Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction

JOHN OVERINGTON, MARK S. JOHNSON, ANDREJ ŠALI AND TOM L. BLUNDELL†

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

SUMMARY

The pattern of residue substitution in divergently evolving families of globular proteins is highly variable. At each position in a fold there are constraints on the identities of amino acids from both the three-dimensional structure and the function of the protein. To characterize and quantify the structural constraints, we have made a comparative analysis of families of homologous globular proteins. Residues are classified according to amino acid type, secondary structure, accessibility of the sidechain, and existence of hydrogen bonds from sidechain to other sidechains or peptide carbonyl or amide functions. There are distinct patterns of substitution especially where residues are both solvent inaccessible and hydrogen bonded through their sidechains. The patterns of residue substitution can be used to construct templates or to identify 'key' residues if one or more structures are known. Conversely, analysis of conversation and substitution across a large family of aligned sequences in terms of substitution profiles can allow prediction of tertiary environment or indicate a functional role. Similar analyses can be used to test the validity of putative structures if several homologous sequences are available.

1. INTRODUCTION

Evolutionally related proteins have sequences that usually adopt similar tertiary structures. We assume that random mutations in the genes have been expressed in the protein during evolution. If the mutant protein folds and functions satisfactorily, then the organism can be either selected for or neutrally accumulated in the population; the mutation is accepted. If the mutation, for example, destabilizes the fold, removes a catalytically active residue, alters substrate binding, affects the protein half-life, etc. then the organism may be selected against; the mutation is likely to be rejected. The structural basis of the acceptance or rejection of a mutation is not fully understood but it may sometimes be appreciated once the tertiary structure and modes of interaction of a protein are known.

The stability of a protein is strongly influenced by exclusion of solvent from non-polar sidechains to give a close-packed hydrophobic core. Consequently solvent-inaccessible residues have a lower rate of acceptance of mutations than those on the surface (see, for example, Hubbard & Blundell (1987); Chothia & Lesk (1986); Lim & Sauer (1989)). Protein stability is also dependent on the formation of inter-residue hydrogen bonds. Highly conserved residues include buried, hydrogen-bonded polar residues (Bajaj & Blundell 1984), for example the conserved threonine of the Asp-Thr-Gly fingerprint for aspartic proteinases (Pearl & Blundell 1984). Secondary structure also provides constraints on sequence variability; α-helices, β-strands and coil regions have preferred composition patterns as described by Chou & Fasman (1974*a*) and Levitt (1978). The turns between secondary structural elements are often formed by residues with a positive main-chain ϕ angle. Because of the β-carbon of most sidechains this is a 'forbidden' conformation, but often occurs for glycine. Under certain conditions asparagine, aspartic acid, and serine can also adopt this conformation (see, for example, Nicholson *et al.* (1989)).

There is a large but rather subjective body of knowledge concerning invariance and conservative variation in the evolution of proteins (see, for example, Taylor (1986)). However this has not been properly characterized in terms of structural parameters. Thus the objective of this study was to establish the nature of the structural constraints that lead to invariance or conservative variation at certain topologically equivalent positions in families of proteins. The analysis depends on a systematic approach to the comparison of sequences and three-dimensional structures, which is now available in a new computer program, COMPARER (Šali & Blundell 1990). This program has been used to compare families of proteins including the globins, serine and aspartic proteinases, phospholipases and immunoglobulin and γ-crystallin domains. These

† To whom correspondence should be addressed.

proteins have coordinates available from the Brookhaven Protein Databank (Bernstein *et al.* 1977) based on high-resolution X-ray analysis.

In this paper we describe distinct patterns of amino acid substitution that characterize specific structural environments. We show that the substitution patterns have several applications. For example, the substitution patterns predict sequence variability at each position in a fold and so allow the construction of a sequence template for a tertiary structure or identify 'key' residues for a motif. In a similar way they allow flexible alignment of homologous sequences and three-dimensional structures. We show how they may be used to predict the local tertiary structure from the pattern of substitution across a family of aligned sequences and to test the validity of models when several homologous sequences are available. For some three-dimensional structures the invariance or conservative variation may not be predicted by the structural environment. In such cases residues involved are likely to express their structural roles in catalytic activity, substrate, cofactor or other ligand binding.

2. COMPARISON OF PROTEIN FAMILIES

We selected families of proteins (see table 1), for which three or more members have three-dimensional structures defined at high resolution by X-ray crystallography and for which there is a range of pairwise sequence identities, typically between 20% and 50%; we did not attempt to segregate the data according to levels of sequence similarity. These families include similar sequences (84% identical for the pair 1BP2 and 1P2P in the phospholipase set) and more distant ones (16% for the pair 1LH1 and 1ECD in the globin set). The selected families have a variety of secondary structures. For example, the globins are α-helical, the immunoglobulins and serine proteinases are mainly antiparallel β-sheet structures, and the aspartic proteinases are a mixture of parallel and antiparallel β-strands. The phospholipases are mainly α-helical with a small amount of β-sheet.

The structural alignment (figure 1), which is the most critical step in the analysis, was achieved with the program COMPARER (Šali & Blundell 1990). In this approach the protein is defined as elements that can exist at several levels in the hierarchical organization of protein structure: residue, secondary structure, supersecondary structure, motif, etc. Every element is associated with several features that may indicate a common fold. At each level of the hierarchy, the compared features can be properties (sequence identity, hydrophobicity, local conformation, solvent accessibility, main-chain dihedral angles, position in space, etc.) or relations (i.e. hydrogen bonds and van der Waals' interaction). Equivalent properties concerning higher levels of structure are also considered. Measures of the differences in these properties are accumulated to define a residue-by-residue mass matrix that is then used in the familiar dynamic programming algorithms to produce an optimal alignment. In addition, the inclusion into this mass matrix of information about specific relations such as hydrogen bonds and local packing is achieved by a simulated annealing alignment for these relations (Šali & Blundell 1990). In this way patterns of hydrogen bonds or van der Waals' interactions are compared and equivalenced.

3. SELECTION OF STRUCTURAL PARAMETERS

The selection of parameters was prejudiced by previous analyses in the laboratory and preconceptions as to the important features of protein folding. They included the following.

(i) Residue type for each of twenty amino acids.

(ii) Main-chain conformation and secondary structure were classified as either α-helical, β-strand, positive ϕ or irregular (coil). Residues with positive ϕ were assigned first; α-helices and β-strands were then defined by using the DSSP program of Kabsch & Sander (1983). Finally, residues as yet undefined were classified as coil.

(iii) Solvent accessibility. A residue was defined as inaccessible if its sidechain had a relative accessibility of less than 7% (Hubbard & Blundell 1987). Accessibilities were calculated by the method of Lee & Richards (1971) by using a probe radius of 1.4 Å†. Calculations were usually done on the entire molecule. However, for the immunoglobulins, accessibilities were calculated for isolated domains, and for the γ-crystallins, they were calculated for globular domains comprising two Greek-key motifs. Prosthetic groups and ligands were omitted from the calculations.

(iv) Hydrogen bonds from a sidechain at position i to residues other than those in the $i-1$, i or $i+1$ positions were examined. These were divided into three classes including hydrogen bonds between two sidechains, between sidechain and main-chain carbonyl (CO) and between sidechain and main-chain amide proton (NH). As sidechain atoms are generally not well positioned by crystallography and not all hydrogen atom positions are fixed by the positions of the heavier atoms, hydrogen-bond formation was defined on the criterion of a donor–acceptor distance $\leqslant 3.5$ Å (Baker & Hubbard 1984); angular criteria were not considered.

4. COMPOSITION OF THE SAMPLE

The relative abundance of each amino acid is similar to that found in other studies. The commonest amino acids are serine and glycine and the rarest, histidine, tryptophan and methionine. The secondary structure shows a sample bias towards β-strands; this is the result of the smaller number of families containing α-helices that have been defined at high resolution by X-ray analysis. The distribution of residues in different secondary structure classes is similar to that found in analyses of non-homologous data sets. Of the residues in α-helices, the most numerous is alanine; those in β-strands show the expected relative abundance of tyrosine and tryptophan; and those in coil include many glycines and prolines.

† 1 Å = 10^{-10} m = 10^{-1} nm.

Table 1. *Structures used in the analysis*

PDB code[a]	description	chain	residue range	resolution/Å
	globins			
2HHB	human haemoglobin, α-chain	A	—	1.7
2HHB	human haemoglobin, β-chain	B	—	1.7
3MBN	sperm whale myoglobin	—	—	2.0
1ECD	erythrocruorin	—	—	1.4
2LHB	lamprey haemoglobin	—	—	2.0
1LH1	leghemoglobin	—	—	2.0
	crystallins			
1GCR	calf γ-II crystallin	—	1–39	1.6
1GCR	calf γ-II crystallin	—	40–87	1.6
1GCR	calf γ-II crystallin	—	88–128	1.6
1GCR	calf γ-II crystallin	—	129–174	1.6
	aspartic proteinases			
4APE	endothiapepsin	—	—	2.1
2APP	penicillopepsin	—	—	1.8
2APR	rhizopuspepsin	—	—	1.8
pep[b]	porcine pepsin	—	—	2.0
chy[b]	calf chymosin	—	—	2.2
	serine proteinases			
1TON	rat tonin	—	—	1.8
2PKA	porcine kallikrein A	A	—	2.0
2PTN	porcine trypsin	—	—	1.5
4CHA	bovine chymotrypsin	A	—	1.7
3EST	porcine elastase	—	—	1.6
3RP2	rat mast cell protease-II	A	—	1.9
1SGT	*S. Griseus* trypsin	—	—	1.7
	immunoglobulin variable domains			
2FB4	FAB (lambda) KOL	H	1–117	1.9
3FAB	FAB (prime) NEW	H	1–116	2.0
1REI	B-J fragment REI	A	1–107	2.0
2HFL	HyHEL-5 FAB	L	1–105	2.5
2RHE	B-J fragment RHE	—	1–111	1.6
3FAB	FAB (prime) NEW	L	1–108	2.0
	immunoglobulin constant domain			
3FAB	FAB (prime) NEW	L	114–214	2.0
1FBJ	FAB (kappa) J539	L	111–213	2.6
1FC1	FC (human)	A	238–340	2.9
2FB4	FAB (lambda) KOL	—	123–221	1.9
1FBJ	FAB (kappa) J539	H	123–218	2.6
	phospholipases			
1P2P	porcine phospholipase A2	—	—	2.6
1BP2	bovine phospholipase A2	—	—	1.7
1PP2	rattlesnake phospholipase A2	—	—	2.5

[a] Bernstein *et al.* 1977.

[b] Courtesy of Jon Cooper (pep) and Matthew Newman (chy), Department of Crystallography, Birkbeck College, London., U.K.

