Journal of Computer-Aided Molecular Design, 8 (1994) 5–27
ESCOM

J-CAMD 232

Structure-based identification and clustering of protein families and superfamilies

Stephen D. Rufino and Tom L. Blundell*

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

Received 10 November 1993
Accepted 20 November 1993

Key words: Protein families; Protein fold; Protein structure comparison; Secondary structure elements; Structure-based drug design

SUMMARY

We describe an approach to protein structure comparison designed to detect distantly related proteins of similar fold, where the procedure must be sufficiently flexible to take into account the elasticity of protein folds without losing specificity. Protein structures are represented as a series of secondary structure elements, where for each element a local environment describes its relations with the elements that surround it. Secondary structures are then aligned by comparing their features and local environments. The procedure is illustrated with searches of a database of 468 protein structures in order to identify proteins of similar topology to porcine pepsin, porphobilinogen deaminase and serum amyloid P-component. In all cases the searches correctly identify protein structures of similar fold as the search proteins. Multiple cross-comparisons of protein structures allow the clustering of proteins of similar fold. This is exemplified with a clustering of α/β- and β-class protein structures. We discuss applications of the comparison and clustering of three-dimensional protein structures to comparative modelling and structure-based protein design.

INTRODUCTION

Comparative modelling of protein structures [1–5] extends the range of application of structure-based drug design to proteins of known sequence for which the structure of at least one homologous protein is known. Several steps in the process of comparative modelling of protein structures rely directly or indirectly on the comparison of three-dimensional protein structures. For example, an important step in modelling is pairwise or multiple alignment of 3D structures of homologues in order to derive a framework for the protein to be modelled. The alignment of more distantly related structures allows a wider range of homologues to be used in comparative modelling.

*To whom correspondence should be addressed.

In some cases, especially where proteins are distantly related, identification of the fold of a protein from its sequence may not be easy, but recognition can often be achieved by the use of structural templates [6], which can be derived from known structures using environment-specific amino acid substitution scores. These are calculated from observed substitutions in structurally aligned protein families [7–9]. The ability to recognise protein structures that are distantly related to a sequence depends on the structural divergence within the protein families used to derive the environment-specific amino acid substitution scores. Thus, procedures for comparing distantly related protein structures should, by improving the substitution scores, increase the range of detection of protein structures that are distantly related to a sequence.

A first approach to automated procedures for the comparison of protein structures relies on the spatial superposition of the protein structures [10,11]. Such procedures search for the pairwise superposition that maximises the number of structurally equivalent residue pairs, typically defined as having inter-C^{α}-atom distances of at most 3.5 Å, and at the same time minimises the mean of the squared distances between these structurally equivalent pairs of C^{α}-atoms [12,13]. An alternative approach involves finding the best superposition of fragments from each structure [11,14,15]. In the case of closely related proteins, these approaches fare well in defining the equivalent residues of the hydrophobic protein core. To assign equivalences to the more structurally variable parts of proteins, such as solvent-exposed loops, dynamic programming algorithms [16–18] can be used to establish topological equivalence of residues, based on amino acid type, secondary structure and local environment [19–23]. Multiple structural comparisons have been used to generate alignments of protein structures within known protein families [7,8].

To be able to compare distantly related proteins or evolutionarily unrelated proteins of similar fold, structural comparison procedures must not be too sensitive to the distortions and deformations that protein folds can accommodate [24–26]. Additionally, the tasks of searching structural databases for protein structures of similar fold and of cross-comparing large numbers of protein structures, so as to cluster those with similar folds, require rapid structural comparison procedures.

Fast structural comparison procedures fall into two categories: those that consider protein structures at the level of their residues and those that consider protein structures at the level of their secondary structure elements.

The optimal structural superposition of two proteins can be found by applying a recursive dynamic programming procedure to the matrix of distances between structurally equivalent C^{α}-atoms and updating the distance matrix after each recursion [27]. The problem of local minima is avoided by using multiple initial superpositions. The detection of pairs of fragments having similar backbone conformations and the clustering of these pairs of fragments into sets with compatible rotation and translation transformations together provide an alternative fast procedure for structural comparison which allows reversal of chain direction and alternative chain connections [28,29]. Protein structures can also be compared on the basis of similarity of the distances between C^{α}-atoms of two residues. However, assumptions about the alignment of one residue pair implicate, through the relations of their C^{α}-atoms, assumptions about the alignment of other residues. Because unmodified dynamic programming algorithms are not suited to solving such interdependent problems, other approaches have been developed. The comparison of weighted inter-C^{α}-atom distance maps is a fast procedure which is insensitive to N- and C-terminal overhangs but sensitive to insertions and deletions [30]. A Monte Carlo procedure allows structural

alignment by maximising a score based on the similarity of inter-C^{α}-atom distances [29,31–33].

Most observed amino acid substitutions, insertions and deletions occur in loop regions on the surface of proteins, rather than within the hydrophobic core. Substitutions within the hydrophobic core of proteins may be accommodated by small deformations or relative rotations and translations of secondary structure elements [24,34,35]. Such deformations and relative movements complicate the structural superposition of distantly related proteins. Procedures that compare protein structures at the level of their elements of secondary structure are less sensitive to such deformations and relative movements [36–45]. Because the number of secondary structures in a protein is relatively small compared to the number of residues, the complexity of the problem of detecting protein structures with similar folds is greatly reduced. Procedures that rely on secondary structures are well suited for database searches and multiple cross-comparisons.

The alignment of the secondary structures of two proteins can be achieved by a two-level dynamic programming procedure [42]. Secondary structures are represented by vectors and are described by a series of features including vector length, accessible surface area, hydrophobicity, cylindrical hydrophobic moment and spherical hydrophobic moment. The environment of a secondary structure is described by its relations with all other secondary structures of the protein. These relations include angles and distances between secondary structure vectors. At the first level of the dynamic programming procedure all pairs of secondary structures, one from each protein, are considered in turn. For each pair of secondary structures a structural alignment is determined by considering the relations of the selected pair with the secondary structures in their respective proteins. The scores along the alignment paths are accumulated in a matrix. At the second level of the dynamic programming algorithm the matrix of cumulated scores is used to determine the structural alignment of secondary structures of the two proteins. Additionally, the scores in this cumulative matrix can be combined with the scores from a residue comparison cumulative matrix [21–23] to generate an improved structural alignment at the residue level.

Alternatively, graph theory algorithms have been used to find the structurally equivalent secondary structures in two protein structures [38–41]. In this approach secondary structures are represented by vectors and a protein structure is represented by these vectors and the distances and angles between them. This representation allows the application of graph theory algorithms to the comparison of protein structures. Given tolerances for variations in inter-vector distances and angles, the largest common substructure of two protein structures can be found [41].

Multiple cross-comparisons of related protein structures can be used to derive dendrograms [12,46,47]. In the case of multiple cross-comparisons of both related and unrelated proteins, those of similar fold are grouped by clustering [31,43,44]. From such clustering a representative, non-redundant set of protein structures can be determined [29,44].

In this paper we present results of the identification of protein structures of similar fold, using an automated procedure for the comparison of protein structures. The procedure compares protein structures at the level of their elements of secondary structure, each of which is described by a set of features and an associated vector. The local environment of a secondary structure is defined as the elements of secondary structure with which it forms contacts. The comparison relies on finding similarities in the features and local environment of secondary structures and in the features of secondary structures defined in local environments. Similarity of features is determined by comparison with variations of features of secondary structures observed in a set of aligned protein structures. Similarity of local environment is determined through the use of

a maximum common subgraph algorithm with defined cutoffs to the variations of inter-vector relations. We show that this comparison procedure is able to identify protein structures of similar fold to porcine pepsin, porphobilinogen deaminase and serum amyloid P-component. In the case of porcine pepsin the procedure finds all other monomeric cellular aspartic proteinases in the database, but also the dimeric retroviral proteinases. In the case of porphobilinogen deaminase the procedure finds all the periplasmic binding proteins of type II. In the case of serum amyloid P-component the procedure finds all the legume lectins in the database, but also 1,3-1,4 β-gluconase of similar fold. Multiple cross-comparisons of the α/β-class protein structures effectively cluster proteins with similar folds, such as type II periplasmic binding proteins. Similarly, multiple cross-comparisons of the β-class cluster the aspartic and viral proteinases. Serum amyloid P-component, the legume lectins and 1,3-1,4 β-gluconase are also clustered. We discuss the application of such comparisons and clustering of protein structures to comparative modelling and structure-based drug design.

METHOD

Definition of elements of secondary structure

Secondary structure was defined from hydrogen-bond patterns, by the method of Kabsch and Sander [48] with the program SSTRUC [49]. Two types of secondary structure elements are considered: helices, including α-, 3_{10}- and π-helices, and extended strands, including parallel, antiparallel and mixed β-strands. 3_{10}-helices were considered only if they are longer than three residues. Short 3_{10}-helices adjacent to α-helices were merged.

For each secondary structure, a vector was determined representing its direction, length and position in the protein structure. The vector is defined so as to minimise the sum of the squares of the distances from the vector to the C^{α}-atoms of the secondary structure [50,51].

Two helices separated by at most one residue were merged if the shorter of the two is not longer than eight residues, the angle between the two is less than 70 degrees and neither of the two helices can be more favourably merged with another helix. Two extended strands separated by at most four residues were merged if the shorter of the two is not longer than five residues, the angle between the two is less than 70 degrees, they both form hydrogen bonds with the same extended strands and neither of the two can be more favourably merged with another extended strand.

We followed the suggestions of Šali and Blundell [19] in defining each secondary structure by a set of features:

- the length in number of residues
- the average number of C^{α}-atoms within 14 Å of the secondary structure's C^{α}-atoms [52]
- the number of secondary structures which form contacts as defined by C^{α}-atom distances with the secondary structure
- the distance between the vector defining the secondary structure and the centre of gravity of the protein structure
- the angle between the vector defining the secondary structure and the vector linking the mid-point of the secondary structure vector to the centre of gravity of the protein structure
- the average percentage accessibility of the residues of the secondary structure
- the spherical accessibility moment

- the angle between the spherical accessibility moment vector of the secondary structure and the vector of the secondary structure
- the angle between the spherical accessibility moment vector of the secondary structure and the vector linking the mid-point of the vector of the secondary structure to the centre of gravity of the protein structure.

