Reprinted from Trends in Biochemical Sciences – *June 1990* *Vol. 15, No. 6*

From comparisons of protein sequences and structures to protein modelling and design

A. Šali, J. P. Overington, M. S. Johnson and T. L. Blundell

A useful approach to modelling proteins exploits knowledge of three-dimensional structures determined by X-ray crystallography together with rules defined by their analysis and comparison.

LET US BEGIN with the assumption that neither quantum mechanical calculations nor molecular dynamics simulations alone will, in the near future, allow us to define which of the multitude of conformers available to a polypeptide will occur in a living organism. The reasons for this assumption are many, but the most important is the immense computational time required to simulate a complex macromolecular system during the folding process.

Many protein sequences, however, adopt the same general fold. 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. If we can use this information from related proteins, then we have a feasible route to modelling a protein from its sequence.

The classical approach to this problem was developed for modelling one protein on the three-dimensional structure of an homologue. The first application was the construction of a model for α-lactalbumin on the basis of lysozyme[1]. Other applications included construction of models for relaxins and insulin-like growth factors (see Ref. 2 for a review) and various serine proteinases[3]. Some rules for modelling protein structure[3-5] have been suggested and the advent of computational techniques have made modelling more straightforward.

In this review we shall consider our own approach to protein modelling which can be completely automated, and in which all decisions are rule based (see Fig. 1). We begin by comparing or aligning protein sequences and three-dimensional structures determined by experiment and organized in databases (see, for example, Ref. 6). We use these comparisons to derive rules about families of protein structures that adopt a common fold. Automation of the approach allows these rules to be tested systematically. The rules are then used to define a template for sequences for each family fold; this can be considered as a projection from three-dimensional structure onto one-dimensional sequence or as a generalized protein sequence summarizing knowledge about the family fold. The sequence of the protein to be modelled is then aligned with the appropriate template, and the alignment is used to extrapolate the three-dimensional features of the known structures to the sequence of interest. A model of the protein is then constructed by making it consistent with the features implied by proteins of known structure. Thus we exploit not only the details of protein structures but also the rules developed from their analysis and comparison; these together form the knowledge base.

The most helpful learning set is that of homologous proteins; 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 also describe new approaches that can be used to model more distantly related protein structures. An alternative rule-

