands which, when positioned adjacent to each other, form a pleated sheet arrangement called β-structure. Between these segments of regular α-structure or strands of β-structure there are regions in the protein chain of non-regular structure where each residue has conformations different from each other. These regions form loops or turns. A globular domain is built up by α- or β-segments which crisscross the domains and thus essentially form the inside. The loop or turn regions between these segments are consequently on the surface of the domains.

These principles give rise to three main classes of domain structure; α where all regular segments are α-helices, β where all regular segments are β-strands and α/β structures where α-helices alternate with β-strands along the chain. A common feature of α-domains is a bundle of four helices packed with their helical axes slightly inclined to each other. In β-domains the strands are arranged antiparallel to each other and usually form a closed barrel. In α/β structures on the other hand the strands are arranged parallel to each other and either form a barrel or a central sheet through the middle of the domain. Fig. 1 which shows schematically the structure of an enzyme determined in our laboratory, alcohol dehydrogenase, illustrates two of these domain classes: an α/β domain and a β-domain.

One very important deduction that can be made from the sample of known structures is that within each class there are only a limited number of topological arrangements of these structures. There are thus only a limited number of theoretically possible basic domain structures that actually occur in Nature. Within each topological arrangement there occur variations in loop size and loop structure but not in the basic topology of the α or β elements.

Protein Folding as a Pattern Recognition Problem

Proteins of similar function in different organisms usually have somewhat different sequences owing to random mutations during evolution from a common ancestral origin. Such proteins have very similar three dimensional structures even when the sequence similarities are so low that they can barely be detected. Thus within a given three dimensional fold there can be a large number of sequence variations. By comparative studies we observe a limited number of these variations and only those that are functionally permitted.

However, it has also been found that domains of different function and with completely different sequences can have the same basic three dimensional fold. Such sequences have been compared and analysed to search for a common basic sequence pattern for a specific domain structure. There is no obvious pattern to be seen even though one would predict the existence of some sort of pattern from logical arguments. A set of rules has been devised to predict the occurrence of turns as well as α and β structure segments, but especially the latter give only slightly higher prediction scores than random guesses. The basic idea is that if these segments could be predicted with high probability, domain structures could be identified from the pattern of segments.

A large number of different protein sequences are now emerging owing to the rapid speed with which genes can be sequenced. This information is, however, of quite limited value at present since the sequences in general tell us nothing about the function of these proteins. Even if we cannot predict the structure from first principles, a pattern recognition method would allow us to identify those domains where we already know the basic structure from a different protein. Such information would be invaluable for intelligent use of the sequence data that is now being collected. My personal guess is that such pattern recognition methods could be developed within the next five years if the field can attract the right people.