From: PROTEIN DESIGN AND THE DEVELOPMENT
OF NEW THERAPEUTICS AND VACCINES
Edited by Jerry B. Hook and George Poste
(Plenum Publishing Corporation, 1990)

Knowledge-Based Protein Modeling and the Design of Novel Molecules

TOM L. BLUNDELL, MARK S. JOHNSON, JOHN P. OVERINGTON, and ANDREJ ŠALI

1. Introduction

Each year protein crystallographers determine about 30 new three-dimensional protein structures, most of which are published in some form, and many of which are deposited in the Brookhaven Protein Databank (Bernstein *et al.*, 1977). Some of these are novel arrangements of a polypeptide chain, but others represent classes of structure that have been observed earlier in other proteins. Thus, each year we learn about an increasing proportion of the folds available to a polypeptide chain.

We are also, at the same time, gaining a better understanding of the principles of a polypeptide architecture. We now understand that proteins have evolved as hierarchical entities consisting of sequence, secondary structure, motif, domain, globular protomer, and oligomer levels (Richardson, 1981). We know that there are many lower-level structures that are compatible with each structure at a higher level. For example, many differing sequences can form an equivalent helix in a globular protein, and helices, strands, and turns of differing lengths and sequences can assemble as topologically similar motifs, such as nucleotide binding domains, alpha-helical bundles, or Greek keys (Rossmann and Argos, 1976; see Bajaj and Blundell, 1984, for a review).

TOM L. BLUNDELL, MARK S. JOHNSON, and JOHN P. OVERINGTON • Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, London University, London WC1E 7HX, England. *ANDREJ ŠALI* • J. Stefan Institute, Ljubljana, Yugoslavia; *present address:* Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, London University, London WC1E 7HX, England.

We can now begin to formulate a powerful set of rules that describe protein construction and architecture at many different levels, for example, the conformations of amino acid side chains at topologically equivalent positions in families of homologous proteins (Summers *et al.*, 1987; Sutcliffe *et al.*, 1987b; McGregor *et al.*, 1987) and the classes and key residues that characterize beta hairpins (Sibanda and Thornton, 1985; Milner-White and Poet, 1986; Efimov, 1986; Edwards *et al.*, 1987). We can consider using these rules, combined with other facts and hypotheses, to indicate which sequences adopt a particular higher-level structure or, in other words, to describe tertiary templates that define all sequences that are consistent with a particular motif, globular domain, or tertiary structure (Ponder and Richards, 1987). We seek to use these facts and rules in a knowledge-based procedure for the modeling and design of proteins.

Knowledge-based modeling can be envisaged as a process concerned with establishing and using rules to generate a model of a protein. One of the most powerful procedures in rules construction is the comparison of related structures, either through an alignment of sequences to identify conserved residues or through a superposition of three-dimensional structures to identify conserved conformations or motifs. Thus, the first step in a knowledge-based modeling procedure is systematic comparison of families of topologically similar structures. This step will lead to establishment of "equivalences" between the structures compared, and to their clustering based on measures of general similarity. The second step involves projection of the results of these 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 equivalenced residues, or as key residues in canonical structures, which are then used to align the sequence of the protein of unknown tertiary structure with the known structures. The third step uses the rules established in the second step to generate a three-dimensional model. The three steps of knowledge-based modeling are shown diagrammatically in Figure 1.

The classical form of knowledge-based modeling is modeling by homology, or comparative modeling. This procedure depends on the knowledge that homologous sequences have similar tertiary structures involving a conserved "framework" of packed helices and strands connected by structurally variable regions that accommodate much of the sequence variation and almost all of the insertions and deletions. The method was first used by Browne *et al.* (1969) to model alpha-lactalbumin on the basis of the known three-dimensional structure of lysozyme. In subsequent years it was used to model insulinlike growth factors and relaxins from the three-dimensional structure of insulin (Bedarkar *et al.*, 1977; Isaacs *et al.*, 1978; Blundell *et al.*, 1978), serine proteinases on the basis of trypsin, chymotrypsin, and elastase (Greer, 1981), renin from the three-dimensional structures of aspartic proteinases (Blundell *et al.*, 1983; Sibanda *et*

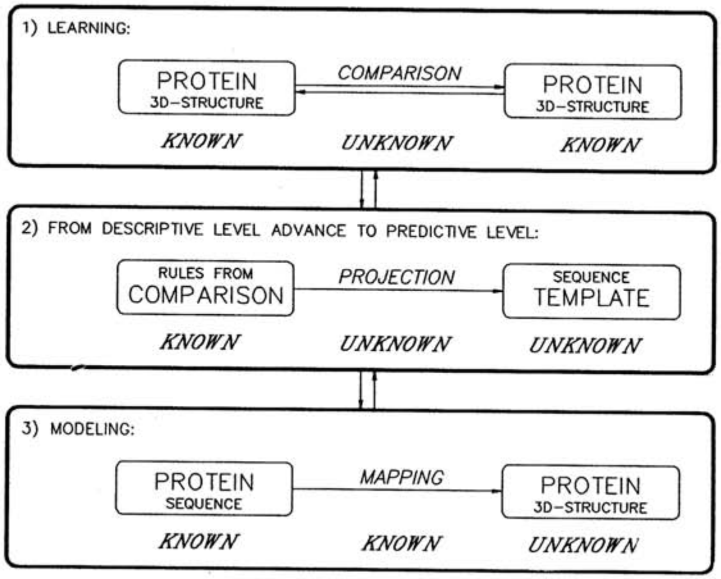

Figure 1. Scheme showing the basic steps of knowledge-based protein modeling. The first uses comparison methods to learn about the structural features of the protein family that contains the protein to be predicted. The second step uses this knowledge to derive the tertiary template and align it with the sequence of the unknown. The third, and last, step uses the constraints imposed on the sequence of the unknown by its alignment with the tertiary template and by general rules of protein structure to map the sequence onto its tertiary structure.

