Molecular Recognition

Host/Hormone Group/Nucleic Acids and Molecular Biology Group Colloquium Organized by S. E. V. Phillips (University of Leeds). 646th Meeting held at the University of Leeds, 30 March–2 April 1993.

Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins

John P. Overington, Zhan-Yang Zhu, Andrej Šali, Mark S. Johnson, Ramanathan Sowdhamini, Gordon V. Louie and Tom L. Blundell*

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

Introduction

Much of our present knowledge of protein three-dimensional structure and function has been derived from the study of single proteins. However, the comparison of the structures of a protein family can provide valuable insights that cannot be obtained from the study of isolated structures. For example, multiple, homologous sequences can greatly enhance the sensitivity and accuracy of sequence searching and alignment techniques [1]. Comparison of many sequences can improve the accuracy of secondary-structure predictions [2] and comparison of three-dimensional structures provides the basis of many comparative-modelling approaches [3–5]. Multiple sequences and three-dimensional structures of homologues can be used to identify invariant features and to reveal functionally important sites [6]. The study of a family of protein structures can also reveal differences between the amino acids in the binding sites that lead to subtle differences in the recognition of substrates, activators and other ligands. Thus, the comparison of related proteins can be very useful in understanding molecular recognition.

The possibilities for comparing related structures increase as the Brookhaven protein databank (PDB) [7] grows in size. The large size of the databank (> 260 MB) and its inherent complexity have led to attempts to process, organize and query the available data; examples are the BIPED and STEP databases of Birkbeck and University College [8, 9] and similar databases of others [10]. These databases generally include data such as main-chain and side-chain torsion angles, hydrogen-bonding interactions, and residue-solvent accessibility that are not directly recorded with the original atomic co-ordinates. Such databases allow the flexible and rapid search for particular structural features that may be of use in studies of molecular recognition.

An important prerequisite of a comparative analysis of three-dimensional structures is the development of powerful and robust methods for protein-structure comparison. Methods that depend on a 'rigid-body' fitting of two structures [4, 11] suffer from the major disadvantage that the fraction of residues in the conserved core rapidly falls as the sequence similarity between the compared structures decreases (as they divergently evolve) [12]. Thus, a strict upper bound on the distance between C_α atoms in the definition of equivalence can leave large regions of the structures unaligned, even though topological equivalence exists. Furthermore, such methods are especially sensitive to rigid-body shifts in sub-structures (for example in allosteric changes or in domain motion). Consecutive fitting of smaller fragments of the compared structures [13] goes some way towards overcoming such problems, but is sensitive to insertions and deletions in one structure relative to another. Although a rigid three-dimensional structure is not conserved between pairs of distantly related proteins, the general organization and the relationships of the elements that make up the fold tend to be maintained [14, 15]. The conserved features of this organization include sidechain packing between secondary structural elements [16] and main-chain hydrogen-bonding patterns [17]. Considerations of these factors may be important in comparisons of very distantly related structures [18, 19].

Comparison methods have been developed that attempt to overcome this rigid-body problem. Taylor and Orengo [20, 21] align protein structures

Abbreviations used: PBGD, porphobilinogen deaminase; PDB, Brookhaven protein databank.

*To whom correspondence should be addressed.

by comparison of patterns of inter-atomic vectors and of other local structural features, such as solvent accessibility and main-chain torsion angles. Šali, Zhu and Blundell, in the program COMPARER [22, 23], include information from all levels of protein organization, to extend the power of the comparison procedures.

Methods of alignment and of comparison of protein sequences and structures have been used by several laboratories to cluster known three-dimensional structures together with each other, and together with the sequences of proteins with a common fold. We have produced environment-dependent amino-acid substitution tables [24, 25] and have used these to create tertiary templates or profiles for each family [26]. We initially concentrated in homologous protein families where all members have very similar topologies, with the majority of sequences having 15–40% sequence identity with each other. Sander and Schneider [5] have quantitatively related sequence similarity, structural similarity and alignment length, using a database of known structures, but have extended the number of proteins that are compared by aligning sequences from proteins that are clearly homologous. Based on an inspection of a structural similarity/sequence similarity scatter plot, a threshold as a function of length has been used as a criterion to identify sequences that have a common fold. Pascarella and Argos [27] have performed rigid-body superpositions between the members of 38 protein families, and have associated these and 45 other individual structures with the primary-structure database. Orengo et al. [28] have extended their original comparisons to more distantly related proteins, by comparing elements of secondary structure rather than individual amino-acid residues, and have clustered all protein structures in the form of a dendrogram.

