PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Recognition of Protein Structure and Function from Sequence: Adding Value to Genome Data [and Discussion]

Alex C. W. May, Mark S. Johnson, Stephen D. Rufino, Hiroshi Wako, Zhan-Yang Zhu, Ramanathan Sowdhamini, Narayanaswamy Srinivasan, Michael A. Rodionov, Tom L. Blundell and G. A. Dover

Phil. Trans. R. Soc. Lond. B 1994 **344**, 373-381
doi: 10.1098/rstb.1994.0076

The recognition of protein structure and function from sequence: adding value to genome data

ALEX C. W. MAY, MARK S. JOHNSON, STEPHEN D. RUFINO, HIROSHI WAKO§, ZHAN-YANG ZHU, RAMANATHAN SOWDHAMINI, NARAYANASWAMY SRINIVASAN, MICHAEL A. RODIONOV‡ AND TOM L. BLUNDELL*

Imperial Cancer Research Fund Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, U.K.

[Plate 1]

SUMMARY

The explosion of DNA sequence data from genome projects presents many challenges. For instance, we must extend our current knowledge of protein structure and function so that it can be applied to these new sequences. The derivation of rules for the relationships between sequence and structure allow us to recognize a common fold by the use of tertiary templates. New techniques enable us to begin to meet the challenge of rule-based modelling of distantly related proteins. This paper describes an integrated and knowledge-based approach to the prediction of protein structure and function which can maximize the value of sequence information.

1. INTRODUCTION

The DNA sequence of yeast chromosome III (Oliver *et al.* 1992), the first determined for an entire chromosome from any organism, revealed the existence of 182 structural genes for putative proteins longer than 100 amino acids and emphasized the need to manage and explore data automatically (Maddox 1992). Sander and coworkers (Bork *et al.* 1992) concluded that as many as 42% of the 182 proteins have a known or probable function, but only 13% of the sequences were similar to a known family fold. However, there are protein superfamilies which include members that share a common topology and often function but lack any significant sequence similarity (Murzin & Chothia 1992). For instance, a polypeptide growth factor superfamily based on a cystine-knot plus β-strands topology but with no detectable sequence similarity has been described (for a review, see Murray-Rust *et al.* 1993). The existence of such superfamilies, with insignificant sequence similarities but often functional relationships, indicates that we can learn more about structure and function if new sequences can be associated with known superfamilies.

Such genome analyses underline the huge disparity between the number of known protein sequences (now about 50000; Johnson *et al.* 1994) and the number of experimentally defined three-dimensional structures (currently more than 1000 are available from the Brookhaven Protein Data Bank; Bernstein *et al.* 1977; Abola *et al.* 1987). Despite recent advances in X-ray crystallography, NMR, computing and recombinant DNA technology which have led to the reporting of new structures 'at the rate of almost one a day' (Short 1993), protein sequences still require less time to determine experimentally than structures. How is it possible to predict protein structure and function from a new sequence?

If a new sequence can be matched to one of the known three-dimensional folds, it is possible to model the three-dimensional structure of the protein and often learn more about function (Blundell *et al.* 1987, 1988). We have recently estimated that there are only between 500 and 700 family folds (Blundell & Johnson 1993). This limit on the number of folds together with the rapid (exponential) increase in the number of new structures appearing each year suggests that soon a representative structure should be available for most of the common folds. Concomitant advances in methods for the identification of a relationship between a new sequence and a family fold (inverted protein structure prediction) should mean that knowledge-based modelling will be able to provide a possible structure for most sequences defined by genome sequencing projects (Blundell *et al.* 1987).

In this paper we emphasize our own approaches to the prediction of protein structure (Šali *et al.* 1990; Blundell & Johnson 1993). We address the broader question of the identification of a common fold and

§ On leave from School of Social Sciences, Waseda University, Tokyo, 169-50, Japan.

‡ On leave from Institute of Bioorganical Chemistry, Belarus Academy of Science, Zhodinscaya st. 5/2, Minsk-141, Republic of Belarus 220141.

* To whom correspondence should be sent.

prediction of protein structure based on knowledge of homologous and other protein structures. We discuss new techniques and their potential applications.

2. KNOWLEDGE-BASED MODELLING OF PROTEINS

The approach consists of two stages: learning and application (Šali *et al.* 1990). The organization of protein sequences and structures into databases has facilitated their comparison. For instance, we have produced a database of aligned three-dimensional structures of related proteins (Overington *et al.* 1993). Analysis of these alignments and individual structural features allows the derivation of rules about families of protein structures that adopt a common fold (for a review, see Thornton 1992). This is the learning stage. These rules can then be used in the application stage to predict sequences that will adopt a common fold, usually by constructing a tertiary template for each family fold (Overington *et al.* 1990). This can be thought of as the projection of restraints from three-dimensional structures onto a one-dimensional generalized sequence summarizing information about the common fold (Šali *et al.* 1990). Inverted protein structure prediction (for a review, see Bowie & Eisenberg 1993) attempts to recognize sequences which might adopt a common fold. The sequence of the protein to be modelled is then aligned with the appropriate template, and the alignment is used to extrapolate the same set of rules derived from the known structures to the 'unknown'. The reconstruction of a three-dimensional structure from the one-dimensional sequence using the same set of restraints is more difficult than the mapping from a structure onto a sequence (Šali 1991).

3. COMPARISON AND CLUSTERING OF PROTEIN SEQUENCES AND STRUCTURES

The comparison of protein sequences to reveal sequence similarity is essential for the interpretation of sequence data. Sequence alignments can identify invariant residues conserved for function or structure in a family. The dynamic programming method of sequence alignment (Needleman & Wunsch 1970) determines the best alignment of two sequences allowing insertions and deletions. The use of single amino acid comparisons in such an approach means that the alignment depends on the nature of the scoring matrix and the gap penalty. We have recently evaluated 14 published amino acid scoring matrices (Johnson & Overington 1993). Unfortunately, regardless of the scoring matrix, alignments between proteins of less than 25–35% sequence identity are often unreliable giving rise to a 'twilight zone' of similarity where sequences might appear quite unrelated (Doolittle 1981). Multiple sequence alignment methods can improve the reliability of sequence alignment.