al., 1984), and many other structures. The method has recently been developed into a systematic approach (COMPOSER) in which several homologous structures can be used simultaneously in modeling the unknown (Sutcliffe *et al.,* 1987a,b; Blundell *et al.,* 1988). Rules are used to establish the relative positions of the framework (Sutcliffe *et al.,* 1987a), to select appropriate fragments for variable regions not only from homologous proteins (Greer, 1981; Chothia *et al.,* 1986), but also from other protein structures (Blundell *et al.,* 1988; Sutcliffe, 1988; Sibanda *et al.,* 1989), and to replace side chains (Sutcliffe *et al.,* 1987b; Summers *et al.,* 1987; McGregor *et al.,* 1987). In a parallel development, Jones and Thirup (1986) have shown that modeling into electron density during protein crystallography can also be aided by selecting conformational fragments from a series of other proteins of known three-dimensional structures.

Approaches, such as COMPOSER, that depend on rigid body superposition of three-dimensional structures are restricted to closely related motifs or homologous structures. Chothia and Lesk (1986) showed that for increasingly divergent structures, the number of topologically equivalent residues obtained by superposition decreases and the root mean square difference increases. This is due mainly to small, but cumulative relative translations and rotations of the packed secondary structural elements. This also affects the core residues (Hubbard and Blundell, 1987) and results in an insufficient framework for modeling. Clearly, a more flexible approach to defining topological equivalence is required.

The problem of defining topological equivalence was addressed more than a decade ago by Rossmann, Matthews, and their colleagues (see Matthews and Rossmann, 1985, for a review), who compared local main-chain direction, conformation, and position to establish topological equivalence. An alternative approach is to simplify the structure to a series of vectors representing the axes of the helices and strands, which are then compared (Murthy, 1984; Richards and Kundrot, 1988). In our approach, we compare properties and relations at each level in the hierarchy of protein structure and derive weight matrices from which the optimal alignment can be deduced using the dynamic programming approach of Needleman and Wunsch (1970). This approach works well for related structures that have little or no significant sequence identity (Šali and Blundell, 1990).

Once the topologically equivalent residues have been defined by comparing properties and relations, templates and key residues can be derived and used to align the sequence of the unknown. However, new approaches are required for the third step in the procedure, whereby a model is generated. The fact that properties and relations have been equivalenced indicates that internal coordinates should be used. In many ways the problem is closely related to that of reconstructing a model from upper and lower bounds on distances obtained from two-dimensional nuclear magnetic resonance (NMR) experiments. Thus, distance geometry (Crippen, 1977; Havel and Wuthrich, 1985) or dihedral angle optimization techniques (Braun and Go, 1985) can be adopted.

In this chapter we discuss the various approaches available in each step of the modeling procedure as they are being developed in our laboratory and then briefly describe applications to protein design.

2. *Learning by Comparison of Structures*

At Birkbeck, we have been concerned with two aspects of the comparison of structures. The first of these is designed to obtain a simultaneous rigid-body superposition of a family of protein structures. The second compares properties and relations.

2.1. *Superposition of Structures*

In order to gain as much information as possible from a family of protein structures, it is advantageous to obtain a multiple superposition that is unbiased by the order in which the proteins are compared. Thus, in the program MNYFIT (Sutcliffe *et al.*, 1987a), the alpha-carbon positions of a set of proteins are compared by least-squares fitting to an average structure or "framework." In this procedure one of the structures is chosen at random as the first approximation to the framework, and all other structures are fitted pairwise, using unit weights. A new framework is then chosen from the average of the fitted structures, using weights that depend on the estimated precision in the molecule (a function of the resolution of the X-ray analysis) and the distance of the atoms from the previous framework. An algorithm by McLachlan (1982) is used for the least-squares-fitting step. The procedure is continued iteratively until stability is obtained in both the number of equivalences and the root mean square distances over the topologically equivalent positions.

This procedure works efficiently for families of closely related protein structures such as the mammalian serine proteinases (Overington *et al.*, 1988), the vertebrate globins (Sutcliffe *et al.*, 1987a), and the constant domains of the immunoglobulins (Sutcliffe *et al.*, 1987a). The result of such a multiple superposition for the mammalian serine proteinases is shown in Fig. 2. With more divergent families the number of topological equivalent residues common to the complete set decreases rapidly, so that the framework is not useful as the basis for modeling.

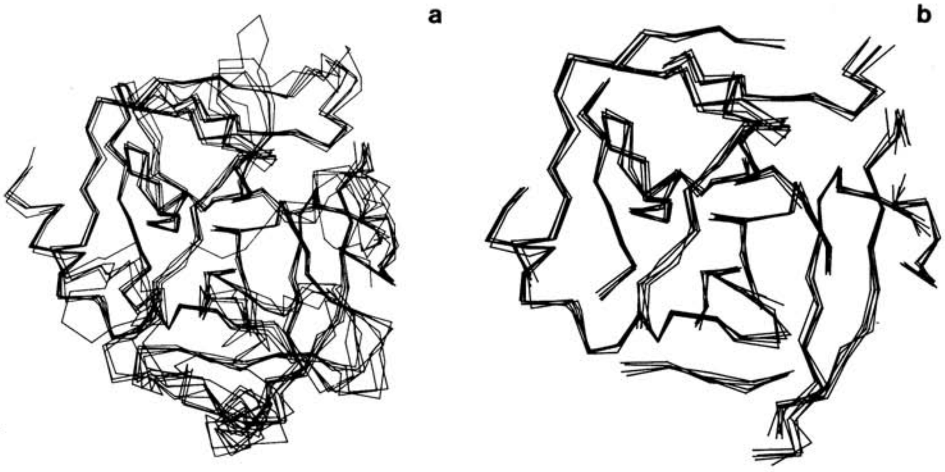

Figure 2. (a) Five mammalian serine proteinases (gamma-chymotrypsin, elastase, kallikrien, trypsin, and rat mast-cell protease type II) optimally superposed using the program MNYFIT; initial equivalences were the three residues in the catalytic triad. Coordinates were taken from the Brookhaven Protein Databank (Bernstein *et al.*, 1977); datasets used were 2GCH, 3EST, 2PKA, 2PTN, and 3RP2. (b) Regions determined to be in the structurally conserved core from the results of the previous fitting.

