J. Mol. Biol. (1994) **244**, 332–350

Structural Features can be Unconserved in Proteins with Similar Folds

An Analysis of Side-chain to Side-chain Contacts Secondary Structure and Accessibility

Robert B. Russell and Geoffrey J. Barton

University of Oxford
Laboratory of Molecular Biophysics, The Rex Richards Building
South Parks Road, Oxford OX1 3QU, U.K.

Side-chain to side-chain contacts, accessibility, secondary structure and RMS deviation were compared within 607 pairs of proteins having similar three-dimensional (3D) structures. Three types of protein 3D structural similarities were defined: type *A* having sequence and usually functional similarity; type *B* having functional, but no sequence similarity; and type *C* having only 3D structural similarity. Within proteins having little or no sequence similarity (types *B* and *C*), structural features frequently had a degree of conservation comparable to dissimilar 3D structures.

Despite similar protein folds, as few as 30% of residues within similar protein 3D structures can form a common core. RMS deviations on core C^{α} atoms can be as high as 3.2 Å. Similar protein structures can have secondary structure identities as low as 41%, which is equivalent to that expected by chance. By defining three categories of amino acid accessibility (buried, half buried and exposed), some similar protein 3D structures have as few as 30% of positions in the same category, making them indistinguishable from pairs of dissimilar protein structures. Similar structures can also have as few as 12% of common side-chain to side-chain contacts, and virtually no similar energetically favourable side-chain to side-chain interactions. Complementary changes are defined as structurally equivalent pairs of interacting residues in two structures with energetically favourable but different side-chain interactions. For many proteins with similar three-dimensional structures, the proportion of complementary changes is near to that expected by chance, suggesting that many similar structures have fundamentally different stabilising interactions.

All of the results suggest that proteins having similar 3D structures can have little in common apart from a scaffold of core secondary structures. This has profound implications for methods of protein fold detection, since many of the properties assumed to be conserved across similar protein 3D structures (e.g. accessibility, side-chain to side-chain contacts, etc.) are often unconserved within weakly similar (i.e. type *B* and *C*) protein 3D structures. Little difference was found between type *B* and *C* similarities suggesting that the structure of similar proteins can evolve beyond recognition even when function is conserved.

Our findings suggest that it is more general features of protein structure, such as the requirements for burial of hydrophobic residues and exposure of polar residues, rather than specific residue–residue interactions that determine how well a particular sequence adopts a particular fold. If detection of similar folds having little in common outside of their core secondary structures is to become a reality, efforts should concentrate on such general principles, and on methods for modelling large loop regions that are likely to differ between similar 3D structures.

Keywords: amino acid side-chain interactions; substitutions; protein evolution; protein structure; correlated mutations

Present address: R. B. Russell, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund Laboratories, Lincoln's Inn Fields, P.O. Box 123, London, WC2A 3PX.
Correspondence: G. J. Barton.

0022-2836/94/480372-19 $08.00/0

1. Introduction

Proteins having no detectable sequence similarity can adopt similar 3D† structures. Similarities are often observed for proteins having no functional similarity, or from different kingdoms or tissues (e.g. Holm & Sander, 1993*a*; Swindells *et al.*, 1993; Russell & Barton, 1993*a*). Despite the possibility for almost infinite variation at the level of the gene, Nature is apparently restricted to a limited number of protein folds.

Currently there are approximately 2000 proteins of known 3D structure, which can be classified further into approximately 150 unique fold families (Orengo *et al.*, 1993). A fold family is a collection of proteins having similar 3D structures, but not necessarily any sequence or functional similarity. Many families contain members with no common features across their sequences (for example, the α/β-barrel, greek key β-barrel and jelly-roll folds).

Here, to simplify discussion, we introduce a three state classification of protein 3D structural similarities. At one extreme (type *A*) are pairs of proteins sharing sequence, structural and (usually) functional similarity. Type *A* similarities include the globins, mammalian serine proteinases, Ig variable domains and cytochromes *c*. In the middle (type *B*) are those proteins having structural and functional similarity, but little sequence similarity, such as the mammalian and bacterial serine proteinases, azurin/plastocyanin, the Rossmann fold dehydrogenases (e.g. lactate, alcohol, etc.), Ig domains and CD4, aspartic proteinase lobes, rhodanese domains, and the heat shock protein/actin fold. Finally, at the other extreme (type *C*), are proteins with only 3D structural similarity, such as the Rossmann fold domains (e.g. lactate dehydrogenase and glycogen phosphorylase), α/β barrels, and greek key β barrels (e.g. Ig domains, azurin, superoxide dismutase, etc.). Families of types *B* and *C* often contain members with some structural differences, and with large insertions required to align structures accurately (e.g. haemocyanin compared to superoxide dismutase; or the α/β barrels from aldolase and rubico). Since functional similarity is difficult to define, the divisions between each type are not discrete, though the three categories provide a convenient means for classifying an observed structural similarity. The frequently used terms "homologous" and "analogous" probably define types *A* and *C*, respectively, with *B* falling somewhere in between. When comparing protein sequences or 3D structures, generally, type *A* and some type *B* similarities are detectable by sequence comparison methods (see Argos *et al.*, 1991 for a review) though many type *B* similarities are undetectable unless one considers 3D structure or functional information for one member of the family (e.g. Barton & Sternberg, 1990; Bowie *et al.*, 1991; Jones *et al.*, 1992). Structural similarities of type *C* are usually only detectable when both 3D structures are considered (Mitchell *et al.*, 1989; Taylor, 1989; Sali & Blundell, 1990; Russell & Barton, 1992), with some notable exceptions (Jones *et al.*, 1992; Godzik *et al.*, 1993). Protein structural families frequently contain similarities spanning types *A* through *C*. Figure 1 shows an example for the family of greek key β barrel structures. The figure shows three similar pairs: (a) two Ig light chain variable domains (*A*), which share functional and sequence similarity; (b) an Ig light chain variable domain and the N-terminal domain of CD4 (*B*), which are both immune system recognition proteins; and (c) an Ig light chain variable domain and poplar plastocyanin (C), which are similar only in that they have a similar arrangement of seven β strands.

Despite dozens of examples of similarities of types *B* and *C*, little is understood as to why different sequences can adopt similar 3D structures. Most studies to date have dealt with specific families of proteins having functional similarity (i.e. types *A* and *B*), such as the globins (Lesk & Chothia, 1980; Bashford *et al.*, 1987; Pastore *et al.*, 1988; Bordo & Argos, 1990, 1991), the Ig domains (Chothia & Lesk, 1982), blue copper (plastocyanin-like) photosynthetic proteins (Lesk & Chothia, 1982; Adman, 1984), nucleotide binding folds (Rossmann *et al.*, 1974; Rossmann & Argos, 1976; Otto *et al.*, 1980), oligonucleotide/oligosaccharide binding folds (Sixma *et al.*, 1993; Murzin, 1993), proteinases (Blundell *et al.*, 1979; Craik *et al.*, 1983) or α/β hydrolases (Ollis *et al.*, 1992). However, some studies have considered more distantly related protein 3D structures (i.e. type *C*), such as greek key β barrels (Hazes & Hol, 1992; Hutchinson & Thornton, 1992), globin/phycocyanin/colicin A (Pastore & Lesk, 1990; Holm & Sander, 1993*b*), α/β barrels (Farber & Petsko, 1990; Farber, 1993), β trefoils (Murzin *et al.*, 1992; Swindells & Thornton, 1993), toxin-agglutinin folds (Drenth *et al.*, 1980) or jelly-roll folds (Chelvanayagam *et al.*, 1992). Such studies generally suggest functional and packing features common to a particular family, though they provide few generalisations that might be applied to other protein structural families. Similarities of type *A* (and some of type *B*) have common features in the protein cores and around common binding or active sites. Similarities of type *C* (and some of type *B*) often have few common features. For example, even the most distantly related oxygen carrying globin folds (type *A* and *B* similarities) share haem binding residues as well as several key hydrophobic core residues (Bashford *et al.*, 1987; Pastore *et al.*, 1988). However, when one adds to the family the structurally similar, but functionally different, phycocyanin and colicin A structures, few common residues can be found (Pastore & Lesk, 1990; Holm & Sander, 1993*b*).

There have been some investigations into the general features of structurally similar proteins. Chothia & Lesk (1986) considered 32 pairs of structures and found that distantly related proteins could have as little as 50% of their structures in a

† Abbreviations used: 3D, three-dimensional; Ig, immunoglobulin; RMS, root mean square; SH2, *src* homology 2; SH3, *src* homology 3; HNF-3, hepatocyte nuclear factor 3; the standard three letter and one letter abbreviations for amino acids are also used throughout.