The definition of topological equivalence requires methods for protein structure comparison (for a review, see Johnson 1991; Orengo 1992; Blundell & Johnson 1993). The requirement of rigid-body methods for an initial set of equivalences means that the superposition is not fully automatic. In fact, the specification of initial equivalences is often not a trivial task. Reliable *a priori* knowledge of such positions is necessary in order to maximize the likelihood of a successful superposition. We decided to develop a procedure for rigid structure comparison in which the assignment of initial equivalences is not required. Our procedure (May & Johnson 1994), in the program GA_FIT, uses a genetic algorithm (GA) to search for a near optimal solution of the rigid-body superposition of two whole protein structures. The definition of topological equivalences in the final structural alignment is by dynamic programming. Finally, a least-squares fitting algorithm is used to optimize the fit between the GA-matched equivalences.

GA_FIT can be used to compare the structures of three aspartic proteinases solved at Birkbeck College. As a means of comparison, we have also used MNYFIT (Sutcliffe *et al.* 1987*a*), a rigid-body superposition procedure developed at Birkbeck. The specification of the initial sets of equivalences for MNYFIT is relatively straightforward: the active site triad Asp-Thr-Gly of the N-terminal lobe was supplied (for a review, see Blundell *et al.* 1991). Table 1 shows a comparison of the structural comparisons obtained using GA_FIT and MNYFIT. GA_FIT was run four times for each of the pairwise comparisons since the procedure does not always converge to the same solution (May & Johnson 1994). GA_FIT defines more topological equivalences than MNYFIT on each occasion. Furthermore, GA_FIT often reports a *lower* root-mean-square distance (r.m.s.d.) between these equivalences than MNYFIT despite the increase in the number of equivalences (table 1).

A superfamily of polypeptide growth factors has recently been described whose members show no significant sequence similarity but share a very distinctive fold (for a review, see Murray-Rust *et al.* 1993). When we reported the structure of nerve growth factor (NGF) in 1991, however, there was no hint as to the existence of a superfamily (McDonald *et al.* 1991). The subsequent crystal structures of transforming growth factor-β2 (TGF-β2) (Schlunegger & Grutter 1992, 1993; Daopin *et al.* 1992) and platelet-derived growth factor-BB (PDGF-BB) (Oefner *et al.* 1992) revealed that all three proteins are based on a cystine-knot plus β-strands topology (see also McDonald & Hendrickson 1993). The use of the six conserved half-cystines of the disulphide knot as the initial equivalences for a MNYFIT superposition of NGF and TGF-β2 allows the definition of only 20 topological equivalences with an r.m.s.d. of 1.29 Å. GA_FIT, however, finds 45 equivalences with an r.m.s.d. of 1.71 Å (figure 1). The conserved β-hairpins are matched much more successfully by GA_FIT than MNYFIT (see figure 1 in Murray-Rust *et al.* (1993)). Having said this, however, GA_FIT misaligns by one position the half-cystine II positions in its structural alignment. This highlights the fact that the best rigid group fitting is not necessarily the best definition of topological equivalence. Our program COMPARER (Šali & Blundell 1990; Zhu *et al.* 1992)

Table 1. *Comparison of aspartic proteinase superpositions using* GA_FIT *and* MNYFIT

protein pair[a]	GA_FIT no. equivalences[b]	GA_FIT r.m.s.d./Å	MNYFIT no. equivalences[b]	MNYFIT r.m.s.d./Å
4CMS:1MPP	265	1.33	262	1.39
4CMS:1MPP	272	1.39	262	1.39
4CMS:1MPP	267	1.35	262	1.39
4CMS:1MPP	275	1.44	262	1.39
4CMS:4APE	270	1.47	256	1.41
4CMS:4APE	258	1.30	256	1.41
4CMS:4APE	260	1.33	256	1.41
4CMS:4APE	264	1.36	256	1.41

[a] Protein codes are those used in the Brookhaven Protein Data Bank (Bernstein *et al.* 1977; Abola *et al.* 1987).
[b] The definition of topological equivalence is operationally defined: a 3.0 Å cut-off distance for equivalent Cα atoms is used here.

shows that the general arrangement of the hydrogen bonding of the β-hairpins is conserved across the superfamily making it possible to align TGF-β2, PDGF-BB and NGF despite the lack of any significant sequence similarity (Murray-Rust *et al.* 1993). This demonstrates the value of using hydrogen bonding interactions in a comparison method.

The application of rule-based comparative modelling to distantly related proteins requires the development of more versatile methods for structure comparison. One limitation to the predictive capability of comparative modelling of homologous proteins is the observation that secondary structure elements in the core undergo rigid-body shifts (for a

