PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Protein Folds: Towards Understanding Folding from Inspection of Native Structures

Janet M. Thornton, David T. Jones, Malcolm W. Macarthur, Christine M. Orengo and Mark B. Swindells

Phil. Trans. R. Soc. Lond. B 1995 **348**, 71-79
doi: 10.1098/rstb.1995.0047

Protein folds: towards understanding folding from inspection of native structures

JANET M. THORNTON, DAVID T. JONES, MALCOLM W. MacARTHUR, CHRISTINE M. ORENGO AND MARK B. SWINDELLS*

Biomolecular Structure and Modelling Unit, Biochemistry and Molecular Biology Department, University College, Gower Street, London WC1E 6BT, U.K.

SUMMARY

Following a short summary of some of the principal features of folded proteins, the results of two complementary studies of protein structure are presented, the first concerned with the factors which influence secondary structure propensity and the second an analysis of protein topology. In an attempt to deconvolute the physical contributions to secondary structure propensities, we have calculated intrinsic ϕ,ψ propensities, derived from the coil regions of proteins. Comparison of intrinsic ϕ,ψ propensities with their equivalent secondary structure values show correlations for both helix and strand. This suggests that the local dipeptide, steric and electrostatic interactions have a major influence on secondary structure propensity. We then proceed to inspect the distribution of protein domain folds observed to date. Several folds occur very commonly, so that 46% of the current non-homologous database comprises only nine folds. The implications of these results for protein folding are discussed.

1. INTRODUCTION: A SUMMARY OF SOME MAJOR DETERMINANTS OF PROTEIN FOLDS AND FOLDING

It is not straightforward to infer information about the folding pathway from an inspection of the final native state of the protein. It is somewhat analogous to the problem facing astrophysicists, who must deduce information about the origins of the universe from the current state of the galaxies. Fortunately we are able to replay folding many times experimentally, but it is still important to use all the information that has been collected on protein structures to improve our understanding of the folding process. Before describing the results of two detailed studies we have recently completed at University College London, it is appropriate to review some of the basic principles which emerge from an inspection of the many structures in the Protein Structure Databank (PDB; Bernstein *et al.* 1977). In this first section we shall highlight principles drawn from work in our laboratory over the past few years, as well as many other groups.

Five essential observations can be made from inspection of protein structures.

1. All proteins exhibit a tightly packed hydrophobic core (Richards 1977; Hubbard *et al.* 1994). In a recent inspection of high resolution structures, Williams *et al.* (1994) found that on average water-sized cavities constitute only 1% of the volume of a protein. The small number of buried waters (about one per 27 residues on average) and cavities suggests that close packing, exclusion of water and burial of hydrophobic groups are major determinants of protein folding.

2. In X-ray derived structures, determined to high resolution, the ϕ,ψ and χ torsion angles are generally confined to low energy conformations (Morris *et al.* 1991). Indeed in such structures almost 90% of ϕ,ψ angles lie in only 14% of ϕ,ψ space. There are some well documented examples of distorted torsion angles (Herzberg & Moult 1991), but these are the exception rather than the rule. Thus, even in the complex interior of a folded structure, the local interactons are sufficiently strong to provide powerful restraints on torsional freedom. During folding, in the absence of strong tertiary interactions, these torsional angles are even more likely to adopt their preferred low energy conformers.

3. Potential hydrogen bond donors and acceptors are nearly always satisfied (McDonald & Thornton 1994). Only about 2% of main chain carbonyls and 6% of main chain amide groups fail to form hydrogen bonds to the protein or the solvent. Thus satisfaction of hydrogen bond potential is clearly an important constraint, which will also apply during folding. It is obviously energetically expensive to bury a potential donor or acceptor without satisfying its potential. As with the torsion angles, there are some exceptions, but these are rare and presumably must be compensated by other favourable interactions. It is this effect which makes the formation of secondary structure an obliga-

* Present address: Department of Molecular Design, Yamanouchi Pharmaceutical Co. Ltd., 21 Miyukigaoka, Tsakuba 305, Tokyo, Japan.

tory feature of compact globular structures, where the main chain is buried in the interior. Only by formation of the regular hydrogen bonds seen in sheets and helices can all the main chain groups be satisfied.

4. Side chain packing varies from random to rather specific, depending on the type of amino acids involved (Singh & Thornton 1990). Interactions between oppositely charged residues (e.g. arginine and aspartic acid) lead to definite preferred orientational patterns, which are clearly seen in proteins. The polar interactions also confer orientational preferences, although these are not so strong as those seen for the formally charged groups. In contrast the apolar interactions appear to be essentially randomly distributed in space. Provided these side chains are shielded from solvent and reasonably well packed they do not have any preferred spatial orientations. Interestingly, the aromatic ring sidechains behave more like polar residues than apolar. There are distinct preferences for negatively charged polar atoms (e.g. oxygens and sulphurs) to eschew the electronegative face of the aromatic ring and prefer to pack against the positively charged edge of the ring (Reid *et al.* 1985). Stacking interactions between aromatics are relatively rare, whereas the energetically more favourable edge–face interactions are much more common (Burley & Petsko 1985; Singh & Thornton 1985). Thus the interactions between side chains are not random, except for the apolar–apolar contacts, which may well dominate the core of the protein. However, specific packing will result from specific polar interactions especially hydrogen bonds, and these may well be vital in the folding process.

5. Protein structures are dominated by the secondary structures adopted on average by 60% of all residues. Hydrogen bonding interactions between strands dictate strand geometry. Similarly specific helix–helix and sheet–sheet interactions are favoured as determined by the requirement for close packing (Chothia & Finkelstein 1990).

These observations reveal much about the energetic forces which drive folding. Folding can be seen as a balance between satisfying the local conformational preferences and the global requirement to bury apolar sidechains, whilst satisfying almost all potential hydrogen bond donors and acceptors. Below we explore two aspects of protein structure with their implications for folding in more detail. First we analyse the preferred ϕ,ψ conformations for the 20 amino acids and then we consider the distribution of currently known structures among the possible protein topologies.

