Protein Science (1992), *1*, 216-226. Cambridge University Press. Printed in the USA.
Copyright © 1992 The Protein Society 0961-8368/92 $5.00 + .00

Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds

JOHN OVERINGTON, DAN DONNELLY, MARK S. JOHNSON, ANDREJ ŠALI, AND TOM L. BLUNDELL
ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College,
University of London, Malet Street, London WC1E 7HX, UK

(Received August 14, 1991; Accepted October 3, 1991)

Abstract

The local environment of an amino acid in a folded protein determines the acceptability of mutations at that position. In order to characterize and quantify these structural constraints, we have made a comparative analysis of families of homologous proteins. Residues in each structure are classified according to amino acid type, secondary structure, accessibility of the side chain, and existence of hydrogen bonds from the side chain. Analysis of the pattern of observed substitutions as a function of local environment shows that there are distinct patterns, especially for buried polar residues. The substitution data tables are available on diskette with *Protein Science*. Given the fold of a protein, one is able to predict sequences compatible with the fold (profiles or templates) and potentially to discriminate between a correctly folded and misfolded protein. Conversely, analysis of residue variation across a family of aligned sequences in terms of substitution profiles can allow prediction of secondary structure or tertiary environment.

Keywords: data-base searching; profiles; residue conservation; sequence alignment; structure prediction; templates

The basis of the acceptance or rejection of amino acid mutations in evolution cannot be fully understood without knowledge of the tertiary structure and function of a protein. A recent study (Overington et al., 1990) established the nature of the structural constraints that led to invariance or restrictive variation at equivalent positions in families of proteins. In this paper we extend our analysis and its applications. In particular we make available the revised substitution tables in computer readable form on the diskette associated with *Protein Science*.

Protein structures and structural parameters

Our analysis relies on a data base of homologous, aligned three-dimensional structures (Table 1); all the alignments were performed with the computer program COMPARER (Šali & Blundell, 1990). An example of an alignment used in the study is shown in Figure 1. In this figure, the standard one-letter amino acid code has been modified to concisely show the local environment and conformation of each residue. Homologous families were selected from the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977) using the results of previous studies in our laboratory (the list presented in Table 1 is in no way meant to be complete and exhaustive). We included some rather poorly refined structures in our data base in order to increase the counting statistics. Analysis of the distribution of unusual environments (for example the impossible $+\phi$ conformation for proline) showed that these tended to occur in the poorly refined data sets. However, the contribution of such structures to the overall substitution tables is very small. The average pairwise sequence identity of these families varies from around 80% for the γ-crystallins to around 25% for the globins. In total the present data base contains 21,651 residues and 79,983 pairwise residue substitutions. Analysis of the number of substitutions as a function of percentage sequence identity of the compared proteins shows that 60% of the substitution data comes from sequences that are 20-40% identical. The families include most known classes of secondary and supersecondary structures and so should be representative of protein structures in general.

The structural features considered in our analysis included:

1. Residue type. Twenty-one amino acids were considered. Cystine and cysteine were considered to be dif-

Reprint requests to: Tom L. Blundell, ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

Table 1. *Protein structures in the Brookhaven alignment data base*[a]

PDB code	Name	n_{residues}	Resolution (Å)	R_{factor} (%)
1. Aspartic proteinase				
2APP	Penicillopepsin (*Penicillium janthinellum*)	323	1.8	13.6
2APR	Rhizopuspepsin (*Rhizopus chinensis*)	325	1.8	14.3
2CMS	Chymosin (*Bos taurus*)	323	2.2	18.8
2REN	Renin (*Homo sapiens*)	313	2.8	19.6
5PEP	Pepsin (*Sus scrofa*)	326	2.0	18.0
4APE	Endothiapepsin (*Endothia parasitica*)	330	2.1	17.8
2. Azurin/plastocyanin				
1AZU	Azurin (*Pseudomonas aeruginosa*)	126	2.7	35
1PAZ	Pseudoazurin (*Alcaligenes faecalis*)	120	1.5	15.9
1PCY	Plastocyanin (*Populus nigra italica*)	99	1.6	17.0
2AZA	Azurin (*Alcaligenes denitrificans*)	129	2.5	15.7
7PCY	Plastocyanin (*Enteromorpha prolifera*)	98	1.8	11.7
3. Calcium-binding protein				
3CLN	Calmodulin (*Rattus rattus*)	143	2.2	17.5
5TNC	Troponin C (*Meleagris gallopavo*)	161	2.0	15.5
4. Carbonic anhydrase				
1CA2	Carbonic anhydrase II (*Homo sapiens*)	259	2.0	17.3
2CAB	Carbonic anhydrase B (*Homo sapiens*)	255	2.0	19.3
5. Cysteine proteinase				
2ACT	Actinidin (*Actinida chinensis*)	218	1.7	16.5
9PAP	Papain (*Carica papaya*)	212	1.6	16.1
6. Cytochrome-*b*				
1FCB	Flavocytochrome b_2 (*Saccharomyces cerevisiae*)	92	2.4	18.8
3B5C	Cytochrome-b_5 (*Escherichia coli*)	106	1.4	16.4
7. Cytochrome-*c*				
1CCR	Cytochrome-*c* (*Oryza sativa*)	111	1.5	19.0
2C2C	Cytochrome-c_2 (*Rhodospirillum rubrum*)	112	2.0	17.2
5CYT	Cytochrome-*c* (*Thunnus alalunga*)	103	1.5	15.9
8. Cytochrome-c_5				
1CC5	Cytochrome-c_5 (*Azotobacter vinelandii*)	83	2.5	29
351C	Cytochrome-c_{551} (*Pseudomonas aeruginosa*)	82	1.6	19.5
9. Dihydrofolate reductase				
1DHF	Dihydrofolate reductase (*Homo sapiens*)	186	2.3	17.6
3DFR	Dihydrofolate reductase (*Lactobacillus casei*)	162	1.7	15.2
4DFR	Dihydrofolate reductase (*Escherichia coli*)	159	1.7	15.5
8DFR	Dihydrofolate reductase (*Gallus gallus*)	186	1.7	18.8
10. Ferredoxin				
1FDX	Ferredoxin (*Peptococcus aerogenes*)	54	2.0	20.6
4FD1	Ferredoxin (*Azotobacter vinelandii*)	106	1.9	21.2
11. Flavin-binding β-barrel				
1GOX	Glycolate oxidase (*Spinacia oleracea*)	369	2.0	18.9
1FCB	Flavocytochrome b_2 (*Saccharomyces cerevisiae*)	402	2.4	18.8
12. Flavodoxin				
1FX1	Flavodoxin (*Desulfovibrio vulgaris*)	147	2.0	–
3FXN	Flavodoxin (*Clostridium* sp.)	138	1.9	21.4
13. γ-crystallin				
1GCR	γ-II crystallin (*Bos taurus*)	174	1.6	23
2GCR	γ-IV crystallin (*Bos taurus*)	173	2.3	14.3
14. Globin				
1ECD	Erythrocruorin (*Chironomus thummi thummi*)	136	1.4	19.0
1MBA	Myoglobin (*Aplysia limnacia*)	146	1.6	19.3
1PMB	Myoglobin (*Sus scrofa*)	153	2.5	18.5
2HHB	Hemoglobin (*Homo sapiens*) α-chain	141	1.7	16.0
2HHB	Hemoglobin (*Homo sapiens*) β-chain	146	1.7	16.0
2LH1	Leghemoglobin (*Lupinus luteum*)	153	2.0	
2LHB	Hemoglobin (*Petromyzon marinus*)	149	2.0	14.2
2MHB	Hemoglobin (*Equus caballus*) α-chain	141	1.7	23.0
2MHB	Hemoglobin (*Equus caballus*) β-chain	146	1.7	23.0
4MBN	Myoglobin (*Physeter catodon*)	153	2.0	17.2
15. Glyceraldehyde phosphate dehydrogenase				
1GD1	Glyceraldehyde 3-phosphate dehydrogenase (*Bacillus stearothermophilus*)	334	1.8	17.7
1GPD	Glyceraldehyde 3-phosphate dehydrogenase (*Homarus americanus*)	333	2.9	
3GPD	Glyceraldehyde 3-phosphate dehydrogenase (*Homo sapiens*)	334	3.5	33.0