2.2. Comparison of Properties and Relations

In order to overcome the problems encountered in global rigid body superposition procedures, we compare proteins for features that indicate a common fold. These features may exist at any level in the hierarchy of protein structure—residue, secondary structure, supersecondary structure, motif, domain, or globular protomer. At each level, the features compared could be properties or relationships. Residue properties include sequence identity, hydrophobicity, size of side chain, and so on, as discussed by Argos (1987). When the three-dimensional structure is included in the comparison, the residue properties include local conformation, orientation of a side chain and main chain relative to the center of mass of the globular structure, accessibility, main chain dihedral angles, and so on, as shown in Table I. Equivalent properties at higher levels include the nature of the secondary structure element i, its accessibility, the orientation of the vector defining a helix or strand compared to the center of mass, and the local dihedral angle formed by secondary structure elements i − 1, i, i + 1. Comparisons of all such properties are included in a residue-by-residue

Table I. Protein Features that Can Be Used in Comparing Proteins[a]

Residues	Segments
Properties	Properties
Identity	Secondary structure type
Physical properties	Amphipathicity
Local conformation	Improper dihedral angle
Distance from gravity center	Distance from gravity center
Side-chain orientation	Orientation relative to gravity center
Main-chain orientation	Side-chain accessibility
Side-chain accessibility	Main-chain accessibility
Main-chain accessibility	Position in space
Position in space	Global orientation
Global direction in space	
Main-chain dihedral angles	
Relations	Relations
Hydrogen bond	Distances to one or more nearest neighbors
Distance to one or more nearest neighbors	Relative orientation of two or more segments
Disulfide bond	
Ionic bond	
Hydrophobic cluster	

[a] Various features are represented by rows; different levels of protein organization by columns. The table can be easily expanded to the right by adding features at higher levels of protein structure. The term "property" is used here for all protein features that imply comparison of only one element from each protein. Conversely, the term "relationship" is used for a feature that implies comparison of at least two elements from each protein.

weight matrix (Šali and Blundell, 1990). Optimal alignment is then obtained using the dynamic programming approach of Needleman and Wunsch (1970).

We also compare relations between different elements at each level of the hierarchy. At the residue level, the hydrogen bonding pattern represents a highly conserved feature of a protein fold. For example, the two lobes of aspartic proteinases, each of about 160 amino acid residues, appear to have evolved by gene duplication. Even though the sequence identity is not significant between the two lobes, about 43 residues are topologically equivalent by pairwise superposition. In fact, when the general arrangement of the beta-strands and their hydrogen bonding arrangements are compared, roughly twice as many residues appear to be equivalent (Blundell *et al.*, 1985). Hence, we include in our comparisons a consideration of hydrogen bond and local packing relationships. However, as such relationships affect more than one element in the sequence, this makes the dynamic programming approach computationally difficult. Instead, a simulated annealing technique is used to provide initial equivalences of relationships that are then directly introduced into the residue by residue weight matrix. Using this approach, 122 residues are found to be equivalent between the two lobes of the aspartic proteinase endothiapepsin. This method also works well for the globins and gives virtually the same set of topological equivalences—and therefore identical alignment—that has been obtained by careful analysis of the packing relationships in this family (Lesk and Chothia, 1980; Bashford *et al.*, 1988).

2.3. *Clustering and Tree Construction*

Phylogenetic relationships can be derived from distances that measure the dissimilarity between structures, and these can be obtained from any of the quantities used in the comparison procedures described earlier (Johnson *et al.*, 1990a,b). For multialigned structures, the root mean square distances for the framework regions can be used. This is not useful for divergent families, and we have calculated pairwise distance scores from a function that combines information from the root mean square distances and the number of topologically equivalent atoms obtained from pairwise superpositions (Johnson *et al.*, 1990a), as suggested by the work of Chothia and Lesk (1986) and Hubbard and Blundell (1987). The topologies and branch lengths of the phylogenetic trees were then constructed using a program KITSCH from the Phylogeny Inference Package (PHYLIP) written by Felsenstein (1985). Figure 3 shows an example of the optimal tree determined using global optimization and excluding negative branch lengths. Although this procedure uses no information from the sequence, it gives a tree that is generally topologically equivalent with the sequence phylogenies based on a residue type. Figure 3 compares the two types of trees; a more general discussion of trees for six homologous sets of proteins (immunoglobulin do-