2. SECONDARY STRUCTURE PROPENSITIES

To facilitate a more thorough understanding of secondary structure propensities, we have developed a novel procedure for deconvoluting the competing factors which govern secondary structure formation. Our work is statistically based, but differs from previous reports (Chou & Fasman 1978) because we calculate the ϕ,ψ preferences in coil regions and separate them from all other structural data (figure 1*a*, *b*). By omitting regular interactions from residues located in helices and strands, it is possible to calculate intrinsic preferences for specific regions of ϕ,ψ space. We refer to these regions as a, b, p and B (see figure 1).

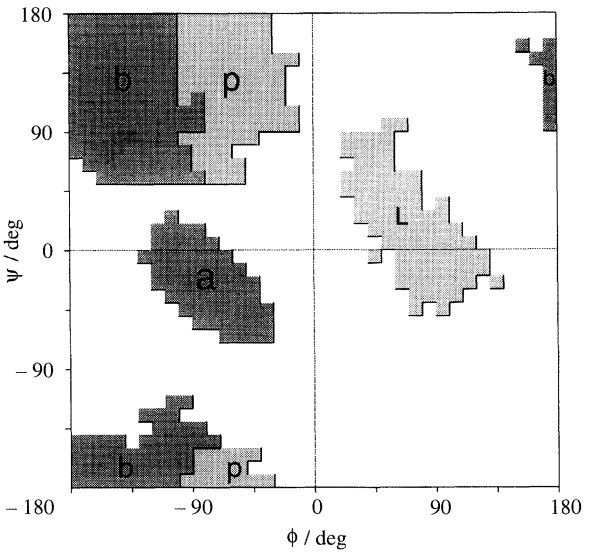

Figure 1. Digitized plot showing the ϕ,ψ regions used in this work: a, b and p. These follow the definitions of Efimov (1980) and Wilmot & Thornton (1990). The b and p regions collectively constitute the B region. Delineation of b and p regions is important because the fixed ϕ angle essentially excludes Pro from the b region. In contrast, other residues can occupy both the b and p regions, although strand residues prefer the b region. For completeness the aL and gL regions are also shown collectively (L).

(*a*) *Methods and data*

Using a dataset of 85 structures from the PDB, propensities for a/coil, b/coil, p/coil, B/coil, helix and strand were calculated (table 1).

Using alanine in the a/coil state as an example, intrinsic propensities for the coil state were calculated in the following manner.

$n(\mathrm{Ala})_{\mathrm{a/coil}}$ = number of alanine residues adopting an a region conformation when in coil,

$n(\mathrm{Ala})_{\mathrm{coil}}$ = total number of alanine residues in coil,

$N_{\mathrm{a/coil}}$ = total number of residues in coil,

N_{coil} = number of residues adopting an a region conformation when in coil,

$$P(\mathrm{Ala})_{\mathrm{a/coil}} = \{n(\mathrm{Ala})_{\mathrm{a/coil}}/n(\mathrm{Ala})_{\mathrm{coil}}\}/\{N_{\mathrm{a/coil}}/N_{\mathrm{coil}}\},$$

where $P(\mathrm{Ala})_{\mathrm{a/coil}}$ is the propensity for Ala to adopt the a/coil conformation.

In this manner, $P(\mathrm{Ala})_{\mathrm{a/coil}}$ measures the propensity for alanine to adopt a ϕ,ψ conformation within the a region, given that it is in the coil state. It gives no indication of the relative preferences for coil, strand and helix. The regular secondary structure propensities are calculated following the standard formalism of Chou & Fasman (1978) using the secondary structure definitions of Kabsch & Sander (1983).

For the analysis of χ_1 angles, we use the formalism: gauche plus $(g^+) = -60°$, trans $(t) = 180°$ and gauche minus $(g^-) = +60°$. A correction is applied for

Table 1. *Intrinsic propensities for a/coil, b/coil, p/coil, B/coil, and Chou & Fasman type propensities for α-helix and β-strand*

	intrinsic ϕ,ψ propensities					regular secondary structure propensities	
acid	a/coil	b/coil	p/coil	B/coil	other	α-helix	β-strand
Gly	0.33	0.34	0.31	0.32	3.80	0.41	0.64
Ala	1.23	0.78	1.32	1.09	0.39	1.47	0.79
Val	0.89	1.83	0.96	1.33	0.36	0.95	1.73
Leu	1.16	0.82	1.40	1.15	0.35	1.32	1.17
Ile	0.98	1.68	0.99	1.29	0.32	1.13	1.76
Phe	0.93	1.63	0.93	1.23	0.55	1.04	1.39
Tyr	0.85	1.46	1.12	1.26	0.59	0.88	1.52
Trp	1.17	0.90	1.24	1.09	0.48	1.05	1.25
Pro	1.00	0.10	2.29	1.35	0.14	0.46	0.42
Cys	0.87	1.34	1.32	1.33	0.41	0.89	1.18
Met	1.07	1.05	1.23	1.15	0.51	1.37	1.32
Ser	1.29	0.95	1.00	0.98	0.56	0.71	0.93
Thr	1.13	1.39	0.96	1.15	0.43	0.71	1.27
Lys	1.20	1.07	0.94	0.99	0.68	1.10	0.92
Arg	1.09	1.40	0.74	1.03	0.77	1.41	0.71
His	0.93	1.37	0.84	1.07	0.95	0.97	0.86
Asp	1.16	1.18	0.82	0.98	0.80	0.85	0.49
Asn	0.79	1.35	0.60	0.92	1.54	0.78	0.56
Glu	1.45	0.84	0.95	0.90	0.50	1.39	0.78
Gln	1.26	1.07	1.00	1.03	0.48	1.36	0.81