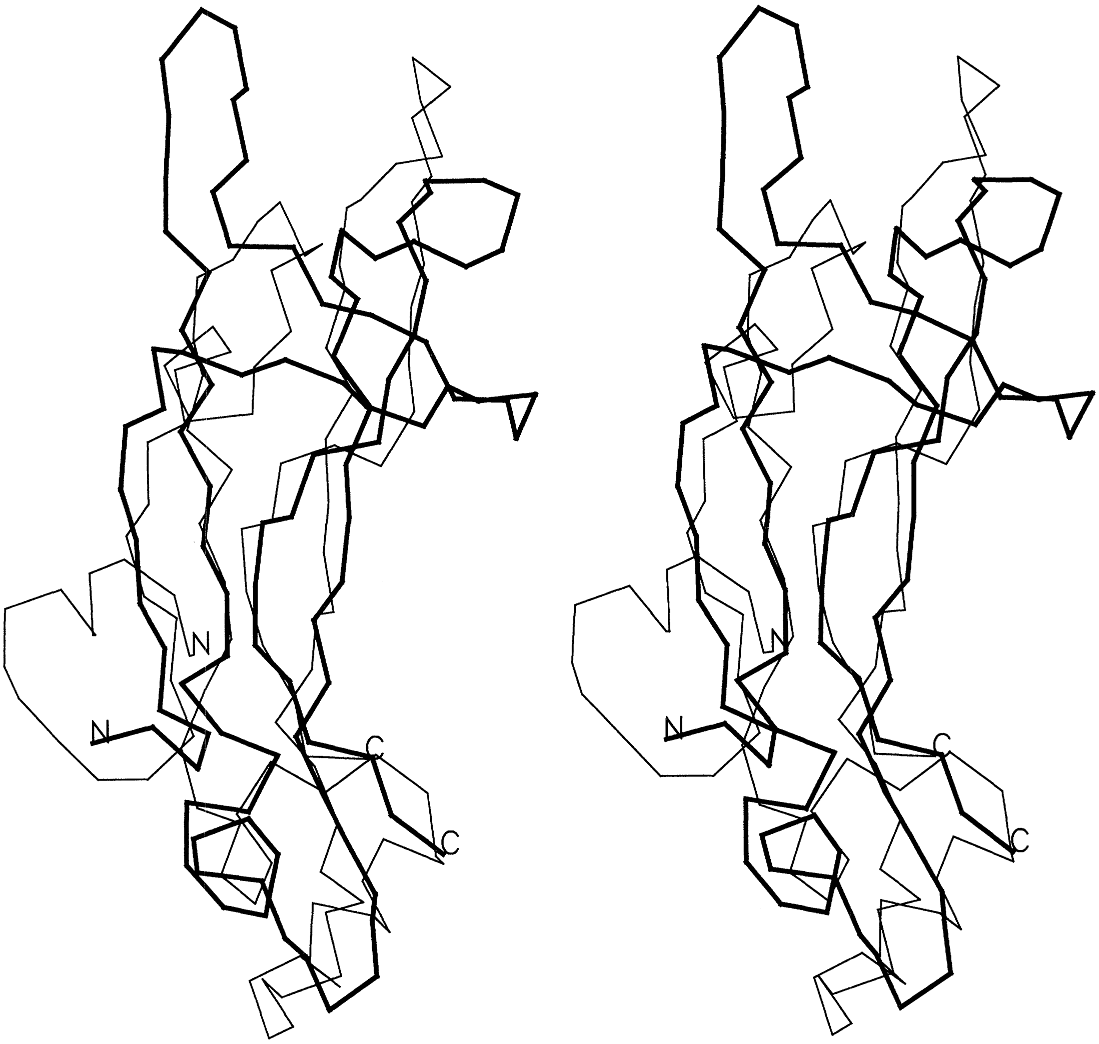


Figure 1. Stereoplot of the Cα-atom trace of protomers of nerve growth factor (bold) and transforming growth factor-β2 after superposition with GA_FIT (May & Johnson 1994).

review, see Reddy & Blundell 1993). This means, for instance, that as more globin structures are superposed the size of the common scaffold or framework decreases. The challenge then is to improve framework extraction. We have now developed a structural feature alignment using dynamic programming for secondary structure elements (Rufino & Blundell 1994). The new procedure, encoded in the program SEA, characterizes each element of secondary structure in terms of a set of features and an associated vector. The secondary structure elements, with which a secondary structure makes contacts, determine the latter's local environment. Secondary structures are compared with respect to their features, their local environments and the features of secondary structures defined in local environments. Similarity of features is established by comparison with variations of features of secondary structures observed in a database of aligned three-dimensional structures of related proteins (Overington *et al.* 1993). Graph theory is used to determine similarity of local environment. A protein structure is represented as a labelled graph, the nodes of which correspond to the secondary structure elements, and the edges to their relationships with respect to length and angle. The use of graph theory in SEA makes the comparison fast enough to compare a target structure against all others in a database. Once secondary structures have been aligned by comparing their features and local environments it is possible to cluster protein folds.

Clustering of representative all-β protein structures by SEA (Rufino & Blundell 1994) (figure 2) shows that families and superfamilies of proteins with a common fold can be automatically defined. For instance, the monomeric pepsin-like aspartic proteinases are separately classified from the dimeric retroviral proteinases. Yet these two families are seen to be sufficiently similar to constitute a superfamily. The same situation arises with the serine proteinases. The percentage identities for pairwise sequence comparisons between the mammalian and bacterial serine proteinases range between 16 and 23% (Johnson *et al.* 1993). Sequence alignments between representative proteins from each group include a number of large insertions/deletions relative to each other making them difficult to align. The structural similarity between the γ- and β-crystallins is recognized.

The determination at Birkbeck of the crystal structure of serum amyloid pentagonal component (SAP), a human pentraxin, revealed that each subunit of the pentamer consists of a jelly roll motif (Emsley *et al.* 1994). Each subunit has two calcium-binding sites involved in binding sugar ligands: SAP exhibits calcium-dependent multi-specific binding. The tertiary fold of SAP is remarkably similar to that of the legume lectins, concanavalin A (Hardman & Ainsworth 1972) and pea lectin (Einspahr *et al.* 1986) and the bacterial enzyme 1,3-1,4 β-gluconase (Keitel *et al.* 1993) (figure 3). This allows the definition of a superfamily of carbohydrate-binding proteins consisting of the mammalian pentraxins and the legume lectins. These multimeric, antiparallel β-structures share a common fold derived from a jelly roll motif and demonstrate sugar binding mediated by metal ions (Emsley *et al.* 1994). The classification of the all-β proteins on the basis of secondary structure elements using SEA (figure 2) shows that it is possible to cluster automatically SAP and the legume lectins into a lectin family sharing a common fold (Rufino & Blundell 1994).

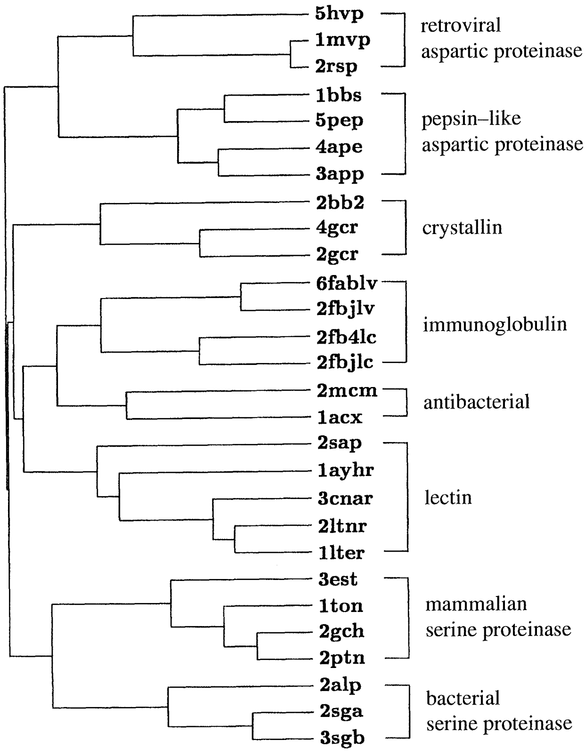


Figure 2. Dendrogram of β proteins on the basis of clustering by SEA (Rufino & Blundell 1994). The families of the clustered proteins are indicated on the right.

4. DERIVATION OF RULES FOR THE RELATIONSHIP BETWEEN SEQUENCE AND STRUCTURE

We have produced a database of homologous protein families (Overington *et al.* 1993) aligned on the basis of three-dimensional structural features using COMPARER (Šali & Blundell 1990; Zhu *et al.* 1992). The comparison of protein families enables us to investigate the constraints that structures place on the sequences that can adopt a common fold. Thus, we can identify what is invariant or conservatively variant in a common fold.

