11

From modelling homologous proteins to prediction of structure

A. Sali, J.P. Overington, M.S. Johnson and T.L. Blundell
ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX

1. INTRODUCTION

Divergent evolution has given rise to families of homologous proteins that differ in sequence but adopt the same general fold. This provides an opportunity to learn about the three-dimensional structures of proteins if their sequences have been defined and at least one other member of the family has a structure defined by X-ray analysis or nuclear magnetic resonance.

The first application recorded in the literature of this procedure was the construction of a model for alpha-lactalbumin on the basis of the 3D structure of lysozyme (Browne *et al.* 1969). Other applications included construction of models for relaxins and insulin-like growth factors (Bedarkar *et al.* 1977, Blundell & Humbel 1980) for review), various serine proteinases (Greer 1981) and aspartic proteinases such as renin (Blundell *et al.* 1983). The advent of computerized techniques, particularly the computer graphics program FRODO (see Jones & Thirup (1986) for references), made the task of replacing sidechains and making insertions and deletions more straightforward. However, modelling was rarely performed applying rigorous rules, although some systematic procedures were suggested, for example for the use of loops from homologous proteins (Greer 1981).

The challenge now is to learn from the experience of these often subjective and usually interactive modelling procedures. We require an automated procedure that uses the known structures and the rules obtained from them — the knowledge base. It has been apparent for some time that such an approach can have very wide applications for prediction of protein structure in general (see Blundell *et al.* 1987). This arises because many protein sequences adopt the same general fold even when

there is no obvious evolutionary relationship. Recent structure determinations suggest that the majority of new structures comprise motifs or domains that have been previously identified in other, often functionally different proteins. For this reason it is important to develop general approaches that can be extended from modelling homologous structures with clear similarities in their sequences to the more difficult problem of protein prediction using weaker analogies between tertiary structures of proteins that may have little sequence similarity and no functional or evolutionary

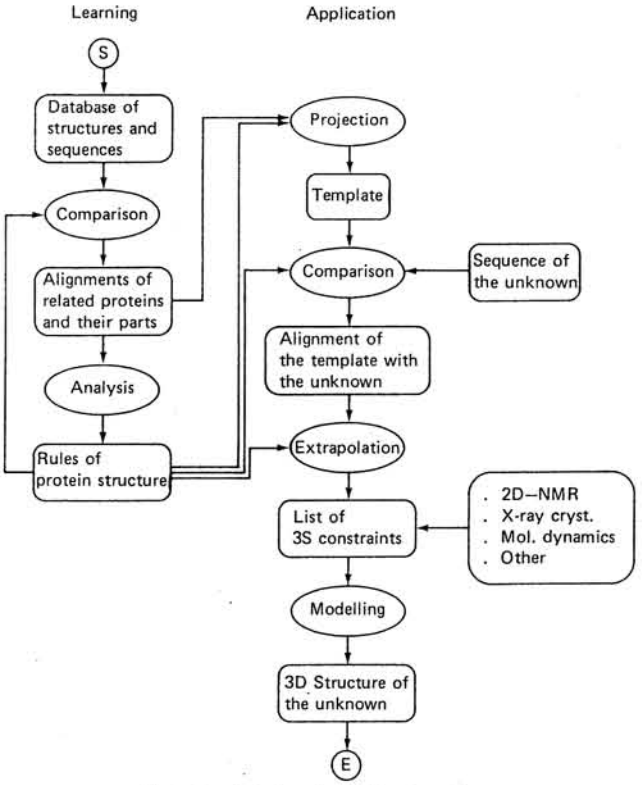


Fig. 1. Knowledge-based modelling of proteins.

relationship. Figure 1 shows a scheme that generalizes such approaches to modelling and provides an outline of our review.

We must first establish and formulate rules. We achieve this not only by analysing individual protein structures as they are defined by experiment, but also by comparing sequences and 3D structures. For this we need methods not only for aligning many

sequences but also for comparing three-dimensional structures in order to establish equivalences within a family of proteins. The rules are then derived by careful, computerized analysis of the compared structures. This is the learning stage and it is summarized on the left-hand side of Fig. 1.

These rules can then be applied in two important ways shown on the right-hand side of Fig. 1. First, they are used to define all those sequences that can adopt each experimentally defined fold; this can be considered as a mapping from a 3D structure onto a 1D sequence. It is usually achieved through construction of a 'template', which summarizes knowledge about the family fold and is presented in a form suitable for comparison with the sequence of the 'unknown'. Secondly, when the alignment of sequence with the template has been achieved, the rules are used to map structural features to the sequence of the protein to be modelled.