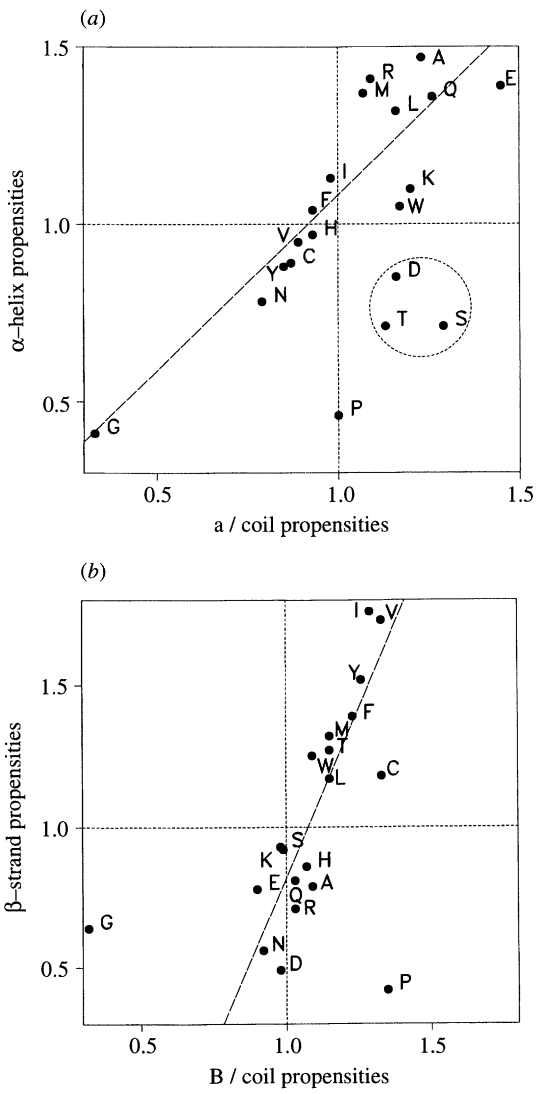

Figure 2. Graphs showing the propensities for (*a*) a/coil versus helix and (*b*) B/coil versus strand.

the anomaly in valine χ_1 classification. Thus valine (t, g^-, g^+) wells are listed as (g^+, t, g^-) in this paper. Calculations of χ_1 angle propensities within each region of the ϕ,ψ plot are implemented in a similar manner to those above. For example, the propensity of threonine to adopt the g^+ conformation, given that it is both in the a region and coil state can be described as:

$$P(\mathrm{Thr})_{\mathrm{g+/acoil}} = \{n(\mathrm{Thr})_{\mathrm{g+/acoil}}/n(\mathrm{Thr})_{\mathrm{acoil}}\}/\{N_{\mathrm{g+/acoil}}/N_{\mathrm{acoil}}\}.$$

Glycine and alanine, which do not have side chains, have been omitted.

(*b*) *Results*

It is clear that each residue type has intrinsic propensities for different regions of ϕ,ψ space, and that these values do not necessarily agree with the associated secondary structure propensities. In figure 2*a*, *b*, comparisons are made between the propensities for a/coil and helix, as well as B/coil and strand. Although it is not meaningful to compare the absolute values of the residue propensities with one another (because the a/coil and B/coil values are only calculated from the coil subset), relative comparisons can be made. The correlation between a/coil and helix values is relatively weak; Pearson correlation coefficients are 0.60 for all residues and 0.44 when Gly and Pro are excluded. In addition, several residues, especially Ser, Thr and Asp show markedly different rank ordering. If Ser, Thr, Asp, Gly and Pro are all excluded, the correlation coefficient increases to 0.81. In contrast, although the correlation coefficient between B/coil and strand propensities is low when all residues are considered (correlation coefficient = 0.53) it becomes much higher when Gly and Pro are excluded (correlation coefficient = 0.86).

These data suggest that the intrinsic ϕ,ψ preferences, and their compatibility with the observed secondary structure propensities, vary with the side-chain concerned. Intrinsic preferences must principally reflect side-chain interactions with the two local peptide units (Ralston & DeCoen 1974; Finkelstein & Ptitsyn 1976*a*, *b*; Zimmerman *et al.* 1977) and these should be evident in the χ_1 distributions. As expected from simple steric considerations, χ_1 values fall into three preferred zones ($+60°\ g^-$, $180°$ trans, $-60°\ g^+$) (Janin *et al.* 1978; McGregor *et al.* 1987; Ponder & Richards 1987; Dunbrack & Karplus 1994). Comparisons of the χ_1 preferences in a/coil and B/coil regions, are significantly different from those occuring in the equivalent regular secondary structures (see table 2).

How then can we rationalize these variations? Glu, Gln, Ser and Asp all have high a/coil propensities (figure 2*a*) and have a polar or charged oxygen acceptor, which can interact favourably with the main-chain NH groups. It is not surprising therefore that Asp and Ser χ_1 angles strongly favour the g^- state, as this conformation facilitates the electrostatic interaction (table 2). This preference is not observed for Glu and Gln because the long, flexible side-chain can form electrostatic interactions in other χ_1 conformers as well. Thr has a slightly lower a/coil propensity than Ser, even though it can stabilize the a/coil conformation via its hydroxyl, and has a χ_1 distribution which strongly favours the required g^+ conformer. This can be attributed to the branched C^β atom which causes steric clashes, similar to those observed in Val and Ile (see below). Consequently, the a/coil propensity is intermediate between the polar Ser and apolar Val and Ile.

Although Ser, Asp and Thr have high a/coil propensities, their helix propensities are low. This is because the g^- rotamer which stabilizes the a/coil conformation, is effectively forbidden in a helix, as the side-chain will interfere with the hydrogen bonds required for helix formation. In contrast, the side-chains of Glu and Gln are presumably sufficiently long and flexible to be displaced as the helix is formed. As a result, their high a/coil propensities are translated into high propensities for helix formation.

Both Leu and Ala exhibit reasonably high a/coil values and very high helix propensities. The most noticeable property of Leu is that, even in the a/coil conformation its χ_1 angle rarely adopts the g^- conformation, due to side-chain main-chain steric clashes (table 1). Because the g^- state is usually forbidden in helices anyway, due to unfavourable interactions with the previous helical turn, Leu can convert from a/coil to helix without the normal loss of side-chain entropy. Leu and Ala (which has no C^γ) are the only two amino acids for which this observation holds, and both have very high helix propensities. One other striking observation is that Leu and Ala are the only two amino acids which have a really strong preference for the p/coil region of ϕ,ψ space, rather than the b/coil region. This may also influence the strong helix propensities observed for Ala and Leu, as only a ψ angle rotation is required to move from the p region to the helix-forming a region.