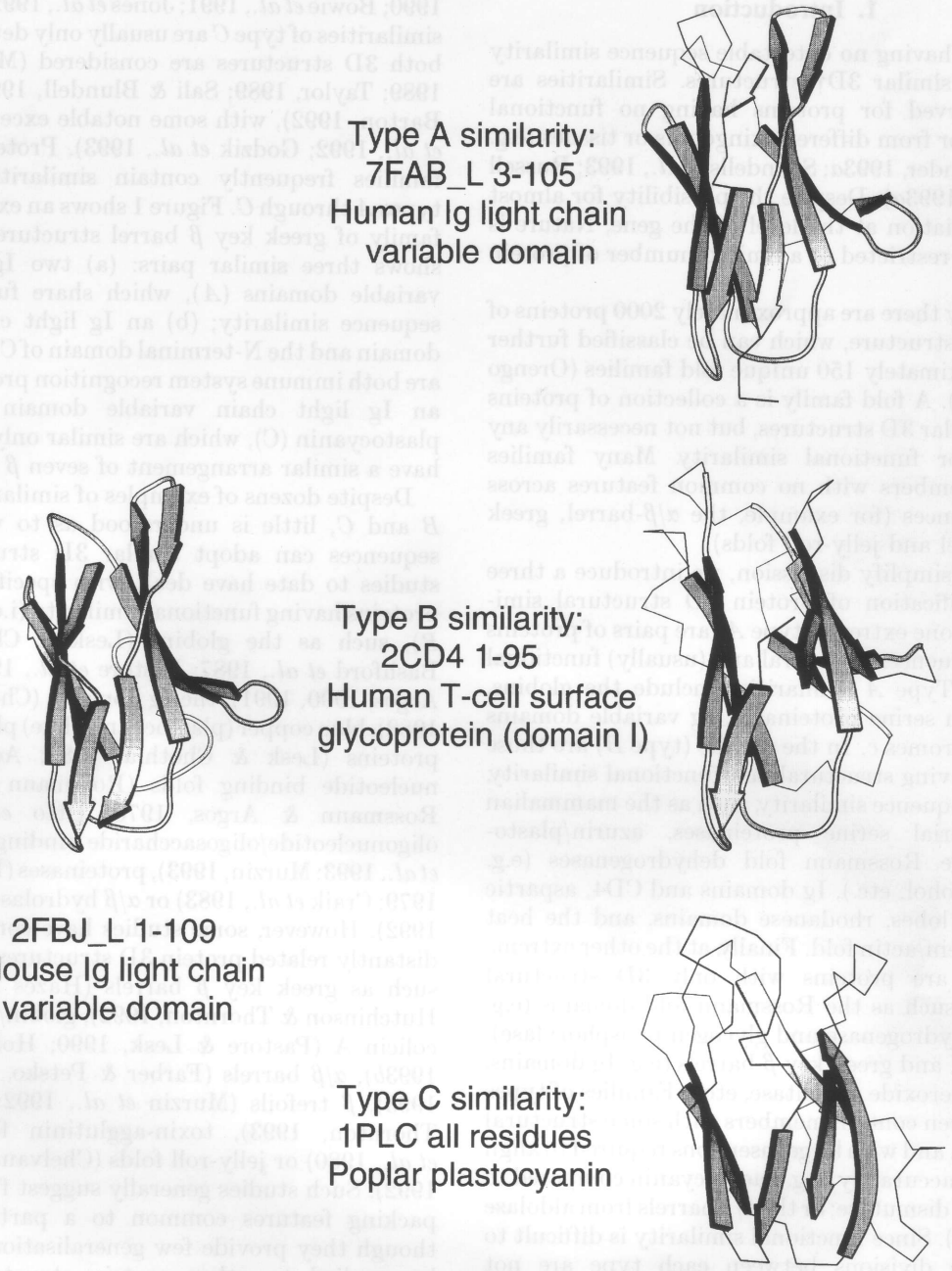


Figure 1. Molscript (Kraulis, 1991) drawings illustrating example of 3 types of 3D structural similarities all with mouse Ig light chain variable domain (2FBJ_L domain 1). The type *A* similarity is between 2FBJ_L domain 1 and a human Ig light chain variable domain; the type *B* between 2FBJL domain 1 and 2CD4 domain 1; and the type *C* between 2FBJ_L domain 1 and popular plastocyanin. Equivalent strand regions within the 3 structures similar to 2FBJ_L are shown as arrows; non-equivalent regions are shown as C^{α} trace.

common core. They also found a logarithmic relationship between sequence identity and RMS deviation on core main-chain atoms; RMS deviation increased exponentially with decreasing sequence identity. Pascarella & Argos (1992) considered families of protein 3D structures and established general rules for the occurrence of insertions and deletions (e.g. that they prefer to be between 1 to 5 residues, and rarely occur within helices or strands). Flores *et al.* (1993) examined how RMS deviation, number of C^{α} to C^{α} contacts, solvent accessibility χ^1 angle and secondary structure behaved as a function of sequence identity for 90 pairs of structurally similar proteins. They found an approximately inverse linear relationship between the variation of all of these properties and sequence identity. For pairs of structures having a similar sequence identity, they found little difference between homologous (i.e. type *A* and some type *B* similarities) and analogous (i.e. some type *B* and type *C* similarities) proteins.

Detection of type *B* and *C* similarities prior to 3D structural determination is of great interest, since

detection and alignment can avail tertiary structure information *via* homology modelling, and can suggest experiments to determine biological function. In an attempt to detect more type *B* and *C* similarities than is possible by sequence comparison, many methods for providing the best fit of a sequence to a structure have been described (Sippl, 1990; Bowie *et al.*, 1991; Luthy *et al.*, 1991; Overington *et al.*, 1992; Jones *et al.*, 1992; Johnson *et al.*, 1993; Bryant & Lawrence, 1993; Godzik *et al.*, 1993; Wilmanns & Eisenberg, 1993; see Bowie & Eisenberg, 1993 or Wodak & Rooman, 1993 for reviews). These methods have been inspired by the earlier work of Novotný *et al.* (1984, 1988), which showed that purposefully misfolded proteins gave rise to favourable energies using CHARMm parameters (Brooks *et al.*, 1983). Novotný *et al.* found that the misfolded proteins had more hydrophobic residues exposed to solvent and more buried ionisable side-chains. Though the details differ, most methods for fitting sequence to 3D structure provide a measure of the quality of the fit based on one or more of: (a) accessibility preferences; (b) loop solvation potentials; (c) secondary structure preferences; and (d) amino acid pair preferences (discussed below). Sippl (1990) first suggested the use of amino acid pair preferences (derived from analysis of known protein 3D structures) for measuring sequence and structure compatibility. Pair preferences provide a measure of how likely each type of amino acid is to interact with every other type, and can be used to assess the quality of the fit of a sequence to a 3D structure if one threads a sequence onto the known structure. Optimal sequence threading involves getting the best alignment of sequence and 3D structure by a consideration of such pair preferences. The use of pair preferences means that threading, unlike most methods of protein sequence alignment, is a 3D problem, since moving residues along the sequence in one region of the structure can affect residues separated by a long length of sequence. Several threading algorithms for protein fold detection have been described (Jones *et al.*, 1992; Godzik *et al.*, 1993; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993).

Methods of protein fold detection have met with some success, being able to detect similarities (and provide accurate sequence alignments) between proteins having little sequence similarity, but which are known to adopt a similar 3D structure. However, most of the success appears to be associated with aligning structures of similarity types *A* and *B*. Many type *B* and *C* similarities remain difficult to detect or align accurately, particularly when pair preferences are not used. For example, the 3D–1D method of Bowie *et al.* (1991) is apparently unable to detect the similarities between hexokinase and actin (Bowie *et al.*, 1991; Thornton *et al.*, 1991), between enterotoxin verotoxin (Sixma *et al.*, 1993), or between various α/β barrels (Pickett *et al.*, 1992). However, the use of pair preferences can enable detection and alignment of several type *B* and *C* similarities. For example, the method of Jones *et al.* (1992) accurately found myohaemerythrin to be a plausible fold for cytochrome B562 by threading the B562 sequence onto each of a database of 102 representative folds, despite the lack of sequence or functional similarity between these proteins. The method of Godzik *et al.* (1993) detected the similarity between the plastocyanin and immunoglobulin structures by using a template derived from the plastocyanin structure to search a sequence database.

An assumption common to fold detection methods is that certain structural features (such as those described by Novotný *et al.* and Sippl) are conserved or shared across proteins having similar 3D structures, even in the absence of sequence similarity. In order for these methods to be successful, secondary structure, accessibility and/or particular side-chain to side-chain interactions must be conserved across similar protein 3D structures. To date, there has been little investigation as to the conservation of particular side-chain properties within distantly related proteins. Studies have concentrated on closely related protein 3D structures (i.e. type *A* or some type *B* similarities), and these have been used to derive environment specific parameters for side-chain substitutions (Overington *et al.*, 1990, 1992; Bowie *et al.*, 1990, 1991; Luthy *et al.*, 1991; Johnson *et al.*, 1993). The high degree of similarity in the proteins used to derive the parameters need not necessarily apply to more distantly related protein pairs, which is, perhaps, why these methods appear to make only a marginal improvement over methods that do not make use of 3D structural data (Taylor, 1986*b*; Gribskov *et al.*, 1987; Lipman & Pearson, 1985; Barton & Sternberg, 1990; Henikoff & Henikoff, 1993).

In this paper, protein 3D structural alignments are used to investigate the conservation of side-chain accessibility, secondary structure and side-chain to side-chain interactions within protein 3D structure pairs having a range of similarities (i.e. types *A*-*B*-*C*). The importance of the results for protein fold detection methods and protein evolution is discussed.

2. Methods

(a) *Multiple alignment of protein 3D structures*

The protein structural families considered are given in Table 1. All structures are refined and of a resolution of 2.5 Å or better, with the exception of the viral coat proteins, which are all refined structures with resolutions between 2.8 and 3.2 Å. As a further test of structural quality, PROCHECK (Morris *et al.*, 1992) was run on all proteins in the dataset using a resolution of 2.5 Å. Those structures showing large deviations (i.e. "WORSE") from typical values for main- and side-chain parameters were not used in the study. Despite poorer resolutions, the viral coat proteins listed in Table 1 were found to have a stereochemical quality comparable to a good 2.5 Å structure, which is expected since molecular averaging greatly improves the quality of these medium resolution structures. All structures were taken from the January 1993 release of the Brookhaven protein

databank (Bernstein *et al.*, 1977), with the exception of sheep 6-phosphogluconate dehydrogenase, *L. mesenteroids* glucose-6-phosphate dehydrogenase (PGD and G6P; kindly provided by Dr M. J. Adams); the human *fyn* SH3 domain (SH3; kindly provided by Dr M. E. M. Noble); chicken *src* SH2 domain (SH2; kindly provided by Dr J. Kuriyan); and human HNF-3 (HNF; kindly provided by Dr S. K. Burley).

Alignments were generated using the STAMP package (Russell & Barton, 1992; Russell, 1994). Pairs of structures that could not be aligned accurately by the method (due to gross structural deviations, etc.) were ignored. It is important to emphasise that the alignments used in this paper are derived from comparison of three-dimensional structures, and thus provide a more accurate set of residue equivalences than alignments created without 3D structural information. Within each family of protein 3D structures, every possible pair of structures were aligned separately to give the most accurate structural alignment. Structurally equivalent regions were defined according to Russell & Barton (1992) by those regions having a residue-by-residue structural similarity index $P'_{ij} \geq 4.5$ for segments of two or more residues.

A total of 607 pairs of aligned protein 3D structures, varying in sequence and 3D structural similarity, were obtained. Type *A* similarities were defined as those proteins having a sequence similarity (%*I*, see below) greater than 20%. The remaining similarities were classified as type *B* or *C* depending on whether the proteins were functionally similar. 137 pairs were classified as type *A*, 292 as type *B* and 178 as type *C*. In Table 1, functionally similar proteins (i.e. types *A* and *B*) are named (e.g. "Oxygen carriers" within the globin fold family). In all the plots that follow, similarities of types *A* and *B* are indicated by the symbol x and type *C* similarities are indicated by the symbol o. Type *A* and *B*/*C* similarities are separated in all plots by a line at $\%I = 20$.

To get a measure of background, 16 pairs of dissimilar structures were aligned using a sequence comparison algorithm (Barton & Sternberg, 1987), with Dayhoff's PAM250 matrix (Dayhoff *et al.*, 1978) and a fixed gap opening penalty of 8. These pairs are