(*continued*)

Table 1. *Continued*

PDB code	Name	n_{residues}	Resolution (Å)	R_{factor} (%)
16. Hemerythrin				
1HMQ	Hemerythrin (*Themiste dyscritum*)	113	2.0	17.3
2MHR	Myohemerythrin (*Themiste zostericola*)	118	1.7	15.8
17. Immunoglobulin constant domain				
1FBJ	Immunoglobulin FAB CL1 (*Mus musculus*)	102	2.6	19.0
1FBJ	Immunoglobulin FAB CH1 (*Mus musculus*)	95	2.6	19.0
1FC1	Immunoglobulin FC CH2 (*Homo sapiens*)	104	2.9	22.0
1FC1	Immunoglobulin FC CH3 (*Homo sapiens*)	103	2.9	22.0
2FB4	Immunoglobulin FAB KOL CL1 (*Homo sapiens*)	100	1.9	18.9
2FB4	Immunoglobulin FAB KOL CH1 (*Homo sapiens*)	103	1.9	18.9
2HFL	Immunoglobulin HY-HEL FAB CH1 (*Mus musculus*)	97	2.5	24.5
18. Immunoglobulin variable domain				
1FBJ	Immunoglobulin FAB LV (*Mus musculus*)	111	2.6	19.0
1FBJ	Immunoglobulin FAB HV (*Mus musculus*)	123	2.6	19.0
1REI	Immunoglobulin Bence-Jones LV (*Homo sapiens*)	107	2.0	23.0
2FB4	Immunoglobulin FAB KOL LV (*Homo sapiens*)	116	1.9	18.9
2FB4	Immunoglobulin FAB KOL HV (*Homo sapiens*)	126	1.9	18.9
2HFL	Immunoglobulin HY-HEL FAB LV (*Mus musculus*)	110	2.5	24.5
2HFL	Immunoglobulin HY-HEL FAB HV (*Mus musculus*)	116	2.5	24.5
2RHE	Immunoglobulin Bence-Jones LV (*Homo sapiens*)	114	1.6	14.9
3FAB	Immunoglobulin λ-FAB LV (*Homo sapiens*)	103	2.0	–
3FAB	Immunoglobulin λ-FAB HV (*Homo sapiens*)	117	2.0	–
19. Kazal-type serine proteinase inhibitor				
1OVO	Third domain ovomucoid inhibitor (*Coturnix coturnix japonica*)	56	1.9	
2OVO	Third domain ovomucoid inhibitor (*Lophura nycthemera*)	56	1.5	19.9
1TGS	Pancreatic secretory trypsin inhibitor (*Sus scrofa*)	55	1.8	18.6
20. Lactate/malate dehydrogenase				
1LDB	Lactate dehydrogenase (*Bacillus stearothermophilus*)	294	2.8	28.6
2LDX	Lactate dehydrogenase (*Mus musculus*)	331	3.0	25.6
4MDH	Malate dehydrogenase (*Sus scrofa*)	334	2.5	16.7
5LDH	Lactate dehydrogenase (*Sus scrofa*)	333	2.7	16.6
6LDH	Lactate dehydrogenase (*Squalis acanthias*)	329	2.8	20.2
21. Lysozyme				
1ALC	α-lactalbumin (*Papio cynocephalus*)	122	1.7	22.0
1LZ1	Lysozyme (*Homo sapiens*)	130	1.5	18.7
1LZT	Lysozyme (*Gallus gallus*)	129	2.0	25.4
2LZ2	Lysozyme (*Meleagris gallopavo*)	129	2.2	19.2
22. Periplasmic binding protein				
2LBP	Leucine binding protein (*Escherichia coli*)	346	2.4	21.3
2LIV	Leu/Ile/Val binding protein (*Escherichia coli*)	344	2.4	17.9
23. Phosphofructokinase				
1PFK	Phosphofructokinase (*Escherichia coli*)	320	2.4	16.5
4PFK	Phosphofructokinase (*Bacillus stearothermophilus*)	319	2.4	16.9
24. Phospholipase A_2				
1BP2	Phospholipase A_2 (*Bos taurus*)	123	1.7	17.1
1P2P	Phospholipase A_2 (*Sus scrofa*)	124	2.6	24.1
1PP2	Phospholipase A_2 (*Crotalus atrox*)	122	2.5	17.8
25. Photosynthetic reaction center				
1PRC	Photosynthetic reaction center (*Rhodopseudomonas viridis*)	854	2.3	19.3
1RCR	Photosynthetic reaction center (*Rhodobacter sphaeroides*)	848	2.4	26.0
26. Potato-type serine proteinase inhibitors				
2CI2	Barley seed chymotrypsin inhibitor (*Hordeum vulgare*)	83	2.0	19.8
1CSE	Eglin (*Hirudo medicinalis*)	71	1.2	17.8
27. Repressors				
2CRO	Cro repressor (phage 434)	65	2.3	19.5
1R69	Repressor (phage 434)	63	2.0	19.3
1LRD	λ repressor (bacteriophage λ)	92	2.5	24.2
28. Rubredoxin				
1RDG	Rubredoxin (*Desulfovibrio gigas*)	52	1.4	13.6
3RXN	Rubredoxin (*Desulfovibrio vulgaris*)	52	1.5	24.2
4RXN	Rubredoxin (*Clostridium pasteurianum*)	54	1.2	12.8