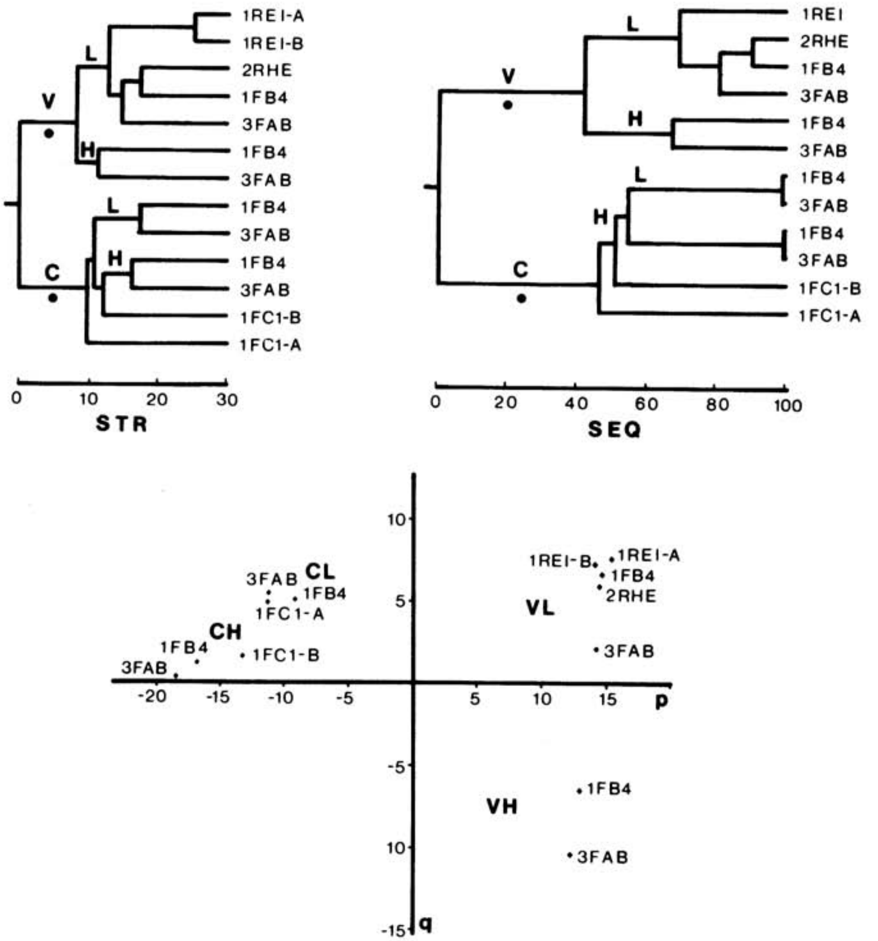

Figure 3. (Upper) Cladograms determined for various immunoglobulin fragments and derived from structural distances (STR) and sequence-based distances (SEQ). The separations between constant (C) and variable regions (V) of the light (L) and heavy (H) chains are indicated. Solid dots indicate branches whose lengths may change depending on the value of an outlier used to construct branches near the root. (Lower) Multidimensional scaling of the structural distances. The variable-domain/light-chain (VL), variable-heavy (VH), constant-light (CL), and constant-heavy (CH) clusters are indicated. The *p* and *q* axes are determined from the two largest eigenvalues and the corresponding eigenvectors of the cross-correlation matrix of distances derived from the structural data. This projection accounts for 72% of the total variance for this dataset. The fragments are designated by the following Brookhaven codes (Bernstein *et al.*, 1977): the chains of the human Bence–Jones dimer Rei—1REI-A and 1REI-B; the human Bence–Jones monomer Rhe—1RHE; the human Fab fragment Kol—1FB4; the human Fab fragment New—3FAB; the two domains of 1FC1—1FC1-A and 1FC1-B.

mains, globins, cytochromes c, serine proteinases, eye lens gamma-crystallins, and dinucleotide binding domains) is described in Johnson *et al.* (1990a). For more divergent families, distances can be obtained from any of the properties or relations used to compare proteins (Šali and Blundell, 1990). Trees based on different properties and relations appear to have very similar topologies (Johnson *et al.*, 1990b).

Multidimensional scaling (Crippen, 1977) can also be used to perform cluster analysis independently of the tree constructing algorithms. With this technique, columns of the cross-correlation matrix of the distances can be considered as points in a higher dimensional space. These coordinates completely describe the relationships among the data. A projection of these points to two dimensions is easily calculated such that the plane depicted is that which displays the maximum variance among the points with only a modest loss of information. Figure 3 also shows the multidimensional scaling analysis for the same set of structural distances between the immunoglobulin fragments that were used for the tree construction.

3. Rules for Tertiary Templates and Key Residues, and Their Comparison with a Sequence of an Unknown

We describe below how comparisons of three-dimensional protein structures can indicate features of sequence that are important for the adoption of a particular conformation. We discuss the derivation of templates that define variation of sequences consistent with a particular tertiary structure and show how these can be used to identify structures that may be useful in modeling.

3.1. Consensus Sequences for Framework Regions

The comparison of three-dimensional structures by multiple superposition described in Section 2.1. leads to a framework, which defines the topologically equivalent residues that are conserved in a family or subfamily of proteins (Sutcliffe *et al.*, 1987a). The consensus sequences for the frameworks can be used to align the sequence of an unknown tertiary structure using a procedure such as that of Taylor (1986). The consensus sequences emphasize those features of the framework on which there are major three-dimensional structural constraints due to conformation or packing. However, by definition they ignore the regions outside the framework which are variable but which can nevertheless contain useful information concerning the extent and the positions of the framework. A comparison of tertiary structures using properties and relations (Section 2.2) can provide an extension of the region of topological equivalence used to derive the consensus sequence template.

3.2. Key Residues from Comparisons of Properties and Relations

The comparison of structures, particularly through comparison of properties and relations, can give clues to the features that are critical for the adoption of a particular three-dimensional structure. For example, we may identify glycine, aspartic acid, and asparagine residues that have positive PHI angles that are not easily adopted by other residues. Alternatively, we may find residues that are conserved, such as hydrophobic aromatic or aliphatic carbon side chains, because they are buried without hydrogen bonding partners. For each position in the alignment, a score that indicates the conservation of properties and relations in the set compared can be used to weight the positions during alignment with the unknown. This provides a general and automatic procedure for identifying key residues in variable regions of homologous proteins, a problem first addressed by Chothia and Lesk (1986) for the immunoglobulin variable regions.