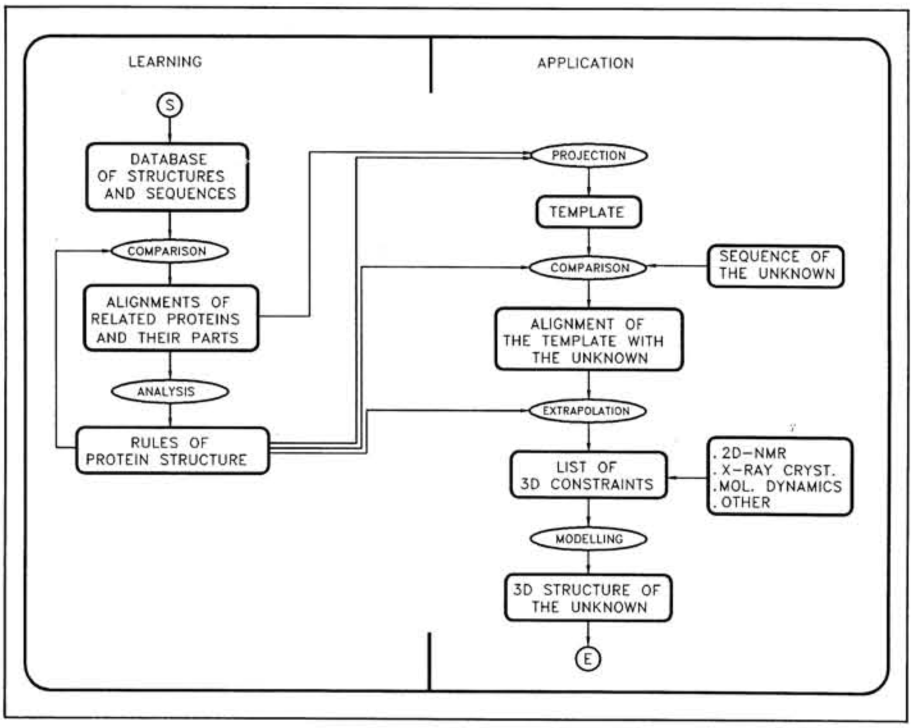

Figure 1
A scheme for knowledge-based modelling of proteins. The approach involves the derivation of rules from the comparisons of sequences and three-dimensional structures, and their use in the generation of a template and the construction of the three-dimensional model. Operations involved in comparison, analysis, projection and extrapolation with modelling are described in detail in the text.

A. Šali, J. P. Overington, M. S. Johnson and **T. L. Blundell** are at the Laboratory of Molecular Biology and ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

```
4APEN   t̃ p a q̃ t̃ L n L d F D̃ T G s̃ S D̃ L W̃ V F S̃ s ẽ T̃ t a s e v
2APPN   - - g t t̃ L n L n F d̃ T G s̃ A D̃ L W V F S t̃ ẽ L p a s q q
2APRN   t̃ p G k k̃ F ñ L d F d̃ T g s̃ S D̃ L W̃ I A S t l C t̃ - ñ C g
PEPN    t̃ p a q d F t̃ V i F d̃ T G S̃ S Ñ L W̃ V P S v y C s s l A C

4APEC   k - s t s I d̃ G I A d̃ T G t̃ t l L y L p a t V V s a - - -
2APPC   g - d g - f s G i A d̃ T G t̃ t l L l L d̃ d̃ s V V s q̃ - - -
2APRC   a - s s - f d̃ G i L d̃ T G t̃ t l L i L p ñ n i A a s - - -
PEPC    a C s g g c q̃ A I V d̃ T g t̃ s l L T G p t̃ s a I a n - - -
```

Figure 2

The multiple alignment of eight aspartic proteinase lobes constructed using COMPARER. Protein codes: 4APE, endothiapepsin; 2APP, penicillopepsin; 2APR, rhizopuspepsin; PEP, hexagonal porcine pepsin. The final N or C of the protein code indicates either the N- or C-terminal lobe. The coordinates of the three-dimensional structures were obtained from the Brookhaven data-bank[29], with the exception of the coordinates of porcine pepsin which were the kind gift of Dr Jon Cooper. The amino acid code is the standard one-letter code formatted using the following convention: italic, positive value of the main chain torsion angle Phi; upper case, solvent-inaccessible residues; lower case, solvent-accessible residues; bold type, hydrogen bonds to mainchain amide; underline, hydrogen bonds to mainchain carbonyl; ~, side chain–side chain hydrogen bonds.