The local environment of residues in each structure is characterized using the program JOY (Overington *et al.* 1990, 1992). The features considered in our analysis are residue type, main-chain conformation and secondary structure, solvent accessibility and side-chain interactions. Environmental-specific substitution tables are constructed on the basis of the classification of observed amino acid substitutions in three-dimensional structures according to the features of the local environment. Classification according to the local environment features produces 64 amino acid substitution matrices. Analysis of the environ-

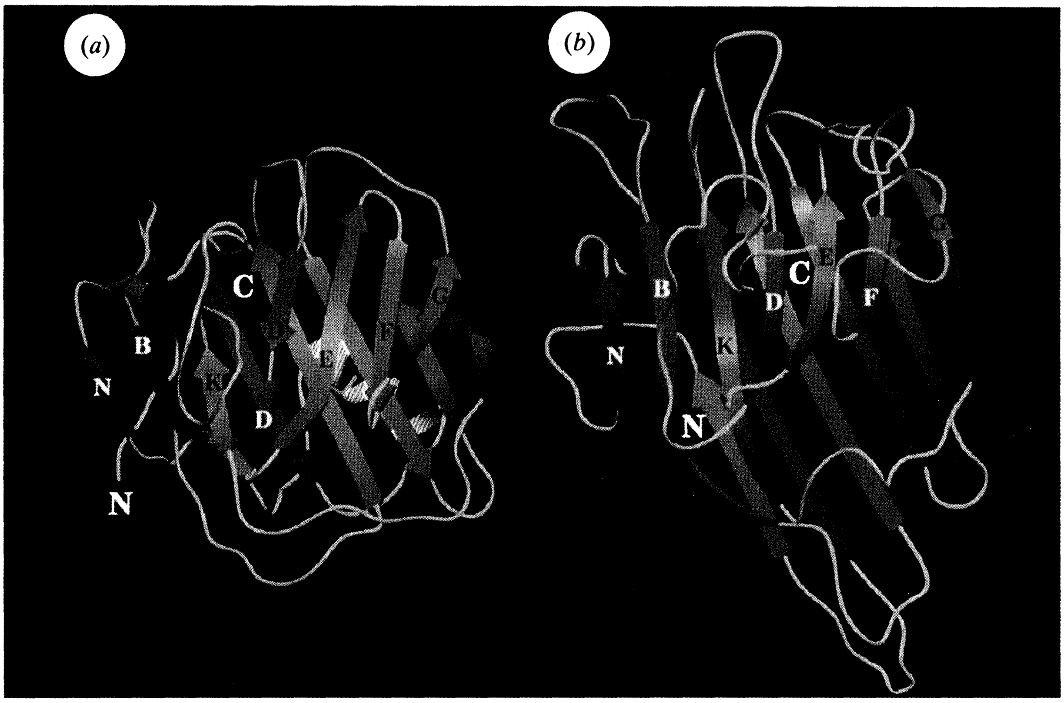


Figure 3. Schematic diagrams of (*a*) serum amyloid P–component (SAP) and (*b*) concanavalin A showing their similar fold. β-strands are drawn as broad arrows, α-helices as ribbons, and coil regions as thin rope. The β-strands (B, D, E, F, G, K and N) of the β-sheets in the foreground are labelled according to their order in SAP. Rearrangement of concanavalin A is needed to achieve identical β-strand connections to those of SAP as the N and C termini are not situated in topologically equivalent positions (Rufino & Blundell 1994).

mental substitution data shows that the substitution patterns for the same residue alter in different environments (Overington *et al.* 1992). For example, hydrogen-bonded polar residues such as aspartic acid and histidine, buried within the protein core, are more conserved than surface polar residues. Yet hydrogen-bonded and inaccessible asparagine residues are rarely invariant. As expected, buried residues are shown to be more conserved than solvent-exposed residues (see also Hubbard & Blundell 1987). One of the most important applications of our substitution pattern data is the construction of tertiary templates which have been used in the recognition of common folds from sequences (Johnson *et al.* 1993). Recently, Topham *et al.* (1993) have used the environment-dependent substitution tables in the selection of loops for rule-based comparative modelling. Conformationally constrained substitution tables have been derived by scoring only those substitutions in which the main-chain conformation is conserved (Topham *et al.* 1993).

Environment-dependent propensities have also been calculated from the alignment database (Wako & Blundell 1994*a*,*b*). We use conformational propensities together with the environment-dependent substitution tables to predict solvent accessibility classes and secondary structures. Buried and exposed residues are predicted by evaluating amino acid substitution patterns and mean propensities for the two solvent accessibility classes with the amino acids at equivalent sites in aligned sequences of homologous proteins. The accuracy of prediction is around 77% when applied to 13 protein families (Wako & Blundell 1994*a*). The method is suitable for the detection of periodicity in the sequence of buried and exposed classes because no averaged properties over neighbouring residues are required. This provides a method for secondary structure prediction which attains a mean percentage accuracy of prediction of 69.0% over the thirteen families. Our method has the advantage that the accuracy does not vary considerably for individual families (Wako and Blundell 1994*b*).

Environment-dependent substitution tables and propensities are incorporated into a procedure for model assessment (Overington *et al.* 1990; Topham *et al.* 1994; R. Sowdhamini, N. Srinivasan, C. M. Topham, J. P. Overington & T. L. Blundell, unpublished results). Eisenberg and colleagues have described a similar approach to the assessment of structures using an extension of their three-dimensional profile method (for a review, see Bowie & Eisenberg 1993).

Analysis of the alignment database has led to a new approach to the derivation of rules for use in comparative protein modelling (Šali 1991; Šali & Blundell 1993). Spatial restraints are derived from the structural alignment and expressed as probability density functions (p.d.f.s) for the features restrained. Associations between protein features are tabulated in the derivation of spatial restraints. In general, every structural feature can be restrained by several knowledge sources. This approach forms the basis of a new protein modelling technique not restricted by a rigid-body model of protein structure (see later).

5. TERTIARY TEMPLATES AND PROFILES

There have been several recent advances in inverted protein structure prediction (for a review, see Bowie & Eisenberg 1993). In our approach, tertiary templates summarize knowledge learnt about a common fold in a form suitable for alignment with the sequence of the 'unknown' (Šali *et al.* 1990). The restraints imposed on the sequence of the unknown by its alignment with the tertiary template and by general rules of protein structure are used to 'map' the sequence onto its tertiary structure. Searching for a common fold with tertiary templates can make use of either propensities or substitution tables. We have constructed tertiary templates from our environment-dependent amino acid substitution tables (Overington *et al.* 1990, 1992; Šali *et al.* 1990; Johnson *et al.* 1993).