In this review we emphasize our own approaches to modelling (Blundell *et al.* 1988; Sali *et al.* 1990). The most helpful learning set is that of homologous proteins; because there are fewer ambiguities in their comparisons, they provide a reliable source for the definition of rules but they remain a challenge for modelling. We describe new approaches that can be used in modelling more distantly related protein structures.

2. PROCEDURES FOR COMPARISON AND CLUSTERING OF PROTEIN SEQUENCES AND STRUCTURES

Our first task is to develop systematic ways of comparing and clustering protein structures.

The comparison of 3D structures often involves rigid-body least-squares superposition of the C_α positions (KenKnight, 1984, for review). Several homologous structures can be aligned (Sutcliffe *et al.*, 1987a) without bias to any one in the set in order to define a 'framework', which comprises a set of helices or strands that are conserved in the family. However, although dissimilar proteins usually retain the general arrangement of strands and helices, the differences in relative orientation and position may preclude their direct superposition (Chothia & Lesk, 1986; Hubbard & Blundell, 1987; Johnson *et al.* 1990a,b).

The extension of comparison methods to more dissimilar tertiary structures was addressed more than a decade ago (Eventoff & Rossmann, 1975; Matthews & Rossmann, 1985). The methods included information about mainchain direction in the alignment or based their comparisons on rigid body superposition of small parts of the whole structure. Our approach to protein comparison (Sali & Blundell, 1990; Z. Zhu, unpublished results), encoded in the program COMPARER, simultaneously employs a large number of protein features that were used individually in previous approaches. COMPARER can be used to obtain alignments of both sequence and three-dimensional structures. We first define the protein as an indexed string of elements that may exist at several levels in the protein hierarchical organization — residue, secondary structure, supersecondary structure, motif or domain. We then associate with every element features, either properties or relationships that indicate

a common fold. Properties include element identity (for example, residue or secondary structure type), hydrophobicity, local conformation and solvent accessibility. Comparison of all such properties can be incorporated in a residue weight matrix and optimal alignment can then be derived using the dynamic programming approach. A similar approach based principally on intra-molecular distances has been described by Taylor & Orengo (1989).

At each level of the hierarchical structure of proteins, specific relationships such as hydrogen bonding interactions or packing relations that tend to be conserved in protein folds can also be used in our alignment procedure (Sali & Blundell 1990). However, a relationship affects more than one element in a sequence and this makes the conventional dynamic programming approach expensive in computer time. Instead we use simulated annealing to provide an initial set of equivalences based on relationships which are then introduced directly into the residue by residue weight matrix.

Figure 2 shows part of the alignment using COMPARER of the two domains of pepsins and the subunits of retroviral proteinases. These have only three residues that are identical in all the structures compared and the sequences vary from 99 to 170 amino acid residues in length. A direct multiple superposition equivalences very few residues and these are mainly in the active site region. In contrast the COMPARER alignment identifies all those strands and helices that have previously been considered equivalent on a more subjective basis and which have been shown to be common with the retroviral proteinases (see, for example, Lapatto *et al.* (1989)).

While extensive methodology for tree construction from protein sequences has been developed for the study of evolution (Doolittle 1989), the clustering of protein three-dimensional structures has been less studied. Rao *et al.* (1975) constructed dendrograms based on structural features alone to describe distant phylogenetic relationships among the mono-nucleotide and di-nucleotide binding proteins. We have shown that a useful structural pairwise distance metric can be defined from fractional topological equivalence and root mean square deviation as calculated by least squares superposition (Johnson *et al.* 1990a,b). This distance measure correlates well with the sequence metric. Recently we have extended this approach by reflecting additional structural and sequence features in the classification (Johnson *et al.* 1990a,b; Sali & Blundell 1990). Moreover, since these features can include relationships such as hydrogen bonding patterns, which are known to be conserved in evolution, structures that bear little similarity in other respects can be compared and classified at statistically significant levels. Figure 3 shows classifications of cytochrome c structures based on sequences and structures (Johnson *et al.* 1990a,b).