Table 1

Protein structural families considered

Family	Code_chain range		
Four helix bundles (*A*, *B*, *C*)			
Cytochromes	2CCY_A 11–125	256B_A all	
Others	1LIG 48–168	2HMQ_A 19–112	1LPE 23–159
α/β Barrels (*A*, *B*, *C*)			
Xylose isomerases	6XIA 8–286	4XIA_A 8–292	
Anthranilate isomerase	1PII 47–236	1PII 254–427	
Triosephosphate isomerases	6TIM_A 6–234	1YPI_A 5–232	
Rubiscos	5RUB_A 159–393	8RUB_L 167–403	
Others	1WSY_A 16–234	1GOX 71–308	1ALD 26–302
	1TMD 22–321	3ENL 147–398	
Aspartic proteinase domains (*A*, *B*)	2CMS – 2–175	4PEP 1–174	1RNE – 1–172
	2APR 1–178	3APP 1–174	
	3CMS 177–326	4PEP 175–326	1RNE 176–323
	2APR 179–325	3APP 175–323	
	1HIV_A all	2RSP_A all	
Cytochromes *c*	1C2R_A all	1YCC all	5CYT_R all
(A)	1CCR all		
Globin folds (*A*, *B*, *C*)			
Oxygen carriers	4MBN all	1PMB_A all	1THB_A all
	1THB_B all	1PXB_A all	1PXB_B all
	1HBG all	1ITH_A all	1ECA all
	1MBA all	2SDH_A all	2LH1 all
Phycoyanins	1CPC_A 30–173	1CPC_B 27–173	
Colicin A	1COL_A		
Greek key β barrels (*A*, *B*, *C*)			
Ig superfamily	2FB4_H 1–120	2FBJ_H 2–120	1FDL_H 1–118
	7FAB_H 1–119	2FB4_L 1–113	2FBJ_L 1–105
	1FDL_L 1–111	7FAB_L 1–107	
	2FB4_H 121–218	2FBJ_H 121–218	1FDL_H 119–218
	2FB4_L 114–214	2FBJ_L 111–212	
	1HSA_A 184–275	1HSA_B 4–97	2CD4 1–96
	2CD4 98–176		
Plastocyanin folds	1PLC all	7PCY all	1AAJ all
	1PAZ all		
Others	1HOE all	2MCM all	
Lysozymes (*A*)	1LZ4 all	2LZT all	1LZ1 all
Ribonucleases (*A*, *B*)	1RNB_A all	1GMP_A all	1RN4 all

continued overleaf

Table 1 *(continued)*

Family	Code_chain range		
Rossmann folds (*A*, *B*, *C*)			
Dehydrogenases	6LDH 20–163	1LLC 20–160	4MDH_A 3–155
	PGD 1–126	8ADH 192–318	G6P 5–175
Others	8ATC_A 154–289	1GPB 560–712	2FX2 2–147
Serine proteinases	2PKA_AB all	1TLD all	3RP2_A all
(*A*, *B*)	3EST all	1HNE_E all	1CHO_E all
	3SGA_E all	4SGB_E all	1P11_E all
Thioredoxin folds (*C*)	1GST_A 1–87	1GP1_A all	2TRX_A all
β trefoils (*B*, *C*)			
Growth factors	2FGF all	8I1B all	
Ricin domains	1AAI_B 10–138	1AAI_B 139–262	
Trypsin inhibitor	1TIE all		
Viral coat proteins	4RHV_1 all	1R1A_1 all	2MEV_1 all
(*A*, *B*)	1TME_1 all	2PLV_1 all	
	4RHV_2 all	1R1A_2 all	2MEV_2 all
	1TME_2 all	2PLV_2 all	
	4RHV_3 all	1R1A_3 all	2MEV_3 all
	1TME_3 all	2PLV_3 all	
	4SBV_A all		
Various pairs			
BirA/HNF-3 (*C*)	1BIA 1–89	HNF all	
BirA/SH2 (*C*)	1BIA 68–244	SH2_A all	
BirA/SH3 (*C*)	1BIA 270–317	SH3_A all	
Enterotoxins (*B*)	1BOV_A all	1LTS_D all	
γ crystallin (*B*)	1GCR 1–80	1GCR 83–174	
Four helix bundles (*C*)	1GMF_A all	1RCB all	
Ribosomal fold (*B*)	2HPR all	1CTF all	
PGD helical domains (*B*)	PGD 304–432	PGD 178–303	
Ubiquitin/ferredoxin (*C*)	1UBQ all	1FXA_A all	

Protein 3D structural families are given in boldface. In parentheses after each family name are the types of 3D structural similarities contained within the family (see the text). Functional (type *A* and *B*) similarities (e.g. Dehydrogenases) are named. 3D structures are specified by their PDB code, chains (if any) are given after an underscore (–). The range of residues considered is given after the PDB code and chain; all, implies that all residues in the specified code/chain were considered.

given in Table 2. In all the plots that follow, dissimilar structural pairs are indicated by the symbol d.

(b) *Sequence similarity*

In this study, sequence identity (%I) is defined as:

$$\%I = 100 \times n_{identical}/l_{max},$$

where l_{max} is the number of amino acids in the shorter of the two sequences or structures being compared and $n_{identical}$ is the number of positions in the alignment that have the same amino acid.

For the 607 pairs considered in this study, the range of %I is 1.1 to 86.2%. For the 16 dissimilar pairs, optimal sequence alignment gives %I between 9.3 and 21.6%. The much higher minimum %I for dissimilar pairs can be explained by the different methods used to align the sequences. For the pairs of similar protein 3D structures, the alignment was derived by a comparison of 3D structures; for the dissimilar pairs the optimal alignment was obtained by comparison of sequences. Randomly aligned unrelated sequences can give values of %I between 5 and 6%. However, if a sequence comparison algorithm (Needleman & Wunsch, 1970; Barton, 1990) is used to optimise the alignment of two unrelated sequences the expected %I is between 16 and 18% (on average; G.J.B., unpublished data). Since many of the pairs of similar 3D structures used in this study have little or no sequence similarity, and since the method used to align 3D structures in this study does not consider sequence information, alignments derived from 3D structure comparison can give very low values for %I. The values for %I for similar and dissimilar protein 3D structures are thus not directly comparable. Accordingly, dissimilar proteins were given %I = 0 for clarity (see points labelled d in the plots that follow).

(c) *Structural similarity*

The structural similarity index from the structure comparison method of Russell & Barton (1992) was calculated for all 607 protein pairs. Briefly, S_c provides an overall measure of global structural similarity. Pairs of proteins having 3D structure and sequence similarity usually have S_c values between 5.5 and 9.8; those having only 3D structure similarity usually have values between 2.5 and 5.5.

Table 2
Dissimilar protein 3D structural pairs

Pair	Code_chain range	
Helix bundle/ubiquitin	1GMF_A all	1UBQ all
Rossmann fold/serine proteinase	1GPB 560–712	3RP2_A all
Globin/greek key β barrel	1HBG all	1PLC all
Helix bundle/greek key β barrel	1LPE 23–159	2FBJ_L 1–105
Lysozyme/Rossman fold	1LZ1 all	4MDH_A 3–155
Ribonuclease/greek key β barrel	1RN4 all	1AAJ all
α/β barrel/serine proteinase	1WSY_A 16–234	1CHO_E all
Helix bundle/aspartic proteinase	256B_A all	2RSP_A all
PGD helical/ferredoxin	PGD 178–303	1FXI_A all
Thioredoxin fold/β trefoil	2TRX_A all	1TIE all
γ Crystallin/ribosomal fold	1GCR 83–174	1CTF all
Globin/lysozyme	1THB_A all	2LZ2 all
Globin/aspartic proteinase	4MBN all	1RNE 176–323
Aspartic proteinase/greek key β barrel	4PEP 1–174	2CD4 98–176
α/β barrel/viral coat protein	8RUB_L 167–403	4RHV_1 all
SH2 fold/β trefoil	1BIA 10–138	1AAI_B 68–244

The names of the folds comprising each pair are as given in Table 1. 3D structures are specified by their PDB chain, code and range as for Table 1.

(d) *Secondary structure and accessibility*

Secondary structure assignments and accessibilities were defined by the Definition of Secondary Structure in Proteins program (DSSP; Kabsch & Sander, 1983). Secondary structure assignments were converted to a three state representation for simplicity: helix = DSSP α-helix, 3_{10}-helix; beta = DSSP β-ladder, β-bridge; coil = DSSP not (α-helix, 3_{10}-helix, β-ladder, or β-bridge). Relative accessibilities were calculated by dividing the DSSP accessibility by the accessibility for a GXG tripeptide given by Rose & Dworkin (1989). When one chain or domain was extracted from a PDB file, DSSP was run on the domain separately, excluding heteroatoms, such as substrates, not integral to the 3D structure.

(e) *Side-chain to side-chain interaction potential*

All contacts, defined as any atom–atom distance of less than 5 Å, were calculated and tabulated for each of the 102 unique 3D structures given by Jones *et al.* (1992). Residues were considered to be in contact if they had at least one shared contact between the atoms of their side-chains.

Several authors have described potentials for the interaction of two residues within protein structures. Some make use of a reduced representative protein structure (Sippl, 1990; Jones *et al.*, 1992; Bryant & Lawrence, 1993), whereas others consider all atoms (Godzik *et al.*, 1993). In this study every atom–atom contact made between protein side-chains was used to derive a simple pseudo-energy term for the interaction of two residues P and Q:

$$\Delta E(P, Q) = E(P, Q) - E^o,$$

where $E(P, Q)$ is defined by:

$$E(P, Q) = -RT \ln \frac{N_o}{N_e} \quad (RT \approx 2.5 \text{ kJ/mol}),$$

N_o is the observed number of contacts between residues of type P and Q, and N_e is the expected number (assuming a random model), and E^o is a reference state energy (discussed below).

Given a database of known 3D structures, a set of pair potentials can be derived by counting the number of times a particular amino acid contact occurs and dividing this number by the number of times expected given the total number of contacts made by each amino acid. For any given amino acid pair, the expected number of side-chain–side-chain contacts under a random model assumption is (Warme & Morgan, 1978):

$$N_e(P, Q) = N(x, x) \frac{N(P, x)}{N(x, x)} \frac{N(Q, x)}{N(x, x)} = \frac{N(P, x) N(Q, x)}{N(x, x)},$$

where x denotes all amino acids, $N(x, x)$ is the total number of side-chain to side-chain contacts in the dataset, and where $N(P, x)$ and $N(Q, x)$ are the total number of side-chain to side-chain contacts made by residues of type P and Q, respectively. Contacts within the database of 102 unique folds were counted, and the observed number of contacts for each pair of amino acids were used to calculate $E(P, Q)$. The reference state energy E^o was calculated by taking the average of all values of $E(P, Q)$, which gives $E^o = 0.033$ kJ/mol. Values of $\Delta E(P, Q)$ were calculated by the equations described above, and are given in Table 3.

The columns/rows of Table 3 can be used to classify amino acids according to their pair preferences. A measure of the difference in pair preference can be obtained by summing the absolute differences between the values in each column for every possible pair of amino acids. Figure 2 shows a complete linkage dendrogram for these data. The clustering of the hydrophobic residues (M, A, V, L, I, W, F) is similar to clustering by side-chain properties (Taylor, 1986*a*), and shows their similar pair preferences. However, unlike other classifications of the amino acids, the charges cluster separately (i.e. R and K do not cluster

Table 3
Residue pair potentials

	A	C	D	E	F	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−4	22	4	26	0	6	−17	22	−16	1	−1	0	1	25	4	−8	−22	−1	0
C		−192	31	62	13	−8	10	56	32	0	11	9	4	23	−1	13	27	1	−2
D			13	5	46	−32	33	−53	38	24	−39	−1	−10	−64	−50	−29	37	28	6
E				8	24	−26	31	−65	26	10	−27	−7	−2	−49	−27	−21	24	29	3
F					−22	2	−8	16	−16	−16	18	2	2	0	22	16	−12	−14	5
H						−36	23	9	21	1	−14	−12	16	−13	−16	−20	14	2	−9
I							−26	20	−20	−15	40	14	20	22	16	3	−20	−5	−3
K								26	18	31	−24	3	−12	62	−4	1	24	−15	−29
L									−24	−9	30	12	9	16	27	12	−15	−5	3
M										−42	1	3	9	14	13	−2	−8	−14	7
N											−38	−10	−29	−5	−25	−24	20	7	−6
P												−7	−10	7	−5	−13	9	−23	−28
Q													−8	−22	−23	−21	18	−8	−13
R														−7	−1	−2	22	−4	−9
S															−53	−26	16	20	−1
T																−26	6	22	7
V																	−27	−11	6
W																		4	2
Y																			−1

Values are $\Delta E(P, Q)$($= E(P, Q) - E^{o}$) in units of kJ/mol as described in the text. All values are multiplied by 100 for clarity.

with E and D), suggesting (as expected) that positive and negative residues, when in contact with other residues, are unlikely to undergo mutations involving a change in sign.