Distances between vectors were calculated as distances between their mid-points. Angles between vectors were calculated as the angles between their direction vectors.

Protein structure description

A protein structure is defined as a string of secondary structures. It is described by the features of its secondary structures and by the relations between its secondary structures, which include:

- the contacts as defined by C^{α}-atom distances
- the distance between the mid-points of the vectors of the secondary structures
- the angles between the direction vectors of the secondary structures
- the dihedral angle formed by the vectors of the secondary structures and the vector linking the mid-points of the vectors of the secondary structures
- the angle between the projections of the vectors of the secondary structures onto a plane that includes the vector linking the mid-points of the vectors of the secondary structures.

Local environment

The local environment of a secondary structure in a protein structure is defined by the secondary structures with which it forms contacts, its relations with them and the features of these contacting secondary structures.

Structural alignment

Two proteins can be aligned structurally by comparing their secondary structures. Each secondary structure and its local environment in the first protein is compared in turn with each secondary structure of the same type and its local environment in the second protein. Two secondary structures are in similar local environments if a subset, greater than some defined minimum, of the secondary structures in their respective local environments has similar relations. The largest set of equivalent secondary structures in the local environments of the two secondary structure elements is determined using a procedure for maximal common subgraph detection [41,53]. If two secondary structures have similar local environments, a score is attributed to their alignment. This score is the sum of the log odd propensities of the variations of their features and of the variations of the features of the corresponding secondary structures, found within their local environments. The log odd propensities of the variations of each feature are determined from the observed variations in a set of 73 aligned protein families [9]. The best structural alignment of secondary structures of two proteins is determined using a dynamic programming algorithm. The algorithm allows for the detection of local similarities [17,54].

Structural alignment scoring

Two different scoring schemes are adopted. The first, DS_{AB}, is used when searching the database with the protein structure *a* for a protein *b* of similar fold, where

$$DS_{AB} = \lambda * \sqrt{\frac{S_{ab}}{S_{aa}}} * \frac{P_{ab}}{L_a}$$

The second scoring scheme, CS_{AB}, is used for cross-comparing several protein structures in search of clusters of proteins with similar folds. For any pair of proteins *a* and *b* the score is:

$$CS_{AB} = \lambda * \sqrt{\frac{S_{ab}}{\min(S_{aa}, S_{bb})}} * \frac{P_{ab}}{\max(L_a, L_b)}$$

S_{ab} is the alignment score of protein *a* with protein *b*. S_{aa} and S_{bb} are the alignment scores of respectively protein *a* with itself and protein *b* with itself. P_{ab} is the number of aligned secondary structures. L_a and L_b are the lengths of protein *a* and protein *b*, respectively. The alignment score, S_{ab}, is first normalised by dividing it by the score S_{aa} for aligning the search structure with itself in the case of a database search and by the smaller of the scores S_{aa} and S_{bb} for aligning the two compared proteins with themselves in the case of a multiple cross-comparison. λ is a constant. The square root is introduced to take account of the dimension of the scoring, which is higher than one due to the local environment scores. To avoid favouring structural similarities involving only a small number of secondary structures, the scores are multiplied by the ratio of aligned secondary structures, P_{ab}, to the length, L_a, of the search structure in the case of a database search and to the maximum of the lengths, L_a and L_b, of the compared structures in the case of a multiple cross-comparison.

In the case of multiple cross-comparisons a matrix containing all the pairwise protein structure similarity or dissimilarity scores is obtained. The dissimilarity score, D, is obtained from the similarity score, S, by:

$$D = \frac{\lambda - S}{\lambda}$$

Analysis of multiple cross-comparisons

The information contained in a dissimilarity matrix is visualised, at least partially, using hierarchical clustering. The KITSCH program [55], part of the Phylogenetic Inference Package PHYLIP, is used to generate a dendrogram from a dissimilarity matrix [12,46,47]. LaTeX dendrograms are obtained through the use of the program DRAWTREE [56]. The dendrograms are constructed so as to minimise the difference between the distances in the dendrogram and those in the dissimilarity matrix [57]. In a dendrogram the distance between two protein structures is represented by the distance along the horizontal axis from the tip of their respective branches to their first common branching point.

RESULTS

Search for protein structures of similar fold to porcine pepsin

The porcine pepsin structure [58–60] is an aspartic proteinase, comprising two structurally similar lobes. The N-lobe, including residues 1–175, is related by a pseudo twofold axis to the C-lobe, including residues 176–327. The sequence similarity between the two lobes is 14%. A rigid-

body superposition of the two lobes identifies 33 equivalent residue pairs with an rms deviation of 1.6 Å. Each lobe consists of a mixed eight-stranded β-sheet packed against a four-stranded antiparallel β-sheet, formed by the curling over of four ends of the strands of the mixed β-sheet, so that the two sheets are orthogonal to each other. Strands from both N- and C-lobes contribute to a fifth six-stranded antiparallel β-sheet which forms a foundation upon which the two lobes sit. The active site is in the cleft between the two lobes, each of which contributes one catalytic aspartate.

The structures of several mammalian and fungal aspartic proteinases have been determined. The known mammalian aspartic proteinase structures include porcine pepsin [58–60], human renin [61,62], mouse renin [62] and bovine chymosin [63,64]. The known fungal aspartic proteinase structures include mucor pepsin [65], endothiapepsin [66], penicillopepsin [67], rhizopuspepsin [68] and yeast proteinase A [69]. The average sequence identity among the pepsin-like aspartic proteinases is 34.7% and the lowest sequence identity, 22.4%, is found between endothiapepsin

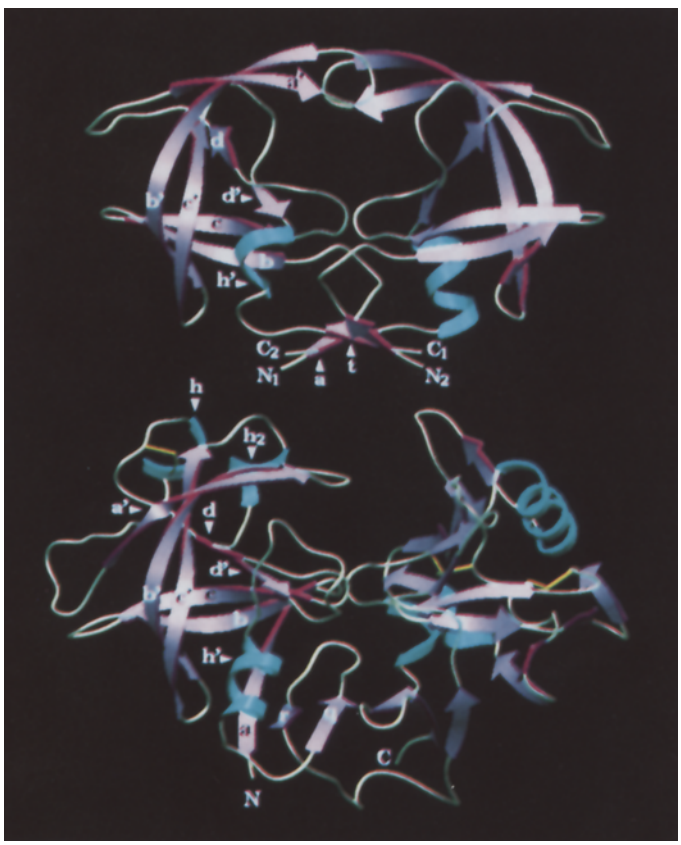

Fig. 1. Similarity of fold of the dimeric retroviral proteinases and monomeric pepsin-like aspartic proteinases. The β-strands (a, b, c, d, a', b', c', d', r, q, t) and the α-helices (h, h_2, h') of the N-terminal lobe of porcine pepsin (5pep [60], lower) and the corresponding subunit of HIV-1 proteinase (5hvp [70], upper) are labelled so as to highlight the similarity in the fold of these two protein families. The two proteins are viewed perpendicular to their twofold axis and along their active site clefts.

and human renin. Rigid-body superpositions reveal 267 (79.7%) structurally equivalent C^{α}-atoms between these two proteins and an rms deviation of 1.57 Å. The highest sequence identity, 59.4%, is found between bovin chymosin and porcine pepsin, which share 315 (95.5%) structurally equivalent C^{α}-atoms with an rms deviation of 0.98 Å.

The retroviral and the pepsin-like aspartic proteinases share a similar function and a conserved Asp-Thr/Ser-Gly tripeptide which includes the catalytic aspartate. On the basis of these similarities it was predicted that the structure of the retroviral proteinase dimer would be similar to that of the pepsin-like aspartic proteinase monomer [70]. This was later confirmed when the structures of the Rous sarcoma virus proteinase [72,73] and the HIV proteinase [74–76] were determined. A third retroviral proteinase structure, the avian myeloblastosis associated virus [77], differs from the Rous sarcoma virus proteinase by only one residue. A rigid-body superposition of the two structures reveals an rms deviation of 0.22 Å for 215 C^{α}-atoms out of 233.

A search of the database for protein structures of similar fold to porcine pepsin identifies, with the highest scores, all the members of the aspartic proteinase family (Table 1). The members of the retroviral aspartic proteinase family are recognised as the second highest scoring family, as would be expected from their related folds (Fig. 1). The scores and the number of aligned pairs of secondary structures for other protein structures are considerably lower than those of the pepsin-like and retroviral aspartic proteinases. Many are nevertheless all-β or mainly-β proteins.

A search for proteins of similar structure to the avian myeloblastosis associated virus retroviral proteinase identifies, with the highest scores, all the retroviral aspartic proteinases (Table 2). The second highest scoring family is that of the pepsin-like aspartic proteinases, of which all members are identified. The scores and number of aligned pairs of secondary structures of the retroviral and aspartic proteinases are much higher than those of the other protein structures identified.