The expected prevalence of glycine in a positive ϕ conformation is observed, (54% of all glycines). All other amino acids have a distinct preference for a negative ϕ angle, but residues that are observed to tolerate a positive ϕ angle include asparagine, aspartic acid and serine (17%, 7% and 3%).

The partitioning of a residue between inaccessible and accessible states follows patterns in previous studies (Lim & Sauer, 1989). For example, the most often buried residues are cystine, valine and isoleucine (85%, 71% and 67% inaccessible), whereas the most accessible residues are glutamic acid, arginine and lysine (18%, 5% and 2% inaccessible).

5. CONSTRUCTION OF SUBSTITUTION TABLES

Each residue in each protein structure is a member of a class defined by a combination of structural features. The features considered were amino acid type (20 possibilities), accessibility (two possibilities), sidechain hydrogen bonding (eight possibilities) and mainchain conformation (four possibilities). Eight amino acids are unable to form hydrogen bonds through their sidechains and most polar residues are unable to act both as donors and acceptors. Considering these and other factors reduces the total number of possible classes to 578, of which 403 are occupied in the present

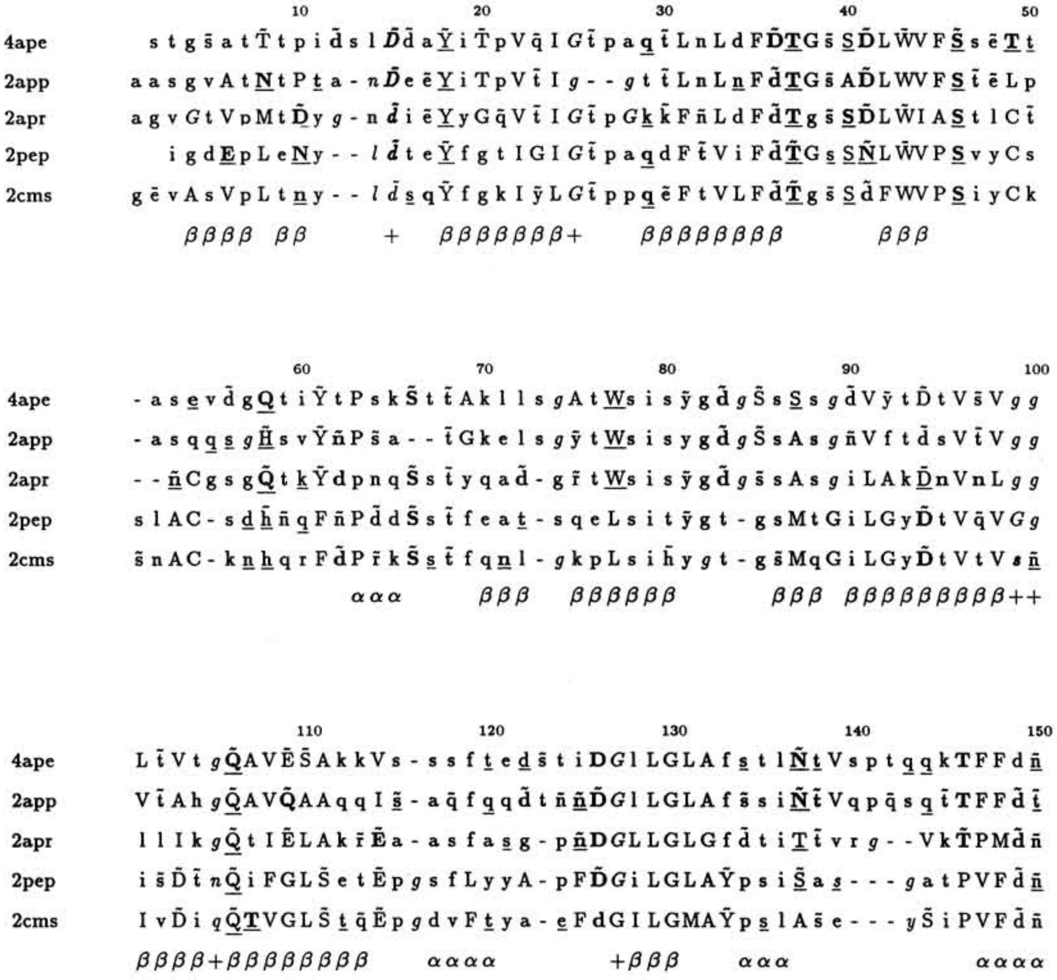
10 20 30 40 50
4ape stgsatTtpidslDdaYiTpVqIGipaqiLnLdFDTGsSDLWVFSseTt
2app aasgvAtNtPta-nDeeYiTpViIg--gtiLnLnFdTGsADLWVFSieLp
2apr agvGtVpMtDyg-ndieYyGqViIGipGkkFnLdFdTgsSDLWIAStlCi
2pep igdEpLeNy--ldteYfgtIGIGipaqdFtViFdTGsSNLWVPSvyCs
2cms gevAsVpLtny--ldsqYfgkIyLGippqeFtVLFdTgsSdFWVPSiyCk
ββββ ββ + βββββββ+ ββββββββ βββ
60 70 80 90 100
4ape -asevdgQtiYtPskStiAkllsgAtWsisygdgSsSsgdVytDtVsVgg
2app -asqqsgHsvYnPsa--iGkelsgytWsisygdgSsAsgnVftdsViVgg
2apr --nCgsgQtkYdpnqSsiyqad-grtWsisygdgssAsgiLAkDnVnLgg
2pep slAC-sdhnqFnPddSsifeat-sqeLsitygt-gsMtGiLGyDtVqVGg
2cms snAC-knhqrFdPrkSsifqnl-gkpLsihygt-gsMqGiLGyDtVtVsn
ααα βββ ββββββ βββ βββββββββ++
110 120 130 140 150
4ape LiVtgQAVESAkkVs-ssftedstiDGlLGLAfstlNtVsptqqkTFFdn
2app ViAhgQAVQAAqqIs-aqfqqdtnnDGlLGLAfssiNtVqpqsqiTFFdi
2apr llIkgQtIELAkrEa-asfasg-pnDGLLGLGfdtiTivrg--VkTPMdn
2pep isDinQiFGLSetEpgsfLyyA-pFDGiLGLAYpsiSas---gatPVFdn
2cms IvDiqQTVGLStqEpgdvFtya-eFdGILGMAYpslAse---ySiPVFdn
ββββ+βββββββββ αααα +βββ ααα αααα

Figure 1. A section of the alignment of sequences of aspartic proteinases achieved by comparing the three-dimensional structures by using COMPARER (Šali & Blundell 1990). The coordinates of the three-dimensional structures were obtained from the Brookhaven Protein Databank (PDB) (Bernstein *et al.* 1977) (PDB codes: 4APE, 2APP and 2APR), with the exception of the coordinates of porcine pepsin (pep) and calf chymosin (chy), which were kind gifts of Jon Cooper and Matthew Newman, respectively. The amino acid code is the standard one-letter code formatted by using the following convention: *italic* for positive ϕ angle; UPPER CASE for solvent inaccessible residues; lower case for solvent accessible residues; **bold type** for hydrogen bonds to main-chain amide; underline for hydrogen bonds to main-chain carbonyl oxygen; tilde (˜) for sidechain-to-sidechain hydrogen bonds. Below the alignment is shown the consensus secondary structure: (α) for α-helical positions; (β) for β-stranded positions (+) for positions in a positive ϕ conformation.

sample. In total, over 27000 residue substitutions were observed.

Under this scheme, substitutions could be considered in terms of tables where there are as many dimensions to the table as there are features included in the analysis of the two positions compared. Every dimension would have as many different values as the feature can assume. All pair-wise comparisons of structures in each alignment are considered in the analysis, and all substitutions implied by pair-wise comparisons were stored in a multidimensional table as a function of the features identified in the three-dimensional structures. Of course this leads to very sparse tables and so some simplifications are required. Our major simplification is to consider the structural features of only one of the two proteins compared. For example, if we consider a residue that is buried, in an α-helix and with particular hydrogen bonding, we consider only the amino acid type of the residues observed at topologically equivalent positions. This corresponds to many applications where only one of the three-dimensional structures of the compared proteins is known. Secondly, to understand the general role of certain structural features in constraining the conservation, we have summed the joint frequencies into selected marginal probability distributions, e.g. all the residues in a particular type of secondary structure irrespective of the accessibility or hydrogen bonding properties. In each case it is convenient to display the data as 20 × 20 probability tables where one dimension refers to the amino acid type restricted to a particular structural environment and the other is simply the residue type. The values are thus the probability of observing any amino acid at a topologically equivalent position in a homologous protein given the residue type and physical environment in one structure; we have not normalized the data at this stage of our analysis for the relative abundance of each amino acid type in the sample. (Probabilities (P) and probability differences (ΔP) are expressed as the frequency of occurrence of an event (x) divided by the sample size (n). Standard errors were calculated as $\sqrt{x(n-x)/n^3}$; quoted errors correspond to one standard deviation, i.e. a confidence interval of approximately 67%. Throughout the paper, P refers to a probability and ΔP to a probability difference.)

To examine the effect of a particular structural feature on conservation and substitution, difference substitution tables were constructed. The values were calculated from the difference between the table for substitutions within a particular environment and the table for all substitutions not in this environment. An increase in the conservation of a residue, or a more favourable substitution due to the environment, will be evident by a positive term in this difference table.

6. ENVIRONMENT-INDEPENDENT SUBSTITUTION TABLE

To compare our results with those previously obtained, we summed the multidimensional tables over all dimensions except amino acid type, giving a 20 × 20 environment-independent substitution table. This substitution table should be comparable to those used in standard sequence alignment and analysis techniques, notwithstanding differences caused by sample bias. This global-substitution table shows the familiar high probabilities for conservation of cystine ($P = 0.78 \pm 0.01$), glycine ($P = 0.57 \pm 0.01$) and tryptophan ($P = 0.52 \pm 0.02$) along the diagonal. The exchange groups (Val, Leu, Ile), (Ser, Thr), and (Phe, Tyr, Trp) are also well defined, as shown previously by Dayhoff and co-workers (1969, 1983), McLachlan (1971) and Risler *et al.* (1988).

7. SUBSTITUTION PATTERNS FOR MAINCHAIN CONFORMATIONS

We begin by considering the role of particular structural features on the substitution properties of residues.