When considering a single pair of interacting residues, the pair potentials provide an approximate test of whether the interaction is favourable (i.e. whether or not it will effect to stabilise or destabilise the overall fold) simply by investigating the sign of $\Delta E(P, Q)$. Negative values (i.e. $\Delta E(P, Q) \leqslant 0$) will be expected to stabilise the fold, whereas positive values will be expected to be disruptive. Although the pair potentials discussed here differ from many of those used previously (Sippl, 1990; Jones *et al.*, 1992; Godzik *et al.*, 1993; Bryant & Lawrence, 1993), the signs of $\Delta E(P, Q)$ are similar, suggesting that this simple test, and the results that follow, would differ little if another pair potential was used.

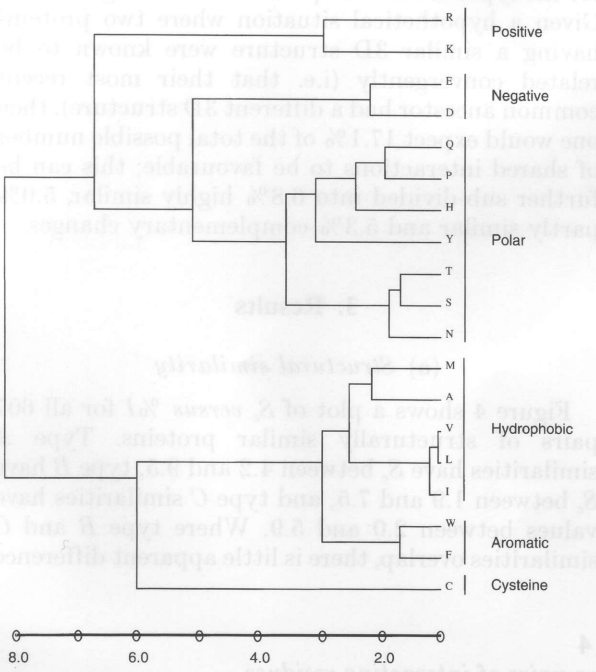


Figure 2. Complete linkage cluster analysis dendrogram derived by a matrix of the sums of the absolute differences between columns/rows in Table 3. The numbers on the X axis correspond to the minimum sum of absolute differences for each cluster (e.g. W and F cluster together with a sum of absolute differences of ≈ 1.8).

(f) *Comparison of interacting residues within protein structure pairs*

Each of the 607 aligned pairs of structures was considered separately. Within an aligned pair of structures (i.e. proteins 1 and 2) two pairs of residues are defined:

i and j (protein 1)
i' and j' (protein 2).

In the alignment, position i is aligned with position i' and position j is aligned with position j'. Figure 3 illustrates these definitions for a pair of simple 3D structures. All pairs of interacting residues were required to have more than four residues between them on the sequence (i.e. $(i-j) \geqslant 5$). All possible $(i, j), (i', j')$ combinations were considered as to whether: (1) the positions are in contact in one or both structures, (2) if in contact whether or not the interaction is favourable (i.e. is $\Delta E(i, j) \leqslant 0$ and/or $\Delta E(i', j') \leqslant 0$), or (3) if both structures are in contact at these positions, whether the interactions are similar.

Tests 1 and 2 provide information about the general nature of interactions within protein structures when considered individually. For example, the data may show how the number of favourable interactions behaves as a function of sequence length. Test 3 provides information as to how different sequences

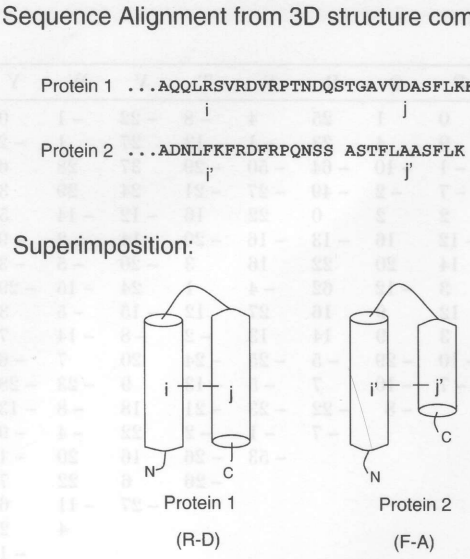


Figure 3. How structurally derived sequence alignments and protein 3D structures are used to find side-chain to side-chain contacts common to a pair of 3D structures. The sequence alignment is shown in the top of the Figure and the 2 structures in the same (i.e. superimposed) orientation are shown in the bottom of the Figure. A pair of residues in contact in protein 1 (i, j) are equivalent to another pair in protein 2 (i', j'). The interaction R–D in protein 1 is replaced in protein 2 by the interaction F–A.

(often with little or no apparent sequence similarity) adopt similar three-dimensional structures, and thus requires further description.

When comparing positions found to be in contact at two aligned positions there are three possibilities: (1) both interactions can be favourable; (2) one interaction can be favourable (and the other unfavourable); and (3) both interactions can be unfavourable. Situations where both interactions are favourable (i.e. when $\Delta E(i,j) \leqslant 0$ and $\Delta E(i',j') \leqslant 0$), can be sub-divided by considering the intermediates involved in mutating one interaction to the other. In other words, if one were to mutate, in two steps, the interaction (i, j) to (i', j'), two mutations would be involved, and the two possible evolutionary paths would be:

$$(i,j) \rightarrow (i,j') \rightarrow (i',j')$$
$$(i,j) \rightarrow (i',j) \rightarrow (i',j').$$

Of course, such evolutionary paths are hypothetical, since the mutation of one pair of residues to another may involve more than one intermediate. However, considering both hypothetical paths enables all shared favourable interactions to be sub-divided by considering the stability of the intermediates (e.g. $\Delta E(i',j)$ and $\Delta E(i,j')$). There are three possible situations:

(1) Highly similar interactions, defined as those positions with both intermediates having a favourable pseudo-energy (i.e., $\Delta E(i,j') \leqslant 0$ and $\Delta E(i',j) \leqslant 0$).

(2) Partly similar interactions, defined as those positions having one favourable intermediate (i.e. either $\Delta E(i,j') \leqslant 0$ or $\Delta E(i',j) \leqslant 0$).

(3) Complementary changes, defined as those positions with both intermediates having an unfavourable pseudo-energy (i.e., $\Delta E(i,j') > 0$ and $\Delta E(i',j) > 0$).

Type 1 describes interactions of a similar character in both structures, and thus suggests features common to the two structures (and perhaps to the fold in general). Type 2 describes less similar interactions, suggesting those positions on the point of diverging away from each other. Pairs of interactions of type 3 are the most interesting, since they are interactions of significantly different character in the two structures, yet which both contribute to the respective stabilities.

By considering the abundancies of the amino acids, it is posible to determine the expected frequency of the interactions described above. All 32,851 unique possible combinations of two residue pairs involving 19 amino acids (excluding glycine, which can make no side-chain to side-chain contacts) were classified by the above definitions. The results are shown in Table 4. The weighted frequencies provide the expected limits for the types of mutated pairs of interacting residues. Given a hypothetical situation where two proteins having a similar 3D structure were known to be related convergently (i.e. that their most recent common ancestor had a different 3D structure), then one would expect 17.1% of the total possible number of shared interactions to be favourable; this can be further sub-divided into 6.8% highly similar, 5.0% partly similar and 5.3% complementary changes.

3. Results

(a) *Structural similarity*

Figure 4 shows a plot of S_c *versus* $\%I$ for all 607 pairs of structurally similar proteins. Type A similarities have S_c between 4.2 and 9.5, type B have S_c between 1.9 and 7.5, and type C similarities have values between 2.0 and 5.9. Where type B and C similarities overlap, there is little apparent difference

Table 4

The expected frequencies of mutating pairs of interacting residues

	Total	Both favourable	Highly similar	Partly similar	Complementary changes
Raw numbers	32,851	7800	2908	2574	2318
Weighted (%)	69.8	17.1	6.8	5.0	5.3

Weighted numbers are those calculated by accounting for the abundancies of the amino acids. The total is not 100%, since glycine residues (having no side-chains) are ignored.

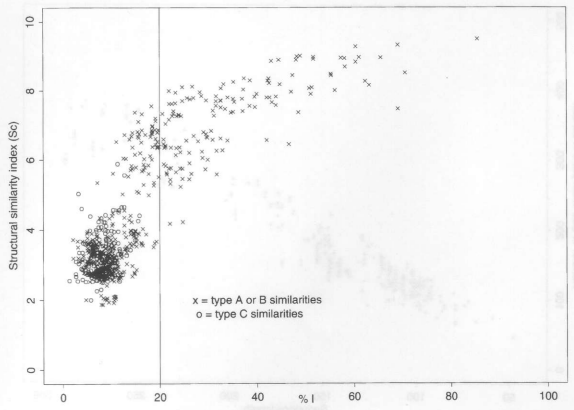


Figure 4. How the structurally similarity index S_c behaves as a function of the percent sequence identity (%I).

between these types of protein pairs; both have a similar range of structural similarity.

(b) *Accessibility*

Three commonly used categories of relative accessibility (A) were defined ($0 \leqslant A < 5\%$, $5 \leqslant A < 25\%$, $A \geqslant 25\%$), corresponding to buried, half-buried and exposed (e.g. see Miller *et al.*, 1987; Bowie *et al.*, 1991). Positions within pairs of aligned structures were defined as conserved with respect to accessibility if structures had accessibilities in the same category. Figure 5 shows the percentage of these positions *versus* %I. Most pairs of structures have $> 40\%$ of such positions in common, though some pairs of structurally similar proteins have a degree of accessibility conservation similar to dissimilar structural pairs, which have between 33.0 and 60.0% conservation of these positions just by chance.

Defining conservation of accessibility after Miller *et al.* (1987) as those positions having an absolute accessibility difference of less than 20 Å^2 gives a similar plot, though with a more steady drop in the percentage of conserved positions with decreasing sequence identity. Both observations suggest that

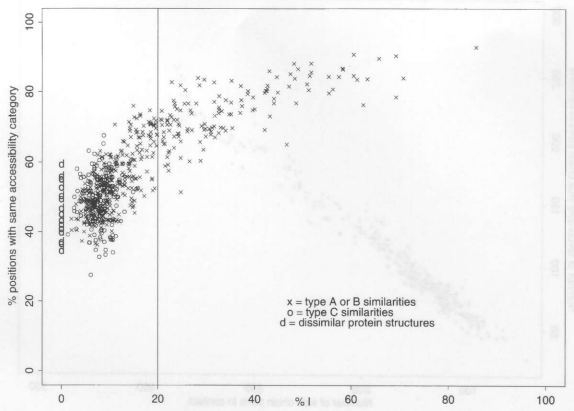


Figure 5. How the percentage of positions having the same accessibility category behaves as a function of the percent sequence identity (%I).