Ile and Val, both have very low a/coil and helix propensities, and clearly cannot provide any electrostatic stabilization for the a/coil conformation. Inevitably this leads to a preference for B/coil. Within the B/coil region, Ile and Val both have a strong preference for b/coil, due to steric effects (between NH and C^γ atoms) which occur in the p/coil region when residues have a branched C^β atom. As a result, the preference for strand formation rather than helix is further enhanced. Although non-polar residues cannot stabilize the a/coil conformation, and in general have a low propensity for the 'aligned' dipole conformation in a/coil, their presence in helices is often obligatory, since helices require a hydrophobic face which can pack against the rest of the hydrophobic core.

As the correlation between B/coil and strand propensities is high (0.86 when Pro and Gly are omitted), it would appear that the intrinsic propensities for B/coil are a major factor in determining those for strand formation. In general, hydrophobic amino acids have the highest B/coil propensities. Of the polar residues, positively charged side-chains such as Arg, Lys and His have slightly higher B/coil propensities than Ser, Asp, Asn and Glu, although the differences are not pronounced. These elevated propensities for positively charged side-chains may reflect their ability to interact favourably with the main-chain CO of residue $i+1$, whereas the lower values for negatively charged residues will inevitably result from their high a/coil propensities. In an extended conformation there are no interactions between sequential side-chains, and their peptide groups lie in a plane which is almost perpendicular to the side-chain C^α–C^β bond. Thus there is also little interaction between side-chain and main-chain. This allows hydrophobic side-chains to be buried and main-chain polar groups to have their hydrogen bonding potentials satisfied through solvation.

(*c*) *Implications for folding*

What are the implications for folding of these empirical ϕ,ψ propensities, which derive from local interactions between the side chain and main chain atoms. In the initial stages of folding, the polypeptide chain will adopt a random coil conformation. However, even at this stage the sequence will influence the coil conformation adopted as driven by the intrinsic ϕ,ψ propensities. Thus Leu will predominantly adopt the a/coil conformer, whereas Ile and Val will prefer the b/coil state. As the chain folds the medium range interactions ($i \ldots . i+3,4$) which occur in a helix will come into effect, modulating the intrinsic ϕ,ψ propensities. Thus Leu, which usually adopts the $g+\chi_1$ conformer favoured in a helix, is most readily 'accepted' into a helix, whereas Ser and Thr have difficulty in forming a helix because of the local side chain–backbone interactions. Similarly, the branched apolar amino acids are further discriminated against by the helix geometry. The close contacts in a helix normally distort the χ_1 angle to alleviate the strain, but this is not possible for the branched C^β sidechains.

As well as influencing secondary structure formation,

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 2. χ_1 *angle distributions and propensities for residues in different regions of* ϕ,ψ *space and secondary structure*

(Distributions are shown in normal type and propensities in **bold**. Ala and Gly have no χ angles and those of Pro are atypical. For χ_1 angles we use the formalism: gauche plus(g^+) = 60 °, trans(t) = 180 ° and gauche minus(g^-) = +60 °. A correction is applied for the anomaly in valine χ_1 classification. Thus, valine(t, g^-, g^+) wells are listed as (g^+, t, g^-) in this paper. Calculations of χ_1 angle propensities within each region of the ϕ,ψ plot are implemented in a similar manner to those in table 1. For example, the propensity for threonine to adopt the g^+ conformation, given that it is both in the **a** region and coil, is described as: $P(\mathrm{Thr})_{g+/acoil} = \{n(\mathrm{Thr})_{g+/acoil}/n(\mathrm{Thr})_{acoil}\}/(N_{g+/acoil}/N_{acoil})$.)