In this paper we describe an expanded database of related three-dimensional structures that have been aligned using COMPARER. The number of families is now 87 and the database includes several families where the average sequence identity is less than 20%. We discuss examples where comparisons of such families of proteins can be useful in understanding molecular recognition, either to define the probable nature of the ligand or to understand its detailed specificity.

Materials and methods

Selection and alignment of protein families

The co-ordinates for the three-dimensional structures were obtained from the February 1993 release of the PDB. Members of families were defined from previous studies within the laboratory or from the work of others (Table 1). 'Bad' or poorly refined structures were identified on the basis of an analysis performed by Morris et al. [29], and were excluded from the analysis. In general we have included into the database only families in which the alignment is relatively unambiguous and where the proteins are probably truly homologous.

Structural parameters used in comparison

Solvent accessibility

Relative sidechain accessibilities were calculated with a program implementing the method of Lee and Richards [30]. A 7% relative-accessibility cut-off was used to define inaccessible residues [31]; in all cases the Cα atom was considered as part of the sidechain. Essential prosthetic atoms, such as the copper atom at the active site of the azurins and the haem groups in the cytochromes and in the globins, were included in the accessibility calculations, since their presence is often essential to the integrity of the fold.

Secondary structure and main-chain conformation

The secondary structural class of a residue was established on the basis of main-chain hydrogen-bonding patterns, using the SSTRUC program written by D. Smith and J. M. Thornton (personal communication). This program implements the algorithm of Kabsch and Sander [32] to define regions of α-helix and of β-strand.

Sidechain van der Waals contacts and hydrogen bondings

Contacts between sidechains were initially defined on the basis of a 4 Å distance cut-off. These contacts were considered to be either van der Waals interactions or hydrogen bonds, depending on the nature of the contacting atoms. Hydrogen bonds were identified using a simple distance-based cut-off of 3.5 Å (4.0 Å for interactions with sulphur atoms) between potential donor and acceptor atoms; we felt that the low resolution, the incomplete refinement of several structures and the difficulty in placing sidechain atoms precluded the application of a more restrictive geometrical definition. For asparagine and glutamine, the identities of the nitrogen and oxygen atoms of the sidechain amide were treated as unknown, as they cannot be directly distinguished by X-ray analysis. The sulphur atoms in methionine and cysteine residues were considered to be potential hydrogen bond donors and/or acceptors [33]. The protonation

Table 1

Protein families in the alignment database

Name of family	$N_{str.}$*	$N_{av.}$†	$ID_{av.}$(%)‡
Small			
Zinc finger (CCHC-type)	2	17	47.06
Zinc finger (CCHH)-type	4	28	32.72
Metallothionein (α-domain)	3	31	93.55
Metallothionein (β-domain)	3	30	83.33
Pancreatic polypeptide	2	36	41.67
E3-binding domain dihydrolipoamide acetyltransferase	2	35	33.33
Rubredoxin	5	51	63.00
Protein G domain	2	63	87.50
Serine-proteinase inhibitor (potato I-type)	2	64	35.48
Ferredoxin (4Fe–4S)	3	72	33.05
Ferredoxin (2Fe–2S)	3	97	73.41
Small — disulphide			
Steroid-binding protein	2	73	55.71
High potential iron protein	2	78	23.19
Serine proteinase inhibitor (squash-type)	2	28	71.43
EGF-like domain	3	47	31.85
Sea-anemone toxin	2	45	27.03
Serine proteinase inhibitor (Bowman–Birk-type)	3	56	74.89
Insulin	3	50	52.16
Serine-proteinase inhibitor (Kazal-type)	5	55	44.26
Serine-proteinase inhibitor (Kunitz-type)	3	57	38.47
Snake toxin	7	64	44.11
Kringle domain	3	86	41.31
All-α			
DNA-binding homeodomain	2	62	50.88
DNA-binding repressor	3	71	32.59
Cytochrome c_5	2	82	25.00
Cytochrome *b*	2	88	29.41
Calcium-binding protein (parvalbumin-like)	4	107	52.10
Cytochrome *c*	7	111	44.56
Cytochrome c_3	2	112	35.42
Haemerythrin	2	115	46.02
Phospholipase A_2	5	122	47.80
Cytochrome *c′*	2	129	21.60
Globin	16	146	27.12
Calcium-binding protein (calmodulin-like)	5	162	33.44
Fe/Mn superoxide dismutase	2	192	36.41
Glutathione S-transferase	2	208	83.57
Membrane-bound all-α			
Photosynthetic reaction centre	2	826	48.49
$\alpha+\beta$			
Lysozyme	4	127	53.44