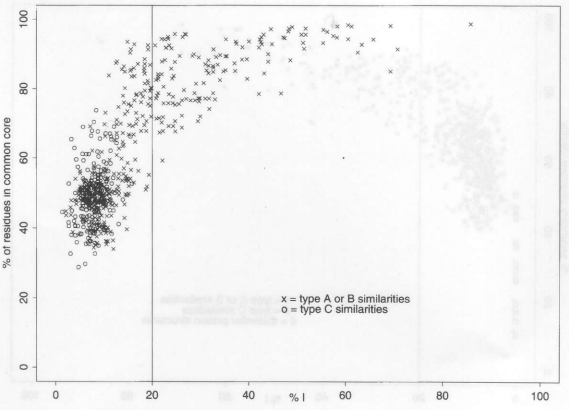


Figure 6. How the percentage of positions within structurally equivalent regions behaves as a function of the percent sequence identity (%I).

many pairs of distantly related proteins have a degree of residue by residue accessibility conservation comparable to that observed for pairs of dissimilar protein structures.

(c) *Core structure*

Figure 6 shows how the percentage of structural equivalence (i.e. the fraction of the smallest structure that lies within structurally similar regions) behaves as a function of %I. Proteins having detectable sequence similarity (i.e. $\%I \geqslant 20\%$) generally have over 60% structural equivalence, or common structural regions. However, as %I drops, the percentage of structural equivalence drops to as low as 28.9%. This is a consequence of distantly related protein structures being similar only within their core secondary structures and having loop/turn regions that differ substantially. This fraction is somewhat less than that reported by Chothia & Lesk (1986), who found a minimum of $\approx 42\%$, though their 32 protein structure pairs were more closely related than the 607 pairs used in this study. For example, the globin fold family studied by Chothia and Lesk contained only haem containing globins and lacked the more distantly related phycocyanin and colicin A structures, which are included in this study.

(d) *Secondary structure*

For pairs of homologous proteins, secondary structure agreement can be as low as 70% (Russell & Barton, 1993*b*; Flores *et al.*, 1993). Figure 7 shows how three-state secondary structure identity (calculated in the same manner as %I) behaves as a function of %I. For type A similarities, secondary structure agreement is between 67.0 and 99.6%. For type B and C similarities, the possible variation in secondary structure content is much greater (range of 40.1 to 87.9%), with the lowest observed pairs having secondary structure identities similar to those for dissimilar structures (between 10.6 and 50.6%). Nearly all pairs having secondary structure identities less than 50% are all β proteins, which could be

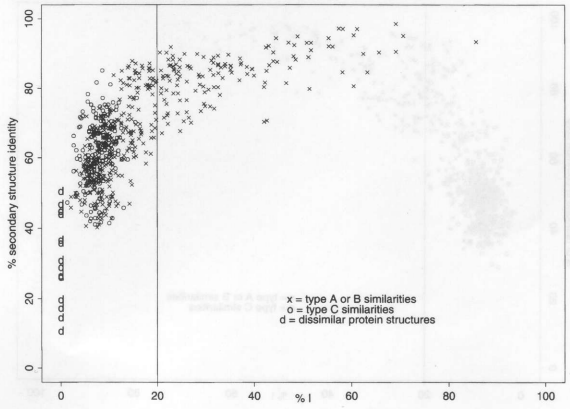


Figure 7. How the percentage of positions having the same 3 state secondary structure assignment behaves as a function of the percent sequence identity (%I).

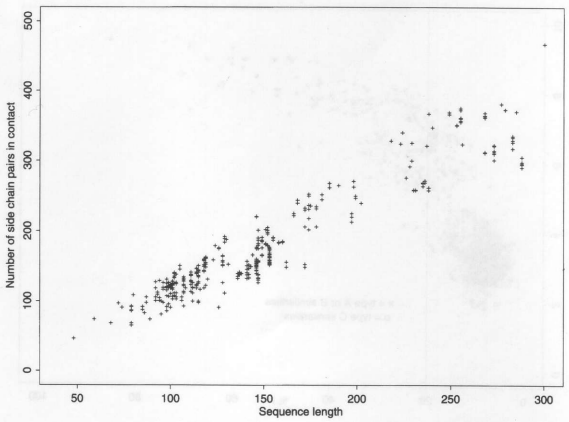


Figure 9. How the number of residue pairs with side-chains in contact behaves as a function of the sequence length.

explained by the shorter (on average) length of β strands compared to α helices making matches of secondary structure strings longer on average for helix containing proteins.

(e) *RMS deviation*

Figure 8 shows RMS deviation, calculated using the method of McLachlan (1979) for pairs of equivalent common core C^{α} atoms *versus* %I. The Figure is similar to that shown by Chothia and Lesk (1986), though with substantially more spread, which one would expect given the greater structural variation of the protein pairs considered here.

Three interesting outliers are labelled on the plot. Two distantly related globin structures (sea hare and leghaemoglobin; 1MBA and 2LH1) show a higher RMS deviation than other structural pairs of a similar %I, which might be explained by the large variations in helix packing angles seen within this family (Pastore *et al.*, 1988). The other two pairs have RMS deviations lower than expected for their respective %I

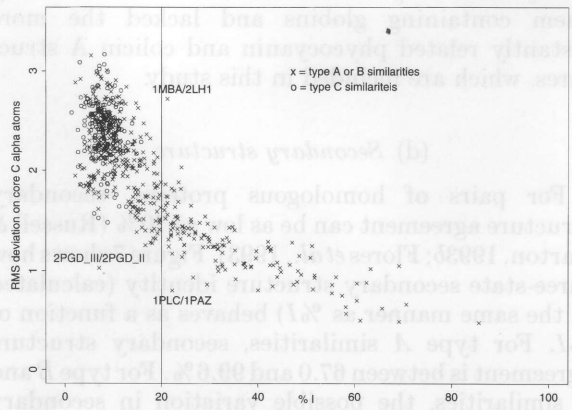


Figure 8. How the RMS deviation for equivalent C^{α} atoms behaves as a function of the percent sequence identity (%I). Interesting outliers are specified by their PDB 4 letter code, their chain letter (if any) and a Roman numeral specifying the number of the domain considered (numbered sequentially from the N terminus).

values. The two helical domains within 6-phosphogluconate dehydrogenase (PGD_II and PGD_III) are closely packed together, and might accordingly be restrained to specific symmetrical conformations in spite of sequence dissimilarity. The two plastocyanin-like structures (1PLC and 1PAZ) have a strong functional similarity, which might also restrict the degree to which their core C^{α} atoms can deviate from one another.

(f) *Residue–residue interactions within single structures*

Figure 9 shows how the number of interacting residue pairs (i.e. those pairs of residues with at least one side-chain to side-chain contact) behaves as a function of the sequence length. The behaviour is approximately linear, and there are ≈ 1.2 interacting pairs per residue. Figure 10 shows how the number of favourable interacting residues pairs (i.e. where $\Delta E(P, Q) \leqslant 0$) behaves as a function of the total number of interacting pairs. The behaviour is also approximately linear, and only about half of the

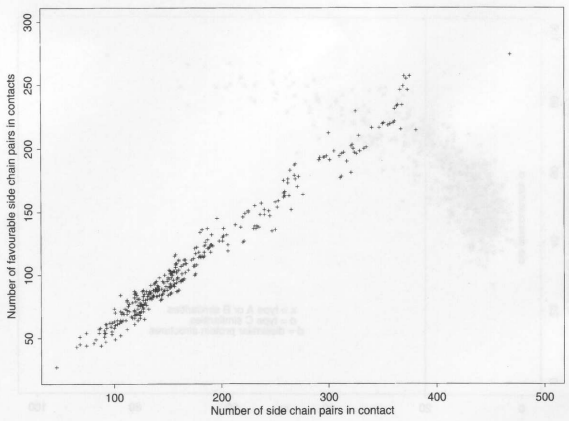


Figure 10. How the number of residues with side-chains in an energetically favourable contact behaves as a function of the total number of residue pairs in contact.

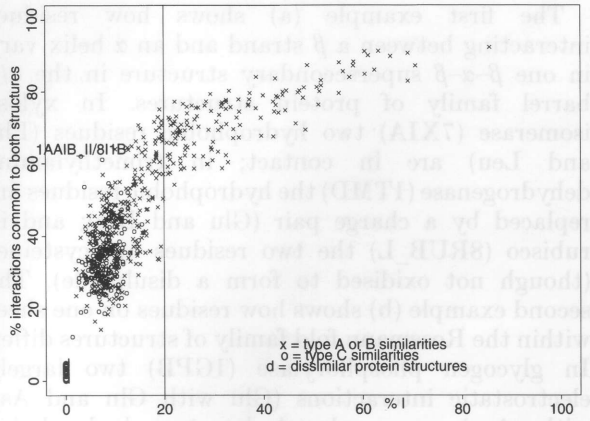


Figure 11. How the percentage of residues with side-chains in contact common to both structures behaves as a function of percent sequence identity (%I). Interesting outliers are specified as described in Figure 8.

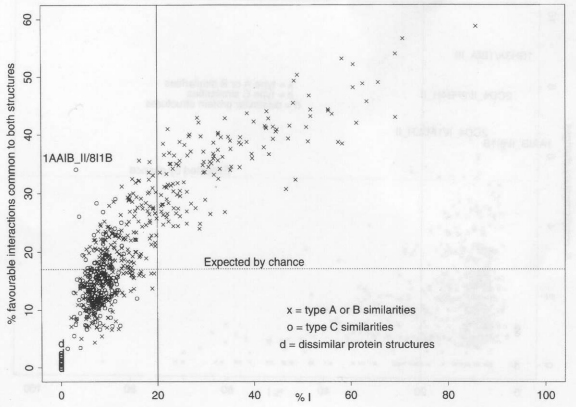


Figure 13. How the percentage of residues with side-chains in an energetically favourable contact common to both structures behaves as a function of percent sequence identity (%I). Interesting outliers are specified as described in Figure 8. The broken line shows the expected percentage of favourable interactions for convergently related proteins of a similar 3D structure.

interactions in proteins are favourable (i.e. contributing a negative pseudoenergy term).

(g) *Shared interactions*

Figure 11 shows the % shared interactions (i.e. the number of residue–residue interactions common to both structures divided by the smallest number of interactions in the pair of aligned structures) *versus* %I. Surprisingly, structurally similar proteins can have as few as 12% of their interactions in common.

The above definition of interacting residues is somewhat strict in that it requires that there be at least one side-chain to side-chain contact (atom–atom contact $\leqslant 5$ Å) between pairs of residues. Relaxing the requirement and defining interacting residues as those residues with C^{β} atoms (or built C^{β} in the case of glycine residues) within 8 Å of each other gives the analogous plot in Figure 12. By this relaxed definition, the percentage of shared interactions tends to be higher, but structurally similar proteins can still have as few as 20% of interactions in common. The greater separation between similar and dissimilar protein pairs suggests that this geometrically less exact consideration of side-chain interactions might be more effective as a tool for protein fold recognition.

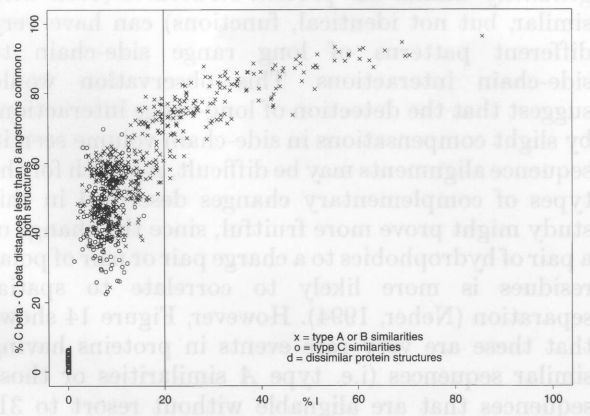


Figure 12. How the percentage of residues with C^{β}–C^{β} distances $\leqslant 8$ Å common to both structures behaves as a function of percent sequence identity (%I).