3. ESTABLISHING RULES FROM FAMILIES OF HOMOLOGOUS PROTEIN STRUCTURES

We have produced a database of alignments of three-dimensional structures obtained by COMPARER (Z. Zhu, A. Sali, J. Overington, T.L. Blundell). We have used this to derive a set of rules useful for modelling. For example, we can quantify the well-

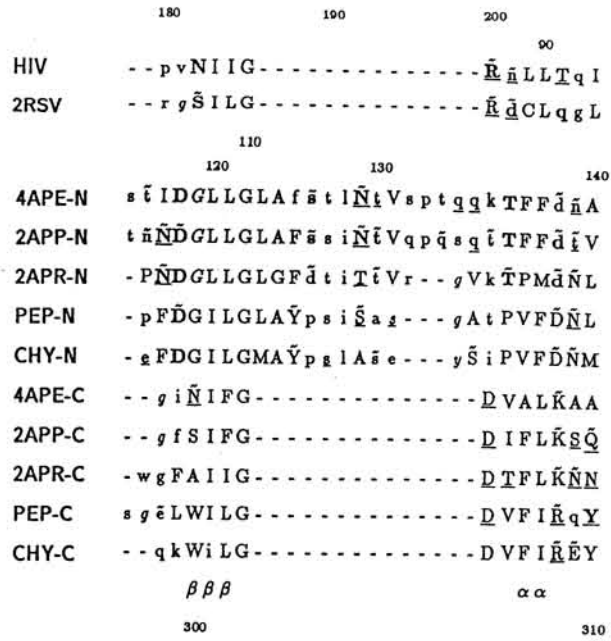


Fig. 2. Two sections of the alignment of sequences of aspartic proteinases achieved by comparing the three-dimensional structures using COMPARER (Sali & Blundell 1989). APE: endothiapepsin; APP: penicillopepsin; APR: rhizopuspepsin; PEP: hexagonal porcine pepsin; CHY: calf chymosin; RSV: Rous sarcoma virus proteinase; HIV: human immunodeficiency virus proteinase. The last letter refers to the amino (N) or carboxy (C) terminal domains of the pepsins. The coordinates of the three-dimensional structures were obtained from the PDB databank (Bernstein *et al.* 1977). The amino acid code is the standard one-letter code formatted using the following convention: *italic*, positive ϕ; UPPER CASE, solvent inaccessible residue; lower case, solvent accessible residue; **bold type**, hydrogen bond to mainchain amide; underline, hydrogen bond to mainchain carbonyl; tilde, sidechain–sidechain hydrogen bond.

known rule that solvent inaccessible residues tend to be among the more conserved residues in a family. Rules obtained from comparisons also include those that correlate an unknown sidechain dihedral angle with those dihedral angles for equivalent positions in related proteins (Summers *et al.* 1987; Sutcliffe *et al.* 1987b; Donnelly *et al.* 1990). Such correlations can be best represented as a multidimensional probability density table. This table has as many dimensions as there are features included in the analysis and every dimension has as many columns as the corresponding feature can assume. Elements in this table are then filled by simply counting the number of occurrences of the corresponding combination of features in the database of alignments

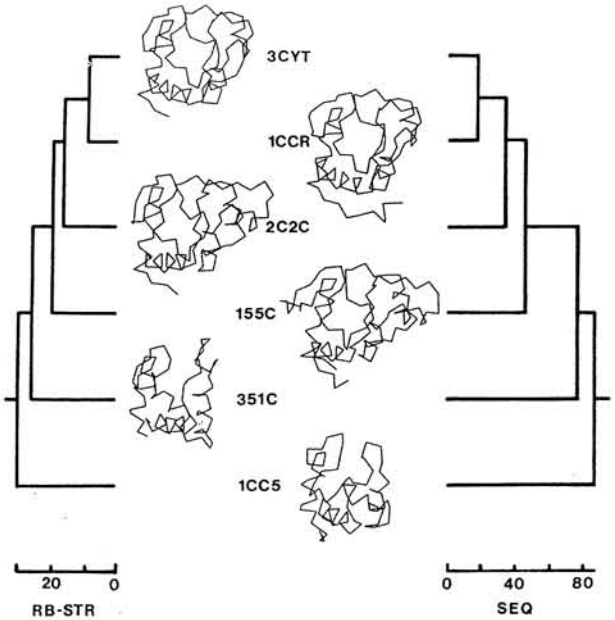


Fig. 3. The classification of cytochrome structures on the basis of sequence and structure. The tree was constructed by the program KITSCH from the PHYLIP package (1985).

where all the features including the 'unknown' are defined. Such rules can be used directly to obtain spatial constraints on the sequence of the unknown.