For each topologically equivalent position in each known structure in a family, we use the substitution tables to predict the variability of amino acid residues (Overington *et al.* 1992). Structural template descriptions of the family alignments or individual structures so constructed represent the projection of constraints derived from three-dimensional structure onto the one-dimensional sequence (Šali *et al.* 1990). We have used structural templates to recognize common folds from sequences by aligning templates with all known amino acid sequences. A databank of structural templates for each protein family has been produced. This allows a sequence or an alignment of sequences expressed as a sequence template to be searched/aligned against all known folds as represented in the structural template databank (Johnson *et al.* 1993). For example, a search of the structural template data bank with the sequence of the DNA-binding human sex determination factor, *Sry* (Harley *et al.* 1992) provides a prediction of its structure (Johnson *et al.* 1993) (figure 4). The highest scoring structural templates are two DNA-binding proteins of phage 434: the cro protein (Cro) and the repressor. Furthermore, two homeodomains achieve the sixth and seventh highest scores. These four proteins share the helix-turn-helix motif which mediates DNA-binding (Johnson *et al.* 1993).

M. A. Rodionov & M. S. Johnson (unpublished results) have constructed residue–residue contact matrices for the three-dimensional structures of each family in the alignment database. The exchange frequencies between pairs of interacting residues are compiled in contact substitution tables which have been used to improve our methods for searching for a common fold.

To increase our ability to recognize a fold from a sequence we are currently compiling templates of globular domains as well as those for whole proteins (R. Sowdhamini, M. S. Johnson and T. L. Blundell, unpublished results). This will hopefully improve the alignment of whole proteins or their parts and thereby extend the usefulness of knowledge-based or comparative modelling by allowing comparisons of distantly related proteins.

Although the preliminary results of fold recognition are encouraging (Bowie & Eisenberg 1993), there is still

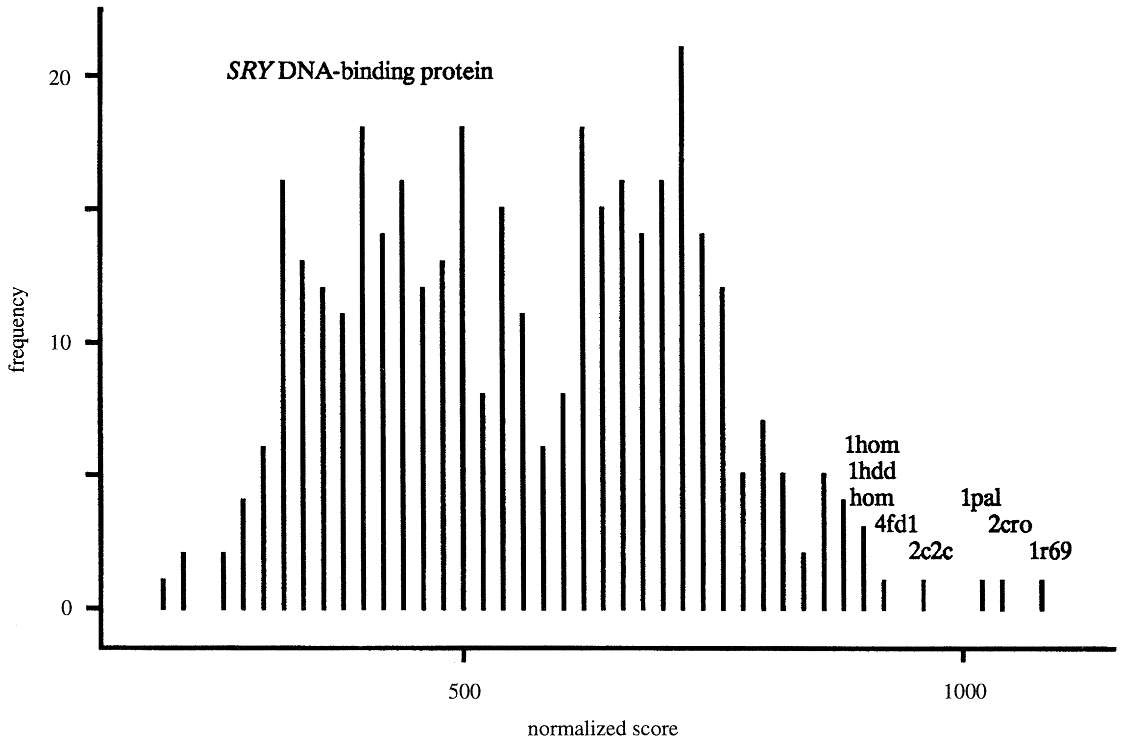


Figure 4. A search of the sequence for DNA-binding sex determination factor, *SRY* (Harley *et al.* 1992), against a structural template data bank (Johnson *et al.* 1993). The top ranked scores include the *cro* bacterial repressors (Brookhaven code, Bernstein *et al.* 1977; Abola *et al.* 1987: 1r69, 2cro), a cytochrome (2c2c), a ferredoxin (4fd1), the homeodomains 1hom and 1hdd as well as the template constructed from their alignment (hom). (From Johnson *et al.* (*J. molec. Biol.* **231**, 735–752 (1993)); used with permission.)

much to be done in impoving these approaches. For example, we believe that the ability of most methods to identify actin on the basis of the 44 kDa fragment of the heat shock cognate protein despite only 10% sequence identity is almost certainly due to their similarity in length (Blundell & Johnson 1993). One area which definitely needs to be improved is the gap penalty for the dynamic programming algorithms. To this end, we have developed structure-dependent gap penalties on the basis of packing density (Johnson *et al.* 1993). The importance of sub-optimal alignments and their evaluation in fold recognition needs to established. As mentioned above, we need to have templates for all levels of protein structure as well as those for individual proteins to improve our ability to catch a common fold.

6. RULE-BASED COMPARATIVE PROTEIN MODELLING

(*a*) *COMPOSER*

Having associated a new sequence with a known fold, we can model the three-dimensional structure of the protein. Most current automated and rule-based approaches depend on the assembly of rigid fragments (for a review, see Bajorath *et al.* 1993). In our approach encoded in the program COMPOSER (Blundell *et al.* 1988) we select three sets of fragments; these define the framework (Sutcliffe *et al.* 1987*a*), the structurally variable, mainly loop regions (Jones & Thirup 1986; Topham *et al.* 1993), and the side chains (Sutcliffe *et al.* 1987*b*). New approaches to automated loop and side-chain model building have been reviewed by Fetrow and Bryant (1993). Finally, the model is refined often by energy minimization to remove small inconsistencies such as steric strain. We have recently improved our procedure (Srinivasan and Blundell 1993). Our modelling procedure is very successful when the known structures cluster around that to be predicted and when the percent sequence identity to the unknown is high (greater than 40%) (Srinivasan & Blundell 1993). 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 which can overcome the problem of rigid-body shifts and is not restricted to the assembly of rigid fragments is essential.

