Advances in Protein Design
International Workshop 1988

Edited by:
H. Blöcker, J. Collins,
P. D. Schmid, D. Schomburg
KNOWLEDGE-BASED PROTEIN MODELLING AND DESIGN

Tom L. Blundell, Devon Carney, Tim Hubbard, Mark S. Johnson,
Alasdair McLeod, John P. Overington, Andrej Sali,
Michael Sutcliffe and Pamela Thomas

Laboratory of Molecular Biology, Department of Crystallography,
Birkbeck College, University of London, Malet Street,
London WC1E 7HX, England.

1. INTRODUCTION

Knowledge-based modelling can be envisaged as a number of steps concerned with the establishment and use of rules to generate a model of a protein. One of the most powerful procedures in learning rules is comparison of related structures either through alignment of sequences to identify conserved residues or superposition of three dimensional structures to identify conserved conformations or motifs. Thus the first step in a knowledge-based modelling procedure is the systematic comparison of families of topologically similar structures. This step will lead to the establishment of "equivalences" between the structures compared and to their clustering based on measures of similarity. The second step involves the projection of the results of the comparisons of three dimensional structures down onto the level of sequence. This step establishes rules relating sequence to structure. These can be expressed as consensus sequences - templates - for topologically equivalent residues, or as key residues in canonical structures, which are then used to align the sequence of the protein of unknown tertiary structure. The third step uses the rules established in the second step to generate a three-dimensional model.

2. MODELLING BY HOMOLOGY OR ANALOGY

The classical form of knowledge-based modelling is modelling by homology or comparative modelling. This procedure depends on the knowledge that homologous sequences
have similar tertiary structures involving a conserved "framework" of helices and strands connected by structurally variable regions that accommodate much of the sequence variation and almost all the insertions and deletions. The method was first used by Browne et al. (1970) to model alpha-lactalbumin on the basis of the three-dimensional structure of lysozyme. In subsequent years it was used to model many proteins including insulin-like growth factors and relaxins from the three-dimensional structure of insulin, serine proteinases on the basis of trypsin, chymotrypsin and elastase, and renin from the three-dimensional structures of aspartic proteinases and many other structures (see Blundell et al., 1987b for a review). The method has recently been developed into a systematic approach (COMPOSER) in which several homologous structures can be simultaneously used in modelling the unknown (Sutcliffe et al., 1987a, b; Blundell et al., 1988).

COMPOSER considers the construction of a model by superposition of rigid three-dimensional structures. We assume initially that there are several homologous analogous structures from which a framework can be generated. The first stage is to select the structures that will be most useful in the modelling exercise. We have developed an automatic procedure for this which depends on combining the phylogenetic tree from sequence - including that of the unknown structure - with that based on the three-dimensional structures alone (Johnson et al., 1989). This enables selection of the set of structures that are clustered around the sequence of the unknown. These structures are then used to produce a framework for the unknown in which contributions of each structure are weighted according to their percentage identity of sequence (Sutcliffe et al., 1987a). The framework so derived is an average structure and it is endowed with real geometry by least squares fitting fragments from the homologous structures for each section of the framework.

The next task is to select the structurally variable regions. This is achieved by first using a geometrical filter in a similar way to that of Jones and Thirup (1986) with a three residue overlap on each end of the fragment. Each fragment selected is then least squares fitted to the framework and the fragments so fitted are clustered using principal components analysis or tree construction (Johnson et al., 1989). Key residues are identified using rules concerning the structural conservation of features such as glycines with a positive phi torsion angle or charged residues that are buried and hydrogen bonded to mainchain function. The fragments are ranked, and the top ranking fragment is then tested for overlap with other parts of the model structure. If it is rejected on these grounds, the next ranking fragment is selected. The optimal fragment is then melded onto the framework.

In fact alternative procedures may need to be adopted. First it may be best to extend the framework using differing subsets of homologous structures at each variable region. Secondly, the rules developed by Sibanda and Thornton (1985); Edwards et al. (1988); Efimov (1986); (1989) may be useful regions within In certain cases a replacement region to be made locally and

The third step from an analogous families (i.e. the framework that relates sidechains within modelling wider groups must be procedures similar all., 1987a; Sali, A.

All knowledge and molecular useful model of 2-D NMR.

3. APPLICATION

Knowledge-based protein enginee
Etimov (1986); Milner-White and Poet (1986); Wilmut et al. (1988); and Sibanda et al. (1989) may be used to select a loop not recognised by the key residues procedure. Thirdly, regions within the framework of the known set may require insertions or deletions. In certain cases where the chain remains the same length but key residues are changed, a replacement conformation may be required. In these cases a further definition of the region to be replaced is required. In general, insertions, deletions, and replacements are made locally and where possible in regions of irregular secondary structure.