3.3. Characteristic Sequences from Features of Tertiary Structure

In many cases there will be only one experimentally defined three-dimensional structure for a particular class of motif, domain, or globular protein. The prediction of sequences that are compatible with this structure—the production of a tertiary template—is central to our understanding of protein diversity and evolution.

Ponder and Richards (1987) have adopted a systematic approach in which a library of amino acid side-chain rotomers is generated and used for testing for combinations that are consistent with the tertiary structure given a variation in the main chain of 0.5 Å. This is a useful approach, but it is limited by the fact that the diversity in real proteins of sequence identity less than 50% involves relative shifts of main-chain residues much greater than 0.5 Å (Clothia and Lesk, 1986). However, identification of the possible sequence variation using this method with an even greater variation of main-chain positions soon becomes computationally infeasible.

An alternative approach is to learn from comparisons of many families of proteins, which may reveal general features of tertiary structure that impose strong constraints on the variation available to the amino acid sequence. This has already been done for many protein families, and some generalizations can be made. For example, it is known that glycines are most conserved in evolution (Bajaj and Blundell, 1984; Blundell *et al.*, 1986) where they adopt positive PHI angles or where packing restrictions are incompatible with the existence of a side chain. It is also well established that residues that are inaccessible to solvent are less variable in evolution. This has been quantified by Hubbard and Blundell (1987), who compared the percentage identity and root mean square distances in several protein families, not only for all topologically equivalent residues, but also for those with side chains inaccessible to solvent. Buried polar, hydrogen-bonded residues tend to be even more strongly conserved in highly divergent families (Bajaj and Blundell, 1984; Blundell *et al.*, 1986). Thus, in the cellular and retroviral aspartic proteinases, which have very low sequence identities, only a buried, hydrogen-bonded threonine is invariant (Pearl and Blundell, 1984) apart from the catalytic aspartates and two glycines. Similarly, in the Greek key motifs of the eye lens crystallins only a buried hydrogen-bonded serine is conserved apart from an invariant glycine. Such broad comparisons can assist in formulating rules that restrict the number of sequences that we need consider for a particular structure.

We use such rules in several ways in selecting structures for modeling. We have used them to construct tertiary templates for a globular domain, for example, in the identification of the crystallin Greek key motif in a bacterial surface protein (Wistow *et al.*, 1985). We are also using them to identify key residues in structurally variable regions, in the following manner. First we select a series of fragments from the data base of protein structures that are geometrically compatible with the framework. These are then clustered, and for each cluster—often containing only one example—key residues are selected on the basis of low solvent accessibility, strong hydrogen bonding, unusual torsion angles, and so forth. These approaches to identifying key residues and their use in selecting conformers in variable regions are being encoded in COMPOSER (Sutcliffe, 1988; Blundell *et al.*, 1988).

4. *Generation of a Model from a Sequence of an Unknown Using Rules from Comparison of Structures*

In the previous sections we have shown that comparing protein three-dimensional structures can define topological equivalence and phylogenies for members of homologous or analogous families. The equivalenced structures provide a basis for deriving rules for the selection and alignment of sequences that adopt the family fold. We will now consider the construction of the protein model using the knowledge derived from these sequence and structural comparisons.

4.1. *Construction of the Model by Assembly of Rigid Groups in Three Dimensions*

We first consider the construction of a model by a superposition of rigid three-dimensional structures. We assume initially that there are several homologous or analogous structures from which a framework can be generated. The first stage is to select the structures that will be most useful in the modeling 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.*, 1990a,b). This enables selection of the set of structures that are clustered around or near the sequence of the unknown. These structures are then used to produce a framework for the unknown (see Fig. 2b) in which contribution of each structure is weighted according to its percentage sequence identity to the unknown (Sutcliffe *et al.*, 1987a). The framework so derived is an average structure, and it is endowed with real geometry by least-squares fitting the 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 multidimensional scaling analysis or tree construction (see Section 2.3). Key residues are identified using procedures discussed in Section 3.3, and the fragments are ranked. 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.

Alternative procedures may, in fact, need to be adopted. First, it may be best to extend the framework using differing subsets of homologous structures at each variable region. Second, the rules developed by Thornton and her colleagues (Sibanda and Thornton, 1985; Edwards *et al.*, 1987; Wilmot and Thornton, 1988; Sibanda *et al.*, 1989) may be used to select a loop not recognized by the key residues procedure. Third, regions within the structurally conserved core of the known set may require insertions or deletions. In certain cases in which the equivalent fragments are of the same length, but the key residues are changed, an alternative conformer may be required. In these cases, a further definition of the region to be replaced is required. In general, insertions, deletions, and main-chain distortions are made locally and, where possible, in regions of irregular secondary structure.

The third step is to replace side chains. 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). 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. When there is no useful prediction, the most probable conformation is chosen, and when there is more than one prediction, the conformation closest to the median of the predictions is chosen.

This procedure for modeling is very successful when the known structures cluster around that to be predicted and when the percentage sequence identity is high (>40%). For example, in a model building of a porcine kallikrein from four other structurally known mammalian serine proteinases, the root mean square difference between the model and the real structure is 0.64 Å for the 150 residues defined to be in the framework by spatial superposition. Where the structure to be predicted lies outside the cluster and the sequence identity is <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 side chains within contact regions (Lesk and Chothia, 1980; J. Overington, unpublished results). For closely related structures, the structurally variable regions tend to be small and the predictions are reasonably good, but for divergent structures, long insertions can lead to poorly modeled conformations. Similarly, 80% of the side-chain conformations are correctly predicted for closely homologous structures, but this decreases for structures that are less similar.