(*b*) *MODELLER*

Šali and Blundell (Šali *et al.* 1990; Šali & Blundell 1993) describe a comparative protein modelling method which generates a model structure by optimal satisfaction of spatial restraints derived from the alignment of the unknown with related structures. These restraints are expressed as probability density functions (p.d.f.s) for the features to be restrained (Šali 1991). For example, if there is a conserved hydrogen bond at an equivalent position in all known three-dimensional structures in a family then we can extrapolate this feature to the unknown structure.

This p.d.f. represents a distance restraint on the atoms involved in the hydrogen bond. In general, every structural feature can be restrained by several knowledge sources. For instance, a distance between two Cα atoms may be restrained by information from several related structures and also by van der Waals criteria. In such cases the p.d.f. for the given feature is obtained as a combination of individual p.d.f.s. Several such p.d.f.s have been obtained from the correlations between structural features in 17 families of homologous proteins which have been aligned on the basis of their three-dimensional structures (Šali & Blundell 1990). A smoothing procedure (adapted from Sippl 1990) is used in the derivation of these relationships to minimize the problem of a sparse database (Šali 1991). The protein three-dimensional structure is uniquely determined if a sufficiently large number of its spatial features are specified. The most probable structure is obtained by optimization of all feature p.d.f.s (molecular p.d.f.) such that the violation of the input restraints by the model is minimized. Optimization of the molecular p.d.f. is performed using a variable target function method that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms (Šali & Blundell 1993). Once the alignment of the unknown with related structures is determined, the method is completely automated. The method has been used to generate a model for the N-terminal lobe of endothiapepsin on the basis of homologous aspartic proteinases (Šali *et al.* 1990), models of four mouse mast cell chymases (Šali *et al.* 1993) and a model of trypsin using two other serine proteinases, elastase and tonin (Šali & Blundell 1993). Preliminary results suggest that MODELLER seems to be at least as accurate as the other manual or automated knowledge-based methods (Šali & Blundell 1993). In particular, it appears that the new approach achieves a slightly higher accuracy of prediction than COMPOSER because it is not restricted by a rigid-body model of protein structure (Šali & Blundell 1993).

7. CONCLUSIONS

Modelling using knowledge of families of proteins with a common fold does not require us to consider the question of convergent or divergent evolution as we are only interested in common features from which we can learn (Blundell & Johnson 1993). For example, using information derived from the known family of pepsin-like aspartic proteinases, we correctly predicted the similarity of the distantly related retroviral and pepsin-like aspartic proteinases folds (Pearl & Taylor 1987; Blundell *et al.* 1988).

Recent structure determinations suggest that most new structures comprise motifs or domains common to other proteins (Overington 1992; Blundell & Johnson 1993). For this reason it is important to be able to model distantly related proteins. Tertiary templates for each protein family can be used to identify a fold for a new protein sequence, which can be used to suggest function. For example, porphobilinogen deaminase (PBGD) has in two of its domains the same fold as the transferrins and the periplasmic binding proteins (for a review, see Louie 1993). These proteins share an active-site cleft located at the interface between two domains. The structural similarity of PBGD to the binding proteins also suggests a functional similarity: the ligands all involve anions (except for maltose-binding protein); even transferrin involves obligatory binding of carbonate before iron. The common fold also correctly predicts a similar hinge-bending mechanism between the domains for ligand recognition and binding.

It is imperative that the methods developed by us and others for each step in the scheme for knowledge-based protein modelling are made widely available so that the value of sequence information derived from genome sequencing projects might be maximized. Prediction of protein structure should now be accessible to non-specialists due to the development of automated, rule-based comparative modelling. If we can identify the relationship between a new sequence and the fold of a protein superfamily we can predict the structure and often learn more about function by analogy.

The figures in this paper were generated with the program SETOR (Evans 1993). We thank the Imperial Cancer Research Fund, the U.K. Science and Engineering Research Council, the U.K. Agricultural and Food Research Council, Waseda University (Japan) and the U.K. Royal Society for financial support. We are grateful to members (past and present) of the modelling group and the Department of Crystallography for their contributions to our current work.

REFERENCES

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. 1987 Protein Data Bank. In *Crystallographic databases information content, software systems, scientific applications* (ed. F. H. Allen, G. Bergerhoff & R. Sievers), pp. 107–132. Bonn, Cambridge & Chester: Data Commission of the International Union of Crystallography.

Bajorath, J., Stenkamp, R. & Aruffo, A. 1993 Knowledge-based model building of proteins: concepts and examples. *Protein Sci.* **2**, 1798–1810.

Bernstein, F.C., Koetzle, T.F., Williams, J.B. *et al.* 1977 The Protein Data Bank: a computer based archival file for macromolecular structures. *J. molec. Biol.* **112**, 535–542.

Blundell, T.L., Carney, D., Gardner, S. *et al.* 1988 Knowledge-based protein modelling and design. *Eur. J. Biochem.* **172**, 513–520.

Blundell, T.L., Cooper, J.B., Šali, A & Zhu, Z.-Y. 1991 Comparisons of the sequences, 3-D structures and mechanisms of pepsin-like and retroviral aspartic proteinases. In *Structure and function of the aspartic proteinases* (ed. B. M. Dunn), pp. 443–453. New York: Plenum Press.

Blundell, T.L. & Johnson, M.S. 1993 Catching a common fold. *Protein Sci.* **2**, 877–883.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. & Thornton, J.M. 1987 Knowledge-based prediction of protein structure and the design of novel molecules. *Nature, Lond.* **326**, 347–352.

Bork, P., Ouzounis, C., Sander, C., Schard, M., Schneider, R. & Sonnhammer, E. 1992 What's in a genome? *Nature, Lond.* **358**, 287.

Bowie, J.U. & Eisenberg, D. 1993 Inverted protein structure prediction. *Curr. Opinion Struct. Biol.* **3**, 437–444.

Daopin, S., Pietz, K.A., Ogawa, Y. & Davies, D.R. 1992 Crystal structure of transforming growth factor-2: an unusual fold for the superfamily. *Science, Wash.* **257**, 369–373.