Rules for the substitution of amino acids in three-dimensional structures are derived in a similar way by counting how many times two residue types occur at structurally equivalent positions. We have constructed a number of residue substitution tables (Overington *et al.* 1990) in which only a subset of residues that have a certain structural environment are considered. For example, 20 by 20 substitution tables were built separately for inaccessible residues (Fig. 4(a)). Other structural features included in our analysis were local mainchain conformation (positive ϕ angle, α-helical, β-strand or other) and sidechain hydrogen-bonding to peptide groups (Fig. 4(b)) or other sidechains. These environment-dependent substitution tables are specific examples of the general multidimensional density table described above.

The environment-dependent substitution tables quantify the importance of individual structural features for the acceptance of amino acid mutations in evolution. For example, Fig. 4 shows that the substitution of polar residues such as aspartic acid, asparagine, glutamine, serine and threonine is strongly influenced by sidechain accessibility and hydrogen bonding. Large differences exist in the mutability pattern of the same residue type in different structural environments. Hydrogen-bonded and

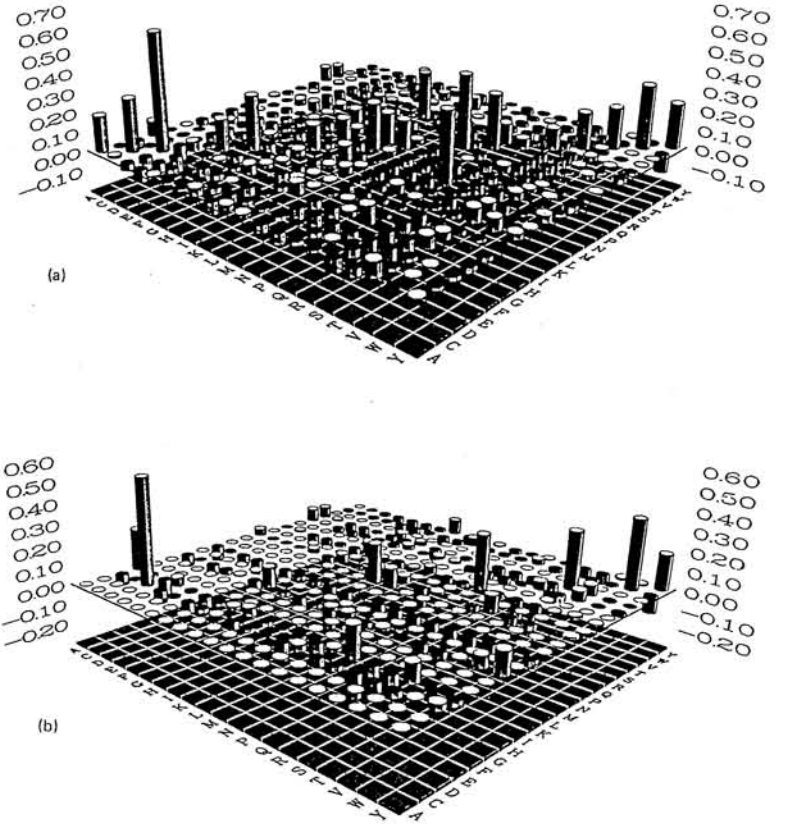


Fig. 4. Difference substitution tables for amino acids that occupy (a) solvent inaccessible and (b) solvent inaccessible and sidechain hydrogen-bonded to a mainchain carbonyl positions in globular proteins. The horizontal axis is that of an amino acid in such an environment in three-dimensional structure of a protein. The vertical axis is the amino acid type in an homologous protein at a topologically equivalent position defined by COMPARER.

inaccessible residues are among the most highly conserved residues in families of proteins. Their structural roles are relatively specific; as a result it is not easy to vary the amino acid type and also retain the important structural role. One specific case is shown in Fig. 2 where Thr 33 and Thr 216 of pepsin are conserved or conservatively varied to serine in all pepsin-like and retroviral proteinases. These buried residues play an important role in holding together the two subunits in retroviral proteinases and the two lobes in pepsins.