Table 1 – *continued*

Name of family	$N_{str.}$*	$N_{av.}$†	$ID_{av.}$(%)‡
Class I histocompatibility antigen-binding domain	4	178	79.49
Cysteine proteinase	3	215	55.35
Carbonic anhydrase	2	256	61.59
β-Lactamase	2	256	43.14
Zinc metalloproteinase	3	310	44.73
Actin/heat-shock cognate	2	377	14.10
Isocitrate dehydrogenase	2	379	28.23
Serine-proteinase inhibitor (serpin-type)	3	380	28.75
Aspartate aminotransferase	2	398	40.40
Disulphide oxidoreductase	5	466	29.96
α/β			
Thioredoxin	4	96	14.53
RNAase H	2	138	25.86
Flavodoxin	5	159	33.16
GTP-binding protein	2	171	15.13
Dihydrofolate reductase	4	172	35.96
Nucleotide kinase	4	202	24.95
Subtilase	7	274	52.10
Thymidylate synthase	2	290	59.85
Periplasmic binding protein (sugar)	3	295	21.33
PBGD/transferrin	4	290	31.47
Phosphofructokinase	2	319	55.35
Lactate/malate dehydrogenase	9	321	33.22
Glyceraldehyde phosphate dehydrogenase	4	339	56.64
Periplasmic binding protein (amino acid)	2	345	79.07
Cholesterol oxidase	2	541	16.40
α/β-barrel			
Tryptophan biosynthesis enzyme	2	226	10.22
Triose-phosphate isomerase	3	247	50.54
Fructose-1,6-biphosphatase aldolase	2	361	70.56
Flavin-binding β-barrel	2	376	41.67
Xylose isomerase	3	390	66.95
Ribulose-1,5-biphosphate carboxylase/oxygenase	3	537	49.57
All-β			
Immunoglobulin (constant domain)	11	98	34.31
Immunoglobulin (cell surface)	4	102	22.97
Retroviral aspartic proteinase	2	107	27.96
Antibacterial protein	2	110	36.11
Azurin/plastocyanin	7	110	32.94
Immunoglobulin (variable domain)	41	118	38.98
Interleukin 1β	3	142	32.67
Cu/Zn superoxide dismutase	2	152	54.67
Glucose permease	2	154	42.28
Lipocalin	6	149	17.76
Plant-virus coat protein	2	186	23.64

Table 1 – *continued*

Name of family	$N_{str.}$*	$N_{av.}$†	$ID_{av.}$(%)‡
Crystallin	4	175	57.96
Serine proteinase (bacterial)	3	188	45.36
Serine proteinase (mammalian)	10	229	36.19
Plant lectin	3	234	43.26
Pepsin-like aspartic proteinase	9	330	34.71
Neuraminidase	3	389	35.68
Picornavirus coat proteins	6	780	33.02

*$N_{str.}$, Number of proteins in the family.
†$N_{av.}$, Average number of residues.
‡$ID_{av.}$, Average sequence identity.

states of histidine, aspartic-acid and glutamic-acid residues were also treated as undefined by the crystal-structure determination.

Due to the highly specific nature of hydrogen bonding, we considered a number of further subclasses involving sidechain functions: (1) those from a main-chain amide, (2) those to a main-chain carbonyl, and (3) those to/from another sidechain or to a non-amino-acid group (excluding water). Interactions with non-amino-acid groupings (for example, with the haem in the globins and in the cytochromes, or with the iron–sulphur clusters in the ferredoxins) were also included in this class.

Results

Alignment database

The alignment database contains 347 protein chains organized in 87 different families (Table 1). One such alignment is shown in Figure 1. The database, which is available from the authors on request, is around 51.5 MB in size, including all the structural data. The database contains fitted three-dimensional co-ordinates for each family; these are useful in homology-modelling studies [4, 34]. The alignments have been used, along with the substitution data [24, 25], to create tertiary templates or profiles for each family [26].