Doolittle, R.F. 1981 Similar amino acid sequences: chance or common ancestry? *Science, Wash.* **214**, 149–159.

Einspahr, H., Parks, E.H., Suguna, K., Subramanian, E. & Suddath, F.L. 1986 The crystal structure of pea lectin at 3.0-Å resolution. *J. biol. Chem.* **261**, 16518–16527.

Emsley, J., White, H.E., O'Hara, B.P. *et al.* 1994 The 3D structure of pentameric human serum amyloid P component defined at 2 Å resolution reveals a lectin-like fold and calcium-mediated ligand binding. *Nature, Lond.* **367**, 338–345.

Evans, S.V. 1993 SETOR: Hardware lighted three-dimensional solid model representations of macromolecules. *J. molec. Graphics* **11**, 134–138.

Fetrow, J.S. & Bryant, S.H. 1993 New programs for protein tertiary structure prediction. *Bio Technology* **11**, 479–484.

Hardman, K.D. & Ainsworth, C.F. 1972 Structure of concanavalin A at 2.4 Å resolution. *Biochemistry* **11**, 4910–4919.

Harley, V.R., Jackson, D.I., Hextall, P.J. *et al.* 1992 DNA-binding activity of recombinant Sry from normal males and XY females. *Science, Wash.* **255**, 453–456.

Hubbard, T.J.P. & Blundell, T.L. 1987 Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171.

Johnson, M.S. 1991 Comparisons of protein structures. *Curr. Opinion Struct. Biol.* **1**, 334–344.

Johnson, M.S. & Overington, J.P. 1993 A structural basis for sequence comparisons: an evaluation of scoring methodologies. *J. molec. Biol.* **233**, 716–738.

Johnson, M.S., Overington, J.P. & Blundell, T.L. 1993 Alignment and searching for common protein folds using a data bank of structural templates. *J. molec. Biol.* **231**, 735–752.

Johnson, M.S., Srinivasan, N., Sowdhamini, R. & Blundell, T.L. 1994 Knowledge-based protein modelling. *CRC C.r. Biochem. molec. Biol.* **29**, 1–68.

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

Keitel, T., Simon, O., Borriss, R. & Heinemann, U. 1993 Molecular and active-site structure of a bacillus 1,3-1,4-beta-gluconase. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5287–5291.

Louie, G.V. 1993 Porphobilinogen deaminase and its structural similarity to the bidomain binding proteins. *Curr. Opinion Struct. Biol.* **3**, 401–408.

Maddox, J. 1992 Ever-longer sequences in prospect. *Nature, Lond.* **357**, 13.

May, A.C.W. & Johnson, M.S. 1994 Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng.* **7**, 475–485.

McDonald, N.Q. & Hendrickson, W.A. 1993 A structural superfamily of growth factors containing a cystine knot motif. *Cell* **73**, 421–424.

McDonald, N.Q., Lapatto, R., Murray-Rust, J., Gunning, J., Wlodawer, A. & Blundell, T.L. 1991 New protein fold revealed by a 2.3-Å resolution crystal structure of nerve growth factor. *Nature, Lond.* **354**, 411–414.

Murray-Rust, J., McDonald, N.Q., Blundell, T.L. *et al.* 1993 Topological similarities in TGF-2, PDGF-BB and NGF define a superfamily of polypeptide growth factors. *Structure* **1**, 153–159.

Murzin, A.G. & Chothia, C. 1992 Protein architecture: new superfamilies. *Curr. Opinion Struct. Biol.* **2**, 895–903.

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

Oefner, C., D'Arcy, A., Winkler, F.K., Eggiman, B. & Hosang, M. 1992 Crystal structure of human platelet-derived growth factor-BB. *EMBO J.* **11**, 3921–3926.

Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L. *et al.* 1992 The complete DNA sequence of yeast chromosome III. *Nature, Lond.* **357**, 38–46.

Orengo, C.A. 1992 A review of methods for protein structure comparison. In *Patterns in protein sequence and structure* (ed. W. R. Taylor) (Springer series in Biophysics, vol. 7), pp. 159–188. Heidelberg: Springer-Verlag.

Overington, J.P. 1992 Comparison of three-dimensional structures of homologous proteins. *Curr. Opinion Struct. Biol.* **2**, 394–401.

Overington, J.P., Donnelly, D., Johnson, M.S., ali, A. & Blundell, T.L. 1992 Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.

Overington, J.P., Johnson, M.S., Šali, A. & Blundell, T.L. 1990 Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond.* B **241**, 132–145.

Overington, J.P., Zhu, Z.-Y., Šali, A. *et al.* 1993 Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans.* **21**, 597–604.

Pearl, L.H. & Taylor, W.R. 1987 A structural model for the retroviral proteases. *Nature, Lond.* **329**, 351–354.

Reddy, B.V.B. & Blundell, T.L. 1993 Packing of secondary structural elements in proteins: analysis and prediction of inter-helix distances. *J. molec. Biol.* **233**, 464–479.

Rufino, S.D. & Blundell, T.L. 1994 Identification of protein families and super-families in structure-based design. *J. comput. aid. molec. Des.* **8**, 5–27.

Šali, A. 1991 Modelling three-dimensional structure of proteins from their sequence of amino acid residues. Ph.D. thesis, University of London.

Šali, A. & Blundell, T.L. 1990 Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. molec. Biol.* **212**, 403–428.

Šali, A. & Blundell, T.L. 1993 Comparative protein modelling by satisfaction of spatial restraints. *J. molec. Biol.* **234**, 779–815.

Šali, A., Overington, J.P., Johnson, M.S. & Blundell, T.L. 1990 From comparison of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.* **15**, 235–240.

Schlunegger, M.P. & Grutter, M.G. 1992 An unusual feature revealed by the crystal structure at 2.2 Å resolution of human transforming growth factor-2. *Nature, Lond.* **358**, 430–434.

Schlunegger, M.P. & Grutter, M.G. 1993 Refined crystal structure of human transforming growth factor-2 at 1.95 Å resolution. *J. molec. Biol.* **231**, 445–458.

Short, N. 1993 The changing shape of structure. *Nature, Lond.* **366**, 203.

Sippl, M. 1990 Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. molec. Biol.* **213**, 859–883.

Srinivasan, N. & Blundell, T.L. 1993 An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng.* **6**, 501–512.

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

Sutcliffe, M.J., Hayes, F.R..F. & Blundell, T.L. 1987*b* Knowledge-based modelling of homologous proteins, part II: rules for the conformations of substituted sidechains. *Protein Eng.* **1**, 385–392.