TABLE 1
LIST OF THE HIGHEST SCORING PROTEIN STRUCTURES AGAINST PORCINE PEPSIN [60]

PDB code	Alignment score	Length of segment	Aligned pairs	Protein structure
5pep	10000	32	32	Porcine pepsin
1bbs	6931	32	31	Human renin
4cms	6238	33	30	Chymosin
2apr	6182	32	30	Rhizopuspepsin
2ren	5525	33	28	Mouse renin
3app	5469	33	29	Penicillopepsin
1ypa	5106	33	30	Yeast proteinase A
1mpp	5035	34	28	Mucor pusillus pepsin
4ape	4991	32	28	Endothiapepsin
1mvp	1404	36	20	Myeloblastosis associated virus proteinase
2rsp2	1233	37	18	Rous sarcoma virus proteinase
5hvp	993	36	16	HIV proteinase
4hvp	963	34	17	HIV proteinase
2bat	225	23	7	Neuramidase N2
2sga	154	20	6	Proteinase A
2alp	124	11	5	α-Lytic proteinase
1bov	121	12	6	Verotoxin
2hsc	115	6	5	Heat-shock cognate
2sns	107	12	5	Staphylococcal nuclease

TABLE 2
LIST OF THE HIGHEST SCORING PROTEIN STRUCTURES AGAINST MYELOBLASTOSIS ASSOCIATED VIRUS PROTEINASE [77]

PDB code	Alignment score	Length of segment	Aligned pairs	Protein structure
1mvp	10000	24	24	Myeloblastosis associated virus proteinase
2rsp2	9311	24	24	Rous sarcoma virus proteinase
5hvp	3234	27	17	HIV proteinase
4hvp	3196	26	18	HIV proteinase
5pep	2497	35	20	Porcine pepsin
2ren	2216	33	19	Mouse renin
3app	2142	37	17	Penicillopepsin
4ape	2012	35	17	Endothiapepsin
4cms	1754	39	16	Chymosin
1ypa	1416	38	15	Yeast proteinase A
1mpp	1413	33	15	Mucor pusillus pepsin
2apr	1342	31	16	Rhizopuspepsin
1bbs	1121	39	15	Human renin
7cat	320	32	7	Catalase
1npc	282	12	6	Neural protease
3bcl	280	18	6	Bacterichlorophyll protein A
2alp	224	10	5	α-Lytic proteinase
3tln	195	13	5	Thermolysin
2mev	156	18	5	Mengo virus coat protein
1bbt3	147	14	5	Foot and mouth disease virus coat protein

Search for protein structures of similar fold to porphobilinogen deaminase

Porphobilinogen deaminase (PBGD) is an essential enzyme in the biosynthetic pathway of tetrapyrroles. It catalyses the multistep polymerisation of four molecules of the substituted pyrrole porphobilinogen, to give an open-chain tetrapyrrole, hydroxymethylbilane, which is a precursor of cyclic tetrapyrroles such as hemes, chlorophylls and vitamin B_{12}. A primer dipyrromethane, comprising two molecules of porphobilinogen, is covalently bound to a cysteine side chain of PBGD.

The structure comprises three α/β-domains [78]. The polypeptide chain crosses from the first domain to the second and then back to the first one before reaching the third domain. The first two domains have a similar fold, consisting of a distorted five-stranded doubly wound parallel β-sheet. Of the five strands, four are parallel while the fifth is antiparallel. The antiparallel strand in each domain corresponds to the strand immediately after the cross-over from the other domain. The third domain consists of a three-stranded antiparallel β-sheet with three helices packed on one of its faces. The active site is situated within a cleft between the first two domains. The cofactor, which is positioned in the cleft, is covalently linked to a cysteine residue of the third domain, and forms many ion pairs between the acid groups on the pyrrole rings and lysines and arginines in the protein.

The first two domains of PBGD were found to be topologically very similar to several known binding proteins, including the transferrins and the group II periplasmic receptors [78]. Transferrin possesses two lobes, which are similar in topology to the first two domains of PBGD. The known transferrin structures include lactoferrin [79,80] and serum transferrin [81]. An example

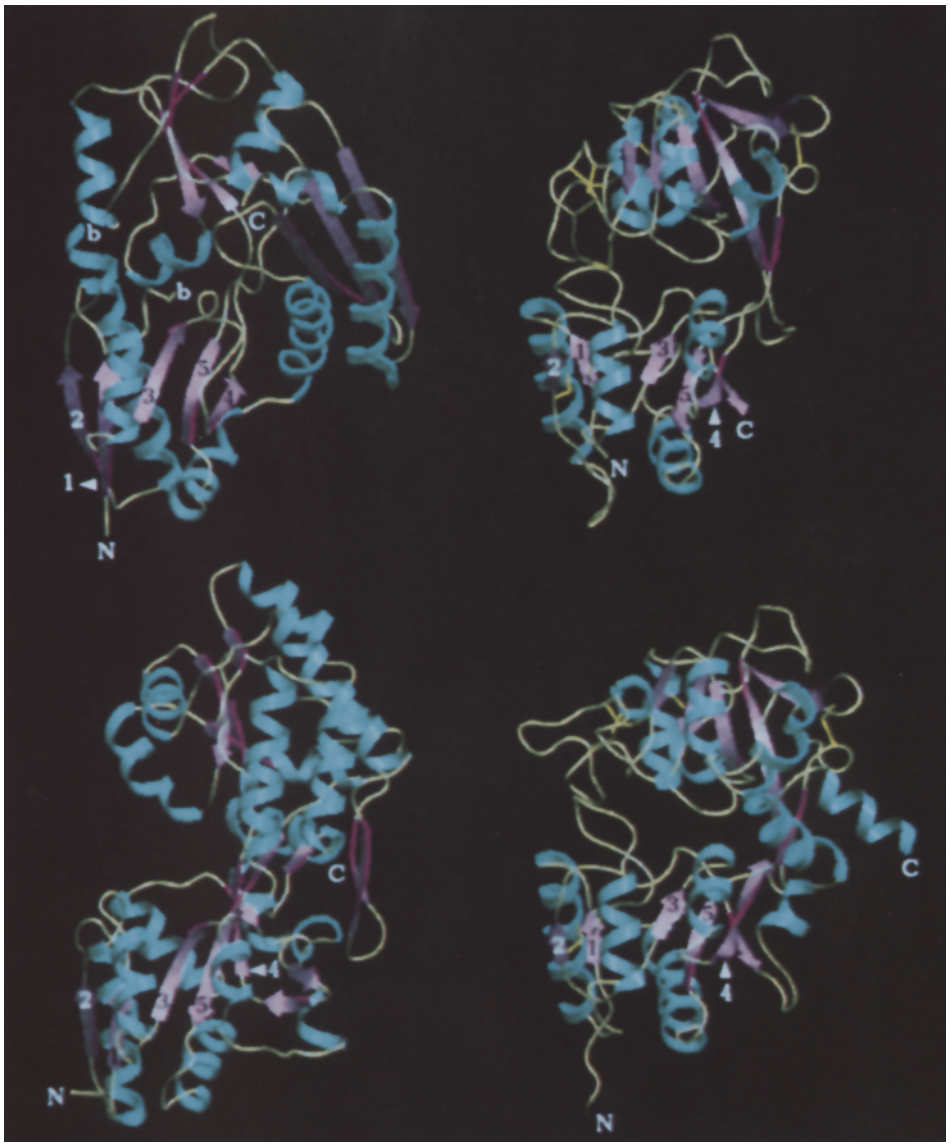

Fig. 2. Similarity of fold of the N-terminal domains of PBGD (1pda [78], upper left) and the type II periplasmic binding proteins transferrin (1tfd [81], upper right), maltose-binding protein (1omp [82], lower left) and lactoferrin (1lfg [79], lower right). The β-strands are numbered 1 to 5, following their sequential order in the N-terminal domain of PBGD. The C-terminal domains have a similar fold, although the different domain orientations do not allow this to be apparent in this figure. A break in the main chain of PBGD is indicated (b).

of a known group II periplasmic receptor structure is the maltose-binding protein [82]. The binding site of these proteins is situated in the cleft between two distorted doubly wound parallel β-sheet domains, as is the active site of PBGD.