Because most of the protein families included in the database are reasonably similar in sequence (Table 1), the structurally conserved cores are extensive for most families. As a consequence, the accuracy of most of the alignments in Table 1 is likely to be high. Our methods reproduce well the 'expert' alignments that are produced by careful manual analysis of a particular protein family, for example, the globins [16]. This accuracy is generally maintained to around the 20% global-sequence-identity level; below this there are often small differences, particularly for the loop regions. Further discussion of the accuracy of the alignments can be found in Zhu, Šali and Blundell [23].

Some of the families that are detailed in Table 1 are themselves related to other families; for example, the retroviral and pepsin-like aspartic proteinases. Additionally, some families contain clear examples of gene-duplication in their evolution, for example the aspartic proteinases [35] and the crystallins [36]. Although we have used COMPARER to align these distantly related structural motifs or domains, we have omitted them from the current database, as the alignments can be less reliable than those that are included here. To extend the alignments to more distantly related proteins, comparison is most usefully carried out at the level of domains. We have now systematically identified domains in proteins that are useful for an analysis of distant relationships between proteins (R. Sowdhamini and T. L. Blundell, unpublished work) and are preparing a database of alignments for these domains.

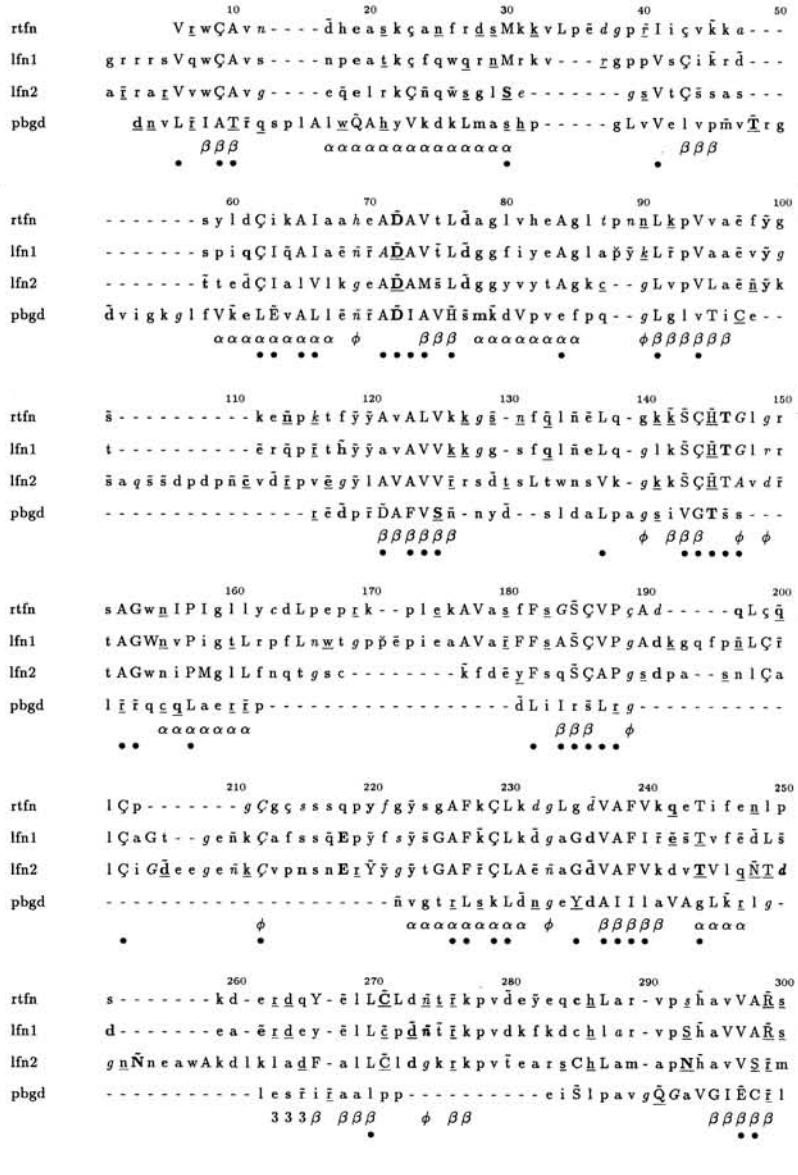
Molecular recognition

The database contains a wealth of information about families of proteins that bind closely related ligands. These include several families of proteinases that recognize polypeptides with different sequences. For example, aspartic-proteinase-inhibitor complexes have been reported for pepsin, endothiapepsin, rhizopuspepsin, penicillopepsin and renin [37 and references therein], but not for chymosin or cathepsin D. The identification of the residues in the specificity pockets of some members of the family, together with the alignments based on the structure in the database, allow a fast assessment of the residues in the binding pockets of