4. DERIVATION OF A SEQUENCE TEMPLATE FROM A 3D STRUCTURE OF A PROTEIN

Traditionally, templates have been constructed from the alignment of many sequences of often quite divergent structures (Doolittle 1989). This may allow identification of sequence fingerprints that are characteristic of the structure or function. Such templates can be used in the form of consensus sequences or mutability profiles to search out distantly related proteins in the sequence database. By defining Venn diagrams describing relatedness of amino acids, Taylor (1986) has increased the versatility of templates when few or only closely related structures are available.

One or more protein three-dimensional structures should also provide a basis for the construction of templates. Ponder & Richards (1987) have suggested such an algorithm for generating all sequences of amino acids and their sidechain conformations that are consistent with a particular fold. The tables described above can also be used to estimate the probability of substituting any amino acid at a particular position in a known three-dimensional structure. For each topologically equivalent position in each known structure, we use the tables to predict the variability of amino acid residues. This allows use of knowledge of the 3D structure to project constraints onto the 1D sequence or to construct the family template (Johnson, M.S., Overington, J. and Blundell, T.L., unpublished results). Such a template expressed in the form of a sequence can be used to align the family fold with the sequence to be modelled.

The templates of all known three-dimensional structures or families of structures including loops, motifs, domains and complete globular proteins should be precalculated so that a new sequence can be compared with them rather than with individual proteins. This will result in a better alignment of whole proteins or their parts and thereby extend the usefulness of knowledge-based or comparative modelling.

5. MODELLING 3D STRUCTURE

5.1 Composer

In the previous sections we have described procedures for comparison of protein three-dimensional structures. We have shown that comparisons of homologous families of proteins can give rise to rules. For example, we have described rules that relate the sidechain dihedral angle with the residue type at equivalent positions in homologous proteins and rules that predict the sequence variability at each position in the tertiary structure of a protein; there are many others. We shall now consider methods for constructing a model on the basis of such rules derived not only from comparison of related structures but also from the analyses of protein structures in general. The use of these rules depends very importantly on the alignment of the sequence of the protein to be modelled with the template for the family fold.

Most current methods depend on the assembly of rigid fragments (Jones & Thirup 1986; Blundell *et al.* 1987, 1988, Claessens *et al.* 1989). In our approach encoded in the program COMPOSER we first select the homologous structures that are most useful for construction of the model; this we do on the basis of the sequence and

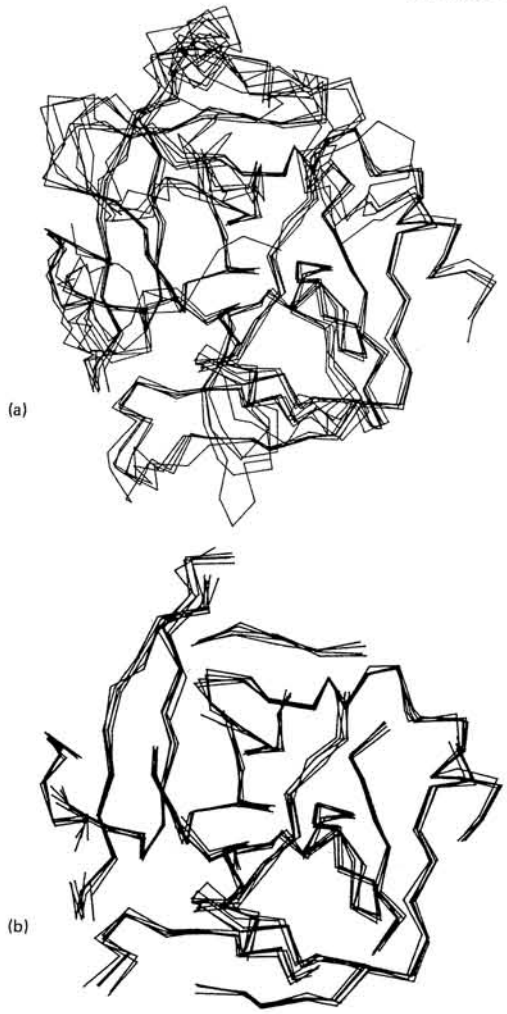
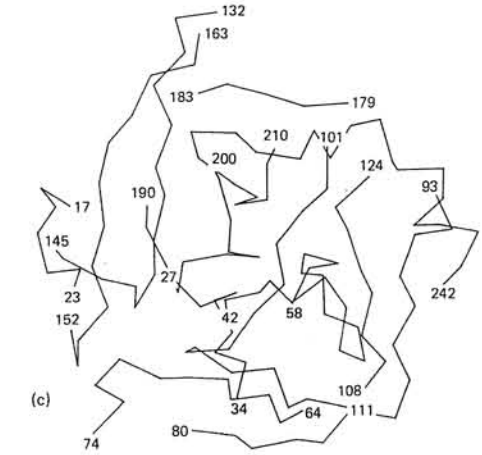


Fig. 5. Modelling the serine proteinase domain of tissue plasminogen activator from homologous serine proteinases by the program COMPOSER. (a) shows the superposition of the structures defined by X-ray analysis. (b) indicates the fragments in the structurally conserved regions that contribute towards generation of the framework shown in (c). Fragments, selected using rules from a broader database of structures, are used to model the structurally variable regions. The C_α atom positions of the complete model are shown in (d). Sidechains, not shown, are also generated by a set of rules derived from comparisons of known structures.



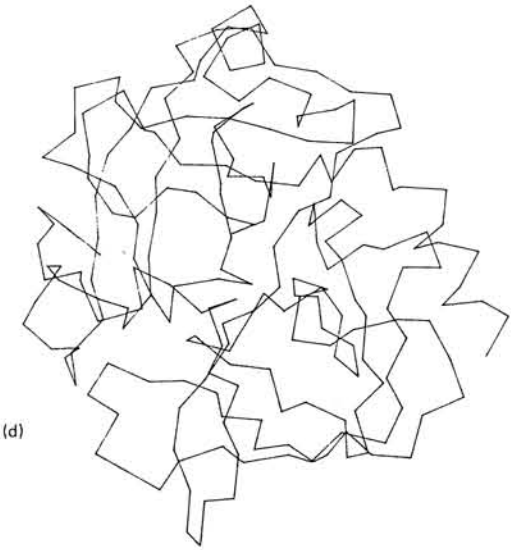


Fig. 5. *continued*

structure classification as described above (Johnson *et al.* 1990a,b). Three sets of fragments are selected:

(1) Fragments from the framework are defined by multiple least-squares superposition of the chosen structures (Sutcliffe *et al.* 1987a).
(2) Fragments for regions outside the framework are selected from the database of loop substructures using a distance filter in a similar way to Jones & Thirup (1986). The sequences of selected fragments are then compared to the sequence of the unknown using the environment-dependent substitution tables (McCleod, Thomas, Topham, Overington, Johnson and Blundell, 1990, work in progress). The top-ranking fragment is annealed onto the core using an optimization procedure (Eisenmenger, unpublished results) and checked for overlap with other parts of the model structure. If it is rejected on these grounds, the next ranking fragment is processed in the same way.
(3) Fragments of sidechains are selected by using a set of rules derived from the analysis of sidechain dihedral angles at topologically equivalent positions in homologous structures (Sutcliffe *et al.* 1987b). The 1200 rules derived from this analysis include one for each of the 20 by 20 amino acid replacements in each of the three secondary structure types (α-helix, β-strand or irregular). Where there is no applicable rule, the most probable conformation is chosen from a rotamer library, and where there is more than one prediction, the one closest to the median of all predictions is chosen. See also Sumners *et al.* (1987) for a related approach.

Finally, the model is energy minimized to remove small inconsistencies such as steric clashes. This modelling procedure is very successful where the known structures cluster around that to be predicted and where the percentage sequence identity to the unknown is high (greater than 40%). For example, in a model building of porcine trypsin from four other structurally known serine proteinases, the root mean square difference between the model and the known structure is 0.60 Å for the 150 residues defined in the framework. Similarly, 80% of sidechain conformations are correctly predicted for closely homologous structures. In all cases the accuracy of the prediction decreases very quickly as the sequence identity between the known and unknown decreases. For these cases a different approach is essential.

5.2 Modeller

New modelling techniques are required that are not restricted by a rigid body model of protein structure. These are best defined in terms of distance constraints in a similar way to the methods of interpreting NMR data like those of Braun & Go (1985) and Havel *et al.* (1983) but which also allow simultaneous inclusion of different types of information and rules into the derivation of the model (Sali *et al.* 1990).

The alignment of the sequence with the template is used to derive a list of spatial constraints, most of which can be expressed as distance constraints. For example, if two equivalent positions in the alignment of known structures are always hydrogen bonded, we can assume that the same hydrogen bond exists in the unknown structure

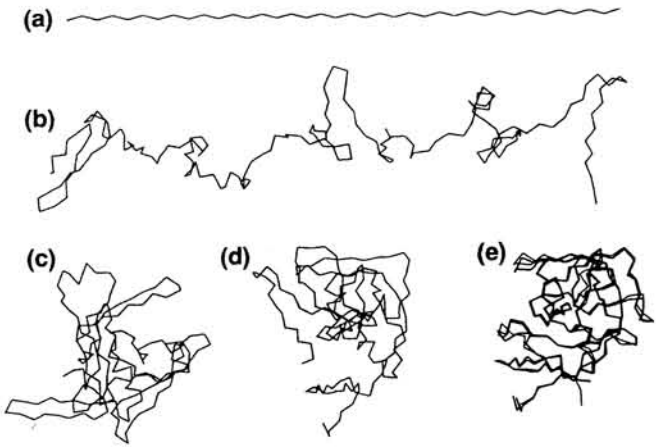


Fig. 6. Generation of a model of a domain of endothiapepsin using rules from homologous aspartic proteinases and from protein 3D structures in general expressed as distance constraints. (a) is the extended chain, (b) shows the influence of mainly local constraints, (c) and (d) are intermediate structures, and (e) is the final structure compared with that experimentally defined.

as well. This represents a distance constraint on the atoms involved in the hydrogen bond.

The most precise description of each constraint on the distance between two atoms treats the distance as a random variable associated with its probability density function. For example, there is a Gaussian probability density function for a length of a chemical bond. Likewise, an estimate of a certain C_α–C_α distance from an equivalent distance in an homologous protein can also be described as a Gaussian probability density function with the mean equal to the known distance and standard deviation proportional to the similarity between proteins and the magnitude of the distance. The probability density function for a sidechain dihedral angle is trimodal with the peaks corresponding to t, g + and g − conformations and their relative magnitudes depending on the particular residue type and the values of equivalent dihedral angles in related known structures (Sali *et al.* 1990).

The goal is then to use the list of spatial constraints on the structure of the protein to construct the three-dimensional model for the protein that will minimize the violations of these constraints. We achieve this by optimization of the molecular probability density function for the whole protein.

In general, every structural feature can be constrained by several knowledge sources. For example, a distance between a particular pair of C_α atoms may be constrained by information from several homologous proteins and also by van der Waals criteria. In such cases we obtain the probability density function for the given feature as a combination of individual probability density functions. The protein

three-dimensional structure is uniquely determined if a sufficiently large number of its spatial features are specified. Obviously, the most probable structure of the molecule as a whole is the one that maximizes the product of all feature probability density functions. So the problem of predicting the structure of the molecule using the knowledge-based approach is transformed into finding the optimum of the complicated function.

The optimization is performed in Cartesian coordinate space using a combination of conjugate gradients and simulated annealing minimization to make the best use of the speed of the former and large radius of convergence of the latter (Sali *et al.* 1990). Additionally, the variable target function (Braun & Go 1985) approach is applied to speed up the program and increase the radius of convergence.

A model for the amino-terminal lobe of endothiapepsin (Fig. 6) obtained by our optimization program has a root mean square deviation to the crystallographically determined structure of 0.76 Å although only C_α–C_α and sidechain dihedral constraints were used.

6. CONCLUSIONS

Knowledge-based or comparative modelling, most often in its simplest form of modelling by homology, is now widely used by biochemists. This reflects the steady advancement in the field including the automation of the algorithms and development of integrated systems synthesizing such diverse tools as databases of sequences and structures (Akrigg *et al.* 1988, Thornton & Gardner 1989), interactive molecular graphics, molecular dynamics and energy minimization together with methods for pattern recognition, comparison and clustering. It also reflects the steady advance in the numbers of sequences and structures defined experimentally.

In this article we have concentrated on a description of our own modelling procedures and those that are closely related. An alternative rule-based approach, which has been developed for predicting protein structures where no obvious homology or analogy is apparent, has been developed by Cohen and his collaborators.

The modelling techniques described here are firmly based on the progress and success of experiment. As a consequence we can expect that the next decade will bring a closer integration of modelling techniques with experimental analyses using crystallography, 2D NMR, image reconstruction in electron microscopy, epitope mapping and cross-linking, which have contributed so much to our understanding of complex protein structures and assemblies. The great challenge will be to unify all techniques for determination or prediction of protein structure into a single protocol making the best use of all available information about the structure of a given protein, regardless of whether it is directly based on experiment, on the broader knowledge base, on empirical force potentials or intuition.

ACKNOWLEDGEMENTS

A.S. was supported by an ORS Awards Scheme, the Research Council of Slovenia, the J. Stefan Institute and Merck, Sharp and Dohme. M.S.J. is funded by the American

Cancer Society and J.P.O. by Pfizer and SERC. We are grateful to the Imperial Cancer Research Fund, the SERC and the EEC for general financial support.

Note. This chapter was written in 1990 and therefore more recent references are not included.

REFERENCES

Argos, P. (1987) *J. Mol. Biol.* **193**, 385–396.
Bedarker, B., Turnell, W.G., Schwabe, C. & Blundell, T.L. (1977) *Nature* **270**, 449–451.
Blundell, T.L. & Humbel, R.E. (1980) *Nature* **287**, 781–787.
Blundell, T.L., Sibanda, B.L. & Pearl, L. (1983) *Nature* **304**, 273–275.
Blundell, T.L., Sibanda, B.L., Sternberg, M.J. & Thornton, J.M. (1987) *Nature* **326**, 347–352.
Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T. & Overington, J. (1988) *Eur. J. Biochem.* **172**, 513–520.
Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C. & Hill, R.L. (1969) *J. Mol. Biol.* **42**, 65–86.
Chothia, C. & Lesk, A.M. (1986) *EMBO J.* **5**, 823–826.
Claessens, M., Cutsem, E.V., Lasters, I. & Wodak, S. (1989) *Prot. Eng.* **2**, 335–345.
Doolittle, R. (1989) *TIBS* **14**, 244–245.
Eventoff, W. & Rossmann, M.G. (1975) *Crit. Rev. Biochem.* **3**, 111–140.
Greer, J. (1981) *J. Mol. Biol.* **153**, 1027–1042.
Havel, T.F., Kuntz, I.D. & Crippen, G.M. (1983) *Bull. Math. Biol.* **45**, 665–720.
Hubbard, T.J.P. & Blundell, T.L. (1987) *Prot. Eng.* **1**, 159–171.
Johnson, M.S., Sutcliffe, M.J. & Blundell, T.L. (1990a) *J. Mol. Evol.* **30**, 43–59.
Johnson, M.S., Sali, A. & Blundell, T.L. (1990b) *Meth. Enzymol.* **783**, 670–690.
Jones, T.H. & Thirup, S. (1986) *EMBO J.* **5**, 819–822.
KenKnight, C.E. (1984) *Acta Cryst.* **A40**, 708–712.
Lapatto, R., Blundell, T.L., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J.R., Whittle, P.J., Danley, D.E., Geoghegan, K.F., Hawrylik, S.J., Lee, S.E., Scheld, K.G. & Hobart, P.M. (1989) *Nature* **342**, 299–302.
Matthews, B.W. & Rossmann, M.G. (1985) *Meth. Enzymol.* **115**, 397–420.
Needleman, S.B. & Wunsch, C.D. (1970) *J. Mol. Biol.* **48**, 443–453.
Overington, J., Johnson, M.J., Sali, A. & Blundell, T.L. (1990) *Proc. Roy. Soc. B.* **241**, 132–145.
Ponder, J.W. & Richards, F.M. (1987) *Proteins*, 775–791.
Sali, A. & Blundell, T.L. (1990) *J. Mol. Biol.* **212**, 403–428.
Sali, A., Donnelly, D. & Blundell, T.L. (1990) unpublished results.
Sibanda, B.L., Blundell, T.L. & Thornton, J.M. (1989) *J. Mol. Biol.* **206**, 759–777.
Summers, N.L., Carson, W.D. & Karplus, M. (1987) *J. Mol. Biol.* **196**, 175–198.
Sutcliffe, M.J., Haneef, I., Carney, D. & Blundell, T.L. (1987a) *Prot. Eng.* **1**, 377–384.
Sutcliffe, M.J., Hayes, F.R.F. & Blundell, T.L. (1987b) *Prot. Eng.* **1**, 385–392.

Taylor, W.R. & Orengo, C.A. (1989) *J. Mol. Biol.* **208**, 1–22.
Thornton, J.M. & Gardner, S. (1989) *TIBS* **14**, 300–304.