(*continued*)

Table 1. *Continued*

PDB code	Name	n_{residues}	Resolution (Å)	R_{factor} (%)
29. Retroviral proteinase				
2RSP	Rous sarcoma proteinase	112	2.0	19.0
5HVP	HIV-1 proteinase	99	2.3	18.9
30. Serine proteinase (bacterial)				
2ALP	α-lytic proteinase (*Lysobacter enzymogenes*)	198	1.7	13.1
2SGA	Proteinase A (*Streptomyces griseus*)	181	1.5	12.6
3SGB	Proteinase B (*Streptomyces griseus*)	185	1.8	12.5
31. Serine proteinase (mammalian)				
1HNE	Neutrophil elastase (*Homo sapiens*)	218	1.8	16.4
1SGT	Trypsin (*Streptomyces griseus*)	223	1.7	16.1
1TON	Tonin (*Rattus rattus*)	227	1.8	19.6
1TRM	Trypsin (*Rattus rattus*)	223	2.3	16.0
2GCH	γ-chymotrypsin (*Bos taurus*)	226	1.7	18.1
2PKA	Kallikrein A (*Sus scrofa*)	232	2.0	22.0
2PTN	Trypsin (*Bos taurus*)	223	1.5	19.3
3EST	Pancreatic elastase (*Sus scrofa*)	240	1.6	16.9
3RP2	Mast cell proteinase (*Rattus rattus*)	224	1.9	19.1
32. Serine proteinase (subtilisin)				
1SBC	Subtilisin Carlsberg (*Bacillus subtilisin*)	274	2.5	20.6
1SBT	Subtilisin BPN′ (*Bacillus amyloliquefaciens*)	275	2.5	44.0
1TEC	Thermitase (*Thermoactinomyces vulgaris*)	279	2.2	17.9
2PRK	Proteinase K (*Trittirachium album limber*)	279	1.5	16.7
33. Snake toxin				
1CTX	α-cobratoxin (*Naja naja siamensis*)	71	2.8	
1NXB	Neurotoxin b (*Laticauda semifasciata*)	62	1.38	24.0
34. Triose phosphate isomerase				
1TIM	Triose phosphate isomerase (*Gallus gallus*)	247	2.5	
1YPI	Triose phosphate isomerase (*Saccharomyces cerevisiae*)	247	1.9	21.0

[a] To save space the references for the tertiary structure determination are not shown; this information is available on request from the authors.

ferent residues as they have distinct preferences for both local environment and differences in the patterns of accepted mutations. The assignment as a cystine was defined on the basis of a 2.5-Å sulfur-to-sulfur atom distance cutoff.

2. Main-chain conformation and secondary structure. Residues with the unusual positive ϕ main-chain angle were assigned first. α-helices and β-strands were then defined using the SSTRUC program of David Smith, which implements the algorithm of Kabsch and Sander (1983). The 3_{10} and α-helices were treated equivalently; in our sample 3_{10} helices make up around 10% of all helical residues. Finally, residues as yet undefined were classified as coil. The *cis*-peptide conformation was also examined; the small number of examples and its almost exclusive restriction to the N-terminal side of proline in the alignment data base made us include it with the coil class.

3. Solvent accessibility. Side-chain accessibilities were calculated by the method of Lee and Richards (1971); residues with side chains of relative accessibility less than 7% (Hubbard & Blundell, 1987) were defined as inaccessible. Calculations were usually performed on the entire molecule. However, for the alignments of subunits (for example, the immunoglobulins), residue accessibilities were calculated for isolated domains. Essential prosthetic groups and ligands were included in the calculations; enzyme substrates, inhibitors, etc. were excluded from the analysis.

4. Side-chain interactions. Specific interactions of a side chain, for example hydrogen bonding, ionic interactions, and covalent bonding, were examined. These were divided into three classes: interactions between two side chains, interactions between a side chain and a main-chain carbonyl, and interactions between a side-chain and a main-chain amide hydrogen (see Kinemage 1 for an example). As side-chain atoms are generally not reliably determined by crystallography, hydrogen bond formation was defined based on the criterion of a donor–acceptor distance $\leq$3.5 Å. Hydrogen and covalent bonds to