The third step is to replace sidechains. This is achieved using a set of rules derived from an analysis of sequence variation at topologically equivalent positions in homologous families (Sutcliffe et al., 1987b; Summers et al., 1987; McGregor et al., 1987). The 1200 rules include one for each of the 20 by 20 amino acid replacements in each of alpha-helical, beta-sheet, and irregular regions. Where there is no useful prediction, the most probable conformation is chosen, and where there is more than one prediction, the conformation closest to the median of the predictions is selected.

This procedure for modelling is very successful where the known structures cluster around that predicted, and where the percentage sequence identity is high. Where the structure to be predicted lies outside the cluster and the sequence identity is less than 40%, a procedure is required to introduce translations and rotations of the elements of the framework relative to each other. For this we are currently exploring algorithms that relate distances between elements of secondary structure to the volumes of the sidechains within contact regions (J. Overington, unpublished results). However, for modelling widely diverged or analogous convergentelyevolved motifs, the assembly of rigid groups must be re-taced by an alternative approach that uses distance geometry approaches similar to those used in constructing models from 2-D NMR data (Sutcliffe et al., 1987a; Sali, A., unpublished results).

All knowledge-based procedures require simulation of the solvent, energy minimisation and molecular dynamics simulations to optimise the final structure and to provide a useful model of the time and space averaged structures determined by X-ray analysis and 2-D NMR.

3. APPLICATIONS OF KNOWLEDGE-BASED DESIGN.

Knowledge-based modelling has applications both to receptor-based drug design and to protein engineering.
3.1 Receptor-based drug design

Although the sequences of many receptors have recently been determined, the three-dimensional structures of very few of pharmaceutical interest are known. In some cases the structure of the receptor from another species or an orthologous protein structure may have been determined by X-ray analysis. For example, in 1984 when the sequence of human renin was determined, no three-dimensional structures of any renin had been determined - this has taken four years! Only structures for homologous fungal aspartic proteases were accurately analysed by X-ray analysis. Modelling by homology produced rough models for mouse renin (Blundell et al., 1983) and for human renin (Sibanda et al., 1984). These have been extended using experimentally determined structures of aspartic proteases complexed with human renin inhibitors to give a model of the human renin - human angiotensinogen (fragment) transition state complex (Blundell et al., 1987a). These models have been used by several pharmaceutical companies as a receptor-based contribution to their design of orally active renin inhibitors for the treatment of hypertension. The model is probably accurate to 0.5Å (comparison of alpha-carbons) close to the active site but will have errors in excess of 1.5Å in the peripheral loops.

3.2 Site-directed mutagenesis

Protein engineering using site-directed mutagenesis involves the introduction of insertions, deletions and replacements in a protein with retention of the three-dimensional structure but modification of catalytic activity, stability to high temperature or non-aqueous solvents or other property in a predictable fashion. The knowledge-based procedures developed for modelling local insertions and deletions and sidechain replacements provide a useful starting point although energy minimisation and molecular dynamics procedures in a simulated aqueous environment will be needed to explore local conformations.

3.3 Chimaeric molecules

For the design of a chimaeric molecule, for example a tissue plasminogen activator serine proteinase domain linked at the COOH-terminus of an Fab fragment of a monoclonal with fibrin specificity, the modelling procedures are first used to model the two domains on the basis of homologous structures (Blundell et al., 1988; Harris, 1988; Overington et al., 1989). Secondly the relative disposition of the two fragments is chosen interactively using computer graphics. A linker region with overlaps in both tPA and Fab domains is then selected from the data base of polypeptide fragments in which the linking residues are small and hydrophilic if the link needs to be flexible. Finally the

ACKNOWLEDGE

We thank the Sci USA, for support. Birkbeck.
contiguous surfaces of the two linked domains are mutated using the sidechain replacement algorithm so that they are compatible with each other and the solvent.

3.4 *Ab initio* design of proteins

Analogous approaches can be used in designing novel proteins. Although this area is presently in its infancy, attempts have been made to design alpha-helical bundles, beta-barrels and other commonly occurring canonical structures. We have used knowledge-based techniques, including COMPOSER, to design a symmetrical two Greek key protein (based on the stable eye lens crystallins and called CRISTANOVA) that is engineered to bind copper in a similar way to superoxide dismutase (Hubbard, 1988). Although such projects are currently academic, protein engineers may in the future be requested to design proteins for example for rare metal ion scavenging or even biochips where no useful parallel is known to exist in Nature.

ACKNOWLEDGEMENTS

We thank the Science and Engineering Council, UK, and the National Cancer Foundation, USA, for support. We are grateful for useful discussions with many of our colleagues at Birkbeck.