amino acid	a/coil			b/coil			p/coil			B/coil (b+p)			α-helix			β-strands		
	g^-	t	g^+	g^-	t	g^+	g^-	t	g^+	g^-	t	g^+	g^-	t	g^+	g^-	t	g^+
Val	46	20	48	58	17	73	26	8	69	84	25	142	32	16	301	100	42	350
	1.24	**1.42**	**0.77**	**1.45**	**0.40**	**1.11**	**1.60**	**0.24**	**1.28**	**1.57**	**0.33**	**1.17**	**1.27**	**0.15**	**1.38**	**1.19**	**0.29**	**1.32**
Leu	1	30	143	2	15	61	1	48	127	3	63	188	1	190	344	13	171	181
	0.01	**1.39**	**1.49**	**0.09**	**0.67**	**1.77**	**0.04**	**0.85**	**1.38**	**0.06**	**0.81**	**1.53**	**0.03**	**1.18**	**1.03**	**0.21**	**1.61**	**0.92**
Ile	35	16	22	25	5	49	8	7	47	33	12	96	19	18	261	46	43	269
	1.47	**1.77**	**0.55**	**1.17**	**0.22**	**1.40**	**0.82**	**0.35**	**1.45**	**1.10**	**0.28**	**1.41**	**0.88**	**0.20**	**1.40**	**0.75**	**0.41**	**1.39**
Phe	7	12	58	16	19	51	3	19	43	19	38	94	3	131	79	54	52	113
	0.28	**1.26**	**1.37**	**0.69**	**0.77**	**1.34**	**0.29**	**0.91**	**1.27**	**0.59**	**0.83**	**1.29**	**0.19**	**2.04**	**0.59**	**1.44**	**0.82**	**0.96**
Tyr	10	13	50	14	24	41	7	42	31	21	66	72	6	98	72	48	62	125
	0.42	**1.44**	**1.25**	**0.66**	**1.06**	**1.17**	**0.55**	**1.64**	**0.74**	**0.62**	**1.36**	**0.94**	**0.47**	**1.85**	**0.65**	**1.20**	**0.91**	**0.99**
Trp	8	4	25	5	6	7	2	15	16	7	21	23	7	39	29	16	15	39
	0.66	**0.87**	**1.23**	**1.03**	**1.16**	**0.88**	**0.38**	**1.42**	**0.93**	**0.64**	**1.35**	**0.93**	**1.27**	**1.71**	**0.61**	**1.34**	**0.74**	**1.03**
Pro	117	0	79	11	0	0	187	0	188	198	0	188	30	0	76	39	0	36
	1.83	**0.00**	**0.73**	**3.40**	**0.00**		**3.16**	**0.00**	**0.96**	**2.40**	01.00	**1.01**	**3.92**	**0.00**	**1.14**	**3.04**	**0.00**	**0.89**
Cys	12	1	30	14	14	14	2	21	32	16	35	46	5	28	54	10	34	45
	0.85	**0.19**	**1.27**	**1.23**	**1.16**	**0.75**	**0.23**	**1.19**	**1.12**	**0.77**	**1.19**	**0.98**	**0.80**	**1.07**	**0.99**	**0.66**	**1.32**	**0.94**
Met	4	3	25	5	6	9	1	11	19	6	17	28	3	34	92	15	29	53
	0.38	**0.76**	**1.42**	**0.93**	**1.04**	**1.02**	**0.20**	**1.11**	**1.18**	**0.55**	**1.10**	**1.14**	**0.32**	**0.87**	**1.13**	**0.90**	**1.03**	**1.01**
Ser	187	12	81	61	45	23	63	71	47	124	116	70	86	53	106	88	89	71
	2.05	**0.35**	**0.53**	**1.74**	**1.20**	**0.40**	**2.19**	**1.22**	**0.50**	**1.87**	**1.22**	**0.46**	**4.86**	**0.72**	**0.69**	**2.08**	**1.24**	**0.53**
Thr	152	3	27	86	20	37	57	9	65	143	29	102	50	6	155	78	39	175
	2.54	**0.13**	**0.27**	**2.23**	**0.49**	**0.58**	**2.76**	**0.21**	**0.95**	**2.45**	**0.35**	**0.77**	**3.28**	**0.09**	**1.17**	**1.56**	**0.46**	**1.11**
Lys	19	34	119	14	21	64	8	37	71	22	58	135	21	135	168	19	94	98
	0.32	**1.53**	**1.20**	**0.51**	**0.72**	**1.42**	**0.43**	**0.98**	**1.15**	**0.47**	**0.87**	**1.27**	**0.88**	**1.36**	**0.81**	**0.53**	**1.53**	**0.86**
Arg	11	26	70	12	23	52	5	19	37	17	42	89	10	149	142	20	35	62
	0.31	**1.96**	**1.19**	**0.51**	**0.92**	**1.35**	**0.52**	**0.97**	**1.16**	**0.54**	**0.93**	**1.25**	**0.46**	**1.64**	**0.75**	**0.99**	**1.02**	**0.98**
His	16	8	35	9	13	33	6	19	20	15	32	53	8	46	56	15	26	34
	0.83	**1.10**	**1.08**	**0.61**	**0.82**	**1.36**	**0.85**	**1.32**	**0.85**	**0.70**	**1.05**	**1.10**	**1.01**	**1.39**	**0.81**	**1.17**	**1.19**	**0.84**
Asp	79	11	142	26	103	23	22	67	51	48	170	74	11	40	197	11	65	36
	1.03	**0.38**	**1.10**	**0.63**	**2.36**	**0.34**	**1.00**	**1.49**	**0.70**	**0.77**	**1.91**	**0.53**	**0.61**	**0.53**	**1.26**	**0.57**	**2.00**	**0.60**
Asn	39	9	84	33	73	37	16	35	32	49	108	69	4	32	152	9	51	43
	0.90	**0.55**	**1.15**	**0.85**	**1.76**	**0.58**	**1.21**	**1.30**	**0.73**	**1.01**	**1.56**	**0.63**	**0.29**	**0.57**	**1.29**	**0.51**	**1.71**	**0.77**
Glu	39	33	112	12	12	43	11	34	54	23	46	97	21	132	240	19	80	72
	0.64	**1.43**	**1.10**	**0.65**	**0.61**	**1.43**	**0.68**	**1.04**	**1.01**	**0.64**	**0.89**	**1.18**	**0.73**	**1.10**	**0.96**	**0.64**	**1.58**	**0.77**
Gln	11	21	67	7	9	37	3	28	36	10	37	73	11	93	137	14	50	47
	0.33	**1.68**	**1.21**	**0.48**	**0.58**	**1.55**	**0.28**	**1.30**	**1.03**	**0.39**	**1.00**	**1.25**	**0.63**	**1.28**	**0.90**	**0.74**	**1.55**	**0.79**

the intrinsic ϕ,ψ propensities also help to determine the conformation of the coil regions. Folding can be seen as a balance between satisfying these local conformational preferences and the global requirement to bury apolar sidechains, while satisfying almost all potential hydrogen bond donors and acceptors.

3. PROTEIN FOLDS: OBSERVED DISTRIBUTION OF PROTEINS AMONG THE FOLD FAMILIES

Having assessed the influence of local interactions on secondary structure and folding, we now consider tertiary structure and an analysis of the global topologies observed for polypeptide chains. It is well established that protein domains, having more than 30% of their amino acid sequences in common, will adopt the same three-dimensional structure. The core of these structures is usually well conserved, whereas their surfaces may incorporate many differences in loop insertions or deletions. Furthermore, as the number of known structures increases, it is apparent that some structures adopt the same fold, even though their sequences are apparently completely dissimilar. We wished to consider these observations in more detail, especially with respect to the implications for the folding process. Therefore we have clustered all the protein structures in the PDB, using an automated procedure to compare all structures, which calculates a quantitative measure of their structural similarity.

(*a*) *Comparison of protein structures*

All the proteins in the April 1993 PDB including prereleases were clustered using the method of Orengo *et al.* (1993), which includes sequence and structure comparisons. Initially all the protein sequences are compared using the standard Needleman and Wunsch algorithm (Needleman & Wunsch 1970) and similarity is measured by % sequence identity. The protein sequences are then clustered into families which show a clear sequence relationship (more than 30% sequence identity to at least one other member of the family). Representative structures of each family are then compared using the SSAP algorithm (Taylor & Orengo 1989), which finds similar regions by comparing the structural environment of each residue. The optimal alignment is achieved by a complex double dynamic programming process, which allows for insertions and deletions between the structures. The program returns a SSAP score, which is normalized to be in the range 0–100 independent of the protein size. The proteins are then clustered into structural families, according to their SSAP scores. Two structures were deemed to adopt the same fold if their SSAP score was > 70 and, to ensure global similarity, at least 70% of the larger protein must be equivalenced against the smaller. It is important to realize that, as with sequence similarities, structural similarities form a continuum and the cutoffs used are necessarily arbitrary. These cutoffs were chosen to reflect the consensus view of when two structures are similar, as found in the literature.