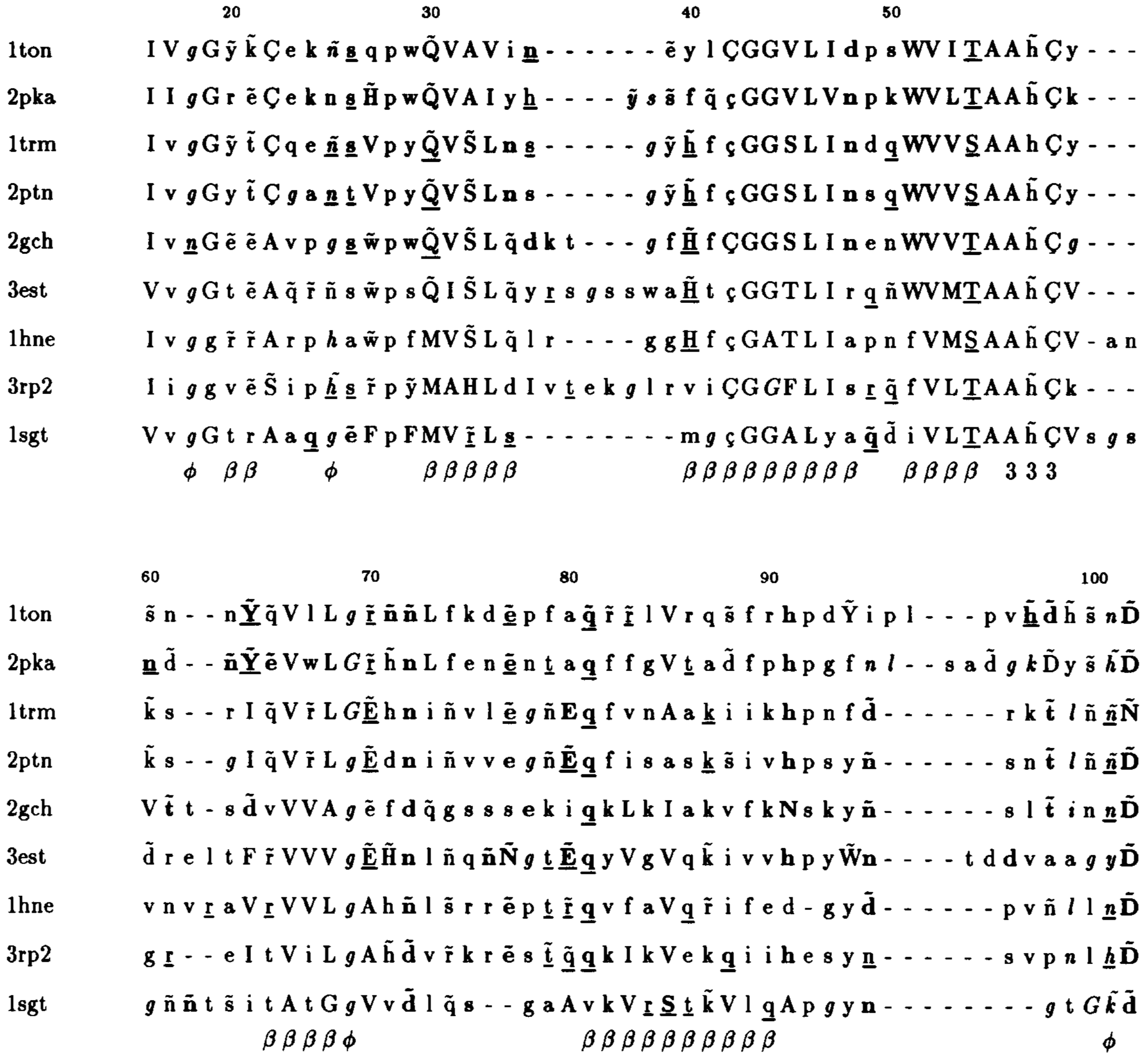

20 30 40 50
1ton I V g G ỹ k̃ Ç e k ñ s q p w Q̃ V A V i n - - - - - - ẽ y l Ç G G V L I d p s W V I T A A h̃ Ç y - - -
2pka I I g G r ẽ Ç e k n s H̃ p w Q̃ V A I y h - - - - ỹ s s̃ f q̃ ç G G V L V n p k W V L T A A h̃ Ç k - - -
1trm I v g G ỹ t̃ Ç q e ñ s V p y Q̃ V S̃ L n s - - - - - g ỹ h̃ f ç G G S L I n d q W V V S A A h Ç y - - -
2ptn I v g G y t̃ Ç g a n t V p y Q̃ V S̃ L n s - - - - - g ỹ h̃ f ç G G S L I n s q W V V S A A h̃ Ç y - - -
2gch I v n G ẽ ẽ A v p g s w̃ p w Q̃ V S̃ L q̃ d k t - - - g f H̃ f Ç G G S L I n e n W V V T A A h̃ Ç g - - -
3est V v g G t ẽ A q̃ r̃ ñ s w̃ p s Q̃ I S̃ L q̃ y r s g s s w a H̃ t ç G G T L I r q ñ W V M T A A h̃ Ç V - - -
1hne I v g g r̃ r̃ A r p h a w̃ p f M V S̃ L q̃ l r - - - - g g H f ç G A T L I a p n f V M S A A h̃ Ç V - a n
3rp2 I i g g v ẽ S̃ i p h̃ s r̃ p ỹ M A H L d I v t e k g l r v i Ç G G F L I s r q̃ f V L T A A h̃ Ç k - - -
1sgt V v g G t r A a q g ẽ F p F M V r̃ L s - - - - - - - - - m g ç G G A L y a q̃ d̃ i V L T A A h̃ Ç V s g s
ϕ β β ϕ β β β β β β β β β β β β β β β β β β 3 3 3
60 70 80 90 100
1ton s̃ n - - n Ỹ q̃ V l L g r̃ ñ ñ L f k d ẽ p f a q̃ r̃ r̃ l V r q s̃ f r h p d Ỹ i p l - - - p v h̃ d̃ h̃ s̃ n D̃
2pka n d̃ - - ñ Ỹ ẽ V w L G r̃ h̃ n L f e n ẽ n t a q f f g V t a d̃ f p h p g f n l - - s a d̃ g k D̃ y s̃ h̃ D̃
1trm k̃ s - - r I q̃ V r̃ L G Ẽ h n i ñ v l ẽ g ñ E q f v n A a k i i k h p n f d̃ - - - - - - r k t̃ l ñ ñ Ñ
2ptn k̃ s - - g I q̃ V r̃ L g Ẽ d n i ñ v v e g ñ Ẽ q f i s a s k s̃ i v h p s y ñ - - - - - - s n t̃ l ñ ñ D̃
2gch V t̃ t - s d̃ v V V A g ẽ f d q̃ g s s s e k i q k L k I a k v f k N s k y ñ - - - - - - s l t̃ i n n D̃
3est d̃ r e l t F r̃ V V V g Ẽ H̃ n l ñ q ñ Ñ g t Ẽ q y V g V q k̃ i v v h p y W̃ n - - - - t d d v a a g y D̃
1hne v n v r a V r V V L g A h ñ l s̃ r r ẽ p t r̃ q v f a V q r̃ i f e d - g y d̃ - - - - - - p v ñ l l n D̃
3rp2 g r - - e I t V i L g A h̃ d̃ v r̃ k r ẽ s t̃ q̃ q k I k V e k q i i h e s y n - - - - - - s v p n l h̃ D̃
1sgt g ñ ñ t s̃ i t A t G g V v d̃ l q̃ s - - g a A v k V r S t k̃ V l q A p g y n - - - - - - - - g t G k̃ d̃
β β β β ϕ β β β β β β β β β β ϕ