Thornton, J.M. 1992 Lessons from analyzing protein structures. *Curr. Opinion Struct. Biol.* **2**, 888–894.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. & Blundell, T.L. 1993 Identification of key residues in structurally variable regions of proteins using conformationally-constrained environmental substitution tables: applications to loop fragment ranking in modelling of protein structure. *J. molec. Biol.* **229**, 194–220.

Topham, C.M., Srinivasan, N., Thorpe, C.J., Overington, J.P. & Kalsheker, N.A. 1994 (In preparation.)

Wako, H. & Blundell, T.L. 1994*a* Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. molec. Biol.* (In the press.)

Wako, H. & Blundell, T.L. 1994*b* Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. molec. Biol.* (In the press.)

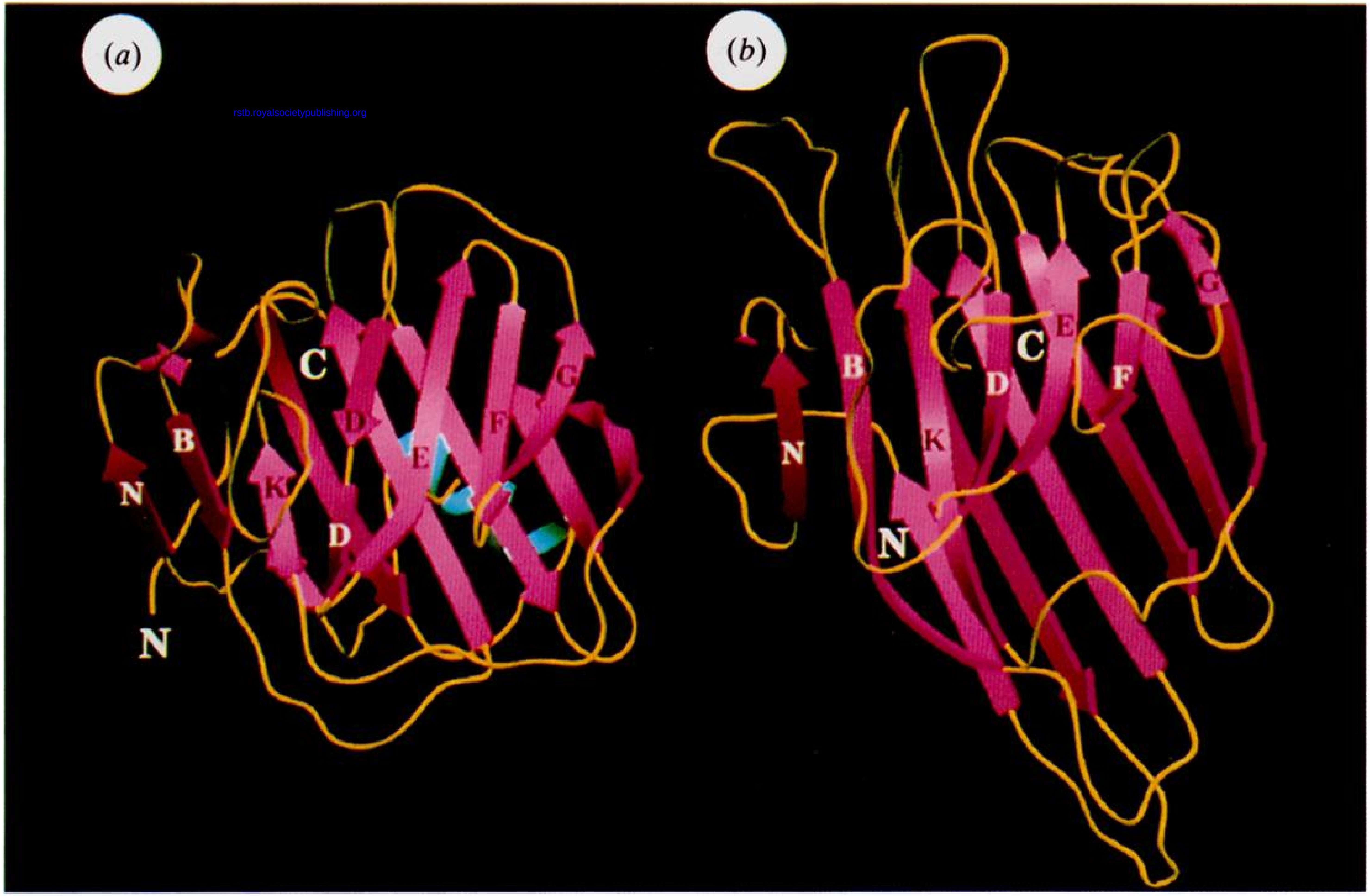
Zhu, Z.-Y., Šali, A. & Blundell, T.L. 1992 A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng.* **5**, 43–51.

Discussion

G. A. Dover (*Department of Genetics, University of Leicester, U.K.*). Professor Blundell has described how structural modules shared by different proteins can be successfully recognized by examining the spacings of key residues that go towards their assembly. Does he also make use of compensatory (complementary) changes to recognize conserved structures? If such changes are frequent, then it could mean that there is structural information to be gained from an examination of residue relationships even in the most divergent regions of the primary sequence, in addition to the key conserved residues in their fixed positions. This sort of exercise has played an important role in tracing the evolutionary history and patterns of sharing of functionally important secondary structural motifs of the ribosomal RNAs. If compensatory mutations are frequent in proteins, then it might mean that widely shared structural motifs reflect divergent rather than convergent evolution.

T. L. Blundell. We initially search for sequences identifying the compatability of residues at each individual position in the new sequence with the putative common fold. Professor Dover is correct that we need to be sure that the sequence changes are complementary and are compatible with a three-dimensional structure of the fold proposed. This we do by constructing a three-dimensional model based on the structure of the known homologue or analogue with a common fold. In fact, sequence changes are rarely complementary in a simple way. The elements of secondary structure differ in their relative positions and orientations within a homologous family and changes of volume of individual residues are easily accommodated. This makes the analysis of compensatory changes difficult and only rarely can it be used to eliminate a sequence.

The colour plate in this paper was printed by George Over Limited, London and Rugby



gure 3. Schematic diagrams of (*a*) serum amyloid P–component (SAP) and (*b*) concanavalin A showing eir similar fold. β-strands are drawn as broad arrows, α-helices as ribbons, and coil regions as thin rope. he β-strands (B, D, E, F, G, K and N) of the β-sheets in the foreground are labelled according to their rder in SAP. Rearrangement of concanavalin A is needed to achieve identical β-strand connections to ose of SAP as the N and C termini are not situated in topologically equivalent positions (Rufino & undell 1994).