There are many examples of conservation being reduced because of an α-helical environment (table 2). Most notable is glycine ($\Delta P = -0.46 \pm 0.04$) with alanine the most often observed substitute; this decrease in conservation in a helix is a consequence of the high main-chain flexibility of glycine that stabilizes the unfolded or less structured state relative to the folded state. Other decreases in conservation are found for proline ($\Delta P = -0.27 \pm 0.05$), tyrosine ($\Delta P = -0.20 \pm 0.05$), serine ($\Delta P = -0.11 \pm 0.03$) and threonine ($\Delta P = -0.16 \pm 0.03$). The reduced conservation of proline is primarily caused by the absence of the amide proton. In a helix it disrupts the hydrogen bonding pattern except at the N-terminus, but a mutation away from a proline introduces another stabilizing hydrogen bond into the structure. The relative rejection of serine and threonine is explained by the presence of the γ-hydroxyl in the sidechain. This has been shown to interact with neighbouring helix carbonyls and weaken the hydrogen bond between the main-chain functions of the helix (Blundell *et al.* 1983). A related feature of serine and threonine was

Table 2. *Difference probabilities (multiplied by 100) for amino acid substitutions involving alpha helical residues*

(Each column in the table represents the difference in probability observed for the named substitutions within the environment compared with all other environments.)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−2.7	+3.3	+7.0	+10.2	+5.3	+19.4	+5.5	+7.8	+6.3	+4.4	−1.6	+6.4	+13.1	+3.5	+9.4	+12.7	+9.1	+4.0	+8.2	+0.8
C	+0.2	−7.1	+0.1	−1.2	−1.8	+0.5	+0.7	−0.8	+0.3	−0.6	−0.4	+1.1	−0.2	+1.3	−0.5	+0.8	−0.3	+0.7	+0.7	−0.5
D	+1.1	+0.4	−5.6	+7.1	−0.9	+3.5	−2.3	−1.6	+2.5	−0.6	+0.4	−2.0	+0.3	+6.1	−1.6	+2.9	−1.1	−0.2	+1.4	+0.1
E	+1.8	−1.6	+7.9	−6.1	−0.4	+6.9	+1.1	−0.8	+1.6	+0.3	−1.9	+5.1	+12.3	+6.8	−0.9	+1.3	−0.1	−1.0	−0.8	+2.6
F	+1.3	−2.0	−0.6	−0.4	+0.6	−0.8	−3.4	−3.1	−1.3	+8.5	+5.0	−0.4	+1.4	+2.7	+0.0	+0.6	−0.1	+4.5	−2.9	−7.9
G	−1.1	+2.6	+0.1	+3.1	−2.3	−46.0	−1.8	+1.3	−1.9	−1.7	−0.8	−4.6	+5.5	−3.2	−0.9	−2.9	+2.6	−0.9	−2.8	+0.6
H	+0.6	+0.9	−0.7	+0.8	−1.2	−0.4	+26.2	−0.3	+0.0	−0.3	+0.8	−1.4	−0.9	+2.2	+0.2	+1.2	−0.4	+1.1	+0.2	+1.1
I	+1.5	−1.4	−1.8	−1.2	−4.6	+3.8	−1.1	+2.5	+2.5	+0.8	+5.4	−1.2	−1.1	−1.1	−0.4	+1.3	−0.6	+0.8	−2.6	+7.5
K	+3.0	+2.7	+3.4	+3.6	−1.1	−4.0	−1.5	+3.6	+0.3	+0.5	+9.7	+4.3	+3.6	+2.0	+14.5	+1.4	+4.9	−0.3	+5.2	−3.0
L	+1.4	−1.6	−0.3	−0.3	+13.0	−1.0	−2.3	−1.1	−1.0	−14.3	+4.3	−0.8	−1.8	−0.2	+1.8	+1.3	+3.0	+5.0	−4.3	+3.0
M	−0.7	−0.2	+0.2	−0.7	+1.8	+0.9	−1.6	+0.9	+0.6	+2.5	−12.4	+0.3	−0.4	−3.2	−0.7	−0.1	−0.5	+0.5	+2.3	+2.4
N	−0.2	+1.9	−1.7	+0.5	−1.0	+0.7	−7.6	−1.5	+0.4	−0.5	+0.4	−5.1	+1.6	+5.1	−0.9	+0.4	+2.9	−0.6	−1.1	+2.7
P	−2.1	−0.3	−1.4	+3.7	+0.0	+1.6	+2.3	+0.1	−2.5	−1.4	−1.6	−0.6	−27.2	−1.8	−2.9	−1.4	−2.8	−0.8	−0.8	−0.5
Q	−1.2	+0.3	+2.0	−2.0	+1.0	−0.9	−3.5	−1.1	−2.6	−1.0	−7.4	+3.9	−1.2	−16.5	−6.8	−1.5	−0.1	−3.0	+1.1	+3.7
R	+0.5	−0.5	−0.9	−1.9	−0.5	+0.0	−1.4	−0.8	+1.7	+0.4	−1.6	−1.2	−0.9	−3.9	+1.5	−0.2	−0.9	−1.4	−0.6	−0.5
S	−1.2	−0.2	−3.0	−3.8	−0.9	+0.3	−1.2	−1.6	−0.6	+1.0	−2.3	−1.5	+0.0	+0.7	−2.0	−11.3	−7.1	−1.5	−1.9	+0.7
T	−1.7	−0.9	−3.9	−6.2	−2.0	+5.2	−2.9	−1.4	−2.7	+0.5	−1.9	−2.2	−4.4	−1.9	−4.0	−5.9	−15.9	+0.9	−0.6	+2.0
V	−1.8	+5.0	−0.1	−2.4	+6.2	+2.0	+0.4	−0.3	−0.6	+3.0	−0.8	−0.8	+1.1	−2.9	−2.9	+0.2	+6.7	−7.5	+0.0	+5.0
W	+1.2	−0.3	+0.1	−0.4	−0.8	+0.0	−1.0	−1.0	+0.9	−1.3	+2.7	−0.4	−0.3	+1.1	−1.3	+0.0	+0.1	+0.1	+1.2	+0.7
Y	−0.3	−0.9	−0.8	−2.4	−10.3	+0.3	−4.8	−0.9	−3.9	−0.4	+3.9	+1.3	−0.5	+3.5	−1.5	−0.7	+0.5	−0.4	−1.8	−20.4

recognized in a previous analysis of secondary structural preferences of amino acids as a function of position in the helix (Argos & Palau 1982).

In general, substitutions of any residue in an α-helix by alanine are more favoured. These results are consistent with the observations made by Padmanabham *et al.* (1990) of the helix-forming tendencies of nonpolar amino acids in peptides. The largest enhancement of conservation because of a helical environment is for histidine ($\Delta P = +0.26 \pm 0.06$). To check that this was not just an artefact of the presence of the functionally required pair of histidines in the globins, the globin data were removed from the analysis. The new difference conservation probability for histidine is $\Delta P = +0.27 \pm 0.18$. Although the confidence level is decreased when the globins are removed from the analysis, the effect does seem to be a genuine one.

Comparison of the observed difference substitution table for residues in β-strands (table 3) with the data for α-helical residues (table 2), reveals that there are fewer negative but more positive terms, indicating that a β-sheet contains more constraints than an α-helix. The most notable of these are tyrosine ($\Delta P = +0.19 \pm 0.04$), glutamine ($\Delta P = +0.22 \pm 0.04$), serine ($\Delta P = +0.14 \pm 0.03$) and leucine ($\Delta P = +0.12 \pm 0.03$). For tyrosine this change is opposite in sign to that observed for a helical position, confirming its strong preference for an extended main-chain conformation. The large enhancements in conservation for leucine and glutamine are surprising in view of their preferences for an α-helical conformation (Chou & Fasman 1974; Levitt 1978).

The probability-difference table for residues with a positive ϕ conformation shows an increase in conservation not only for glycine but also for aspartic acid and asparagine (table 4). (It should be noted that because of the lower sampling of this conformational class, the errors in this table are comparatively high. A full set of data including estimated errors may be obtained from the authors on request.) It also shows that aspartic acid, asparagine, serine and many other residues have a higher probability of being substituted by glycine when they have a positive main-chain ϕ conformation.

The difference table for residues in a coil conformation is more complex than the corresponding tables for α-, β- and positive ϕ positions (table 5). However, the terms are generally smaller indicating fewer constraints caused by the environment itself. Conservation of glutamine, tyrosine and cystine appear to be the most reduced, whereas proline is the only amino acid whose conservation undergoes a significant enhancement ($\Delta P = +0.10 \pm 0.04$). The complex nature of this table may be indicative of the large number of distinct conformations that make up the so-called 'random coil' region. The analysis may benefit from a further subdivision into structural motifs such as β-hairpins, where specific sequence patterns are known to be important (Sibanda *et al.* 1989).