Fig. 1. A section of the alignment of sequences of serine proteinases achieved by comparing the three-dimensional structures using COMPARER (Šali & Blundell, 1990). The coordinates of the three-dimensional structures were obtained from the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977) (PDB codes: 1TON, rat tonin; 2PKA, porcine kallikrein; 1TRM, rat trypsin; 2PTN, porcine trypsin; 2GCH, bovine γ-chymotrypsin; 3EST, porcine elastase; 1HNE, human neutrophil elastase; 3RP2, rat mast cell proteinase-II; and 1SGT, *Streptomyces griseus* trypsin). The alignment is numbered according to the structure of γ-chymotrypsin. The amino acid code is the standard one-letter code formatted using the following convention: Italic for positive main-chain ϕ angle; uppercase for solvent-inaccessible and lowercase for solvent-accessible residues; bold for hydrogen bonds to main-chain amide; underline for hydrogen bonds to main-chain carbonyl oxygen; tilde (˜) for side-chain–side-chain hydrogen bonds; disulfide-bonded cystine residues are shown with a cedilla (̧); *cis*-peptide residues are shown with a breve. The consensus secondary structure is shown below the alignment: α for α-helical positions; 3 for 3_{10} helical positions; β for β-sheet positions; ϕ for positions with a positive main chain ϕ torsion angle. Postscript files for all of the formatted alignments are available from the authors on request.

essential hetatom groups were classed as side-chain–side-chain interactions.

Substitution tables

All the above features (amino acid type [21 classes], accessibility [2 classes], side-chain interactions [8 classes], and main-chain conformation [4 classes]) were used to classify the local environment of a residue. The environment-specific substitution tables were then generated by accumulating substitutions observed in homologous structures. To simplify the analysis, the environment of the replacement residue was not considered; only its amino acid type was considered. This corresponds to many practical applications where only one of the three-dimensional structures of the compared proteins is known. However, for the selection of key residues that characterize a particular local conformation, the conformation of both compared structures must be defined. Also, further subgroups of conformational classes in the nonhelical and strand regions can usefully be considered (C. Topham, A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson, & T.L. Blundell, unpubl. results).

To study the role of structural features in the conservation of amino acids, we have summed the substitution frequencies into selected marginal distributions, e.g., all the residues in a particular type of secondary structure irrespective of the accessibility and hydrogen bonding properties. It is convenient to display the data as 21 by 22 probability tables, where the values are the probability of observing each replacement amino acid type (row titles) given that a residue (column title) was in a particular environment in a homologous protein. The conservation probability ($P_{\text{cons.}}$) is then the probability that a residue will not be substituted by any other residue type.

The summing of this table over all structural dimensions gives a substitution table independent of the environment, analogous to the substitution tables of Dayhoff et al. (1983) and others. Comparison of this global table to other established scoring matrices (M.S. Johnson & J.P. Overington, unpubl. results) shows that it is most similar to the Dayhoff and McLachlan (McLachlan, 1971) matrices, both of which are likewise derived from analysis of naturally observed substitution frequencies.

The tables for main-chain residue classes are all well populated except for the positive ϕ class, which is classically disallowed for residues with side chains (Ramachandran & Sasisekharan, 1968). Examination of the probability tables for α- and β-classes (Tables 2, 3) shows that in general the β-class is more conserved than the α-class; to check that this was not an artifact of sample bias (the α-families contributing heavily to the more distant pairwise comparisons), we examined families in which there are appreciable amounts of both α- and β-secondary structure. In these cases, it is observed that the α-regions are indeed less conserved than the β-regions, often by around 10–15% in sequence identity levels. This is no doubt related to the greater fraction of residues buried in β-sheets than in α-helices, reinforced by common supersecondary structure motifs where parallel β-sheets tend to form a buried core, packed between amphiphilic α-helices.

It is for the accessible/inaccessible classes (Tables 4, 5; Kinemage 2) that the largest differences in patterns of accepted mutations occur; for all residues a buried position is more conserved than a surface position, but not all residues are equal in their response to an inaccessible environment. The residues that undergo the largest increases in conservation (Fig. 2) are generally polar, for example, aspartic acid and histidine. Buried hydrophobic amino acids show generally smaller increases in conservation, but the hydrophobic nature of the replacement residue is strongly conserved. However, there are a number of unexpected features of this comparison. The analogous

Table 2. *Substitution probability table for α residues*[a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.355	0.007	0.090	0.100	0.050	0.177	0.037	0.077	0.096	0.056	0.081	0.103	0.106	0.090	0.088	0.163	0.120	0.098	0.065	0.036	0.252
C	0.001	0.901	0.000	0.000	0.000	0.000	0.000	0.004	0.001	0.000	0.000	0.003	0.000	0.006	0.006	0.004	0.002	0.000	0.007	0.000	0.000
D	0.038	0.000	0.315	0.109	0.006	0.041	0.027	0.009	0.033	0.004	0.009	0.088	0.051	0.089	0.023	0.065	0.048	0.013	0.012	0.011	0.009
E	0.044	0.011	0.111	0.305	0.011	0.048	0.026	0.011	0.059	0.013	0.009	0.068	0.069	0.086	0.053	0.033	0.045	0.017	0.012	0.018	0.000
F	0.017	0.000	0.005	0.007	0.415	0.004	0.009	0.039	0.025	0.097	0.042	0.013	0.006	0.011	0.009	0.009	0.014	0.041	0.053	0.085	0.009
G	0.065	0.000	0.070	0.042	0.006	0.370	0.017	0.022	0.029	0.013	0.015	0.036	0.043	0.031	0.013	0.068	0.049	0.014	0.009	0.021	0.045
H	0.010	0.000	0.012	0.011	0.010	0.007	0.571	0.003	0.022	0.005	0.015	0.043	0.006	0.035	0.021	0.016	0.008	0.017	0.009	0.037	0.009
I	0.029	0.014	0.009	0.008	0.048	0.021	0.004	0.325	0.017	0.076	0.107	0.018	0.007	0.007	0.015	0.014	0.033	0.112	0.016	0.030	0.018
K	0.053	0.007	0.044	0.081	0.020	0.041	0.044	0.026	0.336	0.029	0.059	0.073	0.045	0.094	0.163	0.041	0.054	0.026	0.041	0.028	0.036
L	0.038	0.000	0.006	0.018	0.210	0.019	0.004	0.139	0.033	0.415	0.225	0.033	0.016	0.041	0.028	0.029	0.026	0.133	0.037	0.057	0.036
M	0.013	0.000	0.004	0.003	0.016	0.007	0.000	0.043	0.014	0.053	0.197	0.010	0.000	0.018	0.004	0.003	0.010	0.018	0.021	0.021	0.018
N	0.031	0.007	0.057	0.035	0.010	0.026	0.054	0.012	0.034	0.012	0.013	0.195	0.015	0.066	0.026	0.037	0.046	0.012	0.002	0.048	0.000
P	0.022	0.000	0.036	0.035	0.005	0.026	0.011	0.009	0.020	0.006	0.000	0.013	0.424	0.013	0.016	0.039	0.011	0.009	0.002	0.000	0.000
Q	0.025	0.011	0.045	0.039	0.011	0.021	0.031	0.004	0.045	0.015	0.035	0.059	0.015	0.183	0.029	0.030	0.030	0.008	0.007	0.025	0.009
R	0.019	0.011	0.012	0.023	0.005	0.008	0.019	0.010	0.067	0.009	0.004	0.018	0.013	0.028	0.348	0.030	0.019	0.005	0.007	0.018	0.018
S	0.086	0.021	0.075	0.047	0.012	0.079	0.033	0.020	0.041	0.020	0.009	0.089	0.082	0.069	0.063	0.264	0.096	0.028	0.005	0.020	0.054
T	0.043	0.007	0.039	0.033	0.020	0.038	0.014	0.026	0.032	0.015	0.026	0.057	0.028	0.046	0.035	0.065	0.266	0.037	0.016	0.034	0.000
V	0.055	0.000	0.018	0.021	0.069	0.022	0.044	0.178	0.025	0.111	0.016	0.018	0.025	0.017	0.015	0.029	0.060	0.350	0.012	0.043	0.162
W	0.009	0.000	0.003	0.004	0.022	0.004	0.007	0.006	0.012	0.006	0.020	0.001	0.001	0.006	0.004	0.002	0.007	0.003	0.588	0.064	0.000
Y	0.009	0.000	0.006	0.006	0.046	0.006	0.029	0.014	0.007	0.013	0.031	0.033	0.003	0.020	0.010	0.007	0.017	0.016	0.078	0.377	0.027
J	0.009	0.000	0.001	0.000	0.001	0.004	0.001	0.002	0.002	0.002	0.004	0.000	0.000	0.004	0.003	0.006	0.004	0.010	0.000	0.005	0.297
–	0.028	0.004	0.041	0.074	0.010	0.029	0.017	0.022	0.050	0.031	0.033	0.031	0.045	0.039	0.028	0.047	0.034	0.032	0.002	0.021	0.000