4.2. Construction of a Model Using Optimization Techniques

An alternative approach for modeling divergent protein structures is to adopt distance geometry or optimization techniques such as those used for derivation of protein structures from the two-dimensional NMR data (Braun and Go, 1985; Havel and Wuthrich, 1985). First, protein structures and fragments that are homologous or analogous to the sequence of the unknown are selected using the methodology of Johnson *et al.* (1990a). Second, these known structures and fragments are aligned using the comparison method that takes properties and relationships into account (Šali and Blundell, 1990). Third, this alignment is used to derive the tertiary template, and the tertiary template is then aligned with the sequence of the unknown. From the alignment of the unknown with the tertiary template and from the general rules of protein structure, the set of constraints on the structure of the unknown is obtained. For example, hydrogen bonds, secondary structure, solvent accessibilities, and so on, which are conserved in the alignment of the known structures, constrain the degrees of freedom available to the structure of the unknown. Finally, these constraints are used as the input to the optimization program (Šali and Blundell, unpublished results), similar to the variable target function minimization procedure of Braun and Go (1985), which calculates the structure of the unknown, minimizing the violations of the constraints.

4.3. Refinement of the Model

All knowledge-based procedures require simulation of the solvent, energy minimization, and molecular dynamics simulations to optimize the structure and to provide a useful model of the time-and-space-averaged structures determined by X-ray analysis and two-dimensional NMR.

5. *Applications to Design of Novel Molecules*

Knowledge-based modeling has applications both to receptor-based drug design and to protein engineering (Fig. 4).

5.1. *Receptor-Based Drug Design*

Although the sequences of many receptors have recently been determined, the three-dimensional structures are known for very few of pharmaceutical interest. 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, when the sequence of human renin was determined in 1984, no three-dimensional structures of renins had been determined. This has taken 4 years; only the structures for homologous fungal aspartic proteinases were accurately analysed by X-ray analysis. Modeling 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 proteinases complexed with human renin inhibitors to give a model of the human renin—human angiotensinogen (fragment) transition state complex (Foundling *et al.*, 1987; Blundell *et al.*, 1987). 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 approximately 0.5 Å (comparison of alpha-carbons) close to the active site, but will have errors in excess of 1.5 Å in the peripheral loops.

5.2. *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 nonaqueous solvents, or other properties in a predictable fashion. The knowledge-based procedures developed for modeling local insertions and deletions and side-chain replacements (Section 4.1) provide a useful starting point, although energy minimization and molecular dynamics procedures in a simulated aqueous environment will be needed to explore local conformations.

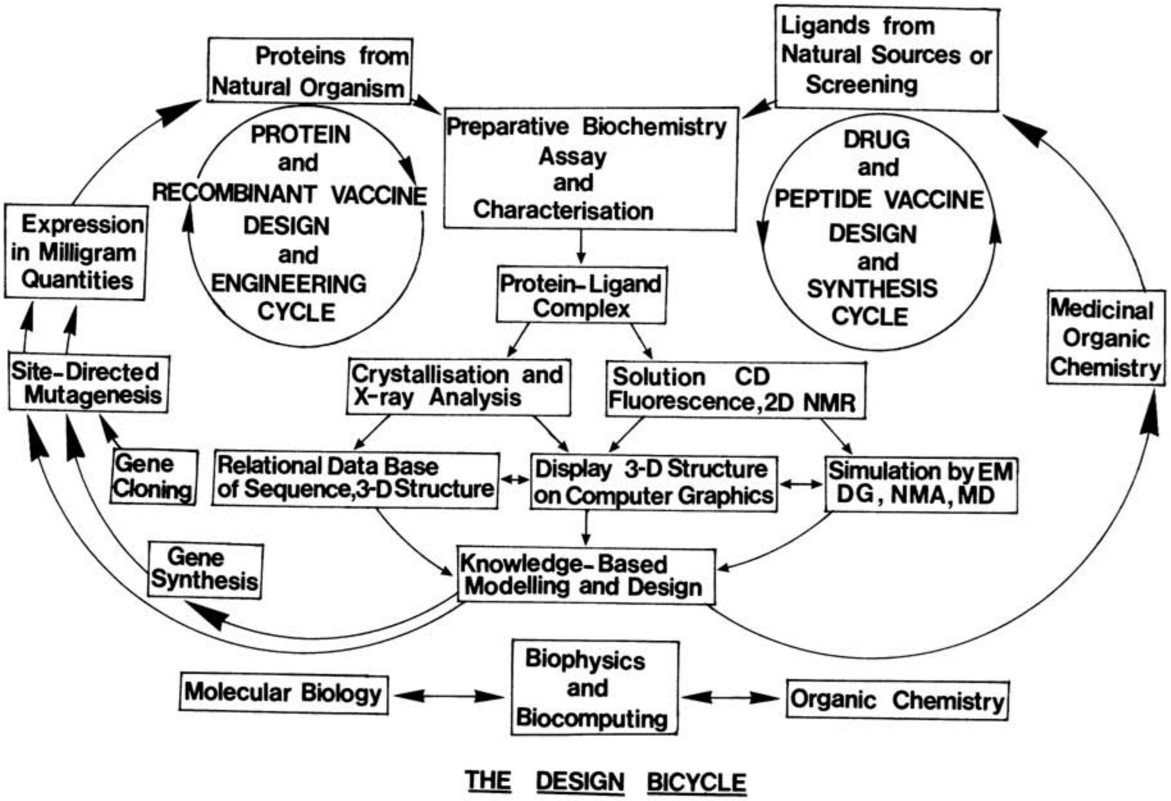
Proteins from Natural Organism
Preparative Biochemistry Assay and Characterisation
Ligands from Natural Sources or Screening
PROTEIN and RECOMBINANT VACCINE DESIGN and ENGINEERING CYCLE
DRUG and PEPTIDE VACCINE DESIGN and SYNTHESIS CYCLE
Expression in Milligram Quantities
Protein-Ligand Complex
Medicinal Organic Chemistry
Site-Directed Mutagenesis
Crystallisation and X-ray Analysis
Solution CD Fluorescence, 2D NMR
Gene Cloning
Relational Data Base of Sequence, 3-D Structure
Display 3-D Structure on Computer Graphics
Simulation by EM DG, NMA, MD
Gene Synthesis
Knowledge-Based Modelling and Design
Molecular Biology
Biophysics and Biocomputing
Organic Chemistry
THE DESIGN BICYCLE