Table 3. *Difference probabilities (multiplied by 100) for amino acid substitutions involving β-sheet residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−0.8	−6.8	−3.6	−7.3	−4.5	−0.1	−4.0	−4.8	−4.1	−1.2	+3.7	−2.6	−5.0	−3.3	+0.2	−3.4	−2.1	+0.4	−3.1	−1.3
C	−1.7	+14.5	−0.3	+1.6	+1.0	−0.2	−0.2	+0.0	+0.5	+0.0	+0.6	−0.2	−0.2	+1.1	+2.1	−0.8	−0.4	−0.9	−0.8	−0.9
D	−2.9	−1.1	−3.2	−6.3	−0.4	−3.2	−3.1	+0.8	−2.6	+0.1	−0.6	+4.3	−0.1	−4.1	−4.0	−4.3	−4.1	+0.1	−0.7	−2.5
E	−3.7	+0.5	−1.8	+2.1	−0.4	+0.8	−1.3	−1.0	+0.5	+0.0	+1.8	−2.0	−0.8	−2.2	−0.3	−2.7	−0.7	−0.5	+0.6	−2.2
F	−0.4	+1.0	+0.0	+0.2	+4.7	−0.5	+2.0	+3.1	+0.6	−1.3	−5.2	−1.7	+0.4	−1.1	+0.1	+1.2	+0.1	−1.6	−4.4	−0.8
G	+1.4	−1.1	−5.2	−1.1	−1.1	+9.6	−5.1	−0.9	−1.5	−0.8	−1.1	−5.8	−4.2	−2.8	−3.1	−2.0	−1.5	−2.0	+5.3	+0.6
H	−1.4	−0.3	−1.5	−0.9	−0.6	−1.2	−21.8	+0.3	+0.1	−0.5	−3.4	−2.0	+0.8	−2.0	+0.3	−0.6	+0.2	−0.5	−1.5	−4.0
I	+1.0	+0.2	+4.1	−0.6	+7.9	+0.4	+3.2	+4.2	−0.8	+0.3	+0.1	+2.2	+1.2	+0.0	+1.0	−0.1	+0.6	+0.5	+2.1	−0.2
K	−2.8	−0.4	−2.9	+2.1	−0.5	−0.3	+4.3	−2.1	+8.2	−2.1	−5.2	−1.4	+1.2	−1.9	+0.1	−3.0	−2.9	−2.6	−3.0	−1.2
L	+2.3	−0.1	+2.8	+2.5	+0.0	+0.8	+1.7	+2.4	+2.6	+11.5	−9.4	+0.6	+1.0	−0.9	+3.1	+0.2	+0.2	+1.8	+7.2	−0.6
M	+2.4	+0.3	+0.9	+2.2	−1.1	+0.1	+2.4	+0.7	−0.8	−2.2	+15.5	+0.3	−0.4	+1.7	+1.2	+0.0	+1.3	+0.4	−0.2	−1.2
N	−2.6	−0.5	+3.7	−1.9	−2.7	−3.1	−1.1	−1.6	−1.1	−2.8	−0.6	−11.2	+0.6	−3.2	−2.4	−2.2	−1.7	−1.9	+1.0	−1.5
P	−3.9	−0.5	−0.6	−2.8	−1.4	−3.0	+2.7	−0.9	−0.5	−1.2	−1.1	+0.8	+1.2	−0.9	−1.5	−4.5	−1.8	−1.3	+0.6	−0.9
Q	+0.6	+1.7	+1.3	+4.6	−1.0	+1.0	+4.3	+0.5	+4.1	−0.2	+4.3	+2.8	+2.2	+22.3	+5.7	−0.1	+3.0	−0.7	−1.0	+0.0
R	+0.7	+1.7	+0.5	+0.9	−0.4	+0.2	+3.6	+0.0	+1.4	+0.1	+2.4	+1.5	+0.8	+0.7	−7.4	−0.6	+0.5	−0.4	−0.1	−0.8
S	−2.6	−3.2	−2.3	−6.3	−0.2	−1.4	+1.9	−1.7	−4.1	−1.3	−1.1	+0.9	−4.4	−5.5	−2.9	+13.7	−1.1	−0.3	−0.1	−2.3
T	+2.8	−1.4	+0.2	+9.1	+0.0	+1.0	+5.3	−0.8	−2.0	−1.7	+5.3	+10.0	+4.7	+3.3	+4.2	+6.6	+6.5	+1.8	+1.6	+1.4
V	+10.5	−2.5	+4.9	+2.1	+1.8	+0.1	+7.1	+2.2	−0.7	+2.8	−0.9	−1.4	+0.7	−0.1	+2.7	+2.5	+3.3	+7.9	+0.8	+1.5
W	−0.2	−0.5	+0.8	+0.7	−0.5	−0.8	−0.9	+0.5	−0.9	+1.2	−0.5	+2.2	+0.3	−0.5	+0.5	+0.0	+0.2	+0.1	+0.3	−2.3
Y	+1.2	−1.4	+2.3	−0.8	−0.6	−0.1	−1.0	−1.1	+1.2	−0.8	−4.6	+2.7	+0.1	−0.6	+0.3	+0.0	+0.5	−0.6	−4.6	+18.9

Table 4. *Difference probabilities (multiplied by 100) for amino acid substitutions involving positive ϕ residues*

	A	C	D	E	F	G	H	I	K	L	M	N	Q	R	S	T	V	W	Y
A	−14.4	+3.5	−4.0	−9.1	−3.8	−6.9	−2.7	−5.9	−8.0	−3.6	−8.2	−1.1	−6.3	−5.3	−4.4	+4.2	−7.0	−3.1	−4.7
C	+11.3	−21.5	+1.0	−1.0	−1.8	−0.2	−0.2	−0.6	−0.8	−0.5	−0.3	−0.2	−1.0	−1.4	−0.6	−0.3	−0.4	+24.8	−0.5
D	+3.7	+6.7	+16.5	+0.8	−0.7	+0.3	−4.9	−1.9	+1.5	−0.7	−0.9	−3.4	−1.4	−0.7	+0.9	+1.0	−1.7	−0.8	+0.9
E	−4.5	−1.4	−4.9	−15.5	−0.7	−1.0	−2.2	−1.3	−4.6	−0.8	−2.6	−4.0	−0.3	−5.6	+3.3	−4.4	−2.1	−0.6	+1.3
F	−2.0	−2.7	−0.8	−0.7	−3.7	−1.3	−2.7	+13.9	−1.4	−7.2	−3.2	−1.7	−1.2	+5.3	+1.4	−1.5	−3.6	−9.0	+4.2
G	−0.5	+6.7	+3.8	+5.2	−1.8	+7.8	+20.0	−1.4	+11.4	−1.9	+47.4	+14.2	+20.8	+9.8	+27.0	+28.1	−2.8	−3.1	+4.9
H	+0.3	−0.2	−0.6	+8.0	−1.3	+1.5	−30.6	+9.5	+8.5	+3.8	+48.5	+2.2	+4.3	+4.5	−1.4	−0.8	−1.4	−0.8	+8.0
I	−2.2	−1.2	−0.7	+7.4	−8.1	−0.6	−1.6	−10.2	−2.1	−5.6	−6.4	−1.1	−1.0	+2.3	+1.3	−2.4	−11.7	−4.0	+5.6
K	+4.0	−1.5	+0.2	+9.1	−1.7	−0.3	+7.1	−2.0	−22.2	−2.4	−3.2	+0.8	−3.6	+2.9	−1.1	−5.0	−2.3	−1.7	−0.2
L	+1.8	+5.9	+1.5	−1.6	−13.8	−0.6	−1.8	−4.5	+9.4	−5.1	−20.2	+0.8	−2.3	−2.2	+0.7	−2.9	+21.4	+18.9	−0.1
M	+0.3	−0.2	−0.2	−1.0	−1.1	−0.6	+1.8	−1.7	+1.7	−3.6	−15.2	−0.3	−2.9	−0.6	−0.4	−0.9	+6.7	−1.0	−1.4
N	+9.1	−0.3	+3.0	−4.6	−1.6	+2.8	+23.1	+17.9	+13.4	+27.6	−0.9	+7.2	+3.3	+2.6	+4.6	+5.2	−1.2	−0.8	+7.6
P	+2.7	−0.3	−2.6	−5.2	−0.8	+0.3	−2.0	−0.9	+0.9	+15.3	−1.2	+1.7	−2.3	−3.5	−3.8	+6.3	−1.1	−0.6	−0.5
Q	+1.9	−1.7	−3.9	−7.8	−1.3	−0.6	+1.0	−0.8	−0.8	−1.3	−9.1	−1.8	−13.1	−6.1	−4.2	−3.4	+22.2	−1.0	+1.4
R	−1.7	−1.4	−0.8	−3.7	−1.2	−0.5	−0.4	−0.6	−2.5	−0.7	−1.2	−0.4	−2.4	−11.4	−2.7	−2.7	−1.4	−1.5	−1.4
S	−1.8	+12.5	−2.4	−9.2	−3.2	+2.2	−4.1	−2.6	−5.0	−2.5	−2.9	−5.5	+10.4	−3.2	−17.5	−3.8	−4.3	−1.5	+3.2
T	−1.0	−0.8	−3.9	+18.8	−2.8	−1.0	−3.1	−3.5	+0.1	−2.9	−5.0	−5.1	−2.8	+2.5	−10.0	−7.6	−6.0	−1.5	−1.0
V	−3.8	−1.4	−1.5	+13.8	+2.9	−1.4	−5.9	−8.0	+1.4	−8.5	−9.9	−2.0	+4.4	−4.5	+4.6	−6.3	+0.1	−1.9	+8.8
W	−0.8	−0.3	−0.3	−0.3	−4.3	+0.4	−0.8	−1.5	−0.7	−1.6	−1.5	−0.4	−0.5	+6.1	+1.1	−0.4	−0.4	−27.6	−3.9
Y	−2.5	−0.8	+0.5	+6.9	+50.6	−0.3	+10.0	+6.2	−0.4	+2.4	−4.1	0.2	−2.0	+8.6	+1.1	−2.4	−2.7	+17.0	−32.5

Table 5. *Difference probabilities (multiplied by 100) for amino acid substitutions involving coil residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	+3.0	+5.3	−2.6	−2.0	+0.6	−5.8	−1.9	−0.8	−1.4	−1.8	−3.3	−2.4	−1.7	+1.7	−5.4	−3.1	−1.5	−3.4	−2.8	+0.9
C	+1.1	−11.3	+0.1	−0.3	+0.4	0.0	−0.4	+0.6	−0.6	+0.5	−0.4	−0.6	+0.2	−1.7	−1.6	+0.3	+0.5	+0.5	+0.3	+1.1
D	+1.2	+0.9	+5.6	−0.5	+1.2	+1.7	+3.7	+0.3	+0.2	+0.3	+0.4	−1.4	−0.1	+1.3	+4.4	+2.2	+4.4	0.0	−0.3	+2.4
E	+1.2	+0.4	−4.2	+2.9	+0.7	−2.7	−0.1	+1.9	−1.6	−0.2	−0.5	−2.0	−4.5	−0.9	+0.7	+1.6	+0.7	+1.3	0.0	+1.1
F	−0.8	+0.1	+0.4	+0.1	−5.7	+0.7	+1.4	−1.2	+0.5	−4.8	+1.8	+1.4	−0.8	−0.1	−0.1	−1.3	−0.1	−1.4	+7.1	+3.9
G	−0.1	−0.3	+3.0	−1.5	+3.1	+5.8	+4.5	−0.1	+2.6	+2.2	+2.2	+6.5	+0.8	+4.4	+3.2	+3.0	+0.5	+2.9	−3.8	−0.8
H	+0.5	−0.2	+1.3	+0.1	+1.6	+1.2	−7.3	−0.1	−0.1	+0.7	+3.8	+2.2	−0.2	+1.0	−0.4	−0.1	0.0	−0.2	+1.6	+3.5
I	−2.2	+0.6	−1.2	+1.3	−4.9	−1.5	−1.1	−7.2	−1.2	−0.9	−5.5	−0.6	−0.4	+0.6	−0.7	−0.5	−0.4	−1.2	−0.3	−2.7
K	−0.6	−1.1	−0.6	−4.1	+1.5	−1.0	−1.4	−0.5	−6.6	+2.0	−2.8	−1.9	−2.4	+1.0	−8.1	+1.9	+0.9	+3.3	−0.6	+2.3
L	−3.0	+1.0	−1.4	−1.6	−10.8	−0.4	+0.8	−2.0	−1.3	−2.6	+8.2	+0.1	0.0	+1.0	−3.7	−0.8	−1.4	−5.7	−4.7	−0.6
M	−1.2	−0.2	−0.7	−1.1	−0.3	−0.4	−0.2	−1.6	+0.2	+0.6	−8.1	−0.4	+0.4	−0.3	−0.6	0.0	−1.1	−0.8	−1.6	+0.2
N	+1.8	−0.5	−1.0	+1.1	+3.8	+2.4	+6.6	+3.2	+0.5	+3.5	+0.4	+10.2	−1.1	+0.9	+2.6	+1.6	+0.5	+2.6	−0.3	+0.5
P	+5.0	+0.8	+1.3	−0.7	+1.5	+2.1	−3.5	+1.0	+2.3	+2.3	+3.0	−0.2	+10.4	+1.8	+2.9	+4.5	+2.8	+2.1	0.0	+1.1
Q	+0.6	−2.0	−2.1	−1.9	+0.3	−0.5	+0.1	+0.3	−1.3	+0.9	+1.6	−4.2	−1.1	−15.2	−1.2	+0.8	−2.9	+2.9	+0.3	−1.4
R	−1.0	−1.5	+0.3	+0.7	+0.9	−0.2	−1.1	+0.6	−2.4	−0.4	−1.6	−0.2	−0.2	+1.1	+5.6	+0.6	−0.1	+1.5	+0.6	+1.0
S	+3.1	+3.6	+3.3	+7.4	+1.0	+1.1	−0.3	+3.5	+3.7	+0.7	+3.7	+0.4	+3.3	+5.4	+3.6	−6.1	+3.9	+1.4	+1.5	+2.0
T	−0.6	+1.9	+2.4	−2.2	+1.6	−2.5	−1.0	+2.2	+3.5	+1.6	−5.1	−4.8	−1.6	−2.5	−1.5	−2.7	−0.1	−2.8	−1.3	−2.2
V	−6.4	−0.3	−2.8	+0.2	−7.1	−0.7	−4.6	−2.5	+1.0	−5.4	+2.0	+1.3	−0.9	+1.4	−0.8	−2.2	−5.8	−3.8	−0.9	−3.5
W	−0.9	+0.8	−0.6	−0.2	+1.2	+0.7	+1.4	+0.2	+0.0	−0.4	−2.0	−1.1	−0.1	0.0	+0.3	0.0	−0.3	−0.2	−1.3	+2.0
Y	−0.6	+1.9	−0.8	+2.3	+9.2	0.0	+4.3	+2.1	+1.9	+1.2	+2.2	−2.5	+0.1	−1.0	+0.5	+0.3	−0.7	+1.0	+6.5	−10.7