(*b*) *Homologous and analogous folds*

In comparing structures it became clear that there were two situations in which we found that a pair of structures looked similar. In the first, the two proteins have the same fold, and similar functions, even though their sequences are dissimilar. For example haemoglobin (PDB code-1eca) and myoglobin (1mbd) have a SSAP score = 85 when compared, even though their sequences show only 15% identity. It is generally agreed that these proteins have arisen by divergent evolution from a common ancestor, and can be considered to be homologous and belong to the same superfamily. When the structures of such proteins were compared we found that their comparison always yielded a SSAP score > 80. In clustering the databank we wished to signify that these structures are homologous, and therefore consider that they belong to the same hyperfamily (i.e. the same superfamily, but extended to include functionally and structurally similar proteins). In contrast there are other examples where two proteins have a similar structure, yet show no sequence or functional similarity. For example, haemoglobin (1eca) and colicin (1colA) have a sequence identity of 5%, no functional similarity (colicin is a protein which penetrates cell membranes and ultimately causes cell death) and yet have the same protein fold and give a SSAP score = 75. Such similarities will be described as analogous folds, and may have arisen by divergent or convergent evolution. Empirically we have found that these proteins yield a SSAP score > 70 but < 80, being in general less similar than functionally related proteins.

(*c*) *Protein superfamilies and fold families*

Thus the proteins in the PDB were clustered by their sequences and structures. Sequence analysis allowed recognition of the superfamilies. These families were then expanded by structural and functional comparisons to include proteins which have the same fold and similar functions, which would generally be considered to be homologous. These clusters were called hyperfamilies. The hyperfamilies could then be clustered into fold families, bringing together proteins which have the same fold but a different function and no obvious sequence similiarity. Thus we have the following hierarchy in the clustering procedure: (i) superfamilies, recognized only from sequence (> 30% identity); (ii) hyperfamilies, very similar structures with SSAP score > 80 and functional similarity; and (iii) fold families, similar folds with SSAP score > 70 but different function.

Using these criteria all the single domain proteins in the PDB were clustered into 392 superfamilies, 274 hyperfamilies and 206 fold familes, as illustrated in figure 3. From this analysis it is apparent that although the PDB is growing rapidly, we still only have a limited number of 'independent' protein structures and still fewer unique domain folds. Furthermore the distribution of the structures between the different fold families is shown in figure 4. This plot is far from a random distribution (Orengo *et al.* 1994). We observe

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

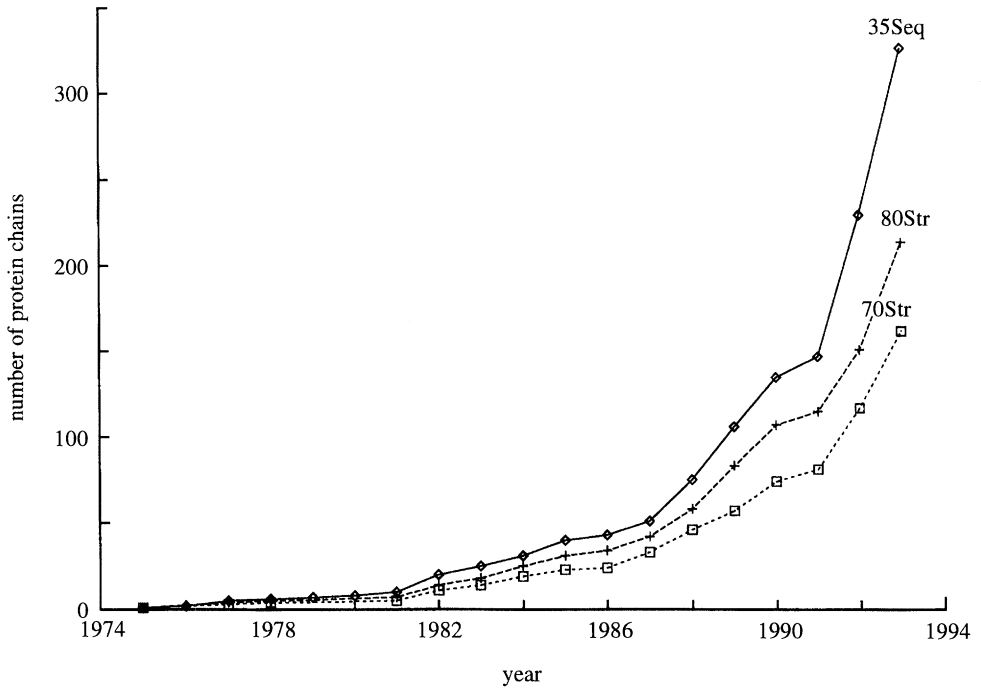

Figure 3. Increase in the number of protein structures deposited each year in the PDB, grouped by sequence family (35SEQ); hyperfamily family (SSAP score > 80) and fold family (SSAP score > 70).

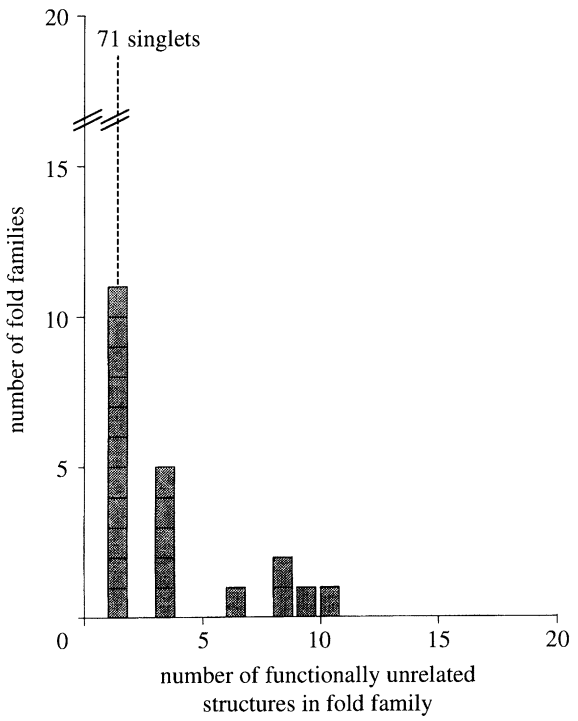

Figure 4. Distribution of current non-homologous structures (< 25% sequence identity and < 70 SSAP score and no functional similarity) among the different folds. The nine superfolds contain between 3–10 representatives, whereas all the remaining folds are singlets.

nine fold families, for which there are between 3–11 examples and 71 folds for which there is only one singlet example. These very common folds, which are termed superfolds, represent 46% of all non-homologous proteins in the PDB.