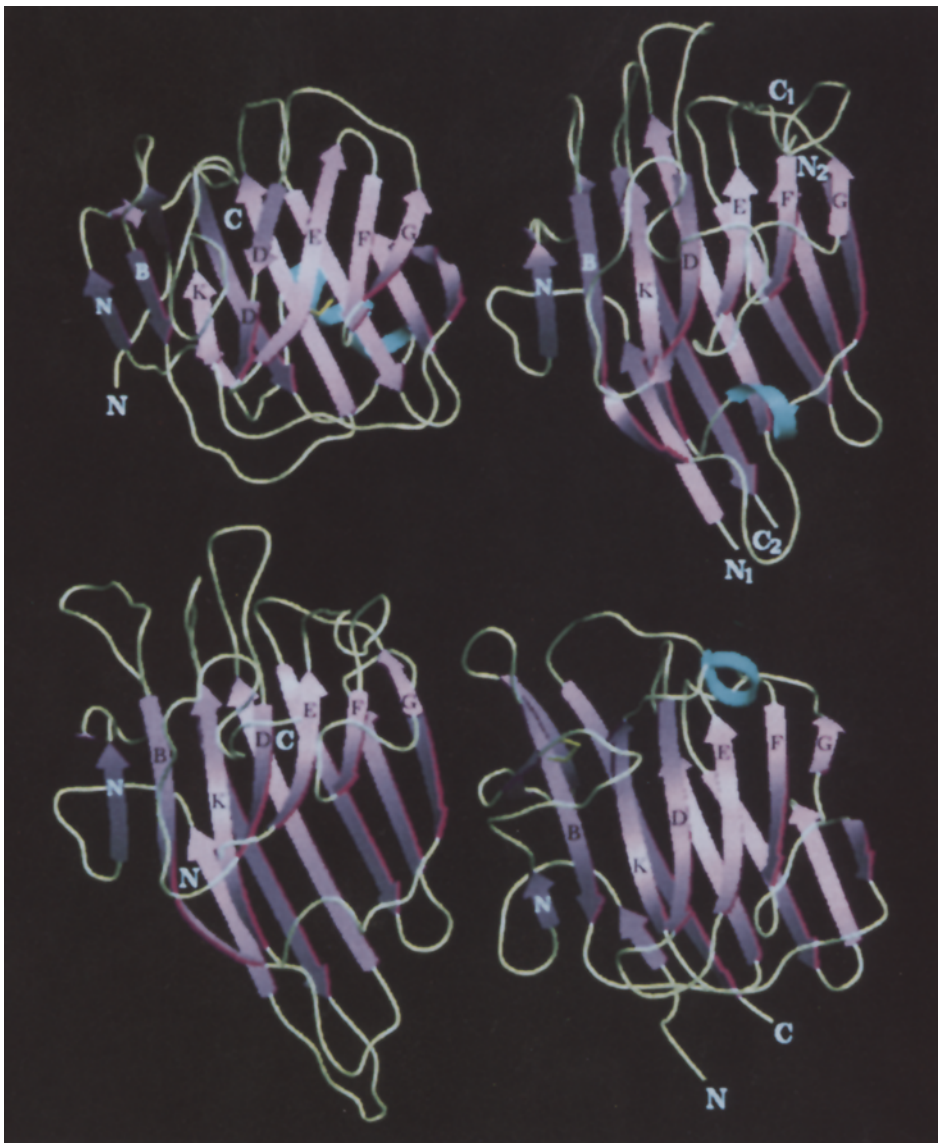

Fig. 3. Similarity of fold of SAP [88], upper left, pea lectin (2ltn [89], upper right), concanavalin A (3cna [90], lower left) and 1,3-1,4 β-gluconase (1ayh [91], lower right). To highlight the similarity of fold of these four proteins, the β-strands (B, D, E, F, G, K, N) of the β-sheets in the foreground are labelled according to their order in SAP.

The group I periplasmic binding proteins have a fold somewhat similar to that of the group II periplasmic binding proteins and PBGD. The two central β-sheets are undistorted doubly wound parallel β-sheets. Therefore, the amino acid chain crosses only once from the first domain to the second. The known group I periplasmic binding protein structures include leucine/isoleucine/valine-binding protein [83], leucine-binding protein [84], L-arabinose-binding protein [85],

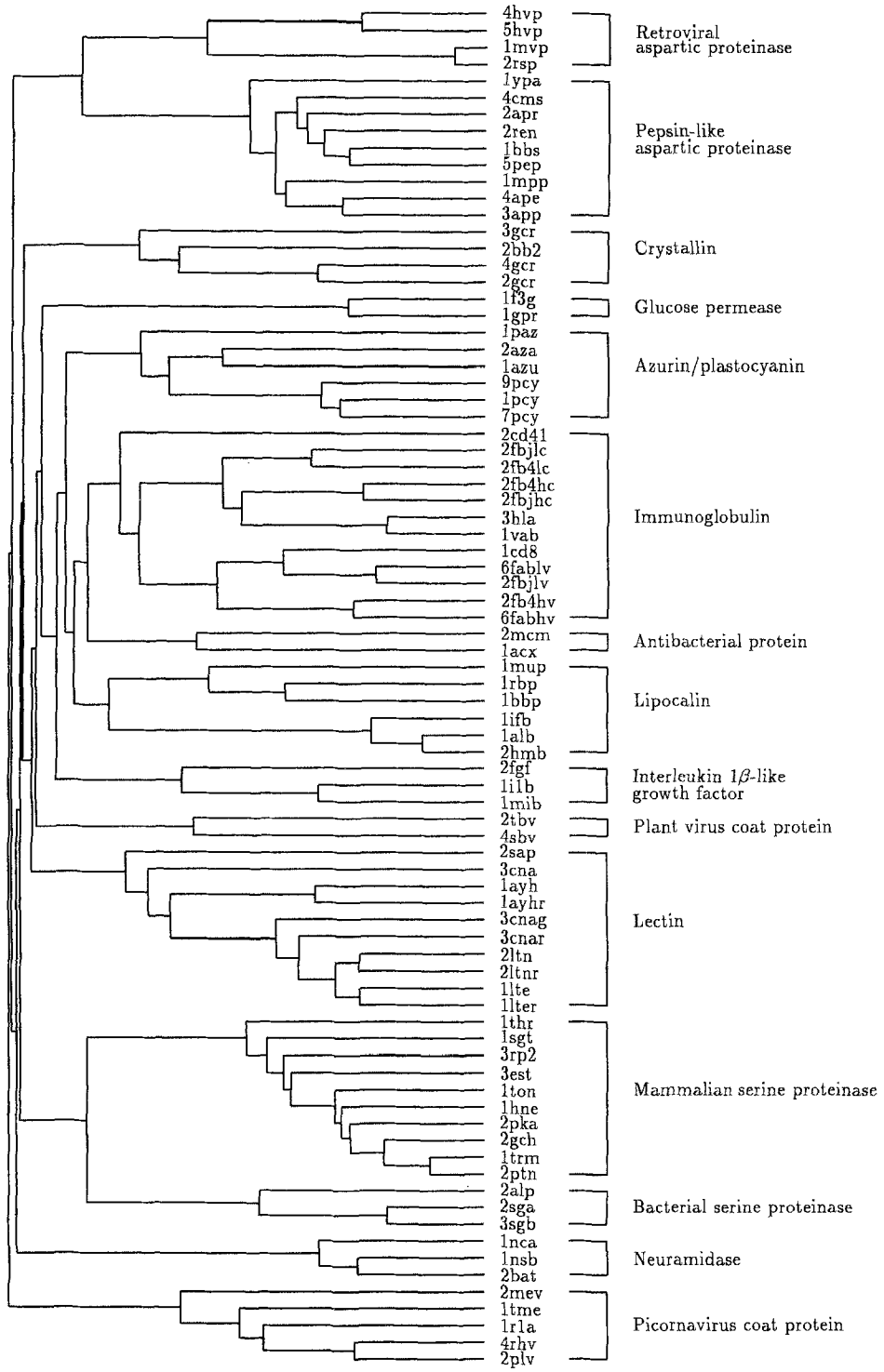

Fig. 4. Dendrogram of all β-proteins. The families of the clustered proteins are indicated on the right.

TABLE 3
LIST OF THE HIGHEST SCORING PROTEIN STRUCTURES AGAINST PBGD [78]

PDB code	Alignment score	Length of segment	Aligned pairs	Protein structure
1pda	10000	23	23	Porphobilinogen deaminase
1lfg_c	2256	21	16	Lactoferrin (C-lobe)
1lfg_n	1950	23	16	Lactoferrin (N-lobe)
1tfd_n	1445	26	12	Transferrin (N-lobe)
2gbp	1408	26	12	Galactose-binding protein
1omp	1246	35	11	Maltose-binding protein
2lbp	1095	28	10	Leucine-binding protein
1dri	1087	19	10	Ribose-binding protein
1pfk	1057	30	11	Phosphofructokinase
2liv	1022	28	10	Leucine/isoleucine/valine-binding protein
3aat	772	23	8	Aspartate aminotransferase
4pfk	581	15	7	Phosphofructokinase
3icd	509	24	7	Isocitrate dehydrogenase
1pgd	498	23	8	Phosphogluconate dehydrogenase
1ama	425	24	6	Aspartate aminotransferase

D-galactose/D-glucose-binding protein [86] and D-ribose-binding protein [87].

A search of the database for protein structures of similar fold to PBGD identifies all the type II and type I periplasmic binding proteins, with the exception of the arabinose-binding protein (Table 3). The transferrin lobes are at the top of the list, but the topologically similar maltose-

TABLE 4
LIST OF THE HIGHEST SCORING PROTEIN STRUCTURES AGAINST SAP [88]

PDB code	Alignment score	Length of segment	Aligned pairs	Protein structure
2sap	10000	16	16	Serum amyloid P-component
2ltnr	3287	16	14	Pea lectin (rearranged)
2ltn	2757	13	12	Pea lectin
1ayhr	2634	20	13	1,3-1,4 β-gluconase (rearranged)
1lter	2535	17	13	Pea lectin (rearranged)
1ayh	2234	14	11	1,3-1,4 β-gluconase
1lte	2222	17	12	Pea lectin
3cnar	1950	15	12	Concanavalin A (rearranged)
3cnag	1903	13	10	Concanavalin A (genomic)
3cna	1224	16	9	Concanavalin A
1nca	1191	14	9	Neuramidase
2pab	1137	22	10	Prealbumin
1gpr	1133	16	9	Glucose permease (domain IIA)
1f3g	1057	21	9	Phosphocarrier II
1nsb	1037	12	9	Neuramidase
1faihv	1019	18	9	Immunoglobulin heavy variable domain
2hla	856	14	8	Histocompatibility antigen
3hlaa	827	8	7	Histocompatibility antigen
2fb4hv	827	16	9	Immunoglobulin heavy variable domain
1hillv	816	12	7	Immunoglobulin light variable domain