Figure 1

COMPARER [22, 23] alignment of rat transferrin (rtfn), N-terminal lobe of lactoferrin (lfn 1), C-terminal lobe of lactoferrin (lfn 2) and PBDG (pbgd) sequences

The PDB co-ordinate entries for the transferrin, lactoferrin and PDBG are 1tfd, 1lfg and 1pda respectively. The alignment is represented by JOY [24, 25].



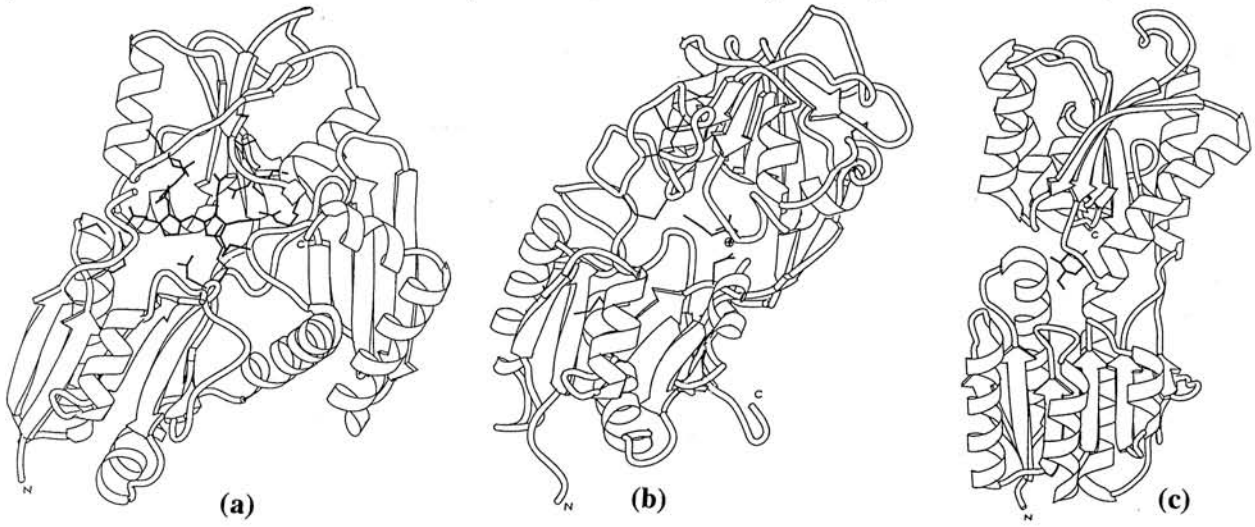
Key to JŏY alignments

solvent inaccessible	UPPER CASE	X
solvent accessible	lower case	x
positive ϕ	*italic*	*x*
cis-peptide	breve	x̆
hydrogen bond to other sidechain	tilde	x̃
hydrogen bond to mainchain amide	**bold**	**x**
hydrogen bond to mainchain carbonyl	underline	x
disulphide bond	cedilla	x̧

Figure 2

Comparison of the folds and of the binding sites of (*a*) PBGD (PDB code, 1pda), (*b*) transferrin (PDB code, 1tfd) and (*c*) maltose-binding protein (PBD code, 1omp) shown in the form of ribbon drawings

Helices are represented as ribbons and β-strand elements as broad arrows, while coil regions are indicated as thin ropes. The three molecules are shown with their domain 1 in the same orientation. Ligand molecules (dipyrromethane cofactor in PBGD, iron and carbonate ions in transferrin, and maltose in maltose-binding protein) and sidechain atoms of residues in direct interaction with the ligand are shown as thick lines. The overall similarity in the folds, as well as in the ligand-binding sites, can be seen.



homologues. This can then be supported by modelling before experimental work; see for example the studies on chymosin [38] and on cathepsin D [39].