The observation that all folds are not equally populated is not new (Ptitsyn & Finkelstein 1980; Finkelstein & Ptitsyn 1987). For example, in the four-helix bundle proteins, the up-down-up-down topology occurs frequently, whereas the more complex up-up-down-down fold is relatively rare. Similarly in the beta-sandwich structures, the immunoglobulin fold is very common. In contrast many other folds have only been observed once, although they were among the first folds to be determined (i.e. all examples of the singlet folds in the database are clearly related, as shown either by sequence or similarity of structure and function, e.g. lysozyme and ribonuclease A).

The existence of superfolds has several possible implications for proteins folding. These folds may represent extra stable folds, which have diverged from a common ancestor and, despite extensive changes to their sequences, are resistant to changes in topology. Any residual sequence patterns are commonly undetectable and the 'original' function may have changed. Alternatively these superfolds may reflect the accidental convergence of disparate protein chains to the same topology and will have diverse functions. Regardless of the origins of these folds, it is apparent that they are compatible with a much larger set of sequences, compared to one of the singlet folds (Finkelstein 1994).

The superfolds are shown in figure 5 and it is immediately apparent that they all exhibit simple topologies, with a high percentage of sequential secondary structures which lie adjacent in the tertiary fold. Thus it is tempting to suggest that these proteins are so common because they can fold up more easily or rapidly than other more complex topologies. In an all sequential structure, such as a TIM barrel, there are in principle no requirements for an obligate pathway, as the structure could nucleate anywhere and fold up in any order. If the relative stability of any part of the structure were altered by mutation, thus affecting the

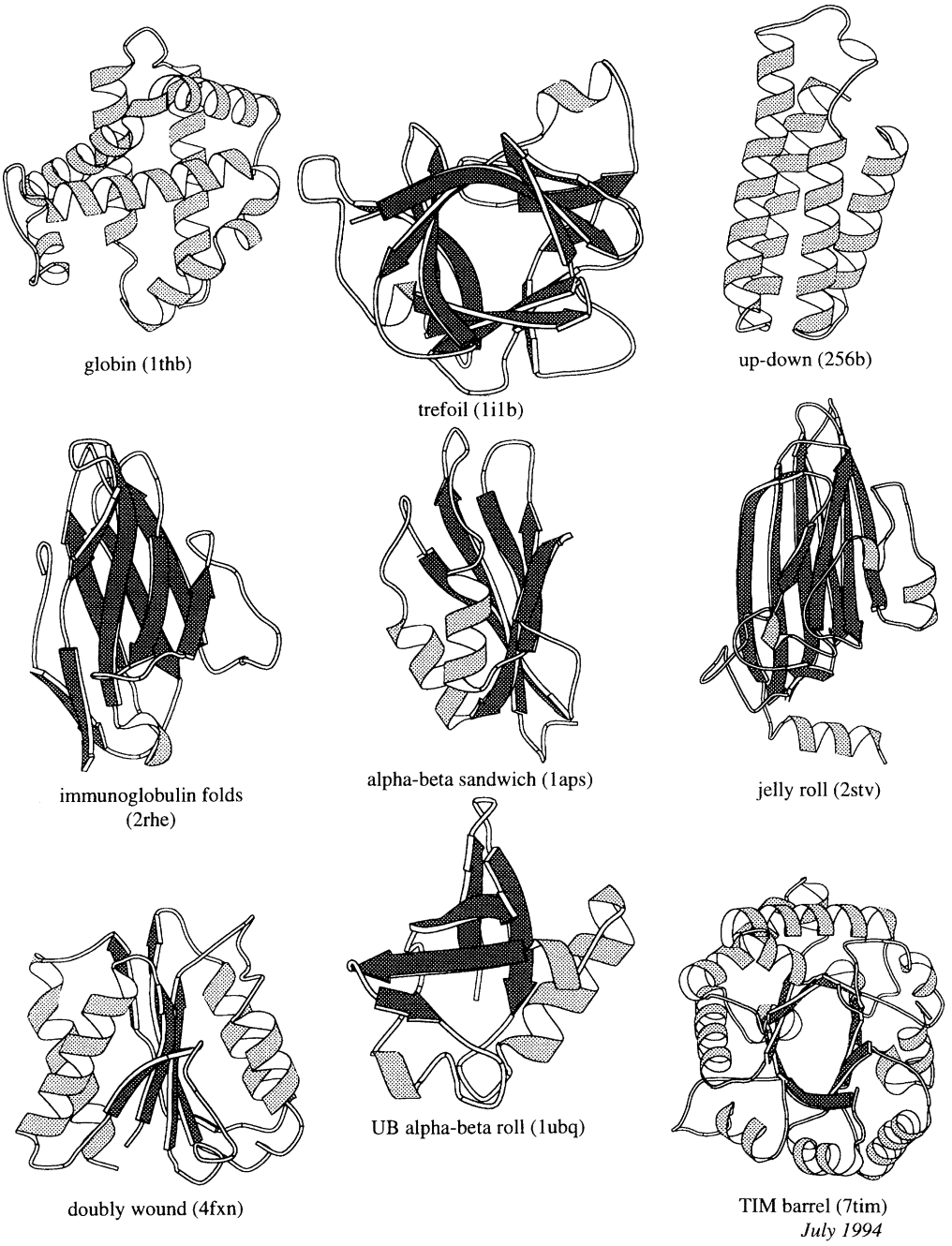

Figure 5. The nine superfold structures. The specific protein shown is indicated in brackets by the PDB code.

local rate of folding, this need not affect the final structure, just the order of folding. If this hypothesis were correct we would expect that the all sequential structures observed to date (the up-down barrels, the β-propeller structures and the β-solenoid fold) would be common. The recently observed 'holy' proteins (Dijkstra & Thunnissen 1994; Hofsteenge 1994) are also all-sequential.