[a] The standard one-letter amino acid code is used with the exception of C for cystine (the disulfide-bonded form) and J for cysteine (the free thiol form). The values in the table give the probability of a substitution of a residue at the top of a column, by all other residues or at the site of an insertion/deletion; thus, the columns sum to 1.0.

Table 3. *Substitution probability table for β residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.275	0.000	0.025	0.047	0.023	0.086	0.007	0.029	0.036	0.031	0.074	0.041	0.035	0.050	0.050	0.057	0.055	0.065	0.014	0.031	0.080
C	0.000	0.910	0.000	0.016	0.014	0.000	0.000	0.003	0.008	0.000	0.000	0.000	0.000	0.008	0.015	0.002	0.001	0.000	0.000	0.000	0.020
D	0.008	0.000	0.350	0.059	0.008	0.011	0.014	0.017	0.018	0.006	0.000	0.095	0.040	0.020	0.010	0.026	0.020	0.013	0.006	0.012	0.000
E	0.018	0.016	0.054	0.192	0.004	0.015	0.021	0.012	0.071	0.009	0.037	0.039	0.028	0.056	0.053	0.018	0.036	0.018	0.002	0.015	0.000
F	0.020	0.022	0.013	0.005	0.398	0.008	0.021	0.049	0.017	0.046	0.023	0.006	0.012	0.006	0.015	0.020	0.012	0.021	0.071	0.096	0.020
G	0.092	0.000	0.021	0.033	0.015	0.623	0.007	0.016	0.018	0.016	0.042	0.019	0.017	0.033	0.017	0.036	0.028	0.013	0.049	0.021	0.020
H	0.003	0.000	0.006	0.010	0.008	0.002	0.332	0.006	0.022	0.004	0.000	0.035	0.014	0.021	0.023	0.009	0.010	0.009	0.000	0.008	0.020
I	0.040	0.010	0.044	0.021	0.089	0.020	0.017	0.358	0.022	0.105	0.077	0.025	0.012	0.010	0.015	0.021	0.026	0.119	0.041	0.034	0.020
K	0.024	0.011	0.021	0.099	0.015	0.007	0.070	0.013	0.299	0.017	0.012	0.060	0.052	0.060	0.139	0.026	0.051	0.010	0.004	0.017	0.040
L	0.057	0.000	0.027	0.031	0.111	0.023	0.031	0.143	0.051	0.459	0.169	0.031	0.038	0.026	0.031	0.025	0.028	0.119	0.096	0.044	0.060
M	0.026	0.000	0.006	0.021	0.011	0.013	0.021	0.021	0.007	0.031	0.244	0.010	0.002	0.031	0.002	0.007	0.013	0.019	0.006	0.006	0.000
N	0.016	0.000	0.092	0.044	0.003	0.018	0.073	0.008	0.038	0.008	0.019	0.261	0.026	0.016	0.011	0.036	0.032	0.004	0.012	0.017	0.000
P	0.014	0.000	0.027	0.020	0.001	0.005	0.021	0.007	0.029	0.019	0.002	0.017	0.504	0.011	0.023	0.010	0.021	0.009	0.018	0.003	0.000
Q	0.038	0.014	0.033	0.098	0.006	0.020	0.073	0.008	0.071	0.014	0.077	0.037	0.019	0.414	0.118	0.025	0.045	0.015	0.002	0.013	0.000
R	0.022	0.011	0.006	0.042	0.007	0.010	0.038	0.008	0.079	0.008	0.002	0.012	0.031	0.055	0.214	0.015	0.027	0.010	0.014	0.017	0.000
S	0.078	0.003	0.085	0.041	0.029	0.056	0.052	0.024	0.048	0.022	0.030	0.112	0.040	0.042	0.065	0.403	0.140	0.028	0.014	0.040	0.040
T	0.081	0.002	0.075	0.111	0.021	0.037	0.052	0.027	0.095	0.022	0.049	0.110	0.073	0.078	0.092	0.153	0.363	0.044	0.008	0.037	0.020
V	0.141	0.000	0.058	0.065	0.074	0.027	0.070	0.202	0.034	0.145	0.123	0.019	0.033	0.039	0.046	0.043	0.062	0.446	0.027	0.059	0.040
W	0.005	0.000	0.008	0.002	0.048	0.000	0.000	0.013	0.001	0.019	0.005	0.015	0.012	0.001	0.013	0.003	0.002	0.005	0.559	0.017	0.000
Y	0.026	0.000	0.027	0.037	0.112	0.011	0.049	0.024	0.020	0.018	0.014	0.033	0.005	0.013	0.032	0.026	0.022	0.023	0.051	0.505	0.000
J	0.003	0.002	0.000	0.000	0.001	0.001	0.007	0.001	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.001	0.000	0.000	0.620
–	0.012	0.000	0.021	0.007	0.002	0.006	0.021	0.012	0.013	0.004	0.002	0.021	0.007	0.008	0.015	0.038	0.007	0.009	0.004	0.009	0.000

