Protein crystallography and protein engineering

E. Subramanian

Protein crystallography is paving the way for rational approaches to drug design and for engineering new proteins with catalytic functions more suited to commercial and technical needs.

The last few decades have seen revolutionary developments in biology, thanks to the intellectual framework provided by physics and chemistry for thinking about biological processes in terms of the three-dimensional structures of the molecules involved. The genetic code has been cracked, the molecular mechanisms of heredity have been elucidated, and genetic engineering is a reality. With the advent of recombinant DNA technology, it is possible now to clone the gene for essentially any protein found in nature, incorporate the gene into bacteria and produce significant quantities of that protein. Since chemical synthesis of DNA is possible, even synthetic genes can be incorporated into bacteria, enabling the production of new proteins not found in nature. This is the methodology of protein engineering which represents the advancing frontier of biotechnology in our efforts to modify the structure of a protein to yield specific and desirable improvements in the function of the protein. Protein engineering thus offers great promise in the commercial, industrial, pharmaceutical and agricultural applications of biotechnology, for the large scale viable economical production of hormones, vaccines, antibiotics, industrial catalysts, pesticides and other useful products[1].

In the matter of engineering proteins, major interest is centered around the enzymes. As catalysts, enzymes are used extensively in the fermentation industry, in chemical and pharmaceutical industry and in food industry. Till recently, the high cost of isolation and purification of the enzymes was a limiting factor in their widespread use in the industry. With the advent of recombinant DNA technology, cost is no longer a limiting factor. The real limiting factor now is the lack of knowledge concerning what changes should be made in a protein in order to bring about the desired changes in the catalytic function of the protein. The desired industrial application of an enzyme may often be far removed from the physiological role played by the enzyme. Industrial applications require enzymes which will remain stable under process conditions. These process conditions may often be non-physiological, ranging to extremes of pH, temperature, concentration and solvents. One would, therefore, like to control the enzyme's properties with regard to thermal stability, pH optimum, stability in non-aqueous solvents and so on. Finding the right set of properties for a particular industrial application is often difficult. An enzyme with the highest activity at the temperature employed may not be the most stable under the given solvent conditions, and so on[1].

In nature, specific catalytic functions are served by different enzymes. If we can find out the structural features of each enzyme that confer a specific desirable property, then we can combine these features by protein engineering techniques to create a totally new enzyme that manifests all of the desirable traits. How does one find out what structural features are responsible for the activity of an enzyme?

A protein is a linear long-chain polymer of amino acids. There are 20 different amino acids. Each protein is a specific combination of these amino acids covalently linked together in a unique sequence. Every protein chain folds into a unique three-dimensional structure which determines the protein's specific catalytic function. Anfinsen[2] discovered that the information regarding this folding is contained in the specific sequence of amino acids constituting the protein chain. Since the sequence of amino acids in a protein chain determines the structure of the protein, if we replace one or more amino acids in the protein chain, this action will introduce subtle or significant changes in the protein's structure. Since structure determines function, such substitutions would result in a protein with altered functional characteristics. The problem is that the rules governing the folding process, that is, the relationship between certain groups of amino acids and certain final structural arrangements are not known yet. As a result, there is no way of guessing what the structure will be for a given sequence of amino acids. Without knowing the structure, the structural features or the reasons behind a particular catalytic function will not be known, and hence we cannot think of engineering proteins. This problem is referred to as the 'folding problem'—the prediction of the three-dimensional structure of a protein purely from its amino acid sequence information—and is one of the central problems in molecular biology.

The resolution of the folding problem has special urgency and relevance for biotechnology. Mass produc-

E. Subramanian is in the Department of Crystallography and Biophysics, University of Madras, Madras 600 025.

tion of commercially viable proteins such as human insulin is now a reality. But the next step of modifying the protein for improved functional properties, or of designing new proteins to suit specific commercial applications will require an understanding of how the proteins fold. The urgent need for this knowledge becomes even more evident as geneticists begin to compile a catalogue of all the genes in the human genome. The gene sequences translate into amino-acid sequences and hence to various proteins. Only by knowing how a given sequence of amino acids might fold, can one predict the biological function of the corresponding protein. Gene sequence data have relevance primarily for the proteins they code for!

Protein crystallography is presently is the major technique that can provide complete information on the three-dimensional structure of a protein molecule. The technique is also free of any preconceived notions as to how a protein molecule should look like. Much of the present day knowledge about the relationship between structure and mechanism of action of enzymes has been deduced from protein crystallographic studies of enzymes. Space does not permit a detailed discussion of protein crystallographic methods here, and the reader is referred to many excellent textbooks and review articles available in the literature on the subject[3,4]. The techniques and methodology have advanced to such an extent[5,6] that under favourable circumstances, it is possible to determine the three-dimensional structure of a protein molecule within the time span of a graduate student's thesis work. The main rate-limiting step is the availability of good protein crystals suitable for X-ray diffraction studies.