Figure 4. Bi-cycle illustrating the common structural analysis and design steps in the engineering of proteins, vaccines, and drugs.

5.3. *Chimeric Molecules*

Let us consider the design of a chimeric molecule that comprises a tissue plasminogen activator (tPA) serine proteinase domain linked to the COOH-terminus of an Fab fragment of a monoclonal antibody with fibrin specificity. The modeling procedures are first used to model the serine proteinase and the immunoglobulin domain on the basis of homologous structures (Blundell *et al.*, 1988; Harris, 1987; Overington *et al.*, 1988). Second, 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 contiguous surfaces of the two linked domains are mutated using the side-chain replacement algorithm so that they are compatible with each other and the solvent.

5.4. *Ab Initio Protein Design*

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 called CRYSTANOVA (based on the stable eye lens crystallins) that is designed to bind copper in a similar way to superoxide dismutase (Hubbard, 1988). Although such projects are currently academic, in the future protein engineers may 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.

ACKNOWLEDGMENTS. We are grateful to all our colleagues in the modeling group for very useful discussion. We thank the American Cancer Society for a fellowship (M.S.J.). J.P.O. is supported by a SERC and CASE studentship with Pfizer. A.S. is funded by the ORS Award Scheme and the Research Council of Slovenia.

References

Argos, P., 1987, A sensitive procedure to compare amino acid sequences, *J. Mol. Biol.* **193:**385–396.

Bajaj, M., and Blundell, T. L., 1984, Evolution and the tertiary structure of proteins, *Annu. Rev. Biophys. Bioeng.* **13:**453–492.

Bashford, D., Chothia, C., and Lesk, A. M., 1987, Determinants of a protein fold: unique features of the globin amino acid sequence, **196:**199–216.

Bedarkar, S., Turnell, W. G., Blundell, T. L., and Schwabe, C., 1977, Relaxin has conformational homology with insulin, *Nature* **270:**449–451.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanovichi, T., and Tasumi, M., 1977, The Protein Data Bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.* **112:**535–542.

Blundell, T. L., Bedarker, S., Rindernecht, E., and Humbel, R. E., 1978, Insulin-like growth factor: A model for tertiary structure accounting for immunoreactivity and receptor binding, *Proc. Natl. Acad. Sci. USA* **75:**180–184.

Blundell, T. L., Sibanda, B. L., and Pearl, L., 1983, Three dimensional structure, specificity and catalytic mechanism of renin, *Nature* **304:**237–275.

Blundell, T. L., Jenkins, J., Pearl, L., and Sewell, T., 1985, The high resolution structure of endothiapepsin, in: *Aspartic Proteinases and Their Inhibitors* (V. Kostka, ed.), pp. 151–161, Walter de Gruyter, Berlin.

Blundell, T. L., Barlow, D., Sibanda, B. L., Thornton, J. M., Taylor, W., Tickle, I. J., Sternberg, M. J. E., Pitts, J. E., Haneef, I., and Hemmings, A. M., 1986, Three-dimensional structural aspects of the design of new protein molecules, *Phil. Trans. Roy. Soc. Lond.* **317:**333–344.

Blundell, T. L., Cooper, J., and Foundling, S. I., 1987, On the rational design of renin inhibitors: X-ray studies of aspartic proteinases complexed with transition-state analogues, *Biochemistry* **26:**5585–5590.

Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M., 1988, Knowledge-based protein modelling and design, *Eur. J. Biochem.* **172:**513–520.

Braun, W., and Go, N., 1985, Calculation of protein conformations by proton–proton distance constraints: A new efficient algorithm, *J. Mol. Biol.* **186:**611–626.

Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L., 1969, A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg white lysozyme, *J. Mol. Biol.* **42:**65–86.

Chothia, C., and Lesk, A. M., 1986, The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5:**823–826.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V., and Poljak, R. J., 1986, The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure, *Science* **233:**755–758.

Crippen, G. M., 1977, A novel approach to calculation of conformation: Distance geometry, *J. Comp. Physiol.* **24:**96–107.

Edwards, M. S., Sternberg, M. J. E., and Thornton, J. M., 1987, Structural and sequence patterns in the loops of beta-alpha-beta units, *Prot. Eng.* **1:**173–181.

Efimov, A. V., 1986, The conformation of beta-turns, *Mol. Biol. (Moscow)* **20:**2250–2260 (in Russian).

Felsenstein, J., 1985, Confidence limits on phylogenies: an approach using the bootstrap, *Evolution* **39:**783–791.

Foundling, S. I., Cooper, S. I., Watson, J., Cleasby, F. E., Pearl, L. H., Sibanda, B. L., Hemmings, A., Wood, S. P., Blundell, T. L., Valler, T. L., Norey, Kay, J., Boger, J., Dunn, B. M., Leckie, B. J., Jones, D. M., Atrash, B., Hallett, A., and Szelke, M., 1987, High resolution X-ray analyses of renin inhibitor–aspartic proteinase complexes, *Nature* **327:**349–352.