Perhaps more interesting than shared interactions are shared favourable interactions. These are interactions common to both structures that both contribute a negative pseudo-energy term (i.e. $\Delta E(P, Q) \leqslant 0$). Figure 13 shows the percentage of shared favourable interactions *versus* %I. As would be expected from Figure 10, proteins having highly similar sequences have about half of their interactions as shared and favourable (50% is the approximate maximum). However, proteins having no detectable sequence similarity have less than 35% of the total possible interactions as shared and favourable, and many distantly related structures have essentially no common favourable interactions. Many distantly related proteins have a proportion of shared favourable interactions near to that expected by chance (see broken line in Figure 13 and Table 4), suggesting that many pairs of structurally similar proteins have completely different stabilising interactions.

A notable outlier in Figures 11 and 13 is the similarity between ricin domain 2 (1AAIB_II) and interleukin 1 β (81IB). This pair of proteins has a % shared interactions greater than others of a similar %I. This might be explained by the conservation of particular amino acids (and their corresponding side-chain to side-chain interactions) within the β trefoil family of proteins (Murzin *et al.*, 1992), though the fact that other β trefoil pairs of a similar %I do not have such a high percentage of shared interactions might suggest that the ricin/interleukin-1-β similarity is fortuitous.

Unlike secondary structure and accessibility, genuine structural similarities tend to have more common interactions than dissimilar structures. In Figures 11 and 12 there is a distinct separation

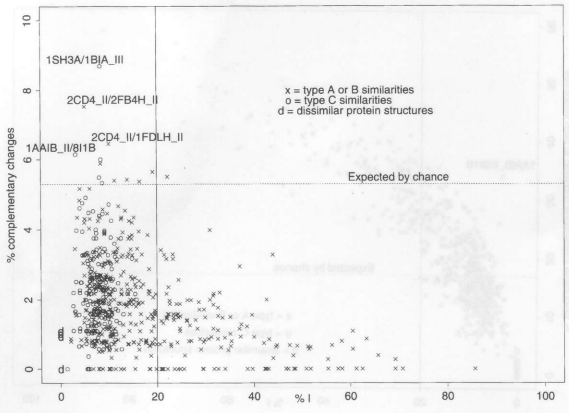


Figure 14. How the percentage of complementary changes behaves as a function of $\%I$. Four pairs with high percentages are shown as described for Figure 8. The broken line shows the expected percent complementary changes for convergently related proteins of a similar 3D structure.

between structurally similar and dissimilar proteins. This difference is less pronounced for favourable pairs (Figure 13).

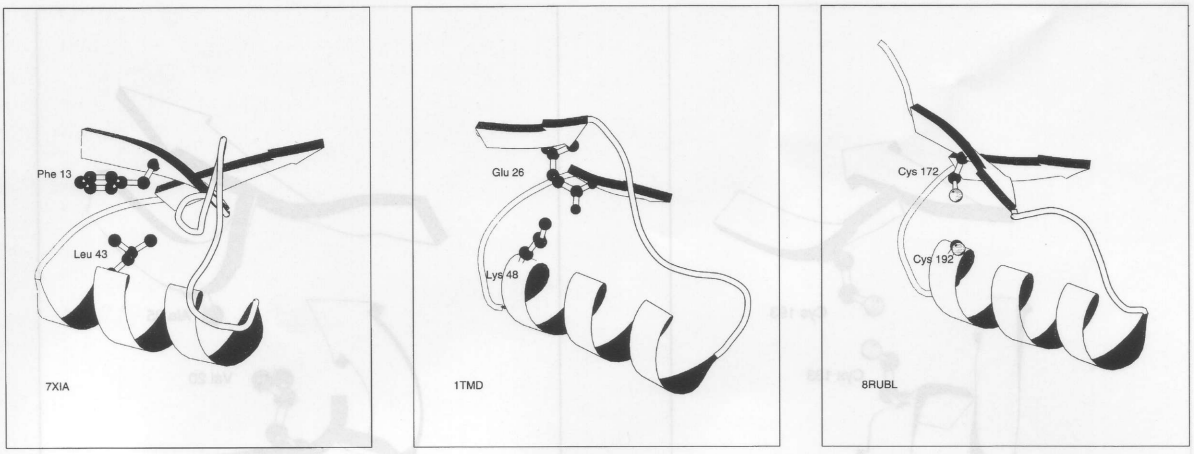
(h) *Complementary changes*

Do different interactions stabilise protein structures having similar folds? Figure 14 shows the distribution of the percentage of complementary changes *versus* $\%I$. For some type *B* and *C* similarities, the proportion of complementary changes is as high as 8% of the total number of possible interacting pairs. Some pairs of 3D structures with an extraordinarily high number of complementary changes are labelled in the Figure. Many similar 3D structures have a proportion of complementary changes similar to that expected by chance (Table 4), suggesting a fundamental difference in stabilisation. The Figure suggests that interactions between residues at equivalent positions in similar 3D structures can differ substantially in character, and has many implications for methods which attempt to use protein 3D structural information to find sequences compatible with a fold. In particular, methods not taking long-range interactions into account (Bowie *et al.*, 1991; Overington *et al.*, 1990, 1992; Johnson *et al.*, 1993) may encounter difficulties in differentiating many genuine structural similarities from noise, since they will be unable to detect such complementary changes.

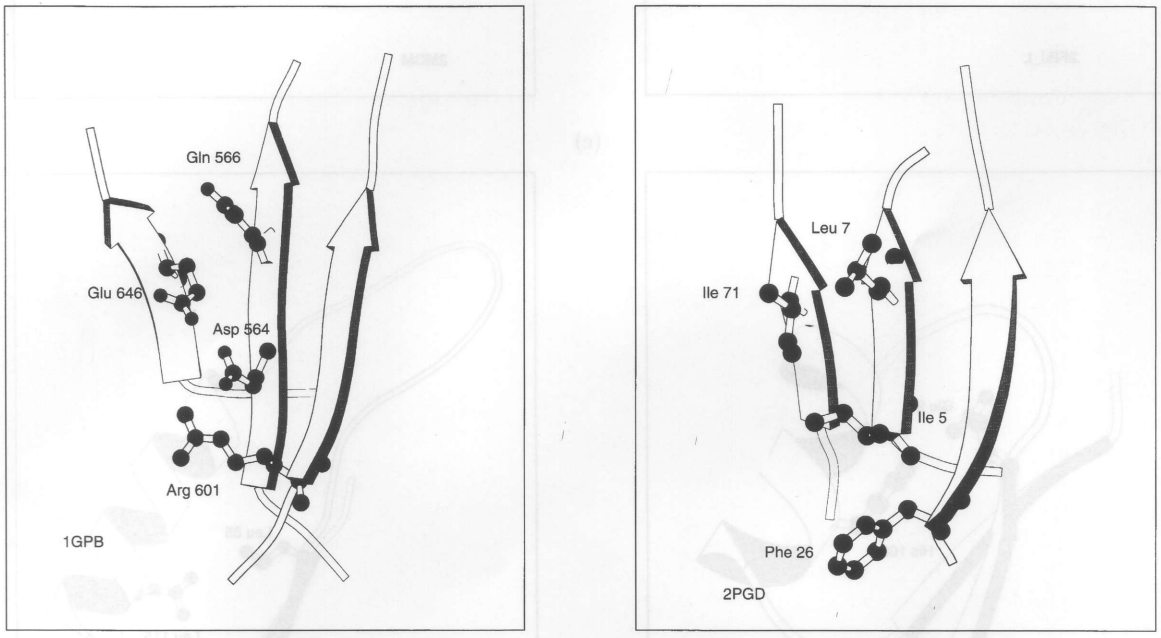
Though comparatively rare, many interesting varieties of complementary changes occur within protein structure pairs. Most involve an interaction changing from a predominantly hydrophobic pair to a charge pair or a pair of polar residues. Five examples are shown in Figure 15. In all of the examples, the regions shown are extracted from a larger structural alignment and superimposition and the residues shown to be in contact fall within core or structurally equivalent regions and all have relative accessibilities of less than 30%.

The first example (a) shows how residues interacting between a β strand and an α helix vary in one β–α–β supersecondary structure in the α/β barrel family of protein structures. In xylose isomerase (7XIA) two hydrophobic residues (Phe and Leu) are in contact; in trimethylamine dehydrogenase (1TMD) the hydrophobic residues are replaced by a charge pair (Glu and Lys); and in rubisco (8RUB_L) the two residues are cysteines (though not oxidised to form a disulphide). The second example (b) shows how residues on one sheet within the Rossmann fold family of structures differ. In glycogen phosphorylase (1GPB) two largely electrostatic interactions (Glu with Gln and Asp with Arg) are replaced by two hydrophobic interactions in 6-phosphogluconate dehydrogenase (PGD; Ile with Leu and Ile with Phe). The third example (c) shows how a disulphide bond around the "pin" in Ig folds (2FBJ_L constant domain in the Figure) is replaced in the antibacterial protein macromycin (2MCM) by a hydrophobic (Val–Ala) interaction. The fourth example (d) shows how a helix–strand interaction differs between two Rossmann fold domains. In glyceraldehyde-3-phosphate dehydrogenase a histidine forms a hydrogen bond with a glutamic acid; in malate dehydrogenase (4MDH) two leucine residues are in contact. The fifth example (e) shows how an electrostatic interaction (Asp–His) within innkeeper worm haemoglobin (1ITH_A) is replaced by a hydrophobic interaction (Phe–Ile) in the bacterial toxin colicin A (1COL_A).

Much recent work has concentrated on attempting to predict 3D contacts in proteins of unknown structures by analysis of complementary changes in multiple protein sequence alignments (Taylor & Hatrick, 1994; Shindylav *et al.*, 1994; Neher, 1994; Göbel *et al.*, 1994). The details of these studies differ, though the general conclusion is that it is not possible to predict such contacts with confidence. Though perhaps not directly comparable, the results of this study shed some light on why these predictive methods are unsuccessful. Although subtle changes to side-chains can have disastrous effects on specific protein function (e.g. Lim & Sauer, 1989), pairs of genuinely similar 3D protein structures (even with similar, but not identical, functions) can have very different patterns of long range side-chain to side-chain interactions. This observation would suggest that the detection of long-range interactions by slight compensations in side-chain volume seen in sequence alignments may be difficult. A search for the types of complementary changes described in this study might prove more fruitful, since the change of a pair of hydrophobics to a charge pair or pair of polar residues is more likely to correlate to spatial separation (Neher, 1994). However, Figure 14 shows that these are very rare events in proteins having similar sequences (i.e. type *A* similarities or those sequences that are alignable without resort to 3D structure comparison), so detection of such sites from multiple sequence alignments is likely to prove difficult.