Table 6. *Difference probabilities (multiplied by 100) for amino acid substitutions involving inaccessible residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	+18.8	−0.3	−3.9	+8.1	−0.1	+0.2	−2.8	−1.7	−7.8	−1.0	−1.0	+1.2	−1.6	−0.6	−5.2	+9.1	+7.7	+0.2	−2.7	−2.0
C	+1.3	+27.5	−0.4	−1.0	−0.5	−0.1	−0.2	+0.2	−0.8	+0.4	−0.7	−0.2	−0.2	+0.5	−1.4	+1.2	+0.8	−0.6	−0.9	−0.2
D	−3.2	−3.1	+60.3	−9.0	−0.9	−3.9	−3.6	−1.8	−3.7	−1.7	−2.2	+3.7	−6.0	−5.1	−4.3	−5.0	−4.7	−2.5	−1.8	−2.3
E	−5.7	−4.2	−7.3	+5.4	−1.3	−1.6	−2.4	−1.7	−7.1	−1.8	−0.5	+4.9	−4.9	−0.8	−0.4	−3.2	−5.1	−2.2	−1.4	−3.1
F	+0.6	−1.6	−0.6	−0.8	+16.6	−0.4	+3.8	+0.4	−1.4	−1.3	+0.5	+1.0	+3.3	−0.9	−1.9	−0.3	−0.6	+1.5	−4.5	+7.4
G	−0.8	−0.2	−6.1	+1.3	−3.2	+28.6	−4.8	−0.9	−4.4	−2.0	−4.9	−4.4	−5.4	−3.5	+5.7	+0.5	+2.0	−4.3	−6.8	−3.1
H	−0.8	−0.8	−2.2	−1.3	−1.6	−1.4	−7.8	−0.4	+4.3	−0.7	−2.0	+8.1	−1.0	+4.7	−2.7	−0.1	−0.9	−0.8	−1.0	−1.8
I	+3.8	−2.4	+0.0	+5.0	+2.1	+0.5	+3.3	+13.2	−2.1	+1.8	+3.3	+2.4	−0.6	−0.1	−1.3	+1.2	+1.7	+5.0	+3.9	+2.1
K	−5.6	−4.0	−3.8	−0.4	−3.1	−3.7	−5.5	−4.4	+16.2	−5.0	−6.8	−4.8	−3.5	−6.6	+26.3	−2.6	−5.6	−4.1	−1.1	−5.2
L	+2.4	−1.3	−0.8	−1.7	+1.5	+0.3	+1.4	+8.3	−3.8	+24.2	+16.4	+4.2	−0.8	−0.5	−2.2	+1.1	+1.1	+7.1	−0.3	+0.9
M	+1.7	−0.8	−0.3	+5.9	+0.4	+0.0	+3.7	+0.6	−0.9	+2.2	+15.4	+2.3	−0.4	+5.2	−0.6	+0.5	+0.9	−0.3	+1.1	−0.1
N	−3.4	−1.6	−4.9	−3.5	−0.9	−4.6	−1.4	−3.4	−5.0	−2.9	−1.0	−13.7	−1.7	−3.5	−4.6	−1.7	−4.5	−2.7	−0.1	−3.3
P	−4.9	−0.6	−4.7	−5.6	+0.9	−2.3	−2.2	−1.4	−4.3	−2.9	−2.9	−0.3	+39.3	−2.2	−3.5	−4.4	−2.5	−1.9	−1.4	−0.7
Q	−1.5	−1.9	−4.6	+2.6	−1.9	−1.7	+4.5	−1.9	−6.0	−2.1	−4.4	+0.4	−2.5	+29.1	+11.0	−3.6	−4.0	−3.3	+1.1	−1.1
R	−0.8	+0.7	−2.5	+8.4	−1.4	−1.0	+1.3	−1.1	+36.7	−0.8	−2.9	−2.9	−2.2	−0.5	+5.8	−0.9	−3.2	−1.3	−2.4	−1.1
S	−5.8	−3.3	−10.1	−7.2	−4.2	−5.3	−1.0	−4.3	−1.2	−4.4	−2.4	+3.7	−7.4	−9.2	−10.2	+11.3	−5.5	−4.7	−0.7	−5.0
T	−3.4	−0.1	−5.7	−9.5	−4.3	−0.8	−1.6	−5.7	−7.7	−4.2	−3.9	−5.0	−3.6	−7.3	−8.2	−4.6	+19.4	−7.6	−3.2	−5.0
V	+7.1	−1.3	−1.4	+6.1	+3.7	−1.8	+7.0	+5.6	−3.8	+3.6	+1.9	+0.1	−0.5	−0.9	+0.7	+2.4	+3.3	+21.9	−1.6	+1.1
W	−0.3	+0.4	+0.6	−0.3	−0.2	−1.1	+5.8	−0.5	−0.6	−0.6	+1.2	−0.4	−0.3	+2.3	−1.1	+0.3	−0.4	+0.4	+33.3	−0.9
Y	+0.4	−1.0	−1.5	−2.5	−1.8	−0.1	+2.6	+1.0	+3.3	−0.9	−2.9	−0.3	+0.1	−0.1	−2.2	−1.2	+0.3	+0.2	−9.4	+23.7

Table 7. *Difference probabilities (multiplied by 100) for amino acid substitutions involving β-sheet residues*

	A	C	D	E	F	G	H	I	K[a]	L	M	N	P	Q	R	S	T	V	W	Y
A	+2.6	−6.1	−0.5	−2.9	−2.8	+0.0	+1.0	−2.6	−7.7	−0.6	+5.2	−2.7	−5.6	−2.2	−5.2	+11.1	+8.4	+1.7	−1.3	+0.9
C	−1.5	+22.2	−0.3	−1.0	+0.9	−0.2	−0.2	+0.5	−0.8	+0.2	−0.5	−0.2	−0.2	+0.9	−1.4	−0.6	−0.3	−0.7	−0.6	−0.6
D	−3.1	−1.0	+18.9	−8.6	−0.4	−3.4	−4.9	−0.7	−3.7	−1.1	−1.4	+11.1	−3.8	−5.4	−4.2	−6.2	−4.1	−1.6	−1.2	−2.3
E	−5.0	−1.0	−6.0	−2.9	−0.9	−0.6	−2.2	−0.5	−7.0	−1.2	+2.3	−4.0	−4.2	−2.3	+0.9	−3.8	−4.8	−1.8	−0.9	−2.6
F	+0.1	−0.7	−0.5	−0.8	+19.6	−0.3	+4.0	+1.9	−1.4	−2.3	−5.0	−1.5	+3.1	−1.4	−1.9	−0.2	−0.4	−1.0	−5.4	+3.3
G	+0.4	−1.0	−5.1	+10.9	−2.3	+19.6	−4.5	−0.3	−4.4	−0.6	−2.0	−0.8	−5.7	−2.4	+8.3	−1.0	+2.0	−3.6	−4.4	−2.6
H	−1.0	−0.2	−1.8	−1.2	−1.2	−1.1	−6.6	−0.2	−2.0	−0.7	−2.7	−3.5	−0.9	+1.6	−2.7	+0.3	−0.9	−0.9	−1.2	−3.0
I	+2.3	−1.3	+5.0	+1.4	+5.6	+1.0	+5.0	+4.3	−2.1	+0.9	+2.8	+0.8	−1.0	−0.5	−1.3	−0.3	+2.3	+2.8	+4.7	−4.4
K	−4.6	−1.2	−3.5	+0.2	−2.2	−2.8	−5.1	−2.9	−21.5	−3.5	−5.0	−5.8	−4.6	−5.8	+17.2	−2.8	−4.9	−3.6	−2.3	−4.3
L	+3.4	−1.0	+1.2	−1.6	+0.8	+1.3	+1.5	+8.0	−3.7	+23.5	−5.7	+10.7	−2.7	+0.1	−2.2	+0.3	+0.5	+5.3	+3.5	−0.6
M	+3.5	−0.2	−0.2	+15.2	−1.4	+0.1	+2.1	+1.4	−0.9	−1.7	+24.8	+3.3	−0.3	+7.3	−0.6	+0.8	+2.1	−1.5	+0.6	−0.9
N	−3.8	−0.5	+5.5	−4.7	−2.1	−4.3	−0.8	−2.5	−4.9	−2.9	−0.1	−14.6	+1.2	−3.2	−4.5	−2.2	−4.1	−1.9	+0.9	−2.1
P	−4.2	−0.5	−0.5	−5.3	−1.0	−3.4	−2.0	−0.6	−4.2	−1.6	−1.8	+4.4	+46.6	−2.0	−3.4	−4.0	−4.8	−1.5	−0.9	−0.6
Q	−1.0	−0.2	−3.8	+8.1	−1.7	+0.0	+4.9	−1.2	−5.9	−1.3	+5.1	+3.7	−2.2	+30.5	+16.1	−4.4	−3.7	−2.8	−0.4	−1.0
R	+0.1	+1.5	−2.1	+5.9	−1.6	−1.2	−0.1	−0.5	−7.6	−0.9	−1.8	−2.7	−1.9	+0.5	+4.5	−1.0	−3.0	−0.8	−2.0	−1.6
S	−1.9	−3.4	−11.0	−6.2	−3.1	−3.4	+2.9	−3.1	−7.4	−2.8	−0.7	+2.8	−8.0	−8.1	−10.2	+9.6	2.7	−2.5	+0.0	−4.5
T	+0.5	−1.2	−4.0	−9.0	−3.1	−0.2	−3.0	−3.7	−7.6	−4.0	−1.3	−2.6	−7.2	−7.0	−8.1	−0.3	3.7	−4.4	−2.0	−5.2
V	+11.3	−2.2	+1.9	+5.0	+2.5	−0.6	+10.8	+3.5	−3.7	+2.4	−6.4	−2.3	−2.0	+0.0	+2.0	+5.4	5.5	+19.1	−0.6	−1.3
W	+0.1	−0.5	+4.3	−0.3	−4.5	−0.8	−0.8	−0.4	−0.6	−0.1	+0.3	−0.4	−0.3	+0.2	−1.1	+0.9	0.4	+0.4	+24.7	−4.5
Y	+1.7	−1.2	+2.6	−2.4	−1.2	+0.1	−1.9	−0.3	−2.9	−1.7	−6.3	+4.3	−0.4	−0.8	−2.2	−1.7	−0.7	−0.7	−11.3	+37.9