Greer, J., 1981, Comparative model-building of the mammalian serine-proteases, *J. Mol. Biol.* **153:**1027–1042.

Harris, T. J. R., 1987, Second-generation plasminogen activators, *Prot. Eng.* **1:**449–450.

Havel, T., and Wuthrich, K., 1985, An evaluation of the combined use of NMR and distance geometry for the determination of protein conformations in solution, *J. Mol. Biol.* **182:**281–294.

Hubbard, T. J. P., 1988, The design, expression and characterization of a novel protein, Ph.D. thesis, University of London.

Hubbard, T. J. P., and Blundell, T. L., 1987, Comparison of solvent inaccessible cores of homologous proteins: Definitions useful for protein modelling, *Prot. Eng.* **1:**159–171.

Isaacs, N., James, R., Niall, H., Bryant-Green, G., Wood, S., Dodson, G. G., Evans, A., and North, A. C. T., 1978, Relaxin and its structural relationship to insulin, *Nature* **271:**278–281.

Johnson, M. S., Sutcliffe, M. J., and Blundell, T. L., 1990a, Molecular anatomy: Phyletic relationships derived from three-dimensional structures of proteins, *J. Mol. Evol.* **30:**43–59.

Johnson, M. S., Šali, A., and Blundell, T. L., 1990b, Phylogenetic relationships from three-dimensional protein structures, *Meth. Enzymol.* **183:**670–690.

Jones, T. H., and Thirup, S., 1986, Using known substructures in protein model building and crystallography, *EMBO J.* **5:**819–822.

Lesk, A. M., and Chothia, C., 1980, How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins, *J. Mol. Biol.* **136:**225–270.

Matthews, B. W., and Rossmann, M. G., 1985, Comparison of protein structures, *Meth. Enzymol.* **115:**397–420.

McGregor, M. J., Islam, S. A., and Sternberg, M. J. E., 1987, Analysis of the relationship between side-chain conformation and secondary structure in globular proteins, *J. Mol. Biol.* **198:**295–310.

McLachlan, A. D., 1982, Rapid comparison of protein structures, *Acta Cryst.* **A38:**871–873.

Milner-White, J., and Poet, R., 1986, Four classes of beta-hairpins in proteins, *Biochem. J.* **240:**289–292.

Murthy, M. R. N., 1984, A fast method of comparing protein structures, *FEBS Lett.* **168:**97–102.

Needleman, S. B., and Wunsch, C. D., 1970, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* **48:**443–453.

Overington, J. P., Sutcliffe, M. J., Watson, F., James, K., Campbell, S., and Blundell, T. L., 1988, The knowledge-based modelling of tissue-type plasminogen activator and its interactions with its inhibitor endothelial cell plasminogen activator inhibitor, *Proceedings of the 8th International Biotechnology Symposium,* Vol. 1, Paris, pp. 279–304.

Pearl, L., and Blundell, T. L., 1984, The active site of aspartic proteinases, *FEBS Lett.* **174:**96–101.

Ponder, J. W., and Richards, F. M., 1987, Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes, *J. Mol. Biol.* **193:**775–791.

Richards, F. M., and Kundrot, C. E., 1988, Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure, *Proteins* **3:**71–84.

Richardson, J. S., 1981, The anatomy and taxonomy of protein structure, *Adv. Prot. Chem.* **64:**167–339.

Rossmann, G. M., and Argos, P., 1976, Exploring structural homology of proteins, *J. Mol. Biol.* **105:**75–95.

Šali, A., and Blundell; T. L., 1990, The definition of general topological equivalence in protein structures, *J. Mol. Biol.* **212:**403–428.

Sibanda, B. L., and Thornton, J. M., 1985, Beta-hairpin families in globular proteins, *Nature* **316:**170–316.

Sibanda, B. L., Blundell, T. L., Hobart, P. M., Fogliano, M., Bindra, J. S., Dominiy, B. W., and Chirgwin, J. M., 1984, Computer graphics modelling of human renin: Specificity, catalytic activity and intron–exon junctions, *FEBS Lett.* **174:**102–111.

Sibanda, B. L., Blundell, T. L., and Thornton, J. M., 1989, The conformation of beta-hairpins in protein structures: A systematic classification with applications to modelling by homology, electron density fitting and protein engineering, *J. Mol. Biol.* **206:**759–777.

Summers, N. L., Carson, W. D., and Karplus, M., 1987, Analysis of side chain orientations in homologous proteins, *J. Mol. Biol.* **196:**175–198.

Sutcliffe, M. J., 1988, An automated approach to the systematic model building of homologous proteins, Ph.D. thesis, University of London.

Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L., 1987a, Knowledge-based modelling of homologous proteins; Part I: Three-dimensional frameworks derived from simultaneous superposition of multiple structures, *Prot. Eng.* **1**:377–384.

Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L., 1987b, Knowledge based modeling of homologous proteins; Part II: Rules for the conformation of substituted sidechains, *Prot. Eng.* **1**:385–392.

Taylor, W. R., 1986, Identification of protein sequence homology by consensus template alignment, *J. Mol. Biol.* **188**:233–258.

Wilmot, C. M., and Thornton, J. M., 1988, Analysis and prediction of the different types of beta-turn in proteins, *J. Mol. Biol.* **203**:221–232.

Wistow, G., Summers, L., and Blundell, T. L., 1985, *Myxococcus xanthus* spore coat protein S may have a similar structure to vertebrate lens beta/gamma crystallins, *Nature* **316**:771–773.