pairs aspartic acid/asparagine and glutamic acid/glutamine are opposite in their behavior; buried aspartic acid residues are highly conserved, whereas the increase in conservation for asparagine is far less marked. The trend for the glutamyl pair is puzzlingly reversed. These conclusions are based on 564 substitutions for buried aspartic acid residues, 329 for buried asparagines, 189 for buried glutamic acids, and 377 for buried asparagines (full data are available in the supplementary material). Proline residues are highly conserved when buried; to replace a buried proline by any other residue, a hydrogen bond must be made to the newly exposed amide proton. Glycine sim-

Table 4. *Substitution table for accessible residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.224	0.013	0.055	0.068	0.031	0.067	0.048	0.053	0.068	0.050	0.087	0.059	0.067	0.073	0.062	0.074	0.059	0.079	0.033	0.035	0.121
C	0.002	0.739	0.001	0.006	0.012	0.000	0.001	0.004	0.003	0.000	0.000	0.001	0.001	0.005	0.008	0.001	0.001	0.000	0.001	0.000	0.008
D	0.044	0.007	0.284	0.091	0.016	0.041	0.056	0.033	0.034	0.012	0.022	0.094	0.047	0.052	0.025	0.054	0.044	0.025	0.014	0.023	0.030
E	0.052	0.029	0.079	0.251	0.016	0.028	0.026	0.026	0.053	0.019	0.031	0.038	0.037	0.071	0.049	0.031	0.044	0.034	0.010	0.027	0.008
F	0.010	0.029	0.006	0.008	0.291	0.004	0.023	0.046	0.011	0.047	0.032	0.012	0.006	0.010	0.009	0.011	0.013	0.018	0.093	0.073	0.000
G	0.079	0.000	0.066	0.047	0.020	0.455	0.042	0.024	0.033	0.028	0.039	0.073	0.054	0.054	0.040	0.064	0.037	0.039	0.041	0.036	0.038
H	0.013	0.003	0.021	0.011	0.024	0.010	0.284	0.008	0.021	0.011	0.020	0.035	0.008	0.020	0.023	0.013	0.012	0.020	0.014	0.025	0.023
I	0.014	0.016	0.017	0.014	0.058	0.006	0.010	0.235	0.015	0.050	0.048	0.018	0.009	0.009	0.015	0.014	0.023	0.075	0.015	0.030	0.008
K	0.062	0.007	0.039	0.072	0.032	0.027	0.068	0.039	0.294	0.050	0.077	0.055	0.045	0.077	0.122	0.043	0.059	0.044	0.037	0.035	0.053
L	0.028	0.000	0.010	0.017	0.097	0.013	0.024	0.094	0.035	0.311	0.141	0.030	0.030	0.028	0.027	0.019	0.029	0.073	0.064	0.033	0.015
M	0.010	0.000	0.003	0.005	0.015	0.005	0.005	0.020	0.011	0.030	0.167	0.004	0.003	0.017	0.005	0.003	0.007	0.013	0.004	0.008	0.015
N	0.041	0.007	0.080	0.041	0.022	0.044	0.087	0.031	0.042	0.035	0.024	0.239	0.019	0.040	0.031	0.050	0.051	0.021	0.008	0.036	0.030
P	0.053	0.000	0.039	0.036	0.006	0.027	0.018	0.017	0.034	0.036	0.014	0.018	0.412	0.021	0.026	0.037	0.031	0.019	0.018	0.008	0.015
Q	0.040	0.013	0.038	0.060	0.018	0.025	0.042	0.017	0.046	0.026	0.075	0.032	0.019	0.231	0.056	0.032	0.042	0.036	0.007	0.015	0.023
R	0.025	0.023	0.015	0.031	0.010	0.017	0.033	0.017	0.062	0.019	0.015	0.022	0.018	0.047	0.248	0.026	0.028	0.022	0.022	0.023	0.000
S	0.100	0.013	0.088	0.059	0.044	0.073	0.057	0.051	0.062	0.043	0.026	0.096	0.070	0.072	0.079	0.290	0.138	0.057	0.025	0.059	0.053
T	0.054	0.010	0.049	0.058	0.042	0.029	0.037	0.059	0.058	0.039	0.049	0.068	0.042	0.066	0.053	0.099	0.266	0.061	0.021	0.041	0.015
V	0.041	0.000	0.021	0.033	0.040	0.020	0.038	0.148	0.031	0.077	0.051	0.018	0.020	0.043	0.033	0.028	0.044	0.269	0.023	0.049	0.091
W	0.005	0.000	0.002	0.002	0.049	0.006	0.006	0.009	0.006	0.017	0.005	0.003	0.004	0.003	0.008	0.003	0.004	0.003	0.421	0.038	0.000
Y	0.014	0.000	0.013	0.018	0.111	0.012	0.034	0.028	0.017	0.023	0.026	0.023	0.005	0.012	0.024	0.018	0.019	0.028	0.109	0.355	0.023
J	0.002	0.000	0.001	0.000	0.001	0.002	0.002	0.001	0.002	0.001	0.003	0.001	0.001	0.001	0.001	0.002	0.001	0.007	0.000	0.001	0.341
–	0.086	0.092	0.072	0.072	0.045	0.089	0.060	0.043	0.061	0.075	0.048	0.061	0.083	0.050	0.056	0.087	0.050	0.057	0.021	0.048	0.091