(a)



(b)

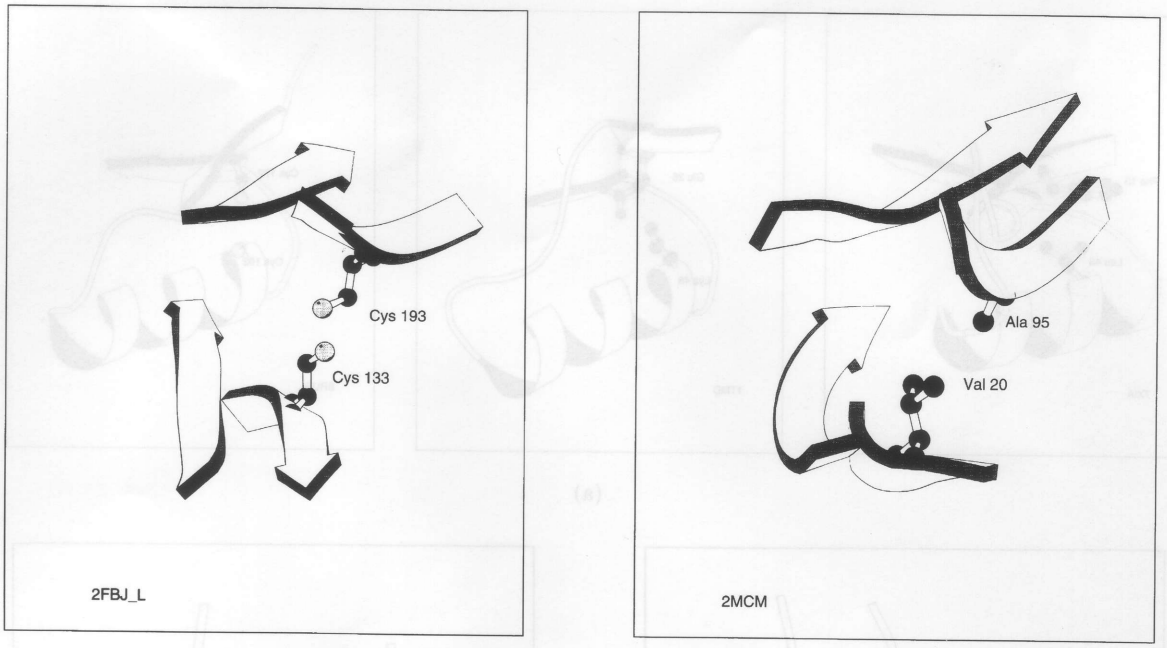
Figure 15 (a)–(b)

(i) *Homologous* versus *analogous proteins*

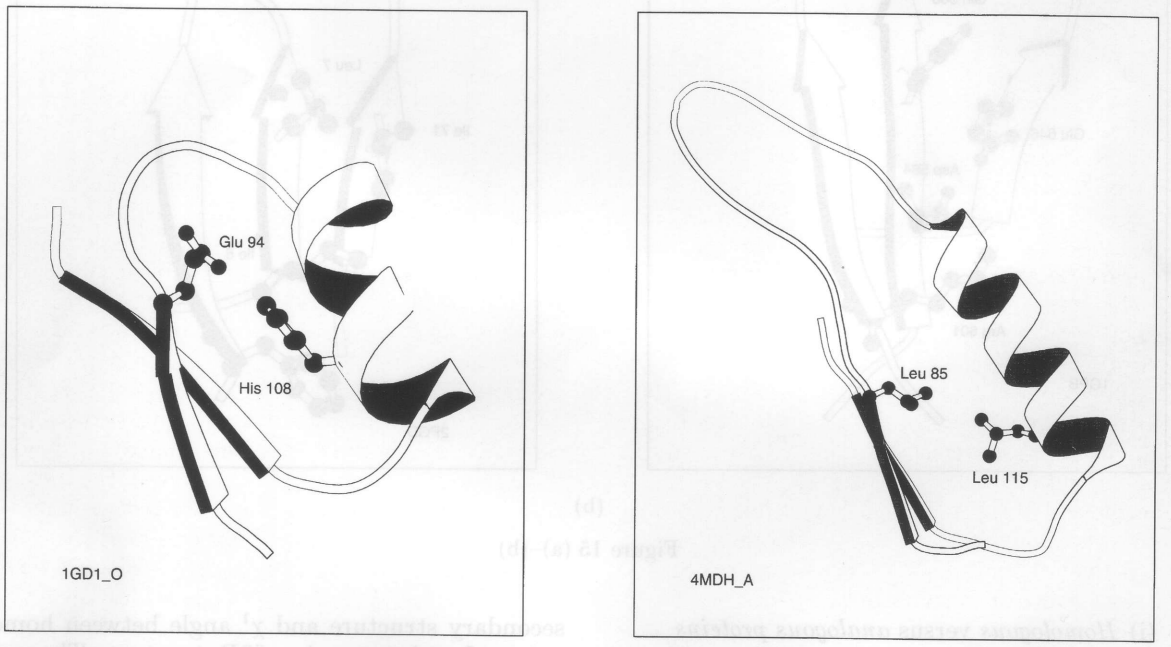
Table 1 shows which proteins were classified as functionally similar (i.e. type A or B), and on all plots described above, a line is drawn to separate type A similarities (i.e. those having $\%I \geqslant 20$) from others. Within all plots, type A and B similarities are labelled as x, whereas type C similarities are labelled o. Table 5 shows the degree of property conservation for the three protein structural pairs (of type A, B and C) shown in Figure 1. Table 6 shows a general summary of the range of structural features for all the types of similarities. For pairs of 3D structures having $\%I<15.8$ (i.e. the range of $\%I$ for type C similarities), no difference can be seen in the Table or any of the plots between those structures with functional similarity and those without. Flores *et al.* (1993) found little difference in the conservation of accessibility, Ooi numbers, secondary structure and χ^1 angle between homologous and analogous pairs of 3D structures. The results presented here agree with their findings. It would seem that even with the restriction of functional similarity, protein structures (and sequences) can vary significantly.

4. Discussion

The results of this study suggest that there is little in common between distantly related protein structures. Within pairs of similar 3D structures, as few as 30% of residues can form a common core, which reinforces previous observations that secondary structure lengths and loops in distantly related structures vary substantially. The degree to which accessibility and secondary structure are conserved on a residue by



(c)

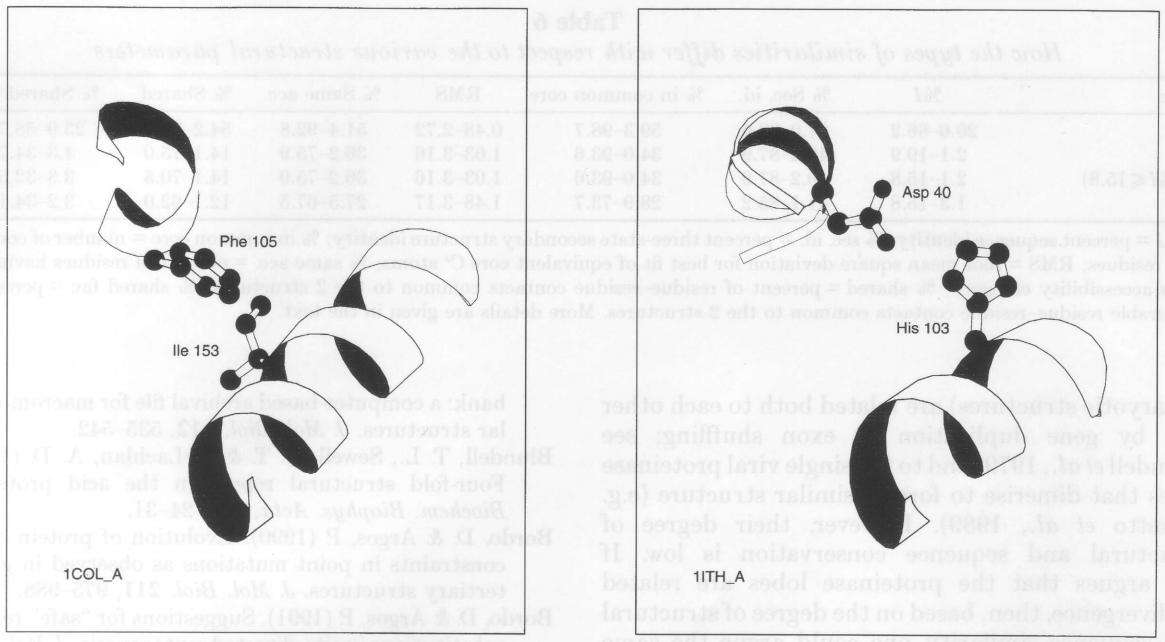


(d)

Figure 15 (c)–(d)

residue basis within structurally similar proteins can be as low as that for dissimilar proteins (i.e. by chance). The fraction of shared interactions (pairs of residues in contact in equivalent positions within two distantly related protein structures) can be as little as 12%, even when a lenient definition of C^{β}–C^{β} distance is used. Structurally similar proteins can have almost no common favourable interactions, or those contributing a negative pseudo-energy term. Finally, regardless of any functional similarity, similar protein 3D structures often have a proportion of complementary changes approaching that expected by chance. All of the results suggest that proteins can adopt very similar folds by using almost completely different interactions, and that proteins having similar 3D structures can have little in common apart from a scaffold of common core secondary structures.

The results presented here have many implications for methods of protein fold detection. The fact that the degree of conservation of secondary structure and accessibility, when considered on a residue by residue basis, is similar to that for structurally dissimilar proteins, and the low proportion of residues in common cores suggests why many methods of fold



(e)

Figure 15. Five examples of complementary changes. Details are described in the text.

detection are often unable to detect genuine 3D structural similarities. In particular, those methods that do not consider long-range interactions (i.e. side-chain to side-chain contacts), are unlikely to detect weak 3D structural similarities, since other residue by residue (i.e. one-dimensional) measures of structural similarity are not well conserved for many genuinely similar proteins.

Methods which thread protein sequences onto 3D structural templates using pair potentials (Sippl *et al.*, 1992; Jones *et al.*, 1992; Godzik *et al.*, 1993; Bryant & Lawrence, 1993), are likely to fare better, though all of these methods require that similar structures should have a reasonable proportion of interacting residues in common. The small fraction of residues common to the core of distantly related proteins (as few as 28.9%), and the even smaller fraction of common interacting residues (as few as 12%) suggests that many protein 3D structural similarities will be undetectable even by threading methods, since key interactions are likely to be modelled incorrectly. Our findings suggest that it is more general features of protein structure, such as having hydrophobic residues buried in the core of proteins, and polar residues on the surface, rather than particular residue–residue interactions that determine how well a particular sequence adopts a particular fold. If detection of similar folds having little in common outside of their core secondary structures is to become a reality, efforts should concentrate on such general principles, and on methods for modelling large loop regions that are likely to differ between similar 3D structures.