For more distantly related and possibly non-homologous proteins, the structural alignments may suggest recognition sites and even ligands. Thus, the unexpected topological similarity between sulphate- and phosphate-binding proteins [40], transferrins [41, 42] and porphobilinogen deaminase (PBGD) [43] reflects not only a similar binding mode, but also some similarities between their ligands. Figure 1 shows the alignment of the transferrins and PBGD on the basis of their three-dimensional structures [44]; the third domain, which covalently binds the dipyrrole primer in PBGD, has no equivalent in transferrin, and is omitted. The sulphate- and phosphate-binding proteins are omitted because co-ordinates are not available in the PDB. The sequences have no significant sequence similarity. Figure 2 shows the three-dimensional structures, and demonstrates the similar positions of the dipyrrole in the PBGD and in the iron complex in transferrin; an equivalent position is adopted by the ligands in the periplasmic anion-binding proteins. The similarity of folds correctly suggests a similarity of the binding position and a similar hinge-bending mechanism between the domains for ligand recognition and binding. The structures also suggest a similarity in the ligands themselves; indeed sulphate, phosphate and porphobilinogen are all anions. Although the iron of transferrin is positively charged, there is an obligatory binding of a carbonate in all transferrins. In all cases, the N-termini of the helices are close to the binding site, and at least one arginine is involved in binding; the alignment in Figure 1 shows that the arginines are not strictly topologically equivalent. As a caution against reading too much into the specific functionality of this bilobal structure, note that Figure 2 also includes maltose-binding protein; this binds maltose in an equivalent position, but, of course, does not involve an anionic ligand.

The value of a database of aligned protein structures is to suggest structural and functional roles by analogy. In each case these are useful to the extent that they can suggest experiments, but they are not useful as firm predictors. Nevertheless, a proper analysis of the structure and function of families of proteins will be of immense value to genome studies. Structural templates or profiles based on one or more members of the family [24–26, 45, 46] can be used to create tertiary templates or profiles for each family [26], and these can be used to identify a fold for a new protein sequence, which can be used to suggest functionality. An understanding of molecular recognition in families of proteins must be essential for the proper exploitation of genome studies of micro-organisms, plants, animals and man.

We acknowledge our colleagues, especially Chris Topham, for many useful discussions. We also thank Louise Morris and Janet Thornton for preprints of their work.

1. Taylor, W. R. (1986) J. Mol. Biol. **188**, 233–258
2. Niermann, T. and Kirschner, K. (1991) Protein Eng. **4**, 359–370
3. Greer, J. (1981) J. Mol. Biol. **153**, 1027–1042
4. Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L. (1987) Protein Eng. **1**, 377–384
5. Sander, C. and Schneider, R. (1991) Proteins: Struct. Funct. Genet. **9**, 56–68
6. Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987) J. Mol. Biol. **195**, 957–961
7. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. (1987) in Crystallographic Databases – Information, Content, Software Systems, Scientific Applications (Allen, F. H., Bergerhoff, G. and Sievers, R., eds.), pp. 107–132, Data Commission of the International Union of Crystallography, Cambridge
8. Akrigg, D., Bleasby, A. J., Dix, N. I. M., Findlay, J. B. C., North, A. C. T., Parrysmith, D., Wooton, J. C., Blundell, T. L., Gardner, S. P., Hayes, F., Islam, S., Sternberg, M. J. E., Thornton, J. M. and Tickle, I. J. (1988) Nature (London) **335**, 745–746
9. Islam, S. A. and Sternberg, M. J. E. (1989) Protein Eng. **2**, 431–442
10. Gray, P. M. D., Paton, N. W., Kemp, G. J. L. and Fothergill, J. E. (1990) Protein Eng. **3**, 235–244
11. Rossmann, M. G. and Argos, P. (1975) J. Biol. Chem. **250**, 7525–7532
12. Lesk, A. M. and Chothia, C. (1986) Philos. Trans. R. Soc. London B: **A317**, 345–356
13. Remington, S. J. and Matthews, B. W. (1980) J. Mol. Biol. **140**, 77–99
14. Rossmann, M. G. and Argos, P. (1977) J. Mol. Biol. **109**, 99–129
15. Bajaj, M. and Blundell, T. L. (1984) Annu. Rev. Biophys. Bioeng. **13**, 453–492
16. Lesk, A. M. and Chothia, C. (1980) J. Mol. Biol. **136**, 225–270
17. Rossmann, M. G., Moras, D. and Olsen, K. W. (1974) Nature (London) **250**, 194–199
18. Murthy, M. R. N. (1984) FEBS Lett. **168**, 97–102
19. Pastore, A., Lesk, A. M., Bolognesi, M. and Onesti, S. (1988) Proteins: Struct. Funct. Genet. **4**, 240–250
20. Taylor, W. R. and Orengo, C. A. (1989) Protein Eng. **2**, 505–519
21. Taylor, W. R. and Orengo, C. A. (1989) J. Mol. Biol. **208**, 1–22
22. Šali, A. and Blundell, T. L. (1990) J. Mol. Biol. **212**, 403–428
23. Zhu, Z.-Y. Šali, A. and Blundell, T. L. (1992) Protein Eng. **5**, 43–51
24. Overington, J. P., Johnson, M. S., Šali, A. and Blundell, T. L. (1990) Proc. R. Soc. London B **241**, 132–145
25. Overington, J. P., Donelly, D., Šali, A., Johnson, M. S. and Blundell, T. L. (1992) Protein Sci. **1**, 216–226
26. Johnson, M. S., Overington, J. P. and Blundell, T. L. (1993) J. Mol. Biol., in the press
27. Pascarella, S. and Argos, P. (1992) Protein Eng. **5**, 121–137
28. Orengo, C. A., Brown, N. P. and Taylor, W. R. (1992) Proteins: Struct. Funct. Genet. **14**, 139–167
29. Morris, A. L., MacArthur, M. W., Hutchinson, E. G. and Thornton, J. M. (1992) Proteins: Struct. Funct. Genet. **12**, 345–364
30. Lee, B. and Richards, F. M. (1971) J. Mol. Biol. **55**, 379–400
31. Hubbard, T. J. P. and Blundell, T. L. (1987) Protein Eng. **1**, 159–171
32. Kabsch, W. and Sander, C. (1983) Biopolymers **22**, 2577–2637
33. Gregoret, L. M., Rader, S. D., Fletterick, R. J. and Cohen, F. E. (1991) Proteins: Struct. Funct. Genet. **9**, 99–107
34. Sutcliffe, M. J., Hayes, F. R. F. and Blundell, T. L. (1987) Protein Eng. **1**, 385–392
35. Tang, J., James, M. N. G., Hsu, I.-N., Jenkins, J. A. and Blundell, T. L. (1978) Nature (London) **271**, 618–621
36. Bax, B., Lapatto, R., Nalini, V., Driessen, H., Lindley, P. F., Mahadevan, D., Blundell, T. L. and Slingsby, C. (1990) Nature (London) **347**, 776–780
37. 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. (1992) Nature (London) **357**, 466–472
38. Strop, P., Sedlacek, J., Stys, J., Kaderabkova, Z., Blaha, I., Pavlickova, L., Pohl, J., Fabry, M., Kostka, V., Newman, M., Frazao, C., Shearer, A., Tickle, I. J. and Blundell, T. L. (1990) Biochemistry **29**, 9863–9871
39. Scarborough, P. E., Guruprasad, K., Topham, C., Richo, G. R., Conner, G. E., Blundell, T. L. and Dunn, B. M. (1993) Protein Sci. **2**, 264–276
40. Quicho, F. A. (1991) Curr. Opin. Struct. Biol. **1**, 922–933
41. Baker, E. N. and Lindley, P. F. (1992) J. Inorg. Biochem. **47**, 147–160
42. Baker, E. N., Rumball, S. V. and Anderson, B. F. (1987) Trends Biochem. Sci. **12**, 350–353
43. Louie, G. V., Brownlie, P. D., Lambert, R., Cooper, J. B., Wood, S. P., Blundell, T. L., Warren, M. J., Woodcock, S. C. and Jordan, P. M. (1992) Nature (London) **359**, 33–39
44. Louie, G. V. (1993) Curr. Opin. Struc. Biol., in the press
45. Bowie, J. V., Luthy, R. and Eisenberg, D. (1991) Science **253**, 164–170
46. Luthy, R., Bowie, J. V. and Eisenberg, D. (1992) Nature (London) **356**, 83–85

Received 26 April 1993