[a] No occurrences of a buried lysine in a β-sheet were observed in the present sample.

8. RESIDUE INACCESSIBILITY

The difference-substitution table for residue inaccessibility (table 6) shows that most terms on the diagonal are positive, indicating an increase in conservation due to inaccessibility from solvent. A large change occurs for proline ($\Delta P = +0.39 \pm 0.04$). This is again a consequence of the lack of a main-chain amide proton; mutation to any other residue would expose a buried amide proton that would require stabilization by a hydrogen bond acceptor from neighbouring residues. The enhancement in conservation of cystine ($\Delta P = +0.28 \pm 0.08$), valine ($\Delta P = +0.22 \pm 0.03$) and leucine ($\Delta P = +0.24 \pm 0.03$) are the expected consequence of the close-packed, hydrophobic environment of inaccessible residues. More surprising are the large increases in conservation observed for aspartic acid ($\Delta P = +0.60 \pm 0.03$), tryptophan ($\Delta P = +0.33 \pm 0.06$) and tyrosine ($\Delta P = +0.24 \pm 0.04$), which we will consider later.

The effect of combining constraints from different structural features can be seen in table 7, which shows the difference substitution table for inaccessible β-positions; this can be compared with table 6 for inaccessible positions. Of particular interest are the substitution patterns of threonine and serine in inaccessible β-positions, where alanine or valine are almost the only accepted substitutes. The accepted mutations for buried residues prefer a small volume change. A similar effect has recently been observed for mutations in the core of the globin fold (Bardo & Argos 1990).

In combination with a positive ϕ angle, local solvent accessibility has a large effect on the conservation of glycine. For example, 68% of all solvent accessible glycines are in a positive ϕ conformation and these are conserved with a probability of 0.46 ± 0.01. However although only 32% of buried glycines have a positive ϕ conformation they are conserved with a probability of 0.82 ± 0.02.

Let us now consider the variation of substitution properties in differing structural environments. Figure 2*a* shows the clustering of the substitution profiles for alanine. Solvent accessibility has a larger effect on substitution than regular secondary structure. For alanine a positive ϕ angle produces very different substitution patterns with glycine as the preferred substitute. In contrast the substitution patterns for glycine (figure 2*b*) in a positive ϕ environment are similar to those for β and coil environment. Solvent accessibility is also a major factor in conservation for glycine as well as alanine. More interesting are the large differences in the substitution patterns for an α-helical environment, as noted above.

9. INACCESSIBLE SIDECHAIN HYDROGEN BONDS

(*a*) *Sidechain to sidechain*

The difference substitution table for inaccessible residues with sidechain-to-sidechain hydrogen bonds (table 8) shows that the largest increases in conservation are observed for sidechains containing oxygen rather than nitrogen. Thus ΔP for tyrosine is $+0.33 \pm 0.04$ and similar changes are found for aspartic and glutamic acid ($\Delta P = +0.49 \pm 0.04$ and $\Delta P = +0.10 \pm 0.03$, respectively).

Inaccessible salt bridges are expected to involve not

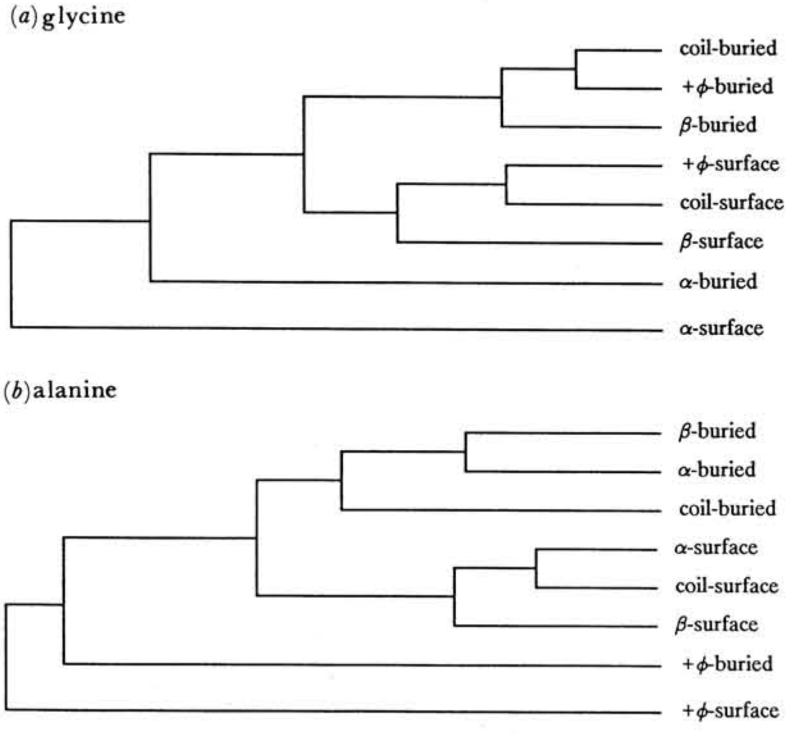

Figure 2. Clustering of the substitution patterns for (*a*) glycine and (*b*) alanine for differing structural environments. The patterns were clustered by using hierarchical cluster analysis based on the χ^2 distance between the respective well-populated probability distributions. Horizontal branch lengths provide a measure of the distances between each environmental grouping.

Table 8. *Difference probabilities (multiplied by 100) for amino acid substitutions involving inaccessible residues hydrogen bonded to another sidechain*

	D	E	H	K	N	Q	R	S	T	W	Y
A	−3.8	+2.7	−5.8	−7.8	−1.1	+2.1	−5.1	+10.6	+5.4	−3.6	−5.5
C	−0.3	−1.0	−0.2	−0.8	−0.2	+1.6	−1.4	+1.8	−0.3	−0.5	+0.1
D	+49.3	−8.7	−2.4	−3.7	+9.3	−5.1	−4.2	−5.9	−4.4	−1.0	−1.7
E	−6.7	+10.3	−2.3	−7.1	+8.4	−1.3	−5.5	−2.9	−4.5	−0.7	−2.6
F	−0.6	−0.8	+7.4	−1.4	+2.0	−1.3	−1.9	−0.4	−1.5	−0.3	+11.7
G	−5.6	+4.7	−4.6	−4.4	−2.7	−2.8	+10.1	+1.0	+6.6	−3.6	−2.6
H	−2.0	−1.2	−6.9	+4.3	−3.6	+0.1	+2.7	−0.1	−0.8	+0.7	−1.7
I	+1.1	+6.3	+5.9	−2.1	+4.4	−0.2	−1.2	+1.5	−2.5	−2.9	−1.7
K	−3.3	−9.7	−5.2	+16.2	−6.0	−6.7	+21.7	−3.1	−5.2	−1.9	−4.3
L	−0.5	−1.7	+3.2	−3.8	+0.0	−2.6	−2.2	+0.8	+0.0	−3.9	−2.2
M	−0.3	+9.2	+6.4	−0.9	+1.5	+5.8	−0.6	+0.1	+0.6	+2.2	−1.6
N	−2.0	−2.7	−2.6	−5.0	−9.8	−2.8	−4.5	−0.4	−4.1	−1.0	−3.5
P	−3.7	−5.4	−2.1	−4.3	+0.8	−2.6	−3.4	−4.1	−1.5	−0.7	−0.6
Q	−4.3	+6.0	+2.4	−6.0	+0.1	+34.2	+19.7	−3.5	−3.5	+0.5	−2.5
R	−2.3	+4.3	−1.0	+36.7	−2.8	−0.8	−0.1	−0.9	−2.8	+0.0	−1.6
S	−10.3	−5.6	−4.7	−1.2	+6.3	−7.8	−10.1	+13.6	+10.8	−1.7	−3.9
T	−4.4	−9.2	−0.6	−7.7	−4.6	−6.7	−8.1	−4.2	+16.9	−1.7	−3.2
V	−0.4	+5.4	+9.2	−3.8	+1.2	−6.1	+2.9	−2.4	−6.5	−0.5	−2.9
W	+1.0	−0.3	+4.2	−0.6	−0.4	+1.3	−1.1	−0.3	−0.4	+18.4	−2.5
Y	−0.9	−2.4	−0.3	+3.3	−2.9	+1.5	−2.2	−1.2	−2.4	+2.4	+32.8

Table 9. *Difference probabilities (multiplied by 100) for amino acid substitutions involving inaccessible residues with a hydrogen bond to a main-chain carbonyl*

	H	K	N	Q	R	S	T	W	Y
A	−5.7	−7.7	+3.7	−0.2	−5.2	+7.0	+6.3	−1.1	+0.0
C	−0.2	−0.8	−0.2	+1.1	−1.4	+0.6	+1.2	−0.5	−0.6
D	−4.9	−3.7	−0.2	−5.3	−4.2	−2.5	−4.3	−1.0	−1.4
E	−2.2	−7.0	−0.1	−3.2	+1.4	−2.5	−4.9	−0.8	−2.5
F	+0.7	−1.4	+2.5	−1.4	−1.9	−0.8	−0.6	−11.0	−7.6
G	−4.5	−4.4	−6.0	−5.0	+9.1	−0.3	−2.3	−3.8	−2.5
H	+10.1	−2.0	+4.5	+6.6	−2.7	−1.5	−0.9	−1.0	−1.4
I	−1.6	−2.1	−0.8	+0.4	−1.3	+2.8	+2.3	−4.8	+8.0
K	−5.1	−21.5	−3.9	−5.5	+39.7	−0.3	−5.7	−2.0	−4.2
L	−1.8	−3.7	+4.5	+0.4	−2.2	+1.6	+0.2	+4.8	+2.0
M	+2.1	−0.9	+3.8	+3.6	−0.6	+1.4	−1.0	−1.3	+0.7
N	+2.5	−4.9	−10.1	−3.0	−4.5	−2.2	−4.6	−1.0	−1.9
P	−2.0	−4.2	+1.3	−1.9	−3.4	−4.0	−4.4	−0.8	−0.6
Q	+1.6	−5.9	+0.6	+27.8	−9.5	−3.2	−3.8	+2.9	+0.7
R	−0.1	−7.6	−2.8	−1.8	+5.8	−1.0	−3.1	−1.8	−0.1
S	−3.8	−7.4	+8.6	−9.2	−10.2	+4.5	−9.6	−1.8	−4.5
T	−3.0	−7.6	−3.9	−6.9	−8.1	−3.2	+30.0	−1.8	−5.1
V	+0.8	−3.7	−2.4	+0.7	+2.4	+5.5	+5.5	−0.9	+1.5
W	+5.8	−0.6	−0.4	+1.8	−1.1	−0.3	−0.4	+36.4	+0.9
Y	+11.4	−2.9	+1.2	+0.9	−2.2	−1.7	−0.1	−8.6	+18.8

only the negatively charged aspartic and glutamic acids, but also the positively charged sidechains of histidine, arginine and lysine. Such inaccessible salt bridges occur rarely within the globular domains on which this sample is based; they are a more common occurrence in inter-domain and inter-protein interactions (Miller *et al.* 1987; Janin *et al.* 1988).

(*b*) *Sidechain-to-main-chain carbonyl*

The difference substitution data for inaccessible residues that are hydrogen bonded to a main-chain carbonyl (table 9) shows that increases in conservation probability occur for tryptophan ($\Delta P = +0.36 \pm 0.07$), glutamine ($\Delta P = +0.28 \pm 0.04$) and tyrosine ($\Delta P = +0.19 \pm 0.05$), respectively. Although glutamine tends to be conserved in this environment, asparagine is not often found conserved when buried and hydrogen bonded to a main-chain carbonyl.

(*c*) *Sidechain-to-main-chain amide*

Figure 3 shows the substitution patterns for buried aspartic acid, asparagine, glutamine, threonine and serine residues that are hydrogen bonded to a main-chain NH via their sidechain. The largest conservation probability is seen for aspartic acid ($P = 0.80 \pm 0.05$). (figure 3*a*). On the infrequent occasions when

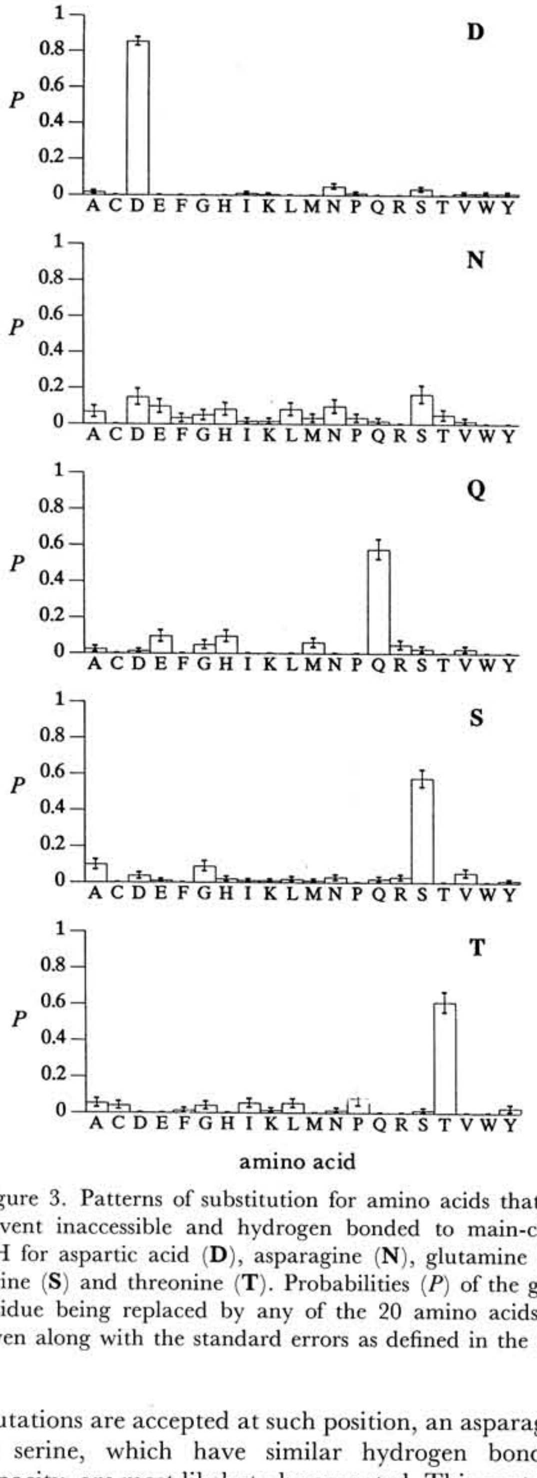

Figure 3. Patterns of substitution for amino acids that are solvent inaccessible and hydrogen bonded to main-chain NH for aspartic acid (**D**), asparagine (**N**), glutamine (**Q**), serine (**S**) and threonine (**T**). Probabilities (P) of the given residue being replaced by any of the 20 amino acids are given along with the standard errors as defined in the text.

mutations are accepted at such position, an asparagine or serine, which have similar hydrogen bonding capacity, are most likely to be accepted. This contrasts strongly with the substitution profile of asparagine (figure 3*b*). Inaccessible asparagines with sidechain-to-main-chain NH hydrogen bonds are highly mutable with a strong tendency to be replaced by aspartic acid or serine rather than remain as asparagine. Surprisingly, glutamine again differs greatly from asparagine but resembles aspartic acid in its high conservation probability ($P = 0.61 \pm 0.08$). Its substitution profile indicates that glutamic acid and histidine are preferred replacements. Similar strong patterns for conservation and variation are shown for buried serine and threonine residues that are hydrogen bonded to a main-chain NH.

From these analyses it is clear that a buried sidechain oxygen hydrogen bonded to a main-chain amide proton plays a larger role in residue conservation than hydrogen bonds to main-chain oxygen or to another sidechain. Such effects have been noted in previous analyses of families of proteins (Hubbard & Blundell 1987), but have not been characterized as a general factor in protein stability. They may reflect the relatively greater importance of satisfying hydrogen bond donor properties of peptide NH compared with the acceptor properties of the peptide carbonyl on removal from aqueous environment. This is usually achieved with a main-chain carbonyl but in some conformations this is not possible; these conformations appear to be characterized by the most conserved pattern of residues that occurs in protein evolution.

Our analysis shows that the most conserved polar residues such as aspartic acid, glutamine, serine or threonine are those that are inaccessible and have at least two hydrogen bonds. In the aspartic proteinase alignment (figure 1), Thr37 (33 in pepsin numbering) is both buried and hydrogen bonded to main-chain NH and CO functions. It is conserved when all sequences of the two topologically similar domains of pepsin-like enzymes are compared. It is also conserved in most of the homologous retroviral proteinases, where this threonine is very occasionally replaced by serine. Asp41 and Ser46 (37 and 42 in pepsin numbering), which are inaccessible with two sidechain hydrogen bonds, are also strongly conserved.

10. APPLICATION OF SUBSTITUTION PATTERNS

We are aware of many potential uses of these environmentally specific amino acid substitution tables. Several applications that we are presently investigating are briefly discussed here.

(*a*) *Construction of templates and identification of key residues*

On the basis of one or more three-dimensional structures one can predict the probability of occurrence of each amino acid for each position. In this way we can construct a simple sequence template for a tertiary structure or identify key residues for a motif.

Figure 4 shows the sequence variability expected on the basis of the three-dimensional structure of four residues in the fourth strand of the third Greek-key motif of γ-II crystallin. This region is a β-strand in which positions 2 and 4 are inaccessible to solvent. Position 4 is a buried serine that is hydrogen bonded to the NH function of a main-chain peptide and so is predicted to be highly conserved. Figure 4 also shows the observed pattern of amino acid substitutions in the equivalent positions of Greek-key motifs of β- and γ-crystallins. This example illustrates how the substitution patterns can provide a remarkably good estimate of sequence variability if the three-dimensional structure of at least one protein is known. This provides a general statistical approach to constructing templates on the basis of the tertiary structure. The method is complementary to the more geometric and analytical approach of Ponder & Richards (1987).

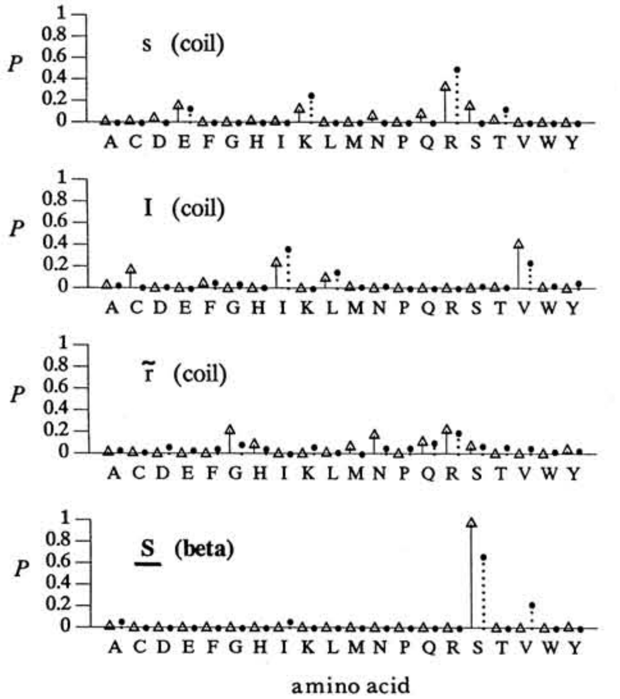

Figure 4. Comparison of probabilities (P) for the predicted pattern (——), (△) of amino acid occurrence with the variability observed among real sequences (....), (●) at strand *d* of motif 3 of eye-lens crystallins. The predicted pattern was derived from the sequence 'Ser-Ile-Arg-Ser' from bovine γ-II crystallin (PDB code: 1GCR). For sequence variability, 155 aligned γ- and β-crystallin domains were used.

A similar approach can be used to identify key residues in loop regions (Chothia *et al.* 1986; Chothia *et al.* 1989) or motifs in knowledge-based modelling. Segments of chain are selected from a database of protein three-dimensional structures so that they overlap either guide points in the electron density (Jones & Thirup 1986) or the framework of the protein modelled from homologous (Blundell *et al.* 1988). Quite often many fragments of acceptable end-point geometry but differing loop conformation are selected. Each of these can be used to generate a template or to identify residues that are likely to be conserved as a result of their inaccessibility, conformation or sidechain hydrogen bonding. The fragments can then be ranked by comparing their templates against the sequence to be modelled. This approach depends on similar structural criteria to those proposed by Chothia *et al.* (1986, 1988) or Sibanda *et al.* (1989) for modelling variable regions, but it is more generally applicable and automated.

(*b*) *Alignment of protein structures*

The substitution tables also allow flexible and sensitive weights to be introduced in the alignment of homologous sequences and structures. The substitution profiles can be used in COMPARER (Z. Zhu & A. Šali, unpublished results) to provide a sensitive residue property that contributes to the accumulated residue-by-residue mass matrix. However, the masses derived in this way take into account the structural context of a substitution, and not just the residue type. For example, the mass for a substitution of asparagine by glutamine would depend not only on the residue types but also on the secondary structure and on whether or not the residues are buried and hydrogen bonded via their sidechains.

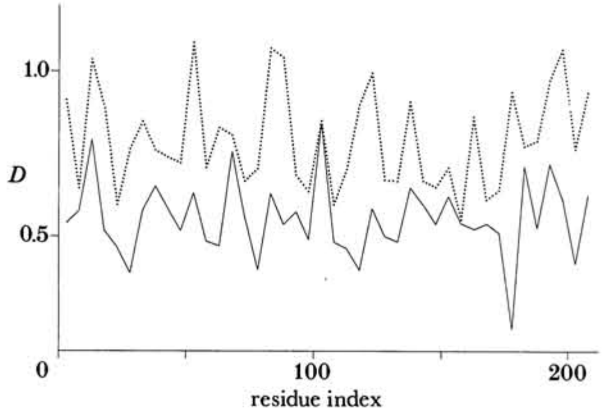

Figure 5. The first 212 of the 223 residues of bovine trypsin, a serine proteinase, were modelled into the fold of papain, (PDB code: 9PAP), a cysteine proteinase. Fifty closely related trypsin-like serine proteinases were used to define sequence variability for each position. The figure shows the difference between expected and observed substitution patterns for the incorrect (....) and correct (PDB code: 2PTN) (——) structures as a function of sequence position.

(*c*) *Comparative modelling*

The results of the analysis of substitution variability can also be used to indicate which features of a protein family are conserved in a homologous or analogous protein. We are incorporating this into a modelling program (Šali *et al.* 1990; A. Šali, unpublished results) in which homologous structures are reconstructed from a complex information function relating known data to interatomic distances. For example, the probability density function for the sidechain hydrogen-bonding distance would depend on whether the equivalent residue in the known homologous structure is hydrogen bonded, on accessibility of the equivalent residue, its residue type, etc.

Similar analyses can be used to test the validity of protein models when several homologous sequences are available. We can test whether the observed pattern of residue substitution is consistent with the proposed structure. Figure 5 shows such a test for a deliberately misfolded structure. The advantage of this technique is that it is sensitive to local errors in conformation, whereas other methods rely on the comparison of a global property to a database of norms, such as those of Baumann *et al.* (1989).

(*d*) *Ab initio structure prediction*

For protein structure prediction the substitution tables allow an analysis of the sequence variability at each position across a family of aligned sequences in terms of structural parameters. For example, they may predict that the invariance of a residue results from its having a particular secondary structure, solvent inaccessibility or sidechain hydrogen bond. Such an application may be understood where the conserved nature of the serine is most probably indicative of an inaccessible, hydrogen-bonded sidechain (figure 4). By matching the observed variability at each position in the alignment with columns of the substitution tables (i.e. figure 4), the most likely environments for each residue may be inferred. This approach provides a secondary/tertiary structure prediction algorithm that could be extended to consider patterns in the sequences in a similar way to that used in many of the standard secondary structure prediction methods, for example Chou & Fasman (1974*b*), Lim (1974) or Garnier *et al.* (1978).

(*e*) *Identification of catalytic/ligand-binding residues*

There is one general class of residues where the substitution patterns are not predicted correctly. These are residues that during function have a varying environment because they interact with a substrate, effector or part of a supramolecular system. The catalytic residues of an enzyme fall into this class. Thus, an analysis of substitution patterns may indicate residues that have roles in catalytic activity. This is evident from the sequences of the aspartic proteinases (figure 1). Asp36 and Tyr80 (32 and 75 in pepsin numbering) are conserved in all aspartic proteinases but neither is both solvent inaccessible and hydrogen bonded through the sidechain. In fact both are involved in substrate binding or catalysis. In complexes of aspartic proteinases with transition state isosteres these residues become both inaccessible and hydrogen bonded to at least two other groups in the complex (see, for example, Foundling *et al.* 1987).

We thank the U.K. Science and Engineering Research Council, Imperial Cancer Research Fund, American Cancer Society, Sloven Research Council, Stefan Institute, Pfizer and Merck Sharp and Dohme for financial support. We are grateful to Janet Thornton, Frank Eisenmenger, Karsten Niefind and our colleagues at Birkbeck College for many stimulating discussions and comments.

REFERENCES

Argos, P. & Palau, J. 1982 Amino acid distribution in protein secondary structures. *Int. J. Pept. Prot. Res.* **19**, 380–392.

Bajaj, M. & Blundell, T. L. 1984 Evolution and the tertiary structure of proteins. *A. Rev. Biophys. Bioeng.* **13**, 453–492.

Baker, E. N. & Hubbard, R. E. 1984 Hydrogen bonding in globular proteins. *Prog. Biophys. molec. Biol.* **44**, 97–139.

Bardo, D. & Argos, P. 1990 Evolution of protein cores. *J. molec. Biol.* **211**, 975.

Baumann, G., Frömmel, C. & Sander, C. 1989 Polarity as a criterion in protein design. *Prot. Engng* **2**, 329–334.

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 macromolecule structures. *J. molec. Biol.* **112**, 535–542.

Blundell, T. L., Barlow, D., Borkakoti, N. & Thornton, J. 1983 Solvent-induced distortions and the curvature of the α-helix. *Nature, Lond.* **306**, 281–283.

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

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

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Maviuzza, R. A., Phillips, S. E. V. & Poljak, R. J. 1986 The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science, Wash.* **233**, 755–758.

Chothia, C. *et al.* 1989 Conformation of immunoglobulin hypervariable regions. *Nature, Lond.* **342**, 877–883.

Chou, P. Y. & Fasman, G. D. 1974*a* Conformational parameters for amino acids in helical β-sheet and random coil regions calculated for proteins. *Biochemistry* **13**, 211–222.

Chou, P. Y. & Fasman, G. D. 1974*b* Prediction of protein conformation. *Biochemistry* **13**, 222–245.

Dayhoff, M. O., Barker, W. C. & Hunt, L. T. 1983 Establishing homologies in proteins. *Meth. Enzymol.* **91**, 524–545.

Eck, R. V. & Dayhoff, M. O. In *Atlas of protein sequence and structure* (ed. M. O. Dayhoff), pp. 33–41. Washington, D.C.: (National Biomedical Research Foundation, Washington, D.C., 1969).

Foundling, S. I. *et al.* 1987 High resolution X-ray analysis of renin inhibitor-aspartic proteinase complexes. *Nature, Lond.* **327**, 349–352.

Garnier, J., Osguthorpe, D. J. & Robson, B. 1978 Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. molec. Biol.* **120**, 97–120.

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

Janin, J., Miller, S. & Chothia, C. 1988 Surface, subunit interfaces and interior of oligomeric proteins. *J. molec. Biol.* **204**, 155–164.

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

Kabsch, W. & Sander, C. 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded geometrical features. *Biopolymers* **22**, 2577–2637.

Lee, B. & Richards, F. M. 1971 The interpretation of protein structures: estimation of static accessibility. *J. molec. Biol.* **55**, 379–400.

Levitt, M. 1978 Conformational preferences in globular proteins. *Biochemistry* **17**, 4277–4285.

Lim, V. I. 1974 Structural principles of the globular organization of protein chains. A steriochemical theory of globular protein secondary structure. *J. molec. Biol.* **88**, 873–894.

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

McLachlan, A. D. 1971 Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome *c*551. *J. molec. Biol.* **6**, 409–424.

Menendez-Arias, L. & Argos, P. 1989 Engineering protein stability: sequence statistics point to residue substitutions in α-helices. *J. molec. Biol.* **206**, 397–406.

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

Nicholson, H., Söderlind, E., Tronrud, D. E. & Matthews, B. W. 1989 Contribution of left-handed helical residues to the structure and stability of bacteriophage T4 lysozyme. *J. molec. Biol.* **210**, 181–193.

Padmanabham, S., Marqusee, S., Ridgeway, T., Laue, T. M. & Baldwin, R. L. 1990 Relative helix-forming tendencies of nonpolar amino acids. *Nature, Lond.* **344**, 268–270.

Pearl, L. & Blundell, T. L. 1984 The active site of aspartic proteinases. *FEBS microbiol. Lett.* **17**, 96–101.

Ponder, J. W. & Richards, F. M. 1987 Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. molec. Biol.* **193**, 775–791.

Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. 1988 Amino acid substitutions in structurally related proteins: a pattern recognition approach. *J. molec. Biol.* **204**, 1019–1029.

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

Šali, A. & Blundell, T. L. 1990 The definition of topological equivalence in homologous and analogous structures: a procedure involving comparison of local properties and relationships. *J. molec. Biol.* **212**, 403–442.

Sibanda, B. L., Blundell, T. L. & Thornton, J. M. 1989 The conformation of β-hairpins in protein structures: a systematic classification with applications to modelling by homology. *J. molec. Biol.* **206**, 759–777.

Taylor, W. R. 1986 The classification of amino acid conservation. *J. theor. Biol.* **119**, 205–218.

(*Received 18 April 1990; Accepted 9 May 1990*)