Protein crystallographic studies have revealed that the chemistry and geometry of the 'active site' of an enzyme are complementary to those of its 'substrate' molecule; that the active site is lined up with amino acids of various chemical character and which are important to the catalytic action of the enzyme; and that the replacement of even one amino acid in the active site can significantly alter the enzymes' structure and hence its catalytic activity. The relevance of these observations to protein engineering studies is obvious: substitution of a new amino acid that differs in size, charge or chemical reactivity from the one it replaces can perturb the active site and hence the catalytic specificity. But the problem is to decide on the type and location of amino acid substitution so as to get the desired specificity and improvement in the catalytic function of the enzyme. How does one make this decision?

Once the three-dimensional structure of a protein has been determined by X-ray crystallographic methods, the analysis is extended to the study of enzyme–substrate interactions, in order to identify the amino-acid residues involved in the catalytic action of the enzymes. In its simpler forms, such studies involve diffusing substrate-analogues through the crystal lattice of the native protein and analysing the diffraction data from such complexes. Using modelling procedures that make use of interactive computer graphics representation of proteins, an analysis of the interplay amongst hydrophobic, electrostatic and hydrogen-bonded interactions between the protein and the substrate–analogue molecule is carried out. Based on these studies, a decision is made as to which amino-acid residues should be altered in order to give rise to improved catalytic function or the desired specificity. The proposed alterations in the amino-acid sequence are then carried out by recombinant DNA techniques. The modified protein is tested for its catalytic properties as well as subjected to X-ray crystallographic analysis. The observed differences in the structures of the native and the modified protein are correlated with the differences in their catalytic properties.

Presently, the three-dimensional structures of more than 300 proteins, not all independent, have been determined by X-ray methods[7]. However, a statistical analysis of these structures has not been successful in developing predictive rules to relate amino-acid sequence to three-dimensional structure. Nevertheless, some valuable insights have been gained into the way protein chains fold. Comparative studies of these protein structures have led to some understanding of the anatomy and taxonomy of protein structures[8]. Certain features of extended structural organization recur among proteins having no apparent sequence or evolutionary relationship. Nature seems to be fond of using certain structural motifs repeatedly in the folding of proteins. Stretches of a protein chain are commonly found to adopt one of three basic structural patterns: alpha helix, beta sheet and sharp chain-reversals. The overall folding appears to be a combination of these motifs in varying proportions.

Several computer programs[9,10] have also been developed to identify common structural features by pattern-matching: that is, searching for patterns in sequence that are consistently associated with defined structural features and then using the pattern information for predicting the folding of the protein chains. Unfortunately, the amount of protein structural data currently available are not sufficient to define an adequate number of patterns with strong predictive powers. This points to the urgent need for more protein crystallographic studies. Since protein structure analysis is a time-consuming affair, and because of the rapid growth in the data-base of protein sequences compared to the rather slow pace in the acquisition of structural knowledge, the need for developing accurate structure prediction rules is also urgent. Consequently, many theoretical biophysicists are busily engaged in 'sequence gazing'[11,12], in order to deduce the predictive rules

governing protein folding through statistical analysis of known sequences and structures. And the protein crystallographers themselves are devising clever strategies to use known structural data to forge ahead with protein engineering studies—they are supplementing the results of protein crystallography with computer-modelling procedures that may be less precise but can provide quick answers in the continuing efforts to engineer proteins that would meet specific commercial requirements[13]. One area where such modelling procedures are useful is in the prediction of the three-dimensional structure of a protein of known sequence, using the known crystal structure of a homologous protein. Such a procedure has been used to predict the unknown three-dimensional structure of mammalian renin[14], using the known structures of homologous fungal enzymes[15]. (Renin is a kidney-enzyme which catalyses the first step in a series of reactions that lead to elevated blood pressure.) Comparison of the crystal structures of several homologous proteins through computer-modelling procedures provides us an opportunity to evaluate the effects on structure and conformation arising from differences in the sequences. Such studies have in fact revealed that the folding pattern of a family of homologous proteins is more conserved in evolution than the amino-acid sequence. Computer-modelling studies of even distantly related proteins can indicate the few invariant amino-acid residues that are essential for catalytic function. These residues can then be targeted for modification by protein engineering techniques.

It may be noted in passing that protein engineering techniques have also been used to speed up X-ray crystallographic studies of proteins! X-ray studies on proteins require in general the availability of *useful* heavy-atom derivatives of the protein in order to solve the 'phase proolem'. The usefulness of a heavy-atom derivative is judged by the conditions that the heavy atom must bind to the protein at a *small number* of specific sites with high occupancy and that the binding must be *isomorphous* that is, there should be no significant alterations to the protein structure or to the crystal packing in the lattice.

The preparation of a useful heavy-atom derivative of the protein is not always easy. Frequently, the heavy atoms bind non-specifically at a large number of sites with relatively low occupancy and the binding may also produce changes in the protein's structure and conformation as well as in crystal packing.

Protein engineering techniques can be used to prepare useful heavy-atom derivatives of a protein. One of the important procedures for preparing a heavy-atom derivative is to use a mercury compound as a reagent that binds to the sulfhydryl group of a cysteine residue in the protein. Even if the protein does not contain a cysteine residue, protein engineering techniques can be used to insert a reactive residue in the protein chain. Such a procedure has been used sucessfully to yield a useful heavy-atom derivative in the high-resolution structure analysis of colicin A (ref. 16).

The next few paragraphs will highlight the role of protein crystallography in the development of drug design and protein engineering studies.

Drug design

Practically all drugs in use today are the result of serendipitous discoveries or trial-and-error processes. Protein crystallography, combined with molecular modelling procedures, offers an entirely new approach for the rational design of drugs. Computer modelling of the active site of a target protein of known crystal structure can indicate the size, shape and chemical character of the molecules that are likely to interact with the protein and bind tightly to its active site, thus leading to the design of suitable inhibitor molecules which can then be tested clinically for their therapeutic effectiveness. It should be emphasized that interactive computer graphics plays an important part in these efforts.

Development of vaccines

Our body's immune response to infections is mediated by antibody molecules (immunoglobulins) which specifically recognize certain surface features (called antigenic determinants) on the invading organism's surface proteins. These antigenic sites often correspond to discrete surface regions comprising short peptide fragments, also called 'epitopes'. Using recombinant DNA techniques, synthetic oligo-peptides which mimic an epitope can be coupled to a carrier protein. The resulting hybrid protein may be used like a vaccine to provoke the immune system. This approach opens up the possibility for developing 'supervaccines' in which epitopes from the surface proteins of various infectious agents can be incorporated into a single polypeptide chain by recombinant DNA techniques. Such a hybrid synthetic protein as a vaccine could be expected to confer simultaneous protection against several diseases[13].

Protein engineering studies

As examples of the significant role played by protein crystallography in protein engineering studies, the following reports are illustrative. Greer[17] compared the X-ray structures of chymotrypsin, trypsin and elastase to model the construction of 'new' serine protease based on structurally conserved and variable regions of these proteins. In the efforts to find a cure for sickle-cell

anaemia, Beddel *et al.*[18] used the X-ray structure of human haemoglobin to design compounds that would modulate the protein's function. Their modelling studies suggested an allosteric site at which compounds designed for this site would modulate the protein's oxygen affinity. These compounds are being tested in clinical trials as anti-sickling agents.

The most illustrative example of the power of protein crystallography in drug design aspects of protein engineering concerns the studies on the structure of dihydrofolate reductase (DHFR), an enzyme isolated from many different organisms. Much of the interest on this enzyme stems from the importance of the enzyme as a therapeutic target. The anti-cancer drug methotrexate, the antibiotic drug trimethoprim and the anti-malarial drug pyrimethamine are all targeted towards inhibition of DHFR. The most important point is that these different drugs display differential inhibition of DHFR from different sources. For example, trimethoprim is highly selective for *E. coli* DHFR, but little effective against malarial infections or against vertebrate DHFR, pyrimethamine is selective for DHFR of the malarial protozoan *Plasmodium berghei* and so on. Since the DHFR enzymes from various sources constitute a homologous class of enzymes, it is clear that structural differences between the various DHFR enzymes must exist. These structural differences must arise because of differences in sequences, and must be responsible say, for trimethoprim's highly selective inhibition of *E. coli* DHFR, relative to say the vertebrate DHFR. Crystallographic studies of trimethoprim complexed with both *E. coli* DHFR and chicken DHFR show that while the gross three-dimensional structures are conserved in the two DHFR enzymes, the small structural differences in the active site, arising from amino-acid sequence differences between the two enzymes are responsible for making trimethoprim bind to each enzyme with a different conformation, thus explaining the difference in specificity. Using such studies, Kuyper *et al.*[19] have designed inhibitors of DHFR more potent than trimethoprim.

Studies on pepsin-like enzymes

Among the proteins that are currently the targets for protein engineering studies and for the design of new therapeutic drugs are the class of enzymes commonly referred to as acid-proteases. Mammalian digestive enzyme pepsin is a well-known member of this class, hence these enzymes are also referred to as pepsin-like enzymes. Other members include chymosin, the milk-clotting enzyme from calf stomach used in making cheese; renin, the kidney enzyme involved in the control of blood pressure; and cathepsin-D, a lysosomal enzyme involved in tissue and intra-cellular protein turnover. Several species of fungi also produce pepsin-like enzymes.

The pepsin-like enzymes have many important clinical and commercial uses, and hence are targets for the design of inhibitor compounds. Renin is involved in the control of blood pressure. Cathepsin-D has been implicated in diseases of the joints and in the release of vasocative peptides which cause pathological conditions such as shock, inflammation and diarrohea; it has also been reported to be found in certain tumour cells[20]. Pepsin has been implicated in the etiology of peptic ulcer. The fungal pepsin-like enzymes are used in the cheese industry to produce cheeses of different flavours and consistency. The understanding of the mechanism of catalytic action of these enzymes and the identification of the amino-acid residues involved in substrate binding should lead to the design of a variety of inhibitors for specific uses. Such studies can also aid in the understanding of fungal growth in industrial or municipal waste system treatment; to engineer new proteins which can be more effective in waste disposal; and for the inhibition of fungal growth on proteinous foods. Since fungal acid proteases are used in the cheese industry to produce cheese of different flavour and consistency, protein crystallographic studies on the fungal enzymes will enable us to understand the molecular basis of flavour, and to engineer the production of new proteins to improve the flavour and consistency of cheese to suit popular tastes.

Much of the protein crystallographic studies on the pepsin-like enzymes have been carried out on the fungal enzymes, mainly during the last decade. The present author has been instrumental in determining the three-dimensional structure of a fungal pepsin from *Rhizopus chinensis*[15,21]. This work is among the first to establish the three-dimensional folding for the whole class of homologous pepsin-like enzymes. Presently, structural studies on the fungal enzymes had been used to model the structure of the homologous mammalian enzyme, renin[14]. More recently, the AIDS virus has been found to code for a protease which is homologous with the pepsin-like enzymes. The three-dimensional structure of this AIDS virus protease has also been determined recently by protein crystallography[22] and found to be similar to the three-dimensional structure of the microbial pepsin-like enzymes. The similarities extend to the active sites of these enzymes and open up the possibilities for extrapolating the structural studies on the fungal enzymes for the design of specific AIDS protease inhibitor compounds which can be used as safe and effective drugs in the treatment and control of the AIDS virus infection[23]. Towards this objective, the determination of the three-dimensional structures of several pepsin-like enzymes from other fungal and mammalian species is being actively pursued by protein crystallographers, including those at Madras.

In conclusion, it may be no exaggeration to state that protein crystallography is an important armour in the arsenal of biotechnology in our efforts to harness the power of genetic engineering to improve the welfare of the human species.

1. Ulmer, K. M., *Science*, 1983, **219**, 666.
2. Anfinsen, C. B., in *New Perspectives in Biology*, (ed. M. Sela), Elsevier, Amsterdam, 1964, p. 42.
3. Blundell, T. L. and Johnson, L. N., *Protein Crystallography*, Academic Press, 1976.
4. Lattman, E. E. and Amzel, L. M., *Methods Exp. Phys.*, 1982, **20**, 229.
5. McPherson, A., *Prepartion and Analysis of Protein Crystals*, John Wiley, 1982.
6. Oxender, D. L. and Fred Fox, C. (eds), *Protein Engineering*, Alan R. Liss Inc, New York, 1987.
7. Brookhaven Protein Data Bank.
8. Richardson, J. S., *Adv. Protein Chem.*, 1981, **34**, 167.
9. Chou, P. Y. and Fasman, G. D., *Adv. Enzymol.*, 1978, **47**, 45.
10. Lim, V. I., *J. Mol. Biol.*, 1974, **88**, 857.
11. Jasny, B., *Science*, 1983, **240**, 722.
12. Thornton, J. M., *Nature*, 1988, **335**, 10.
13. Hol, W. G. J., *TIBTECH*, May 1987, p. 137.
14. Blundell, *et al.*, in *X-ray Crystallography and Drug Action*, (eds. Horn and De Ranter), Clarendon Press, 1984, pp. 427–440.
15. Subramanian, E. *et al.*, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 556.
16. Tucker, A. D. *et al.*, *Protein Engg.*, 1989, **2**, 399.
17. Greer, J., *J. Mol. Biol.*, 1981, **153**, 1027.
18. Beddel, C. R. *et al.*, *Br. J. Pharmacol.*, 1976, **57**, 201.
19. Kuyper, L. F. *et al.*, *J. Med. Chem.*, 1982, **25**, 1120.
20. Greenbaum, L. M., *Trends Biochem. Sci.*, 1977, **2**, 41.
21. Subramanian, E., *Trends Biochem. Sci.*, 1978, **3**, 1.
22. Navia, M. A. *et al.*, *Nature*, 1989, **337**, 615.
23. Blundell, T. and Pearl, L., *Nature*, 1989, **337**, 596.