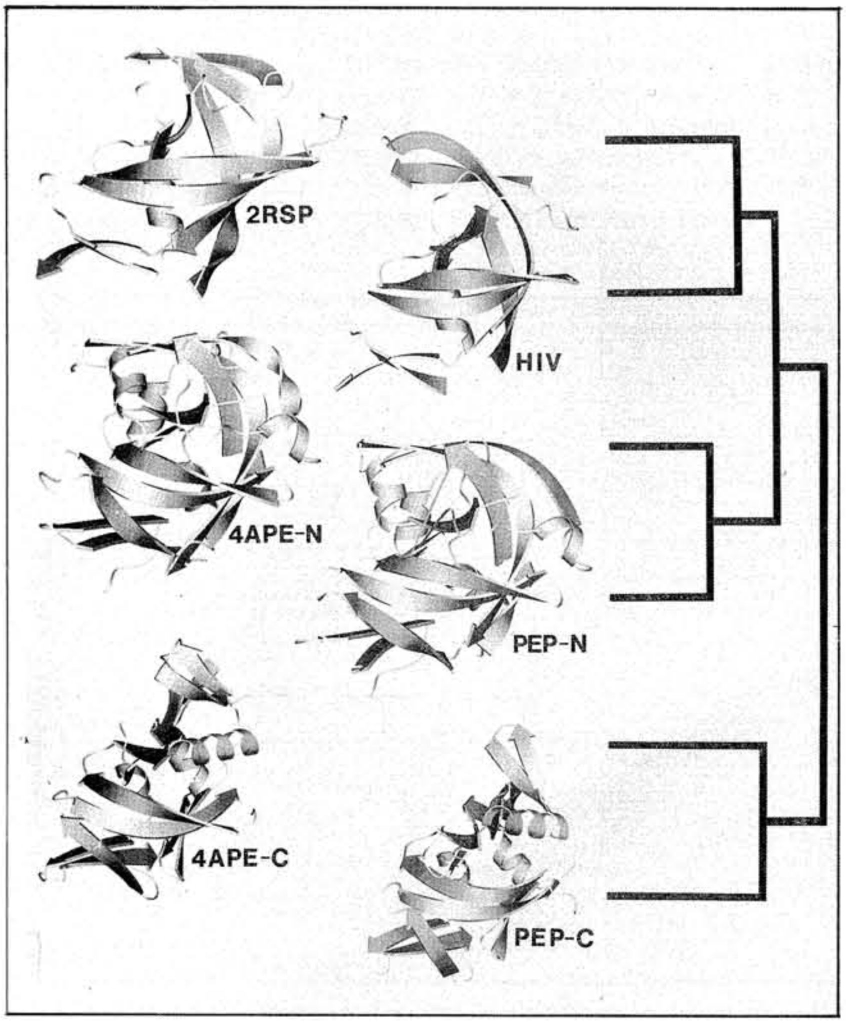

Figure 3

The clustering of the N-terminal and C-terminal lobes of the pepsin-like aspartic proteinases and the subunits of the retroviral aspartic proteinases on the basis of structures using methods of Johnson *et al.*[17,18]. HIV, human immunodeficiency virus proteinase; 2RSP, Rous sarcoma virus proteinase; PEP, pepsin; 4APE, endothiapepsin. HIV structure was kindly provided by Lapatto *et al.*[30]

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[7].

Comparison and clustering

The first task in knowledge-based modelling is to develop ways of comparing protein structures and of clustering them into related subgroups. The most familiar comparison methods are those for sequences, which use dynamic programming procedures based on the algorithm of Needleman and Wunsch[8] for pairwise or multiple sequence alignments[9]. These methods usually consider the mutation rates of amino acid residues to derive optimal comparison scores and corresponding alignments, but other properties such as physicochemical parameters can also be included[10,11].

Comparison of protein tertiary structures has often involved their superposition in three-dimensional space (see Ref. 12 for a review). If several homologous structures are available[13,14] we usually find that many helices and strands are conserved in the family. This is called the 'framework'. However, in more distantly related proteins differences in the positions and orientations of these strands and helices may preclude their superposition[15–18].

The problem of defining topologically equivalent residues in polypeptides that have little sequence similarity but adopt similar folds was addressed more than a decade ago by Rossmann, Matthews and co-workers. They either included information about mainchain direction in the alignment or based their comparisons on superposition of small parts of the whole structure (see Ref. 19 for a review). Others used inter- or intramolecular distances or relationships between secondary structure elements.

In our approach to comparing distantly related proteins[20], we consider features of both sequence and three-dimensional structure simultaneously. Many features depend only on local properties such as hydrophobicity, local conformation and solvent accessibility. These can be aligned using the same approach used for sequence alignment. Furthermore, a similar comparison can be made at each level in the protein hierarchical organization – residue, secondary structure, supersecondary structure, motif or domain. A related approach based principally on intramolecular distances has been described by Taylor and Orengo[21]. Specific relationships such as hydrogen bonding or packing interactions tend to be conserved in protein folds and the patterns of interactions can also be used to align structures[20].

Figure 2 shows part of the alignment of the two domains of pepsins using our programme COMPARER. The alignment identifies all those strands and helices that have previously been considered equivalent on a more subjective basis and which have recently been shown to be shared with the retroviral proteinases.

Once protein structures have been aligned, clustering methods using a matrix of similarity scores between all pairs of proteins can be used to construct a tree that describes relationships between them. Whilst there has been much discussion of methods for construction of evolutionary trees from sequences (see Ref. 9 for review), the clustering of protein three-dimensional structures has been less studied. Eventoff and Rossmann[22] constructed dendrograms based on structural features alone to describe distant phylogenetic relationships among the mono- and dinucleotide binding proteins. We have shown that a useful measure of structural difference can be obtained from differences between superposed structures[17,18]. Figure 3 shows an example of a tree constructed for the domains of the aspartic proteinases on the basis of the three-dimensional structures. We have extended this approach by considering additional structural and sequence features in the comparisons[18,20] so that proteins with little sequence similarity can be compared and classified. Such structural trees can be used in conjunction with the sequence-based trees to provide automatic selection of the best structures and fragments for modelling from sequence.

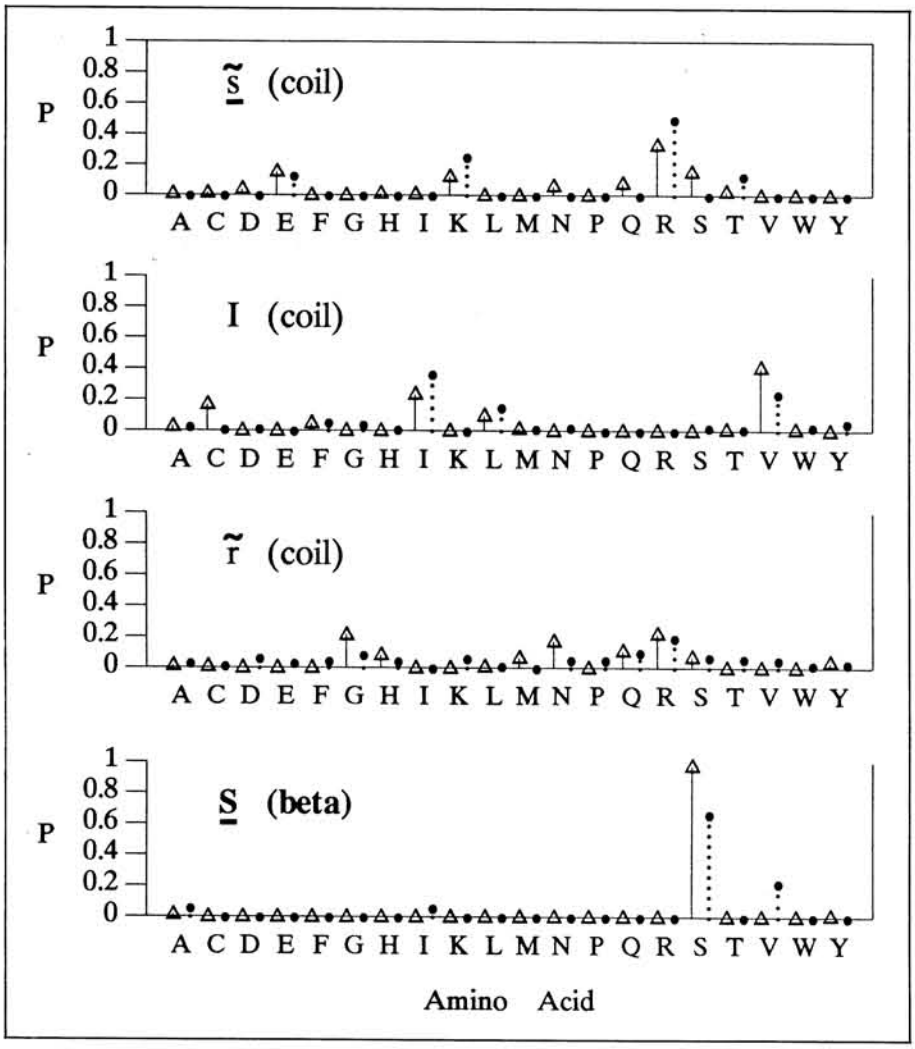

Figure 4

Comparison of the predicted pattern of amino acid substitution with the variability observed among real sequences at strand d of motif 3 of γ-crystallins which has the sequence Ser-Ile-Arg-Ser. Residues 2 and 4 are inaccessible to solvent. Residue 4 is serine which is hydrogen-bonded to the amide function of a main chain peptide and so is predicted to be highly conserved.

Derivation of rules

The alignment of three-dimensional structures allows derivation of rules useful for protein modelling. For example, rules can be obtained that correlate an unknown side chain dihedral angle with the dihedral angles at equivalent positions in related proteins[13,14,23].

Rules for the substitution of amino acids are derived by counting how many times two residue types occur at structurally equivalent positions. We have constructed specific substitution tables in which only a subset of residues that have a certain structural environment are considered. For example, 20 by 20 tables were built separately for solvent-accessible and -inaccessible residues; in this way we can quantify the well-known rule that buried residues tend to be among the more conserved residues in a family. Other structural features included in our analysis were local mainchain conformation and side chain hydrogen-bonding to peptide groups or other side chains.

These tables quantify the influence of structural features on the acceptance of amino acid substitutions in evolution. Large differences exist in the substitution patterns of the same residue type in different structural environments. Hydrogen-bonded and inaccessible polar residues, such as aspartic acid, serine or threonine, are amongst 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. For example, as shown in Fig. 2, Thr33 and Thr216 are conserved or conservatively varied to serine in all pepsin-like and retroviral proteinases. These residues play an important role in holding together, through buried hydrogen bonds, the two subunits in retroviral proteinases and the two lobes in pepsins.

Projection from structure to sequence

The substitution tables described above can also be seen as sets of rules that relate the structural environment of an amino acid to the probability for the acceptance of any other of the 20 residue types. This allows us to use three-dimensional structure to project constraints onto the one-dimensional sequence or, in other words, to construct the template representing the sequence of the family of proteins with a particular tertiary fold.

For each topologically equivalent position in each known structure, we use the tables to predict the substitution of amino acid residues. Figure 4 shows both the predicted substitution of four residue positions in a protein on the basis of its three-dimensional structure, and the observed pattern of amino acid substitutions in the equivalent positions of 155 aligned sequences. The example demonstrates how the environment-dependent substitution tables can provide a remarkably good estimate of sequence variation if the three-dimensional structure of at least one protein is known. Likely places for insertions and deletions can also be predicted. This provides a general approach to the generation of templates. An alternative approach has been suggested by Ponder and Richards[24] who have suggested an algorithm for systematically constructing all sequences of amino acid types and their side chain conformations that are consistent with a particular fold.

These templates constructed on the basis of one or more three-dimensional structures are complementary to those constructed from the alignment of many sequences[9,11]. Both kinds of template can define sequence fingerprints that are essential to structure or function. They can be used in the form of consensus sequences or substitution tables to search out distantly related proteins in the sequence database.

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 in a better extrapolation of spatial features from known structures onto the sequence of the unknown in knowledge-based modelling.

Extrapolation from sequence to structure

We have shown that analyses of aligned structures of homologous proteins can give rise to simple rules, and shall now consider methods for using these rules to produce a three-dimensional model from a sequence.

Let us first consider the use of such rules to model homologous structures. Most approaches depend on the assembly of fragments of three-dimensional structures[25–27]. In the computer program COMPOSER we select three sets of fragments. The first set is derived from the framework defined by multiple superposition of the chosen structures[14] (Fig. 5). A second set of protein fragments for regions outside the framework is selected from the database of loop substructures using a distance filter in a similar way to that of Jones and Thirup[25]. The third set of fragments, the side chains, is selected using rules derived from the analysis of homologous structures[14]. These 1200 rules 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). The templates of selected fragments are clustered and ranked using the methods described above, and the top-ranking fragments are annealed together. The model is checked for serious overlaps between fragments; where this occurs the next ranking fragment is used. The final model is energy minimized to remove minor inconsistencies.

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 (>40%). For example, in building a model of porcine trypsin from four other structurally known serine proteinases, the root mean square distance difference between the model and the known structure is 0.64 Å for the 150 residues defined in the framework. Similarly, 80% of side chain conformations are correctly predicted for closely homologous structures. In all cases, the accuracy of

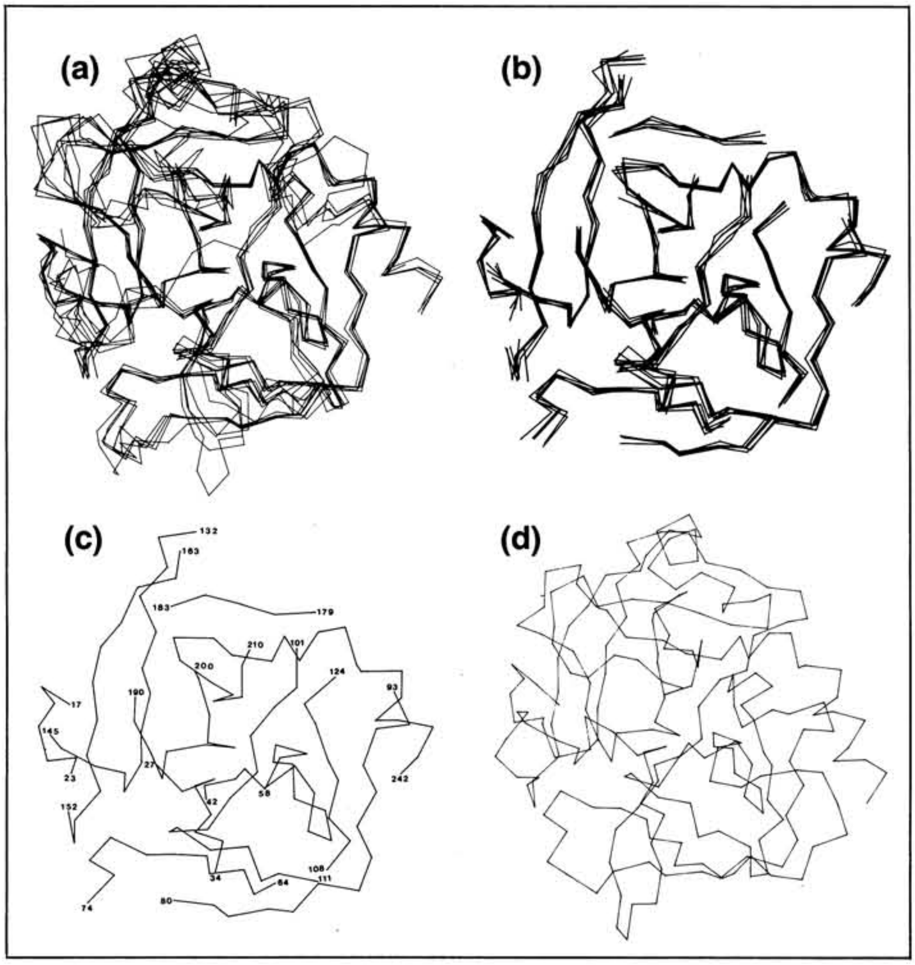

Figure 5

Modelling human plasma kallikrein 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 data base of structures, are used to model the structurally variable regions. The C$^\alpha$ atom positions of the complete model are shown in (d). Side chain positions (not shown) are also generated by a set of rules derived from comparisons of known structures.

the prediction decreases very quickly as the sequence identity between the known and unknown decreases. For these cases, a new modelling technique is required that is not restricted by the idea of assembling rigid fragments of protein structure. In this procedure we construct a model from distances between atoms, similarly to methods for structure analysis using two-dimensional NMR data[28,29].

The sequence to be modelled is first aligned with sequences of known related structures to derive a list of 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 as well. This represents a distance constraint on the atoms involved in the hydrogen bond. In general, we use the associations between protein features to predict main chain and side chain dihedral angles, C^α–C^α distances and hydrogen-bonding distances from the known structures aligned with that being modelled.

The predicted distances are expressed as Gaussian probability functions. For non-bonded atoms this often involves taking a mean distance from an homologous protein and a standard deviation proportional to the similarity between the proteins and the magnitude of the distance. For side chain dihedral angles the probability functions are usually trimodal with the relative magnitudes depending on the particular residue type and the values of equivalent dihedral angles in related known structures. 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 combine the individual probability functions.

The goal is now to construct the three-dimensional model that will satisfy these constraints. Obviously, the most probable structure of the molecule as a whole is the one that maximizes the product of all feature probability functions. The optimization is performed in Cartesian coordinate space using a variable target function[28] approach and a combination of conjugate gradients and simulated annealing minimization to make the best use of the speed of the former and the large radius of convergence of the latter.

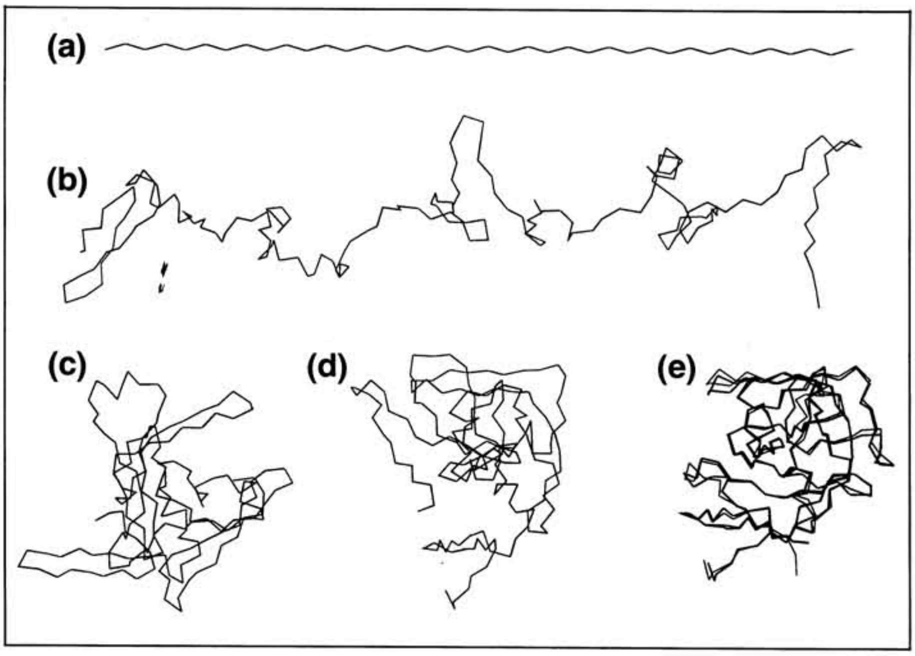

Figure 6
Stepwise generation of a model of a domain of endothiapepsin using information from homologous aspartic proteinases and from protein three-dimensional structures in general expressed as distance constraints. (a) is the extended chain, (b) shows the influence of mainly local constraints, (c) and (d) are intermediate structures, and (e) is the final structure compared with that experimentally defined.

Figure 6 illustrates the generation of a preliminary model for the N-terminal lobe of endothiapepsin; it has a distance root mean square deviation from the crystallographically determined structure of 0.76 Å although only C^α–C^α constraints were used.

Concluding remarks

Knowledge-based 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, 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.

Knowledge-based modelling techniques 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, two-dimensional NMR, image reconstruction in electron microscopy, epitope mapping and crosslinking, all of 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 on intuition.