Many of the singlet folds are stabilized by disulphide bridges, and denature if the disulphides are reduced. This adds sensitive hot-spots to the structure, with the requirement for conservation of the disulphide, if the native fold is to be maintained. Without these additional constraints these folds may not be sufficiently stable to withstand random mutagenesis. Members of these fold families can often be identified as related as the pattern of cystines is conserved.

4. CONCLUSION

In this paper we have concentrated on two aspects of protein folding, from the details of the ϕ,ψ distribution to the gross topology of a protein structure. As we discover more about the details of protein structures at all levels of the structural hierarchy, we are better able to understand the folding process and make hypotheses which can be tested experimentally. Clearly the folds we observe reflect the complex interplay of the weak non-covalent interactions between the side chains and the intrinsic conformational preferences of the 20 amino acids in a water environment. It is still not clear whether we need to understand the folding pathway in order to be able to predict protein structure from sequence. However, with current progress in determining folding pathways, it may soon be possible to

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

understand the folding process in much more detail and perhaps model this complex macromolecular process and predict its endpoint.

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. 1977 The Protein Data Bank: a computer based archival file for macromolecular structures. *J. molec. Biol.* **122**, 535–542.

Burley, S.K. & Petsko, G.A. 1985 Aromatic-aromatic interactions: a mechanism of protein structure stabilisation. *Science,Wash.* **229**, 23–29.

Chothia, C. & Finkelstein, A. 1990 Tha classification and origins of protein folding patterns. *A. Rev. Biochem.* **59**, 1007–1039.

Chou, P.Y. & Fasman, G.D. 1978 Predictions of the secondary structure of proteins from their amino acid sequence. *Adv. Enzmol.* **47**, 45–148.

Dijkstra, B.W. & Thunnissen, A.-M.W.H. 1994 'Holy' proteins II: the soluble lytic transglcosylase. *Curr. Opin. Struct. Biol.* **4**, 810–813.

Dunbrack, R. & Karplus, M. 1994 Conformational analysis of the backbone dependent rotamer preferences of protein side-chains. *Nature Struct. Biol.* **1**, 334–339.

Efimov, A.V. 1980. Standard conformations of polypeptide chains in irregular regions of proteins. *Molec. Biol. Moscow* **20**, 208–216.

Finkelstein, A.V. & Ptitsyn, O.B. 1976*a* Theory of protein molecule organisation. I. Thermodynamic parameters of local secondary structures in the unfolded protein chain. *Biopolymers* **16**, 469–495.

Finkelstein, A.V. & Ptitsyn, O.B. 1976*b* A theory of protein molecule self organisation. IV. Helical and irregular local structures of unfolded protein chains. *J. molec. Biol.* **103**, 15–24.

Finkelstein, A.V. & Ptitsyn, O.B. 1987 Why do globular proteins fit the limited set of folding patterns. *Prog. Biophys. molec. Biol.* **50**, 171–190.

Herzberg, O. & Moult, J. 1991 Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins* **11**, 223–229.

Hofsteenge J. 1994 'Holy' proteins I: ribonuclease inhibitor. *Curr. Opin. Struct. Biol.* **4**, 807–809.

Hubbard, S.J., Gross, K.H. & Argos, P. 1994 Intramolecular cavities in globular proteins. *Protein Eng.* **7**, 613–626.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. 1978 Conformation of amino acid side-chains in proteins. *J. molec. Biol.* **125**, 357–386.

Kabsch, W. & Sander, C. 1983 Dictionary of protein secondary structure. *Biopolymers* **22**, 2577–2637.

McDonald, I.K. & Thornton, J.M. 1994 Satisfying hydrogen bonding potential in proteins *J. molec. Biol.* **238**, 777–793.

McGregor, M.J., Islam, S.A. & Sternberg, M.J.E. 1987 Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. molec. Biol.* **198**, 295–310.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. & Thornton, J.M. 1992 Stereochemical quality of protein structure coordinates. *Proteins* **12**, 345–364.

Needleman, S.B. & Wunsch, C.D. 1970 A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. molec. Biol.* **48**, 433.

Orengo, C.A. & Taylor, W.R. 1993 A local alignment method for protein structure motifs. *J. molec. Biol.* **233**, 488–497.

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.

Orengo, C.M., Jones, D.T. & Thornton, J.M. 1994 Protein superfamilies and domain superfolds. *Nature, Lond.* **372**, 631–634.

Ponder, J.W. & Richards, F.M. 1987 Tertiary templates for proteins. *J. molec. Biol* **193**, 775–791.

Ptitsyn, O.B. & Finkelstein, A.V. 1980 Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding. *Q. Rev. Biophys.* **13**, 339–386.

Richards, F.M. 1977 Areas, volumes, packing and protein structures. *A. Rev. biophys. Bioeng.* **6**, 151–176.

Ralston, E. & DeCoen, J.-L. 1974 Folding of polypeptide chains induced by the amino acid side-chains. *J. molec. Biol.* **83**, 393–420.

Reid, K.S.C., Lindley, P.F. & Thornton, J.M. 1985 Sulphur aromatic interactions in Proteins. *FEBS Lett.* **190**, 209–213.

Singh, J. & Thornton, J.M. 1990 SIRIUS An automated method for the analysis of the preferres packing arrangements between protein groups. *J. molec. Biol.* **211**, 595–615.

Singh, J. & Thornton J.M. 1985 The Interaction between phenylanine rings in proteins. *FEBS Lett.* **191**, 1–6.

Taylor, W.R. & Orengo, C.A. 1989 Protein structure alignment. *J. molec. Biol.* **208**, 1–22.

Wilmot, C.M. & Thornton, J.M. 1990 β-turns and their distortions: a proposed new nomenclature. *Protein Eng.* **3**, 479–493.

Williams, M.A. Goodfellow J.M. & Thornton J.M. 1994 Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **3**, 1224–1235.

Zimmerman, S.S. Pottle, M.S. Némethy, G. & Scheraga, H.A. 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules* **10**, 1–9.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES