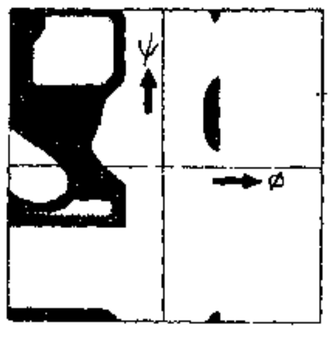

Special issue:
The Ramachandran map and
protein and peptide structures

Applications of environment specific amino acid substitution tables to identification of key residues in protein tertiary structure

John Overington, Mark Johnson, Chris Topham, Alasdair McLeod, Andrej Sali, Zhan-yang Zhu, Lynn Sibanda and Tom Blundell*

Imperial Cancer Research Fund Unit of Structural Molecular Biology, and Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, London University, Malet Street, London WC1E 7HX, UK

Amino acid substitution tables have been calculated for families of homologous proteins of known three-dimensional structures. Amino acids are classified according to residue type, secondary structure, accessibility of the side-chain, and existence of hydrogen bonds from side-chain to other side-chains or peptide carbonyl or amide functions. Distinct patterns of substitution characterize most classes especially where amino-acid residues are both solvent inaccessible and hydrogen-bonded through their side-chains. These tables can be used to identify key residues that are critically important to the three-dimensional structure. They can also be used to identify patterns of amino acids in proteins of unknown three-dimensional structures that are characteristic of globular domains, super-secondary structures or structural motifs.

IT is well established that highly divergent proteins can display little sequence identity but still assume very similar tertiary structures. Nevertheless, a few amino acids are usually strictly conserved or conservatively varied. Some of these may be critical for the function of the protein; for example, they may play an important role in the catalytic mechanism of a class of enzymes. However, one or two amino-acid residues can usually be identified that are conserved because they are crucially important to the integrity of the family fold. These key residues may be formulated as sequence templates which can be used to scan sequence data bases to identify distant phylogenetic relationships.

The invariance of such key residues is a consequence of strong structural constraints that can be met only by a particular amino-acid residue at a particular position in the tertiary fold. These constraints appear to arise from a combination of structural features, which individually lead to conservative variation, but together favour invariance. For example, solvent-inaccessible residues, whose side-chains give a close-packed core, have a relatively lower rate of acceptance of mutations than those on the surface (see for example refs. 1–4). The requirement for an inter-residue hydrogen bond especially with peptide NH functions can also act as a constraint on the substitution of amino ácids. However, the combination of solvent inaccessibility with a hydrogen bond leads often to the invariance of polar residues[5, 6]. Secondary structure also provides strong constraints on sequence variability; α-helices and β-strands have preferred compositions as described by Chou and Fasman[7], and positive main-chain ϕ angle favours glycine[8] and in certain conformations accepts aspartic acid, asparagine or serine (see for example Nicholson *et al.*[9]).

Although there has been a large but rather subjective body of knowledge concerning conservative variation and invariance in the evolution of proteins, this has only recently been characterized and quantified in terms of structural parameters[10]. Comparison of three-dimensional structures has been used to align sequences for families of proteins. On the basis of these alignments substitution tables have been calculated for amino-acid residues classified by secondary structure, solvent accessibility and hydrogen bonding.

In this paper we describe distinct patterns of substitution that characterize specific combinations of structural features. We show that the substitution tables have applications for predicting the variability at each amino acid position in a protein. If the tertiary structure of a protein is known, the substitution tables

*For correspondence.

allow an automatic identification of amino acids that will be invariant or conservatively varied. Key residues can be identified for protein domains, for super-secondary structural motifs such as Greek keys and for individual loops such as those formed by beta-hairpins.

Calculation of environment specific substitution tables

We used the systematic approach encoded in the computer program COMPARER to compare three-dimensional structures[11, 12]. As described previously[10] this has been used to align sequences for families of proteins including the globins, serine and aspartic proteinases, phospholipases, immunoglobulins and gamma-crystallins, for which there are several high-resolution X-ray analyses and coordinates in the Brookhaven Protein Databank[13].

Each residue in each protein structure was considered a member of a class defined by a combination of features. The features considered were residue type (20 values), accessibility (2 values), side-chain hydrogen bonding (8 values) and main-chain conformation (4 values). This gave a maximum of 20×64 classes of amino acids. In fact several amino acids are unable to form hydrogen bonds through their side-chains and most polar residues are unable to act both as donors and acceptors except at extreme pH values. Furthermore inaccessible ion pairs rarely occur except at domain or subunit interfaces which were largely omitted from the study. As a result of these factors the effective number of classes is about three hundred.

All pairwise comparisons of structures in each alignment produced by COMPARER were considered in the analysis, and all substitutions implied by pairwise comparisons were stored in tables as a function of the features identified in the three-dimensional structures. In order to avoid very sparse tables, we considered the structural features of only one of the two proteins compared. For example, if we considered an amino acid that is inaccessible, in an alpha-helix and with a particular hydrogen bonding, we recorded only the residue type of the amino acids observed at topologically equivalent positions. In fact this corresponds to the situation of many applications where only one of the three-dimensional structures of the compared proteins is known. Secondly, in order to understand the general role of certain structural features in constraining the mutability, we accumulated the values across various features. In each case it is convenient to display the data as 20 by 20 matrices where one dimension refers to the amino acid type restricted to a structural environment and the other is simply residue type.

The raw scores were normalized for each of the matrices by division by the respective column sums. Standard errors were associated with each of these probabilities by a formula which accounts for sampling errors[10]. To examine the effect of an environmental feature on the conservation and mutation properties for a given residue, difference mutation matrices were constructed. The values were calculated from the differences between probability matrices for amino acids with and without certain features or combinations of features. An increase in the conservation of a residue, or a more favourable mutation due to the environment of the residue, will be evident by a positive term in this difference matrix.

In order to compare our results with those previously obtained, we calculated a twenty-by-twenty environment-independent mutation matrix, which included all residues in the sample. This is equivalent to summing the matrices for all classes of features. This matrix is directly comparable to those used in standard sequence alignment and analysis techniques, notwithstanding differences due to sample bias. We did not include the time of divergence of the pairs of proteins in the derivation of the matrices because this would have increased the problems arising from sparse matrices. However, most of the data are derived from proteins with between 20% and 40% pairwise sequence identity.

The difference mutation matrix for inaccessibility (Figure 1) shows most terms on the diagonal as positive indicating an increase in conservation due to the inaccessibility from solvent and the packing in the core of the protein. One of the largest positive changes occurs for proline. This is probably a consequence of the lack of a main-chain amide hydrogen; substitution by any other residue would reveal a buried amide hydrogen that would require stabilization by a hydrogen bond acceptor in the neighbouring structure. There is also a large enhancement in conservation of cystine, valine and leucine, which is expected as they often close-packed in a hydrophobic environment of inaccessible residues. More surprising is the large positive values for aspartate, tryptophan and tyrosine, which we will consider later. Difference mutation matrices for main-chain conformations: alpha-helical, beta-strand, positive ϕ and coil were also calculated[10], they reflect the expected preference of amino acids for differing secondary structures reported in earlier analyses of protein structures[7] and also shown by recent experimental work on peptides with stable secondary structure[14]. The difference mutation matrix for residues involved in side-chain–side-chain hydrogen-bonds shows that tyrosine is most affected by the presence of this feature, followed by aspartic and glutamic acids. The changes are of the same order of magnitude as those observed for the effects of secondary structure.

The most characteristic substitution tables occur when combinations of features are considered. Figure

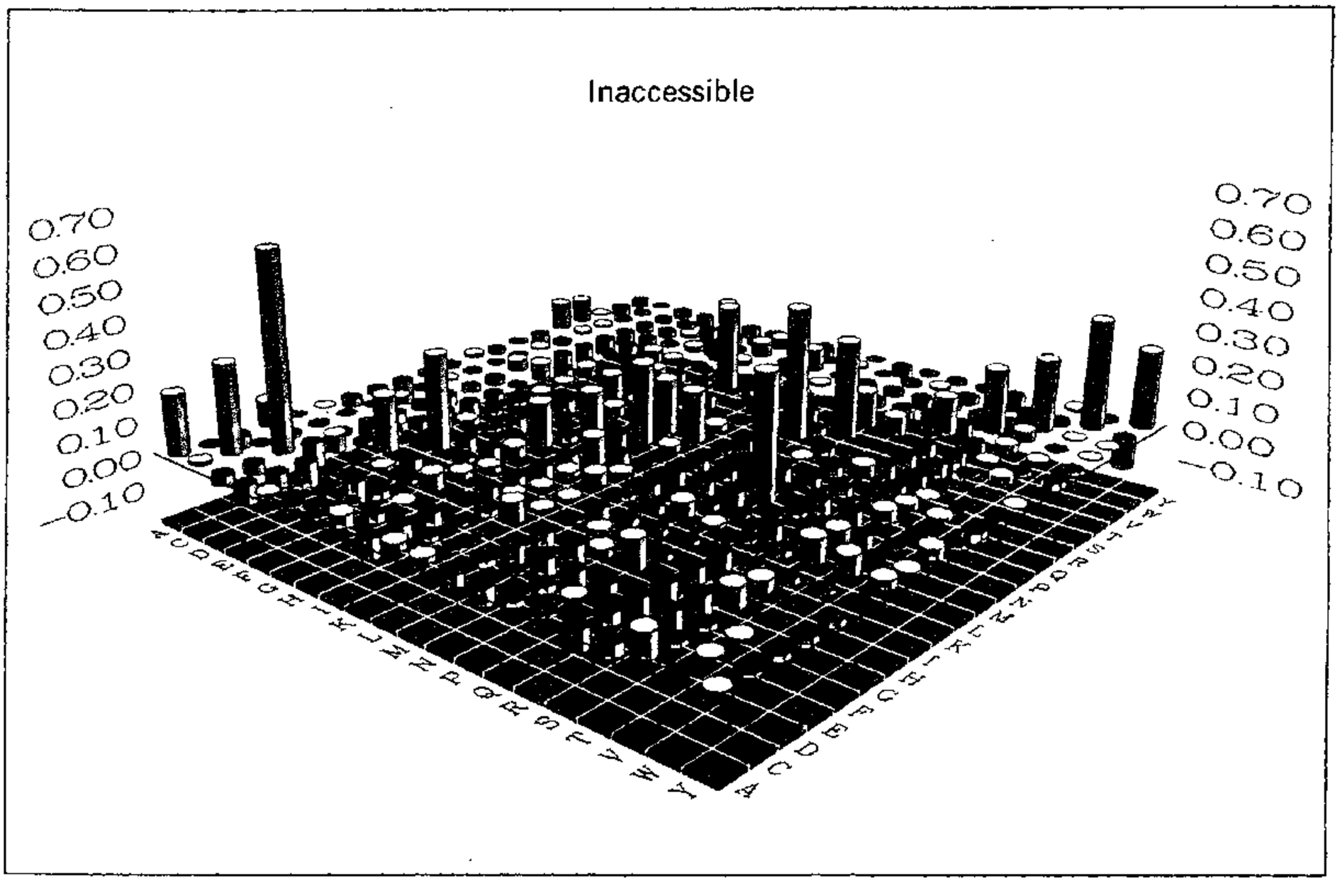

Figure 1. A difference substitution table for amino acids that occupy solvent inaccessible positions in globular proteins. The horizontal axis is that of an amino acid in such an environment in the three-dimensional structure of a protein. The vertical axis is the amino acid type in an homologous protein at a topologically equivalent position defined by COMPARER.

2, *a* shows the difference substitution table for inaccessible residues with side-chain–side-chain interactions. The largest effects observed are for side-chains containing oxygen rather than nitrogen, and most are between a charged amino acid and neutral one rather than in salt bridges. Inaccessible salt bridges would be expected to involve not only oxygen-containing negatively-charged residues such as aspartate and glutamate, but also the positively charged histidine, arginine and lysine, which involve nitrogen. In fact such inaccessible salt bridges rarely occur within globular domains on which the accessibility calculations are largely based; they occur more often at inter-domain and inter-protein interfaces.

The difference substitution data for inaccessible amino acids that have a side-chain to main-chain carbonyl hydrogen bond (Figure 2, *b*) show that tryptophan is the residue whose substitution is most affected by such a hydrogen bond, followed by glutamine and tyrosine. It is surprising that, although glutamine occurs in this group, asparagine which has a similar side-chain amide function is not often found conserved forming a solvent inaccessible hydrogen bond to a carbonyl.

Figure 3 shows the substitution of Asp, Asn, Gln, Thr and Ser residues where there is a side-chain to main-chain nitrogen hydrogen bond. The largest value for conservation is seen for aspartic acid (Figure 3, *a*), which exceeds others attributable to hydrogen bond interactions. On the relatively infrequent occasions when substitutions are accepted at such positions, an asparagine or serine, which have similar hydrogen bonding capacity, are most likely to occur. This contrasts strongly with the substitution patterns of asparagine (Figure 3, *b*). Inaccessible asparagines with side-chain to main-chain NH hydrogen bonds are rarely homomutated but are more often substituted with aspartate or serine; leucines, alanines and many other residues are accepted. Surprisingly glutamine differs greatly from asparagine but resembles aspartate in its relatively high conservation. Its substitution profile indicates that glutamate and histidine are preferred substituents. Similar strong preferences for conservation are shown for solvent inaccessible serine and threonine.

From these analyses it is clear that a side-chain oxygen hydrogen bond to a main-chain nitrogen is a larger factor in residue conservation than hydrogen bonds to main-chain oxygen or to another side-chain. Such effects have been noted in previous analyses of families of proteins[5] but have not been characterized as a general factor in protein stability. The origin of the effect undoubtedly lies in the relatively greater importance of satisfying hydrogen bond donor properties of peptide NH compared to the acceptor properties

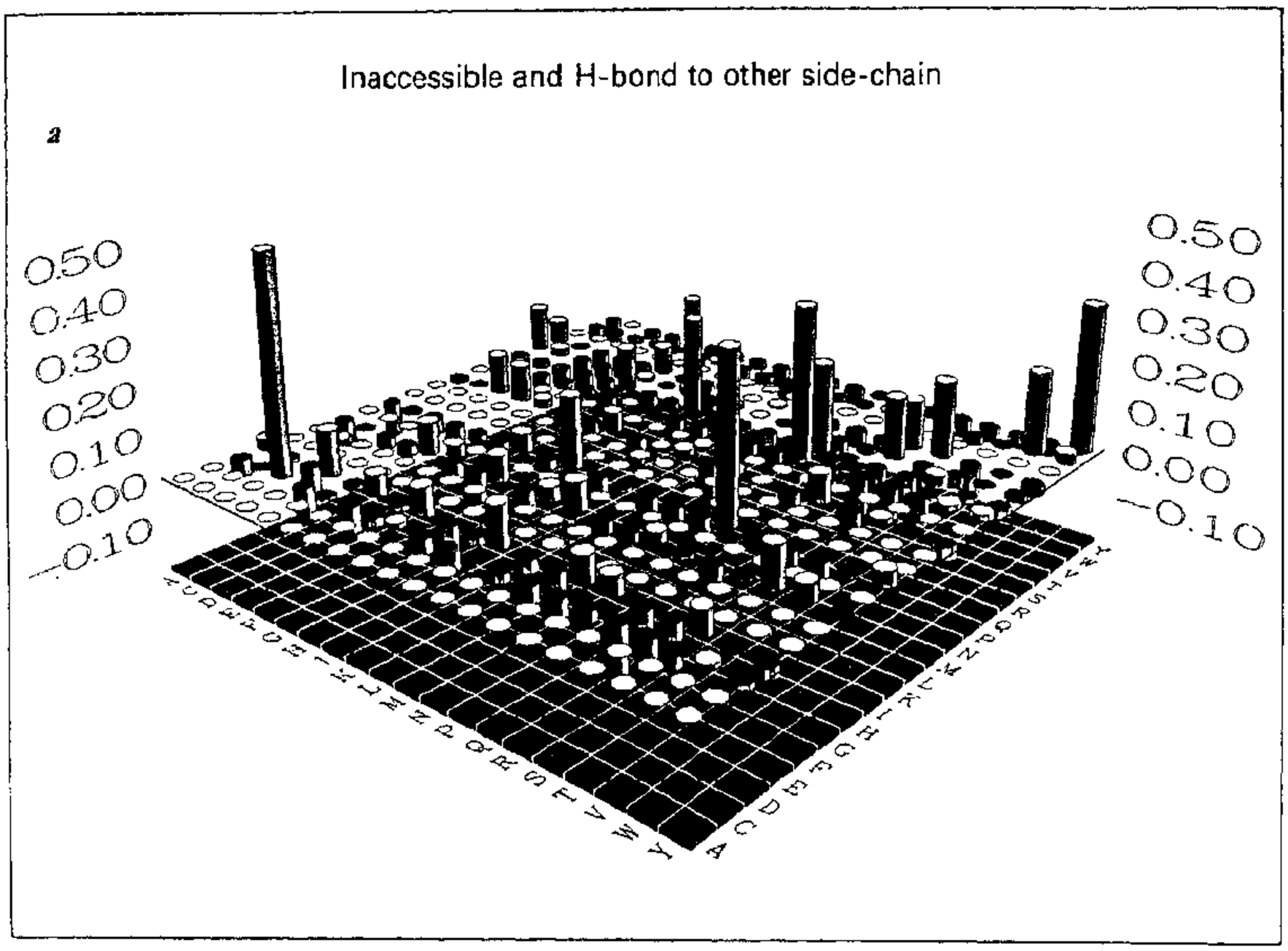

Figure 2. Difference substitution tables for (*a*) inaccessible side-chain–side-chain and (*b*) inaccessible side-chain–main-chain carbonyl amino acids.

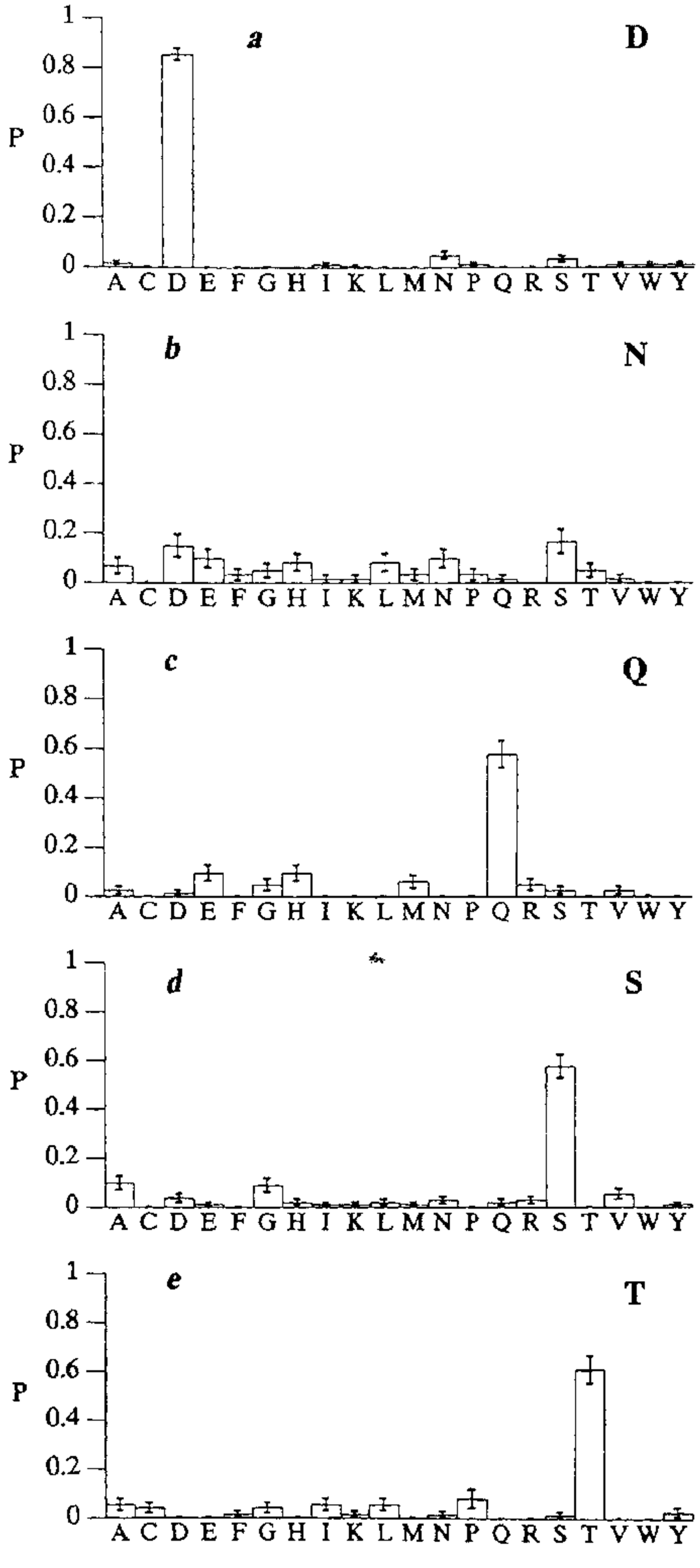

Figure 3. Patterns of substitution for amino acids that are solvent inaccessible *and* hydrogen-bonded to main-chain NH for (*a*) Asp, (*b*) Asn, (*c*) Gln, (*d*) Ser, (*e*) Thr. Probabilities (P) of a given residue being replaced by any of the 20 amino acids are given with standard errors.

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.

Key residues

The analysis shows that the most conserved polar residues such as aspartate, glutamine, serine or threonine are those that are inaccessible and have at least two hydrogen bonds. Let us illustrate this for globular proteins by the aspartic proteinases. Figure 4 shows the alignment of the sequences based on comparison of the three-dimensional structures. Thr-37 (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 it is very occasionally varied to serine. Figure 4 shows that Asp-41 (37 in pepsin numbering) and Ser-46 (42 in pepsin numbering) which are inaccessible with two side-chain hydrogen bonds are also strongly conserved.

Key residues can also be observed in motifs. One such example is found in the beta/gamma crystallins of the vertebrate eye lens. Five independent X-ray analyses have shown that each protomer comprises four similar Greek-key motifs[15, 16]. Each motif consists of four antiparallel beta-strands (a, b, c, d) organized so that the beta hairpin between strands a and b is folded over a beta sheet. Figure 5 shows the alignment of four motifs of gamma-II crystallin based on comparisons of their three-dimensional structures using COMPARER. It can be seen that Ser-39 is completely conserved, and is inaccessible with the side chain hydrogen bonded to main chain CO and NH functions; in this way it is responsible for holding the folded hairpin onto the beta sheet. Also conserved is Gly-13; this residue has a positive ϕ value for its main chain, a conformation that is required to allow the beta hairpin to fold over. Comparisons of all 20 motifs of crystallins defined by X-ray analyses confirm that these structural features are conserved. Comparison of nearly two hundred sequences shows that the residues are invariant with only one exception. Remarkably, the substitution tables allow this conservation to be inferred from the three-dimensional structure of just one motif. Figure 6 shows the substitution pattern predicted in this way for four residues including the invariant serine in the d strand of the third Greek-key motif of gamma-II crystallin. Figure 6 also shows the observed pattern of amino acid substitutions in the equivalent positions of Greek-key motifs of beta and gamma-crystallins. Note the inaccessible, hydrogen bonded serine is correctly predicted as virtually invariant whereas the broader substitution of a neighbouring serine is also quite well

```
                    10        20        30        40        50
4APE     s t g s a t T t p i d s l D d a Y i T p V q I G t p a q t L n L d F D T G s S D L W V F S s e T t
2APP   a a s g v A t N t P t a - n D e e Y i T p V t I g - - g t t L n L n F d T G s A D L W V F S t e L p
2APR   a g v G t V p M t D y g - n d i e Y y G q V t I G t p G k k F n L d F d T g s S D L W I A S t l C t
pep      i g d E p L e N y - - l d t e Y f g t I G I G t p a q d F t V i F d T G s S N L W V P S y y C s
chy    g e v A s V p L t n y - - l d s q Y f g k I y L G t p p q e F t V l F d T G s S D F W V P S i y C k
       ββββ ββ     +   βββββββ+     βββββββββ     βββ

                    60        70        80        90        100
4APE   a s e v d g Q t i Y t p s k S t t A k l l s g A t W s i s y g d g S s S s g d V y t D t V s V g g L
2APP   a s q q s g H s v Y n P s a - - t G k e l s g y t W s i s y g d g S s A s g n V f t d s V t V g g V
2APR   - n C g s g Q t k Y d p n q S s t y q a d - g r t W s i s y g d g s s A s g i L A k D n V n L g g l
pep    s l A C s d h n q F n P d d S s t f e a t - s q e L s i t y g t - g s M t G i L G y D t V q V G g i
chy    s n A C k n h q r F d P r k S s t f q n l - g k p L s i h y g t - G s M q G i L G y D t V t V s n I
       ααα     βββ   βββββββ     βββ  βββββββββ++β

                    110       120       130       140       150
4APE   t V t g Q A V E S A k k V s - s s f t e d s t i D G l L G L A f s t l N t V s p t q q k T F F d n A
2APP   t A h g Q A V Q A A q q I s - a q f q q d t n n D G l L G L A f s s i N t V q p q s q t T F F d t V
2APR   l l k g Q t I E L A k r E a - a s f a s g - p N D G L L G L G f d t i T t v r g - - V k T P M d n L
pep    s D t n Q i F G L S e t E p g s f L y y A - p F D G I L G L A Y p s i S a s - - - g a t P V F d n L
chy    v D i q Q T V G L S t q E p g d f F t y A - e F D G I L G M A Y p s l A s e - - - y S i P V F d n M
       βββ+βββββββ     ααα     +βββ     ααα     ααααα
```

Figure 4. A section of the alignment of sequences of aspartic proteinases achieved by comparing the three-dimensional structures using COMPARER (Sali and Blundell[11],). APE: endothiapepsin; APP: penicillopepsin; APR: rhizopuspepsin; PEP: hexagonal porcine pepsin; CHY: calf chymosin. The coordinates of the three-dimensional structures were obtained from the PDB databank[13]. The amino acid code is the standard one-letter code formatted using the following convention (J. Overington, unpublished results): *Italic* for positive ϕ; UPPER CASE for solvent inaccessible residues; lower case for solvent accessible residues; **bold type** for hydrogen bonds to main-chain amide nitrogen; underline for hydrogen bonds to main-chain carbonyl oxygen; tilde ~ for side-chain–side-chain hydrogen bonds.

```
                10        20        30        40
G2m1     G k I T F y E d r g f q g h c y e C s s - d C p n L q p y F - - s - - r C N S I r V d s - - - - -
G4m1     G k I T F y E d r G f q G r h y e C s s - d h s n L q p y F - - s - - r C N S I r V d s - - - - -
G2m3     f r M r I y e r d d f r G q m S e I t d - d C p S L q d r f h l s - - e V h S L N V l e - - - - -
G4m3     h r L r I Y e r e d y r G q m v e I t e - d C s S L q d r f h f s - - d I h S F h V m e - - - - -
G2m2     G c W m L Y E r p n y q G h q Y f L r r G d y p D y q - q W m G f n d S I r S C r l i p q h t g t
G4m2     G c W m L Y e q p n f t G c q Y f L r r g d y p d y q - q W m G f s d S V r S C r l i p h t - s s
G2m4     G s W v L Y E m p s y r G r q y l L r p g e y r r y l - d W g A m n A k V G S L r r v m d f y - -
G4m4     G y W v L Y E m p n y r G r q y l L r p g d y r r y l - d W g A a n A r V G S L r r a v d f y - -
```

Figure 5. Alignment of sequences of four Greek-key motifs of gamma-II (G2) and γ-IV (G4) crystallins obtained by comparing their three-dimensional structures. One-letter code as in Figure 4.

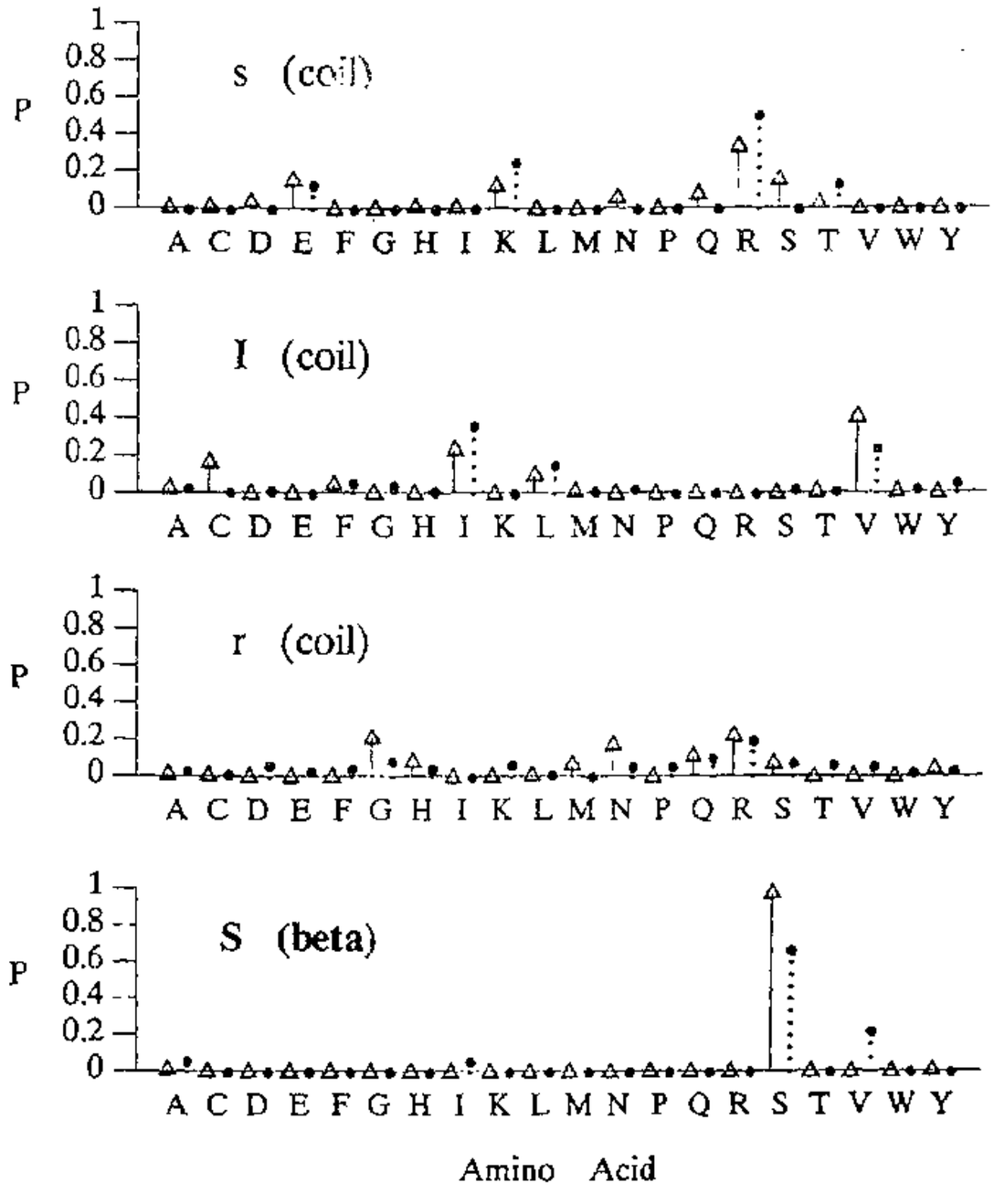

Figure 6. Comparison of predicted residues on the basis of substitution tables and the variability observed at strand d of motif 3 of beta/gamma crystallins. Predicted (●); observed (▵).

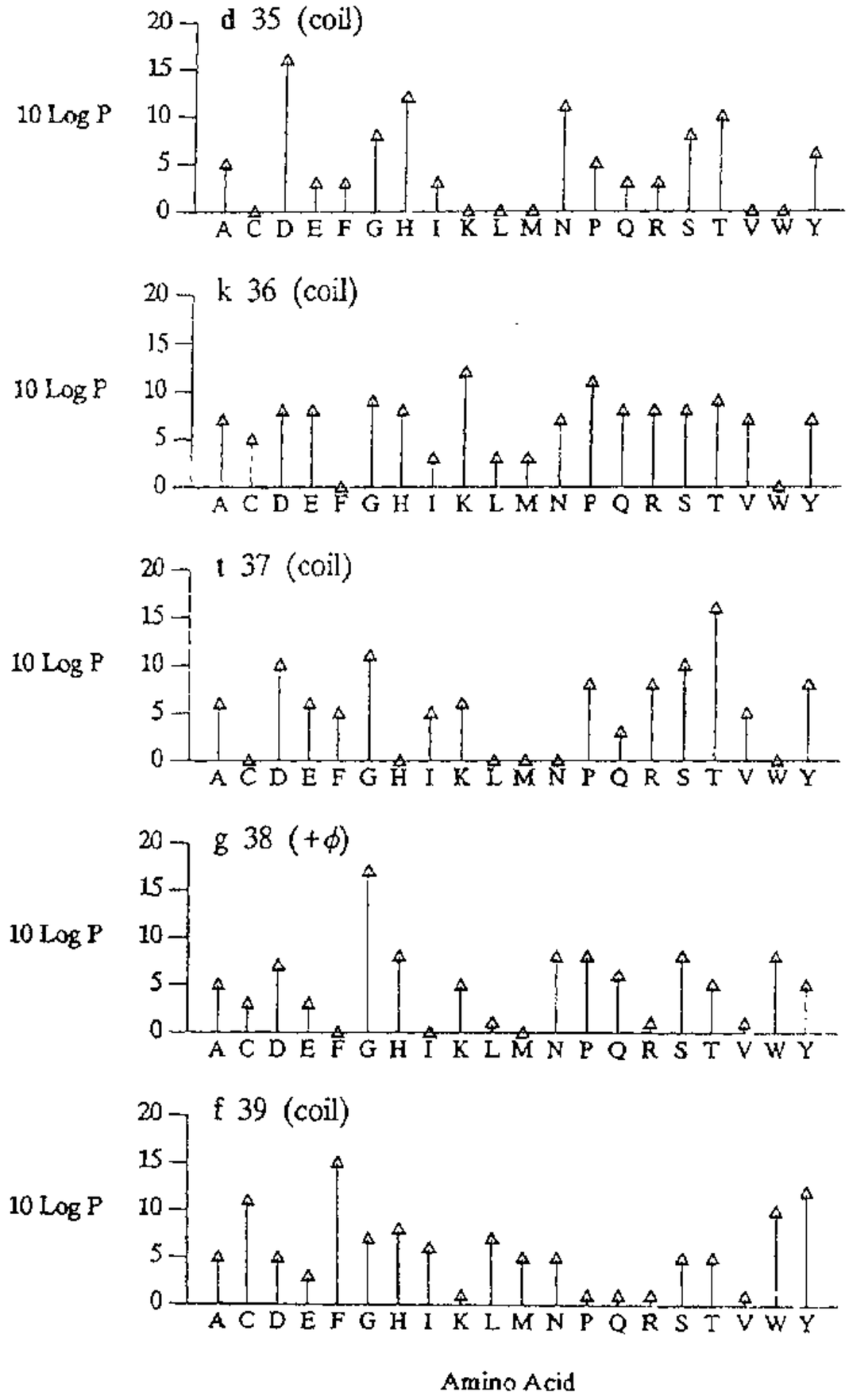

Figure 7. Predicted substitution pattern for a 3:5 beta-hairpin structure (chymotrypsin, sequence 35–39, Asp–Lys–Thr–Gly–Phe), comprising a type I beta-turn followed by a residue at position 4 with a positive main-chain ϕ torsion angle (Sibanda *et al.*[18]).

predicted. The gamma crystallins were of course omitted from the calculation of the substitution tables used in these experiments. The example shows how the mutation matrices can provide a remarkably good indication of the key residues that are necessary for a structural motif provided 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 and Richards[17].

A similar approach can be used for identifying key residues in loop regions or motifs. Let us consider a characteristic 3:5 beta-hairpin, the conformation of which comprises a type I turn followed by a residue with a positive ϕ torsion angle at position 4 (ref. 18). Figure 7 shows the sequence of one such turn along with the predicted substitution pattern. The relative conservation of residue 4 and the patterns for each position in the hairpin are well predicted. Such a procedure can be very useful in rule-based modelling procedures where segments of chain are selected from a database of protein three-dimensional structures so that they overlap either guide points in the electron density[19] or the framework of the protein modelled from homologues[20]. Quite often more than ten fragments of acceptable geometry 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 side-chain hydrogen bonding, and the fragments can then be ranked by comparing their templates against the sequence to be modelled.

1. Miller, S., Janin, J., Lesk, A. M. and Chothia, C., *J. Mol. Biol.*, 1987, **210**, 181.
2. Hubbard, T. J. P. and Blundell, T. L., *Protein Eng.*, 1987, **1**, 159.
3. Lim, W. A. and Sauer, R. T., *Nature*, 1989, **339**, 31.
4. Bardo, D. and Argos, P., *J. Mol. Biol.*, 1990, **211**, 975.
5. Bajaj, M. and Blundell, T. L., *Annu. Rev. Biophys. Bioeng.*, 1984, **13**, 453.
6. Blundell, T. L., *Chem. Scr.*, 1986, **B26**, 213.
7. Chou, P. Y. and Fasman, G. D., *Biochemistry*, 1974, **13**, 211.

8. Ramachandran, G. N. and Sasisekharan, V., *Adv. Protein Chem.*, 1968, **23**, 283.
9. Nicholson, H., Söderlind, E., Tronrud, D. E. and Matthews, B. W., *J. Mol. Biol.*, 1989, **210**, 181.
10. Overington, J., Johnson, M., Sali, A., Blundell, T. L., *Proc. R. Soc. London*, 1990 (in press).
11. Sali, A. and Blundell, T. L., *J. Mol. Biol.*, 1990, **212**, 403.
12. Zhu, Z., Sali, A. and Blundell, T. L. 1990 (manuscript in preparation).
13. Bernstein, F. C., *et al.*, *J. Mol. Biol.*, 1977, **112**, 535.
14. Padmanabhan, S., Marqusee, S., Ridgway, T., Laue, T. M. and Baldwin, R. L., *Nature*, 1990, **344**, 268.
15. Blundell, T. L., Lindley, P. F., Miller, L., Moss, D. S., Slingsby, C., Tickle, I. J., Turnell, W. G. and Wistow, G., *Nature*, 1981, **289**, 771.
16. Bax, V., Lapatto, R., Driessen, H., Nalini, V., Blundell, T. L., Slingsby, C., *Nature*, 1990 (submitted).
17. Ponder, J. W., and Richards, F. M., *J. Mol. Biol.*, 1987, **193**, 775.
18. Sibanda, B. L., Blundell, T. L. and Thornton, J. M., *J. Mol. Biol.*, 1989, **206**, 759.
19. Jones, T. H. and Thirup, S., *EMBO J.*, 1986, **5**, 819.
20. Blundell, T. L. *et al.*, *Eur. J. Biochem.*, 1988, **172**, 513.
21. Sibanda, B. L. and Thornton, J. M., *Nature*, 1985, **316**, 170.

ACKNOWLEDGEMENTS. We thank the UK Science and Engineering Research Council, Imperial Cancer Research Fund, American Cancer Society, Slovenian Research Council, The J. Stefan Institute, the Royal Society, the European Commission, Pfizer, Merck Sharp and Dohme for financial support. We are grateful to our colleagues Frank Eisenmenger, Janet Thornton, Devon Carney, Pam Thomas, Karsten Niefind and Dan Donnelly for many stimulating discussions.