Acknowledgments

As was supported by an ORS Awards Scheme, the Research Council of Slovenia, the J. Stefan Institute and Merck, Sharp and Dohme. MSJ is funded by the American Cancer Society and JPO by Pfizer and SERC. We are grateful to the Imperial Cancer Research Fund, the SERC and the EEC for general financial support. We also thank our colleagues at Birkbeck for many helpful discussions.

References

1 Browne, W. J., North, A. C. T., Phillips, D. C., Brew,K., Vanaman, T. C. and Hill, R. L. (1969) *J. Mol. Biol.* 42, 65–86
2 Blundell, T. L. and Humbel, R. E. (1980) *Nature* 287, 781–787
3 Greer, J. (1981) *J. Mol. Biol.* 153, 1027–1042
4 Sibanda, B. L., Blundell, T. L. and Thornton, J. M. (1989) *J. Mol. Biol.* 206, 759–777

5 Chothia, C., Lesk, A. M., Tramonto, A., Levitt, M., Smith-Gill, S., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. A., Colman, P. M., Spinelli, S., Alzari, P. M. and Poljak, R. J. (1989) *Nature* 342, 877–883

6 Thornton, J. M. and Gardner, S. (1989) *Trends Biochem. Sci.* 14, 300–304

7 Presnell, S. R. and Cohen, F. E. (1989) *Proc. Natl Acad. Sci. USA* 86, 6592–6596

8 Needleman, S. B. and Wunsch, C. D. (1970) *J. Mol. Biol.* 48, 443–453

9 Doolittle, R. (1989) *Trends Biochem. Sci.* 14, 244–245

10 Argos, P. (1987) *J. Mol. Biol.* 193, 385–396

11 Taylor, W. R. (1986) *J. Mol. Biol.* 188, 233–258

12 KenKnight, C. E. (1984) *Acta Crystallogr.* A40, 708–712

13 Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L. (1987) *Protein Eng.* 1, 377–384

14 Sutcliffe, M. J., Hayes, F. R. F. and Blundell, T. L. (1987) *Protein Eng.* 1, 385–392

15 Chothia, C. and Lesk, A. M. (1986) *EMBO J.* 5, 823–826

16 Hubbard, T. J. P. and Blundell, T. L. (1987) *Protein Eng.* 1, 159–171

17 Johnson, M. S., Sutcliffe, M. J. and Blundell, T. L. (1990) *J. Mol. Evol.* 30, 43–59

18 Johnson, M. S., Šali, A. and Blundell, T. L. (1989) *Methods Enzymol.* 783, 670-690

19 Matthews, B. W. and Rossmann, M. G. (1985) *Methods Enzymol.* 115, 397–420

20 Šali, A. and Blundell, T. L. (1990) *J. Mol. Biol.* 212, 403–428

21 Taylor, W. R. and Orengo, C. A. (1989) *J. Mol. Biol.* 208, 1–22

22 Eventoff, W. and Rossmann, M. G. (1975) *Crit. Rev. Biochem.* 3, 111–140

23 Summers, N. L., Carson, W. D. and Karplus, M. (1987) *J. Mol. Biol.* 196, 175–198

24 Ponder, J. W. and Richards, F. M. (1987) *Proteins* 193, 775–791

25 Jones, T. H. and Thirup, S. (1986) *EMBO J.* 5, 819–822

26 Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L. and Sutcliffe, M. J. (1988) *Eur. J. Biochem.* 172, 513–520

27 Claessens, M., Cutsem, E. V., Lasters, I. and Wodak, S. (1989) *Prot. Eng.* 2, 335–345

28 Braun, W. and Go, N. (1985) *J. Mol. Biol.* 186, 611–626

29 Havel, T. F., Kuntz, I. D. and Crippen, G. M. (1983) *Bull. Math. Biol.* 45, 665–720

30 Lapatto, R. *et al.* (1989) *Nature* 342, 299–302