The results provide little insight as to whether structurally similar proteins have evolved by divergence or convergence. However, the fact that there is no detectable difference between pairs of structures that are functionally similar and those that are not (at a similar $\%I$) suggests that it may be impossible to discern divergence from convergence. Those proteins that were defined as type *B* similarities are often thought to have a common ancestor. For example, it seems very likely that the aspartic proteinase lobes (i.e. N and C-terminal domains in the

Table 5

How conservation of structural features varies across the type A, B *and* C *similarities shown in Figure 1*

Pair	Type	Equiv.	$\%I$	% Sec. id.	N_{core}	RMS	% Same acc.	% Shared	% Shared fav.
2FBL_L 1–109 & 7FAB_L 3–105	*A*	10S/6L	46.6	81.6	81	1.060	65.0	71.1	30.1
2FBJ_L 1–109 & 2CD4 1–95	*B*	9S/2L	16.6	67.7	67	1.253	61.6	47.5	16.7
2FBJ_L 1–109 & 1PLC all	*C*	7S/0L	5.1	49.5	38	2.657	46.5	28.7	7.4

Domains compared are given by their Brookhaven code, chain identifier and the range of residues used. Type gives the type of structural similarity as defined in the text. Equiv. gives the number of strands (S) and loops (L) common to the two structures. $\%I$ = percent sequence identity; % sec. id. = percent three-state secondary structure identity; N_{core} = number of common core residues; RMS = root mean square deviation for best fit of equivalent core C^{α} atoms; % same acc. = percent of residues having the same accessibility category; % shared = percent of residue–residue contacts common to the 2 structures; % shared fav. = percent of favourable residue–residue contacts common to the 2 structures. More details are given in the text.

Table 6

How the types of similarities differ with respect to the various structural parameters

Type	%*I*	% Sec. id.	% in common core	RMS	% Same acc.	% Shared	% Shared fav.
A	20.0–86.2	67.0–98.6	59.3–98.7	0.48–2.72	51.4–92.8	54.2–92.4	23.9–58.7
B	2.1–19.9	40.2–87.9	34.0–93.6	1.03–3.16	36.2–75.9	14.1–75.0	4.3–34.7
B (%$I \leqslant 15.8$)	2.1–15.8	40.2–87.9	34.0–93.6	1.03–3.16	36.2–75.9	14.1–70.5	3.3–32.5
C	1.3–15.8	41.1–85.2	28.9–73.7	1.48–3.17	27.5–67.5	12.1–62.0	3.2–34.1

%*I* = percent sequence identity; % sec. id. = percent three-state secondary structure identity; % in common core = number of common core residues; RMS = root mean square deviation for best fit of equivalent core C^{α} atoms; % same acc. = percent of residues having the same accessibility category; % shared = percent of residue–residue contacts common to the 2 structures; % shared fav. = percent of favourable residue–residue contacts common to the 2 structures. More details are given in the text.

eukaryotic structures) are related both to each other (i.e. by gene duplication or exon shuffling; see Blundell *et al.*, 1979) and to the single viral proteinase lobes that dimerise to form a similar structure (e.g. Lapatto *et al.*, 1989). However, their degree of structural and sequence conservation is low. If one argues that the proteinase lobes are related by divergence, then, based on the degree of structural and sequence similarity, one could argue the same for the quite obviously functionally dissimilar plastocyanin and Ig light chain variable domain shown in Figure 1. It would seem that both the sequence and structure of similar proteins can evolve beyond recognition even when function is conserved.

We thank Professor L. N. Johnson for providing a stimulating working environment. We are grateful to the Royal Commission for the Exhibition of 1851 and the Royal Society for support. We thank András Fiser for numerous helpful discussions and a critical review of the manuscript, and Drs J. Kuriyan, S. K. Burley, M. J. Adams and M. E. M. Noble for providing co-ordinates prior to deposition.

References

Adman, E. T. (1984). Structure and function of small blue copper proteins. In *Metallo-proteins: Metal Proteins with Redox Rules* (Harrison, P. M., ed.), pp. 1–42, Verlag Chemie, Basel.

Argos, P., Vingron, M. & Vogt, G. (1991). Protein sequence comparison: methods and significance. *Protein Eng.* **4**, 375–383.

Barton, G. J. (1990). Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.* **183**, 403–428.

Barton, G. J. & Sternberg, M. J. E. (1987). A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327–337.

Barton, G. J. & Sternberg, M. J. E. (1990). Flexible protein sequence patterns—a sensitive method to detect weak structural similarities. *J. Mol. Biol.* **212**, 389–402.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold, unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanovichi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Sewell, B. T. & McLachlan, A. D. (1979). Four-fold structural repeat in the acid proteases. *Biochem. Biophys. Acta*, **580**, 24–31.

Bordo, D. & Argos, P. (1990). Evolution of protein cores: constraints in point mutations as observed in globin tertiary structures. *J. Mol. Biol.* **211**, 975–988.

Bordo, D. & Argos, P. (1991). Suggestions for "safe" residue substitutions in site-directed mutagenesis. *J. Mol. Biol.* **217**, 721–729.

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

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257–264.

Bowie, J. U., Luthy, R. & Eisnenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading a protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Chelvanayagam, G., Heringa, J. & Argos, P. (1992). Anatomy and evolution of protein displaying the viral capsid jelly roll topology. *J. Mol. Biol.* **228**, 220–242.

Chothia, C. & Lesk, A. M. (1982). Evolution of proteins formed by β-sheets I. Plastocyanin and azurin. *J. Mol. Biol.* **160**, 309–323.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Craik, C. S., Rutter, W. J. & Fletterick, R. (1983). Splice junctions: association with variation in protein structure. *Science*, **220**, 1125–1129.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. Matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 345–358, National biomedical research foundation, Washington, DC.

Drenth, J., Low, B. W., Richardson, J. S. & Wright, C. S. (1980). The toxin-agglutinin fold. A new group of small protein structures organized around a four-disulphide core. *J. Biol. Chem.* **255**, 2652–2655.

Farber, G. K. (1993). An α/β barrel full of evolutionary trouble. *Curr. Opin. Struct. Biol.* **3**, 409–412.

Farber, G. K. & Petsko, G. A. (1990). The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* **15**, 228–234.

Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.

Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.

Godzik, A., Kolinski, A. & Skolnick, J. (1993). Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Hazes, B. & Hol, W. G. J. (1992). Comparison of the hemocyanin β-barrel with other greek key β-barrels: possible importance of the "β-zipper" in protein structure and folding. *Proteins: Struct. Funct. Genet.* **12**, 278–298.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Funct. Genet.* **17**, 49–61.

Holm, L. & Sander, C. (1993*a*). Globin fold in a bacterial toxin. *Nature (London)*, **361**, 309.

Holm, L. & Sander, C. (1993*b*). Structural alignment of globins, phycocyanins and colicin A. *FEBS Letters*, **315**, 301–306.

Hutchinson, E. G. & Thornton, J. M. (1992). The greek key motif: extraction, classification and analysis. *Protein Eng.* **6**, 233–245.

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

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature (London)*, **358**, 86–89.

Kabsch, W. & Sander, C. (1983). A dictionary of protein secondary structure. *Biopolymers*, **22**, 2577–2637.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.

Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., Whittle, P. J., Danely, D. E. & Geoghegan, K. F. (1989). X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homologue among retroviral enzymes. *Nature (London)*, **342**, 299–302.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.

Lesk, A. M. & Chothia, C. (1982). Evolution of proteins formed by β-sheets II. The core of the immunoglobulin domains. *J. Mol. Biol.* **160**, 325–342.

Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature (London)*, **339**, 31–36.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Luthy, R., McLachlan, A. D. & Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins: Struct. Funct. Genet.* **10**, 229–239.

McLachlan, A. D. (1979). Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.

Mitchell, E. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1989). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166.

Morris, A. L., MacArthor, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereo-chemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* **12**, 345–364.

Murzin, A. G., Lesk, A. M. & Chothia, C. (1992). β-trefoil fold patterns of structure and sequence in the Kunitz inhibitors interleukins-1β and 1α and fibroblast growth factors. *J. Mol. Biol.* **223**, 531–543.

Murzin, A. Z. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12**, 861–867.

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. Mol. Biol.* **48**, 443–453.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Nat. Acad. Sci., U.S.A.* **91**, 98–102.

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D–1D compatibility and detecting weak homologies. *Protein Eng.* **8**, 811–820.

Novotny, J., Bruccoleris, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* **177**, 787–818.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19–30.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, S. M., Ranken, S. M., Harel, M., Remington, S. J., Silman, I. & Schrag, J. (1992). The alpha/beta hydrolase fold. *Protein Eng.* **5**, 197–211.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.

Otto, J., Argos, P. & Rossmann, M. G. (1980). Prediction of secondary structural elements in glycerol-3-phosphate dehydrogenase by comparison with other dehydrogenases. *Eur. J. Biochem.* **109**, 325–330.

Ouzounis, O., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* **232**, 805–825.

Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Roy. Soc. London ser. B*, **241**, 132–145.

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

Pascarella, S. & Argos, P. (1992). A data bank merging protein structures and sequences. *Protein Eng.* **5**, 121–137.

Pastore, A. & Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins: Struct. Funct. Genet.* **8**, 133–155.

Pastore, A., Lesk, A. M., Bolognesi, M. & Onesti, S. (1988). Structural alignment and analysis of two distantly related proteins: aplysia limacina myoglobin and sea lamprey globin. *Proteins: Struct. Funct. Genet.* **4**, 240–250.

Pickett, S. D., Saqi, M. A. S. & Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction. Discrimination of the β/α-barrel fold. *J. Mol. Biol.* **228**, 170–187.

Rose, G. D. & Dworkin, J. E. (1989). The hydrophobicity profile. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), chapter 15, pp. 625–634, Plenum Press, New York, NY.

Rossmann, M. G. & Argos, P. (1976). Exploring structural homology in proteins. *J. Mol. Biol.* **105**, 75–95.

Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature (London)*, **250**, 194–199.

Russell, R. B. (1994). *STAMP User Manual.* University of Oxford.

Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.

Russell, R. B. & Barton, G. J. (1993*a*). An SH2-SH3 domain hybrid. *Nature (London)*, **364**, 765.

Russell, R. B. & Barton, G. J. (1993*b*). The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* **234**, 951–957.

Sali, 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. Mol. Biol.* **212**, 403–428.

Shindylav, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structure be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358.

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

Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.

Sippl, M. J., Hendlich, M. & Lackner, P. (1992). Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin β_4. *Protein Sci.* **1**, 625–640.

Sixma, T. K., Stein, P. E., Hol, W. G. & Read, R. J. (1993). Comparison of the b-pentamers of heat-labile enterotoxin and verotoxin-1: two structures with remarkable similarity and dissimilarity. *Biochemistry*, **32**, 191–198.

Swindells, M. B. & Thornton, J. M. (1993). A study of structural determinants in the interleukin-1 fold. *Protein Eng.* **6**, 711–715.

Swindells, M. B., Orengo, C. A., Jones, D. T., Pearl, L. H. & Thornton, J. M. (1993). Recurrence of a binding motif? *Nature (London)*, **362**, 299.

Taylor, W. R. (1986*a*). Classification of amino acid conservation. *J. Theoret. Biol.* **119** 205–218.

Taylor, W. R. (1986*b*). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Taylor, W. R. (1989). A template based method of pattern matching in protein sequences. *Prog. Biophys. Mol. Biol.* **54**, 159–252.

Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng.* **4**, 853–870.

Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348.

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1991). Prediction of progress at last. *Nature (London)*, **354**, 105–106.

Warme, P. K. & Morgan, R. S. (1978). A survey of amino acid side-chain interactions in 21 proteins. *J. Mol. Biol.* **118**, 289–304.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequence with the β/α-barrel fold. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 1379–1383.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247–259.

Edited by F. E. Cohen

(Received 23 May 1994; accepted 5 September 1994)