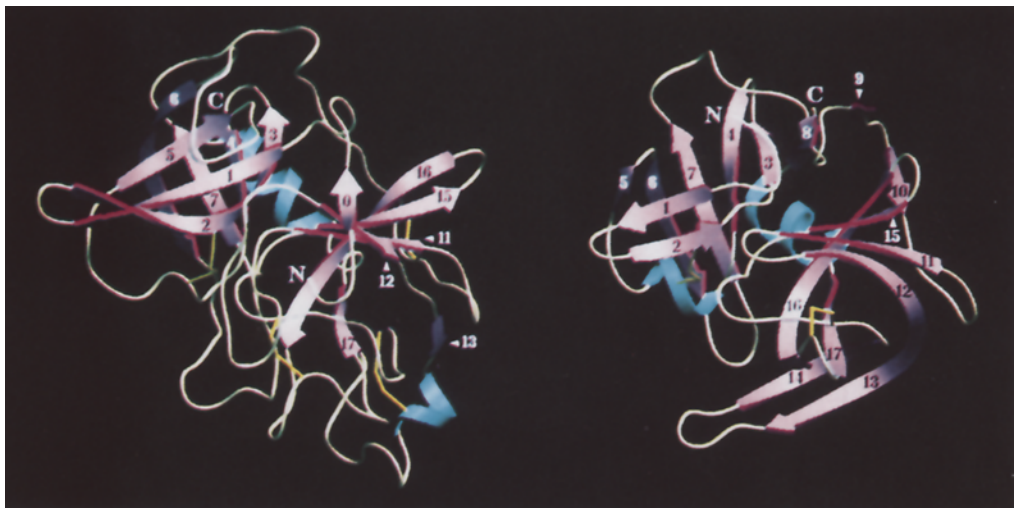

Fig. 5. Similarity of fold of the mammalian and bacterial serine proteinase families. Porcine pancreatic elastase (3est [93], left), a mammalian and proteinase A (2sga [94], right), a bacterial serine proteinase, are shown. To highlight the similarity between these two serine proteinase families, β-strands are labelled 1 to 17, following their sequential order in proteinase A. The β-strands numbered 1 to 7 form the N-terminal antiparallel β-barrel and those numbered 11 to 17 form the C-terminal antiparallel β-barrel of the serine proteinase fold.

binding protein scores slightly lower than the D-galactose/D-glucose-binding protein which has an all-parallel β-sheet. Maltose-binding protein is the type II periplasmic binding protein which is structurally most dissimilar to PBGD (Fig. 2). It has several insertions, including a long loop before the fifth strand in the N-domain sheet. Most of the other proteins identified are α/β-structures. These include members of the phosphofructokinase family, of the aspartate amino transferase family and of several dehydrogenase families.

Search for protein structures of similar fold to serum amyloid P-component

Serum amyloid P-component (SAP) is a member of the pentraxin family of pentameric plasma proteins that adopt a cyclic arrangement of their subunits. Although the precise role of SAP is not known, it has several well-characterised calcium-dependent functions. SAP is the major DNA- and chromatin-binding protein of the plasma. It binds to amyloid deposits including those associated with Alzheimer's disease, probably through the carbohydrate moieties, and it is also a calcium-dependent lectin.

The SAP protomer consists of two seven-stranded β-sheets [88], based on a jelly-roll formed by four N-terminal and four C-terminal strands. The remaining six central strands extend the two antiparallel β-sheets on the same side. The two sheets pack against each other with their strands roughly parallel.

Suprisingly, Emsley et al. [88] found the fold of SAP (Fig. 3) to be similar to that of the legume lectins pea lectin [89] and concanavalin A [90] and the bacterial enzyme 1,3-1,4 β-glucon-ase [91]. The relationships between SAP and the legume lectins are distant. The N- and C-termini are not situated in topologically equivalent positions, and rearrangements of both legume lectins

and 1,3-1,4 β-gluconase are necessary to achieve identical β-strand connections to SAP. SAP and concanavalin A, after rearrangement, share only 9% sequence identity, while SAP and 1,3-1,4 β-gluconase, after rearrangement, share only 8.5% sequence identity [92].

A search of the database for protein structures of similar fold to SAP identifies the legume lectins and 1,3-1,4 β-gluconase (Table 4). Rearranged legume lectins and 1,3-1,4 β-gluconase having the same β-strand topology as SAP are identified with higher scores and more aligned secondary structures than the original proteins. Additionally, a rearranged concanavalin A corresponding to its genomic sequence is identified. Several other protein families containing two antiparallel sheets are also identified. These include the immunoglobulin family, glucose permease family, and prealbumin. The similarity with the neuramidase β-sheet propeller family resides in six antiparallel β-strands of SAP.

Clustering of all-β protein structures

Multiple cross-comparisons of all-β protein structures (Fig. 4) reveal a coherent clustering of protein structures into families [7–9]. Families identified include pepsin-like aspartic proteinases, retroviral aspartic proteinases and lectins (Fig. 3).

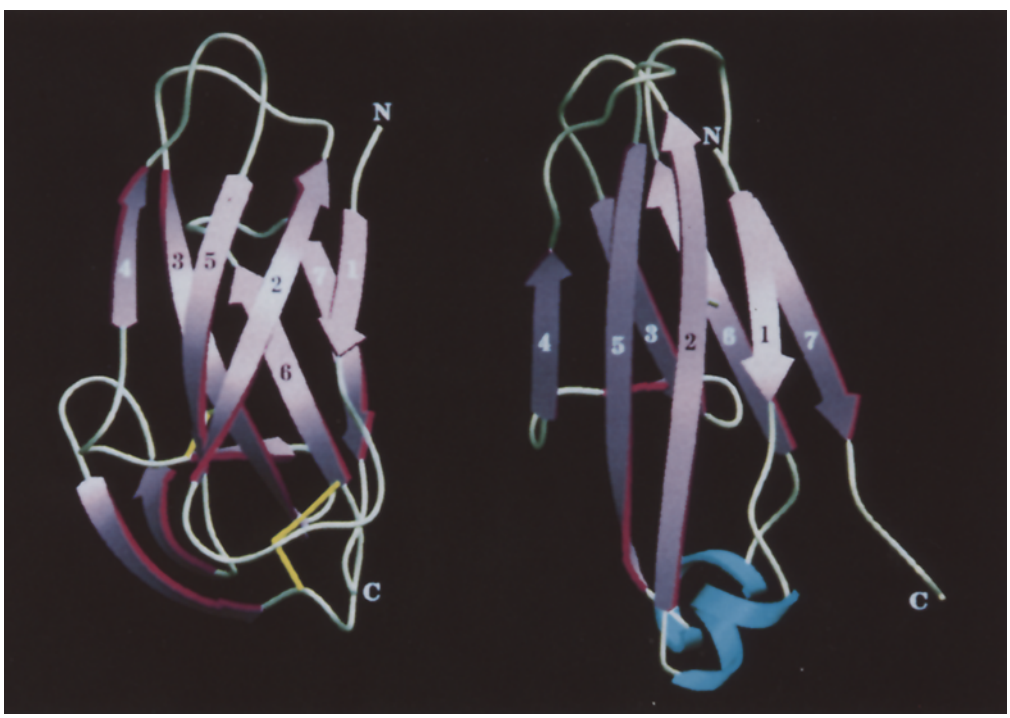

Fig. 6. Similarity of fold of the antibacterial and the immunoglobulin families. Macromycin (2mcm [95], left) and the constant domain of an immunoglobulin light chain (2fbj [97], right) are shown. To highlight the similarity between these two protein families, the β-strands of the packed jelly-roll β-sheets are labelled 1 to 7, following their sequential order in the immunoglobulin constant domain. The topologies of these β-sheets differ in that the β-strand labelled 4, which belongs to the N-terminal β-sheet in the immunoglobulin fold, is hydrogen bonded to the C-terminal β-sheet in macromycin.

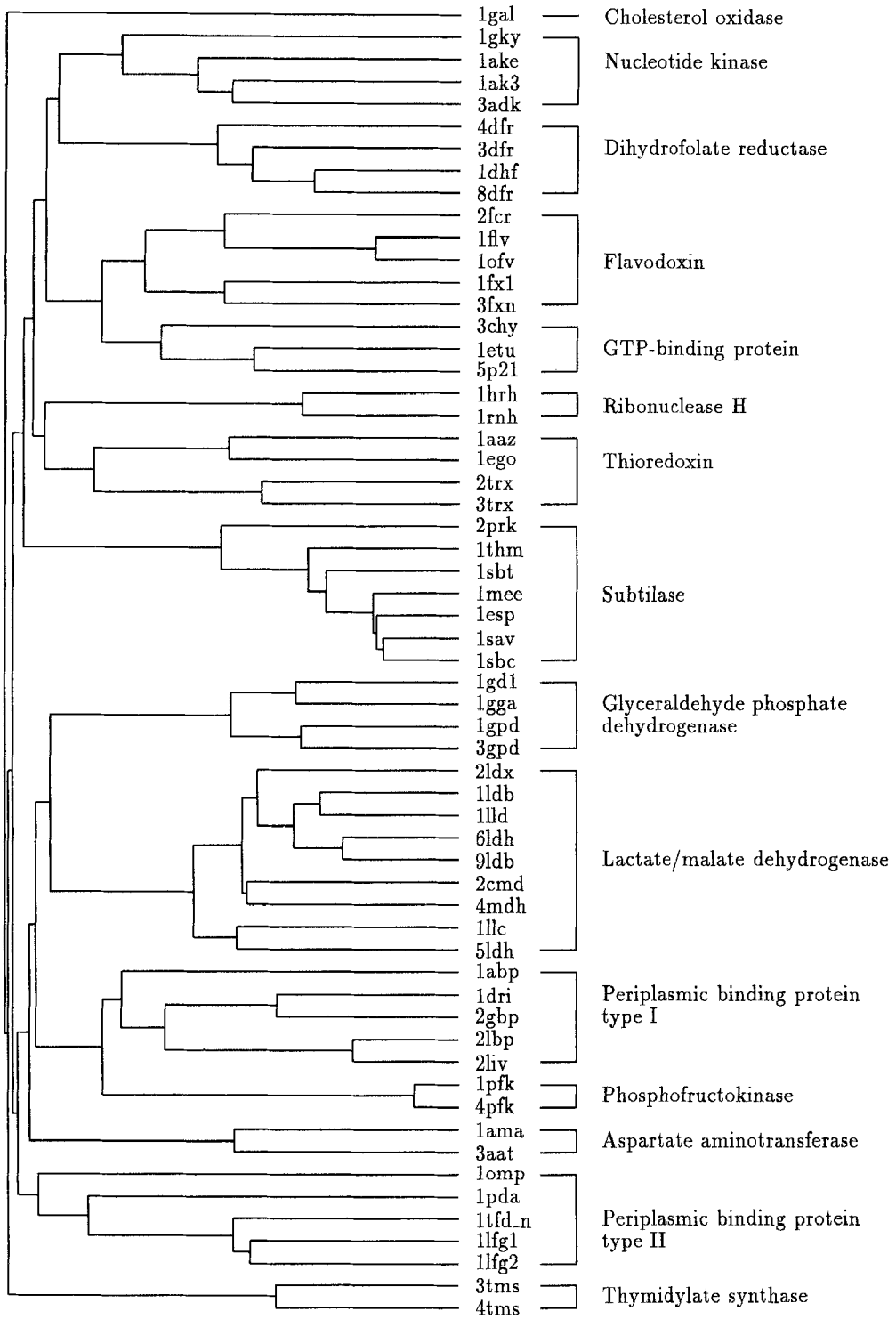

Fig. 7. Dendrogram of α/β-proteins. The families of the clustered proteins are indicated on the right.

Superfamilies of proteins with different functions and highly dissimilar sequences but similar fold can also be detected. The superfamilies identified include the dimeric retroviral and the monomeric pepsin-like aspartic proteinases (Fig. 1), the mammalian and bacterial serine proteinases (Fig. 5), the immunoglobulins, and antibacterial proteins (Fig. 6). The antibacterial proteins macromycin [95] and actinoxanthin [96] contain two packed jelly-roll β-sheets and a third small β-sheet perpendicular to the first two. The topology of the two packed β-sheets can be described as having an immunoglobulin constant-domain fold where the last strand (Fig. 6, β-strand 4) of the first β-sheet has been switched over to the second β-sheet.

Clustering of α/β protein structures

Multiple cross-comparisons (Fig. 7) group α/β protein structures into coherent families [7–9]. PBGD, transferrin, lactoferrin and maltose-binding protein are clustered together (Fig. 2). The type I periplasmic binding proteins are also identified as a family.

Several superfamilies are clearly identifiable from the clustering patterns. The classical doubly wound β-sheet topology of the NAD-binding domains of the glyceraldehyde phosphate dehydrogenase and lactate/malate dehydrogenase families is revealed by the clustering of these two families (Fig. 8). The clustering of the GTP-binding protein family and the flavodoxin family identifies the similarity in the topology of the parallel β-sheets of the GTP-binding domain of chemotaxis Y protein [100], p2l ras protein [101], elongation factor Tu [102] and the flavodoxins (Fig. 9). The clustering of type I periplasmic binding protein family and the phosphofructokinase family (Fig. 10) derives from the similarity in topology of their N-terminal domains. The N-terminal domain of phosphofructokinase [104,105] has a classical doubly wound β-sheet topology with an additional C-terminal β-strand; that of the type I periplasmic binding protein family is similar but lacks a β-strand on the edge of the β-sheet (Fig. 10, β-strand 3). The nucleotide kinase and dihydrofolate reductase families show a local similarity, consisting of a four-stranded doubly wound β-sheet topology (Fig. 11).

DISCUSSION

Comparative modelling and structure-based drug design

The approach described here automatically detects and compares proteins with a common fold, in a way that could previously be achieved only by an expert with detailed knowledge of the protein structures. This allows a broader range of protein structures to be included in the process of comparative modelling.

In general, the framework of a protein to be modelled is best obtained by using a weighted set of structures that are the closest homologues [108]. Where no close homologues are available, clustering of protein structures, as in Figs. 4 and 5, allows a systematic choice of structures. Thus, if new type II periplasmic binding proteins are to be modelled, the dendrogram in Fig. 4 indicates that the porphobilinogen deaminase and transferrin structures may be of value. This will be particularly so when the modelling is carried out by a procedure that allows for the flexibility of the framework, for example as in the approach to modelling using spatial constraints developed by Šali and Blundell [109], rather than by a procedure that involves assembly of rigid fragments such as COMPOSER [4,108,110].

If the sequence similarity is in the 'twilight zone' of less than 20%, the crucial step is the re-

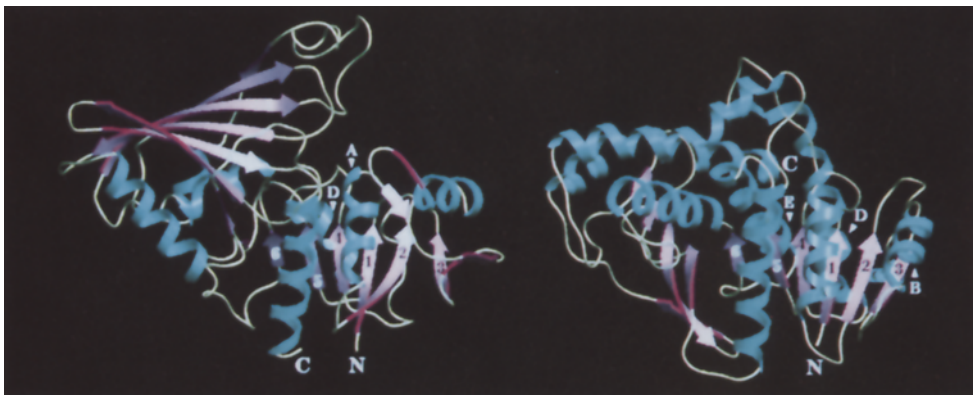

Fig. 8. Similarity of fold of the NAD-binding domain of D-glyceraldehyde-3-phosphate dehydrogenase (1gd1 [98], left) and malate dehydrogenase (2cmd [99], right). The β-strands (1, 2, 3, 4, 5, 6) and α-helices (A, B, C, D, E) of the NAD-binding domains are labelled according to their sequential order to highlight the fold similarity. The N-terminal domain of the malate dehydrogenase possesses a classical doubly wound β-sheet topology, whereas the N-terminal domain of the D-glyceraldehyde-3-phosphate dehydrogenase has deletions of α-helices B, C and E and insertions after β-strands 2 and 5.

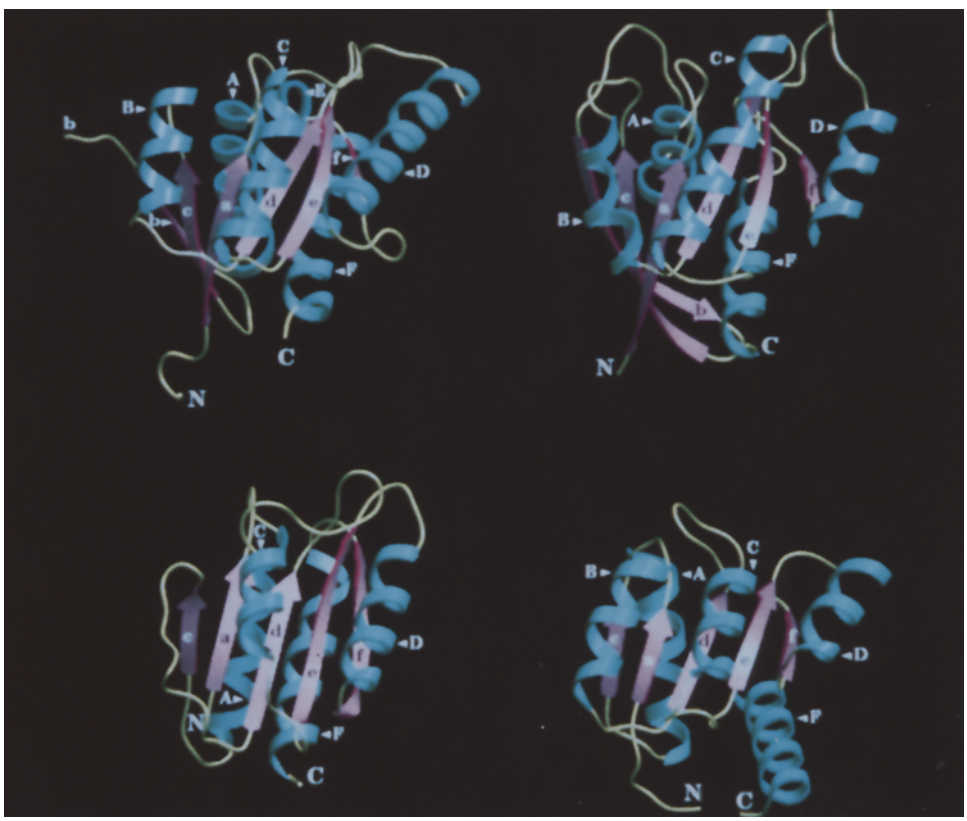

Fig. 9. Similarity of fold of the GTP-binding domain of elongation factor Tu [102] (upper left), p2l ras protein [101] (upper right), the flavodoxin family [103] (lower left) and chemotaxis Y protein [100] (lower right). To highlight the similarity between these two protein families, β-strands (a, b, c, d, e, f) and α-helices (A, B, C, D) are labelled according to their sequential order in elongation factor Tu. A break in the main chain of elongation factor Tu is indicated (b).

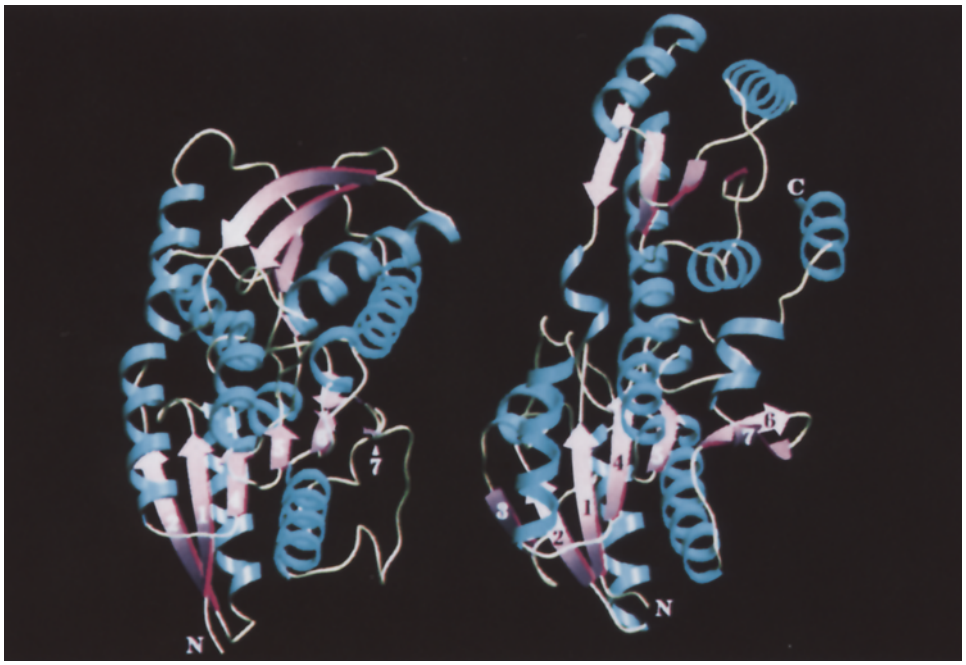

Fig. 10. Similarity of fold of the N-terminal domains of the type I periplasmic binding protein and the phosphofructokinase families. D-galactose/D-glucose-binding protein (2gbp [86], left) and phosphofructokinase (1pfk [104], right) are shown. To highlight the similarity between the topology of the N-terminal domains of these two protein families, β-strands are labelled 1 to 7, according to their sequential order in the N-terminal domain of phosphofructokinase.

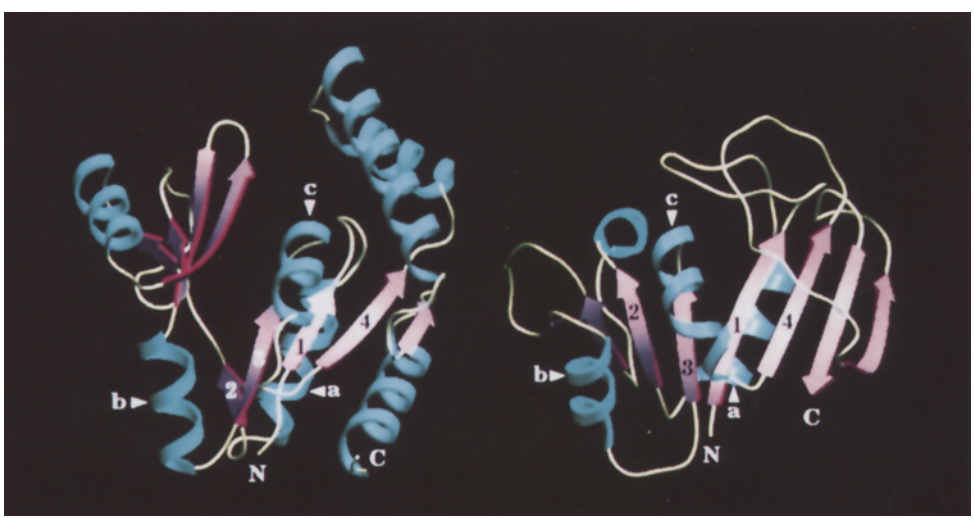

Fig. 11. Similarity of fold of the nucleotide kinase and the dihydrofolate reductase families. Guanylate kinase (1gky [106], left) and dihydrofolate reductase (4dfr [107], right) are shown. To highlight the similarity between these two protein families, β-strands (1 to 4) and α-helices (a, b, c) common to both protein families are sequentially labelled.

cognition of the common fold. This has been approached by 'threading' sequences onto a known structure [111], by comparing profiles [112] and by searching with tertiary templates derived from environment-dependent amino acid substitution tables [6–9]. Such approaches consider the local structural constraints on the acceptance of amino acid substitutions. In the approach of Johnson et al. [6] comparison of 3D structures is used to define topological equivalence, and the probabilities of substitutions are systematically evaluated by consideration of many families of proteins, each adopting a common fold. In our earlier analysis [7–9] we have increased the number of families to ~100 but have restricted the analysis to members with significant (> 20%) pairwise sequence similarities. The new clustering and alignment procedures now allow us to identify proteins that are more distantly related than was possible with COMPARER [19], although once identified, the alignments produced by our new approach can be refined using local amino acid properties and relationships rather than those of secondary structures. The comparison of more distantly related protein structures allows the development of amino acid substitution tables [113] that are more suited to the recognition of distantly related folds.

One of the most interesting contributions to drug discovery in the past decade has been the recognition of the common fold adopted by the dimeric retroviral proteinases and the monomeric pepsin-like aspartic proteinases. This allowed the chemistry of design, developed for the anti-hypertensive inhibitors of renin, to be transferred to the design of HIV proteinase inhibitors that might be useful AIDS antivirals. The clustering of structures described here correctly identifies this structural homology and allows binding sites in one protein, such as renin, to be used as the basis of hypothesis for a second, such as HIV proteinase.

Surprisingly many of the clustered proteins have similar functions. For example, the sulphate and phosphate binding proteins (coordinates not available in the data bank) have a structure similar to that of porphobilinogen deaminase and both bind anions at similar positions between two equivalent domains. Even transferrins, further members of the same family, have anion (carbonate) binding which is obligatory for iron binding. Such analogies are very suggestive in modelling and design.

A further example highlighted here is the structural similarity between the pentraxins, i.e. mammalian lectins, and the legume lectins, such as concanavalin A. In this case sugar binding is mediated by calcium and manganese ion binding in the legume lectins and two calcium ions in the pentraxins. The binding occurs on the same face of the common fold, but in different positions. Again, understanding such subtle binding relationships has much to teach us about using protein structures in drug discovery.

ACKNOWLEDGEMENTS

We thank Dr. Gordon Louie for helpful discussions on porphobilinogen deaminase and Drs. Helen White, N. Srinivasan, Steve Wood and Mr. Jonas Emsley for advice on the pentraxins. We thank Dr. R. Sowdhamini, Mr. Zhan-Yang Zhu and Dr. Mark Johnson for help on computational approaches to the comparison of protein structures. We would also like to thank Dr. Judith Murray-Rust for her help. The figures in this paper were generated with the program SETOR [114], written by S. Evans.

REFERENCES

1 Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C. and Hill, R.L., J. Mol. Biol., 42 (1969) 65.
2 Greer, J., J. Mol. Biol., 153 (1981) 1027.
3 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M., Nature, 326 (1987) 347.
4 Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L., Protein Eng., 1 (1987) 377.
5 Sutcliffe, M.J., Haneef, I. and Blundell, T.L., Protein Eng., 1 (1987) 385.
6 Johnson, M.S., Overington, J.P. and Blundell, T.L., J. Mol. Biol., 231 (1993) 735.
7 Overington, J.P., Johnson, M.S., Šali, A. and Blundell, T.L., Proc. R. Soc. London, Ser. B, 241 (1990) 132.
8 Overington, J.P., Donnelly, D., Johnson, M.S., Šali, A. and Blundell, T.L., Protein Sci., 1 (1992) 216.
9 Overington, J.P., Zhu, Z.-Y., Šali, A., Johnson, M.S., Sowdhamini, R., Louie, G.V. and Blundell, T.L., Biochem. Soc. Trans., 21 (1993) 597.
10 McLachlan, A.D., J. Mol. Biol., 128 (1979) 49.
11 Matthews, B.W. and Rossmann, M.G., Methods Enzymol., 115 (1985) 397.
12 Rossmann, M.G. and Argos, P., J. Mol. Biol., 105 (1976) 75.
13 Rossmann, M.G. and Argos, P., J. Mol. Biol., 109 (1977) 99.
14 Remington, S.J. and Matthews, B.W., Proc. Natl. Acad. Sci. USA, 75 (1978) 2180.
15 Remington, S.J. and Matthews, B.W., J. Mol. Biol., 140 (1980) 77.
16 Needleman, S.B. and Wunsch, C.D., J. Mol. Biol., 48 (1970) 443.
17 Smith, T.F. and Waterman, M.S., J. Mol. Biol., 147 (1981) 195.
18 Argos, P., Vingron, M. and Vogt, G., Protein Eng., 4 (1991) 375.
19 Šali, A. and Blundell, T.L., J. Mol. Biol., 212 (1990) 403.
20 Zhu, Z.-Y., Šali, A. and Blundell, T.L., Protein Eng., 5 (1992) 43.
21 Taylor, W.R. and Orengo, C.A., Protein Eng., 2 (1989) 505.
22 Taylor, W.R. and Orengo, C.A., J. Mol. Biol., 208 (1989) 1.
23 Orengo, C.A. and Taylor, W.R., J. Theor. Biol., 147 (1990) 517.
24 Lesk, A.M. and Chothia, C., J. Mol. Biol., 136 (1980) 225.
25 Tramontano, A., Chothia, C. and Lesk, A.M., Protein Struct. Funct. Genet., 6 (1989) 382.
26 Chothia, C., Levitt, M. and Richardson, D., J. Mol. Biol., 105 (1977) 1.
27 Subbiah, S., Laurents, D.V. and Levitt, M., Curr. Biol., 3 (1993) 1441.
28 Vriend, G. and Sander, C., Protein Struct. Funct. Genet., 11 (1991) 52.
29 Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G., Protein Sci., 1 (1992) 1691.
30 Yee, D.P. and Dill, K.A., Protein Sci., 2 (1993) 884.
31 Holm, L. and Sander, C., J. Mol. Biol., 233 (1993) 123.
32 Holm, L. and Sander, C., FEBS Lett., 315 (1993) 301.
33 Holm, L. and Sander, C., Nature, 361 (1993) 309.
34 Lesk, A.M. and Chothia, C., J. Mol. Biol., 160 (1982) 325.
35 Chothia, C. and Lesk, A.M., J. Mol. Biol., 160 (1982) 309.
36 Murthy, M.R.N., FEBS Lett., 168 (1984) 97.
37 Richards, F.M. and Kundrot, C.E., Protein Struct. Funct. Genet., 3 (1988) 71.
38 Mitchell, E.M., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Mol. Biol., 212 (1989) 151.
39 Artymiuk, P.J., Rice, D.W., Mitchell, E.M. and Willett, P., Protein Eng., 4 (1989) 39.
40 Artymiuk, P.J., Grindley, H.M., Park, J.E., Rice, D.W. and Willett, P., FEBS Lett., 303 (1992) 48.
41 Grindley, H.M., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Mol. Biol., 229 (1993) 707.
42 Orengo, C.A., Brown, N.P. and Taylor, W.R., Protein Struct. Funct. Genet., 14 (1992) 139.
43 Orengo, C.A., Flores, T.P., Jones, D.T., Taylor, W.R. and Thornton, J.M., Curr. Biol., 3 (1993) 131.
44 Orengo, C.A., Flores, T.P., Taylor, W.R. and Thornton, J.M., Protein Eng., 6 (1993) 485.
45 Koch, I., Kaden, F. and Selbig, J., Protein Struct. Funct. Genet., 12 (1992) 314.
46 Johnson, M.S., Sutcliffe, M.J. and Blundell, T.L., J. Mol. Evol., 30 (1990) 43.
47 Johnson, M.S., Šali, A. and Blundell, T.L., Methods Enzymol., 183 (1990) 670.
48 Kabsch, W. and Sander, C., Biopolymers, 22 (1983) 2577.
49 Smith, D.K. and Thornton, J.M., unpublished results.

50 Chou, K.-C., Nemethy, G. and Scherega, H.A., J. Am. Chem. Soc., 106 (1984) 3161.
51 Sowdhamini, R., Srinivasan, N., Ramakrishnan, C. and Balaram, P., J. Mol. Biol., 223 (1992) 845.
52 Oobatake, M. and Ooi, T., J. Theor. Biol., 67 (1977) 567.
53 Bron, C. and Kerbosch, J., Commun. Assoc. Comput. Machinery, 16 (1973) 575.
54 Fredman, M.L., Bull. Math. Biol., 46 (1984) 553.
55 Felsenstein, J., Evolution, 39 (1985) 783.
56 Zhu, Z.-Y., unpublished results.
57 Fitch, W.M. and Margoliash, E., Science, 155 (1967) 279.
58 Andreeva, N.S., Fedorov, A.A., Gustchina, A.E., Schutzkever, N.E. and Safro, M.G., Mol. Biol. (Moscow), 12 (1978) 704.
59 Cooper, J.B., Khan, G., Taylor, G., Tickle, I.J. and Blundell, T.L., J. Mol. Biol., 214 (1990) 199.
60 Abad-Zapatero, C., Rydel, T.J. and Erickson, J., Protein Struct. Funct. Genet., 8 (1990) 62.
61 Sielecki, A.R., Hayakawa, K., Fujinaga, M., Murphy, M.E.P., Fraser, M., Muir, A.K., Carilli, C.T., Lewicki, J.A., Baxter, J.D. and James, M.N.G., Science, 243 (1989) 1346.
62 Dhanaraj, V., Dealwis, C.G., Frazao, C., Badasso, M., Sibanda, B.L., Tickle, I.J., Cooper, J.B., Driessen, H.P.C., Newman, M., Aguilar, C., Wood, S.P., Blundell, T.L., Hobart, P.M., Geoghegan, K.F., Ammirati, M.J., Danley, D.E., O'Connor, B.A. and Hoover, D.J., Nature, 357 (1992) 466.
63 Gilliland, G.L., Winborne, E.L., Nachman, J. and Wlodawer, A., Protein Struct. Funct. Genet., 8 (1990) 82.
64 Newman, M., Safro, M., Frazao, C., Khan, G., Zdanov, A., Tickle, I.J., Blundell, T.L. and Andreeva, N., J. Mol. Biol., 221 (1991) 1295.
65 Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S.P., Tickle, I.J. and Blundell, T.L., J. Mol. Biol., 230 (1993) 260.
66 Blundell, T.L., Jenkins, J.A., Sewell, B.T., Pearl, L.H., Cooper, J.B., Tickle, I.J., Veerapandian, B. and Wood, S.P., J. Mol. Biol., 211 (1990) 919.
67 James, M.N.G. and Sielecki, A.R., In Jurnak, F. and McPherson, A. (Eds.) Biological Macromolecules and Assemblies, Wiley, New York, NY, 1983, pp. 43–60.
68 Suguna, K., Padlan, E.A., Smith, C.W., Carlson, W.D. and Davies, D.R., Proc. Natl. Acad. Sci. USA, 34 (1987) 7009.
69 Aguilar, C., Badasso, M., Cooper, J.B., Wood, S.P. and Blundell, T.L., in preparation.
70 Fitzgerald, P.M.D., McKeever, B.M., Van Middlesworth, J.F., Springer, J.P., Heimbach, J.C., Leu, C.-T., Herber, W.K., Dixon, R.A.F. and Darke, P.L., J. Biol. Chem., 265 (1990) 14209.
71 Pearl, L.H. and Taylor, W.R., Nature, 329 (1987) 351.
72 Miller, M., Jaskolski, M., Rao, J.K.M., Leis, J. and Wlodawer, A., Nature, 337 (1989) 576.
73 Jaskolski, M., Miller, M., Rao, J.K.M., Leis, J. and Wlodawer, A., Biochemistry, 29 (1990) 5889.
74 Navia, M.A., Fitzgerald, P.M.D., McKeever, B.M., Leu, C.-T., Himbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L. and Springer, J.P., Nature, 337 (1989) 615.
75 Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J. and Kent, S.B.H., Science, 245 (1989) 616.
76 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., Lees, S.E., Scheld, K.G. and Hobart, P.M., Nature, 342 (1989) 299.
77 Olendorf, D.H., Foundling, S.I., Wendoloski, J.J., Sedlacek, J., Strop, P. and Salemme, F.R., Protein Struct. Funct. Genet., 14 (1992) 382.
78 Louie, G.V., Brownlie, P.D., Lambert, R., Cooper, J.B., Blundell, T.L., Wood, S.P., Warren, M.J., Woodcock, S.C. and Jordan, P.M., Nature, 359 (1992) 33.
79 Baker, E.N., Rumball, S.V. and Anderson, B.F., Trends Biochem. Sci., 12 (1987) 350.
80 Anderson, B.F., Baker, H.M., Norris, G.E., Rice, D.W. and Baker, E.N., J. Mol. Biol., 209 (1989) 711.
81 Sarra, R., Garratt, R., Gorinsky, B., Jhoti, H. and Lindley, P., Acta Crystallogr., B46 (1991) 763.
82 Spurlino, J., Lu, G.-Y. and Quiocho, F.A., J. Biol. Chem., 266 (1991) 5202.
83 Sack, J.S., Trakhanov, S.D., Tsigannik, I.H. and Quiocho, F.A., J. Mol. Biol., 206 (1989) 193.
84 Sack, J.S., Saper, M.A. and Quiocho, F.A., J. Mol. Biol., 206 (1989) 171.
85 Quiocho, F.A. and Vyas, N.K., Nature, 310 (1984) 381.
86 Vyas, N.K., Vyas, M.N. and Quiocho, F.A., Science, 242 (1988) 1290.
87 Mowbray, S.L. and Cole, L.B., J. Mol. Biol., 225 (1992) 155.

88 Emsley, J., White, H.E., O'Hara, B.P., Oliva, G., Srinivasan, N., Tickle, I.J., Blundell, T.L., Pepys, M.B. and Wood, S.P., Nature, (1994) in press.
89 Einspar, H., Parks, E.H., Suguna, K., Subramanian, E. and Suddath, F.L., J. Biol. Chem., 261 (1986) 16518.
90 Hardman, K.D. and Ainsworth, C.F., Biochemistry, 11 (1972) 4910.
91 Keitel, T., Simon, O., Borriss, R. and Heinemann, U., Proc. Natl. Acad. Sci. USA, 90 (1993) 5287.
92 Srinivasan, N., White, H.E. and Blundell, T.L., in preparation.
93 Meyer, E., Cole, G., Radhakrishnan, R. and Epp, O., Acta Crystallogr., B44 (1988) 26.
94 Moult, J., Sussman, F. and James, M.N.G., J. Mol. Biol., 182 (1985) 555.
95 Van Roey, P. and Beerman, T.A., Proc. Natl. Acad. Sci. USA, 86 (1989) 6587.
96 Pletnev, V.Z., Kuzin, A.P. and Malinina, L.V., Bioorg. Khim., 8 (1982) 1637.
97 Suh, S.W., Bath, M.A., Naiva, G.H., Cohen, G.H., Rao, D.N., Rudikoff, S. and Davies, D.R., Protein Struct. Funct. Genet., 1 (1986) 74.
98 Skarzynski, T., Moody, P.C.E. and Wonacott, A.J., J. Mol. Biol., 193 (1987) 171.
99 Hall, M.D., Levitt, D.G. and Banaszak, L.J., J. Mol. Biol., 226 (1992) 867.
100 a. Volz, K. and Matsumura, P., J. Biol. Chem., 266 (1991) 15511.
b. Stock, A., Mottonen, J.M., Stock, J. and Schutt, C.E., Nature, 344 (1989) 745.
101 Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W. and Wittinghofer, A., EMBO J., 9 (1990) 2351.
102 la Cour, T.F.M., Nyborg, J., Thirup, S. and Clark, B.F.C., EMBO J., 4 (1985) 2385.
103 Smith, W.W., Burnet, R.M., Darling, G.D. and Ludwig, M.L., J. Mol. Biol., 117 (1977) 195.
104 Shirakiahara, Y. and Evans, P.R., J. Mol. Biol., 204 (1988) 973.
105 Evans, P.R., Farrants, G.W. and Hudson, P.J., Phil. Trans. R. Soc. London, Ser. B., 53 (1981) 53.
106 Stehle, T. and Schulz, G.E., J. Mol. Biol., 224 (1992) 1127.
107 Bohn, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13560.
108 Srinivasan, N. and Blundell, T.L., Protein Eng., 6 (1993) 501.
109 Šali, A. and Blundell, T.L., J. Mol. Biol., 234 (1993) 779.
110 Šali, A., Overington, J.P., Johnson, M.S. and Blundell, T.L., Trends Biochem. Sci., 15 (1990) 235.
111 Jones, D.T., Taylor, W.R. and Thornton, J.M., Nature, 358 (1992) 86.
112 Bowie, J.U., Lüthy, R. and Eisenberg, D., Science, 253 (1991) 164.
113 Sowdhamini, R. and Rufino, S.D., in preparation.
114 Evans, S.V., J. Mol. Graphics, 11 (1993) 134.