Table 5. *Substitution probability table for inaccessible residues*

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.426	0.022	0.025	0.095	0.036	0.075	0.009	0.039	0.000	0.029	0.061	0.036	0.041	0.053	0.015	0.149	0.103	0.071	0.020	0.032	0.112
C	0.007	0.879	0.000	0.000	0.005	0.000	0.000	0.003	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.008	0.004	0.000	0.003	0.001	0.004
D	0.012	0.000	0.775	0.042	0.002	0.008	0.004	0.007	0.000	0.002	0.001	0.112	0.004	0.011	0.000	0.016	0.005	0.005	0.000	0.006	0.000
E	0.007	0.009	0.011	0.460	0.004	0.008	0.000	0.005	0.000	0.003	0.009	0.036	0.000	0.064	0.015	0.009	0.001	0.006	0.000	0.001	0.000
F	0.022	0.014	0.004	0.011	0.493	0.006	0.015	0.045	0.000	0.075	0.036	0.006	0.016	0.005	0.000	0.012	0.013	0.033	0.058	0.143	0.017
G	0.070	0.000	0.016	0.037	0.006	0.732	0.000	0.013	0.000	0.007	0.015	0.009	0.004	0.013	0.023	0.064	0.032	0.008	0.005	0.004	0.043
H	0.005	0.000	0.002	0.000	0.003	0.003	0.705	0.003	0.029	0.003	0.006	0.058	0.000	0.061	0.008	0.008	0.001	0.006	0.005	0.014	0.000
I	0.043	0.007	0.009	0.032	0.061	0.017	0.009	0.402	0.029	0.110	0.096	0.015	0.018	0.008	0.008	0.013	0.036	0.135	0.043	0.038	0.017
K	0.014	0.010	0.000	0.005	0.002	0.004	0.002	0.003	0.412	0.003	0.004	0.021	0.014	0.008	0.145	0.013	0.007	0.004	0.000	0.006	0.004
L	0.048	0.005	0.005	0.037	0.162	0.015	0.022	0.163	0.015	0.524	0.250	0.033	0.031	0.032	0.038	0.027	0.022	0.130	0.048	0.052	0.030
M	0.019	0.000	0.004	0.037	0.014	0.004	0.019	0.027	0.000	0.048	0.262	0.027	0.000	0.072	0.000	0.009	0.018	0.020	0.014	0.017	0.009
N	0.011	0.000	0.071	0.011	0.005	0.007	0.028	0.006	0.000	0.002	0.013	0.398	0.007	0.005	0.008	0.035	0.012	0.003	0.006	0.009	0.009
P	0.014	0.002	0.005	0.000	0.011	0.013	0.004	0.005	0.000	0.002	0.000	0.015	0.764	0.008	0.008	0.005	0.012	0.007	0.000	0.001	0.000
Q	0.017	0.009	0.002	0.101	0.002	0.009	0.039	0.001	0.088	0.003	0.039	0.018	0.000	0.560	0.122	0.008	0.001	0.007	0.006	0.005	0.004
R	0.013	0.009	0.000	0.058	0.003	0.004	0.015	0.004	0.294	0.003	0.001	0.000	0.002	0.040	0.588	0.013	0.002	0.004	0.003	0.007	0.009
S	0.074	0.010	0.023	0.016	0.011	0.039	0.017	0.009	0.015	0.007	0.016	0.103	0.020	0.013	0.000	0.450	0.111	0.017	0.005	0.004	0.052
T	0.044	0.006	0.012	0.000	0.006	0.018	0.006	0.015	0.015	0.009	0.025	0.024	0.025	0.005	0.008	0.081	0.504	0.025	0.002	0.005	0.034
V	0.116	0.001	0.012	0.053	0.073	0.015	0.054	0.205	0.000	0.134	0.118	0.033	0.031	0.027	0.015	0.043	0.079	0.476	0.014	0.047	0.082
W	0.006	0.003	0.007	0.000	0.024	0.001	0.011	0.011	0.000	0.009	0.013	0.003	0.000	0.005	0.000	0.003	0.001	0.005	0.731	0.031	0.000
Y	0.015	0.001	0.005	0.000	0.071	0.006	0.034	0.021	0.029	0.013	0.023	0.024	0.007	0.008	0.000	0.014	0.012	0.017	0.037	0.561	0.009
J	0.013	0.002	0.000	0.000	0.002	0.002	0.000	0.002	0.000	0.002	0.003	0.006	0.000	0.003	0.000	0.008	0.007	0.003	0.000	0.004	0.547
–	0.006	0.010	0.012	0.005	0.007	0.013	0.006	0.011	0.074	0.010	0.009	0.018	0.016	0.000	0.000	0.011	0.018	0.016	0.002	0.012	0.017

ilarly undergoes a large change in conservation when buried; the most conserved glycine residues tend to be those that are both buried and in the unusual $+\phi$ conformation.

Applications of substitution pattern data

Tertiary templates and the inverse folding problem

Given at least one three-dimensional structure one can estimate the frequency of occurrence of each amino acid for each position in a fold, as suggested by Overington et al. (1990). In this way we can construct a simple sequence template for a tertiary structure, a tertiary template. For an alternative definition see Ponder and Richards (1987).

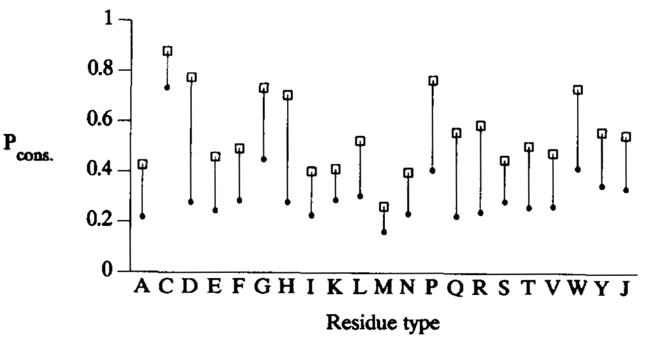


Fig. 2. Comparison of conservation probabilities ($P_{\text{cons.}}$) for surface (circle) and buried (open square) positions. The standard one-letter amino acid code is used with the exception of C for cystine and J for cysteine.

Figure 3 shows the sequence variability expected on the basis of the three-dimensional structure for 15 residues from the C-terminus of the G helix of the globin terminus erythrocruorin (1ECD) (Steigemann & Weber, 1979). The residues at positions 97, 100, 103, and 104 are buried (the classic i, $i + 3$, $i + 4$ spacing for an α-helix). As can be seen, the residues that are most conserved in this region are the nonpolar residues making up the buried face of this amphipathic helix. It is these residues that are of the greatest utility in data-base searching and alignment studies.

Figure 4A shows a conventional search for sequences related to 1ECD, whereas Figure 4B shows the same search using the environmental-specific tables for scoring. As can be seen, the profiling method is more discriminating in the identification of true globin sequences. The standard jumbling test for significance in alignment studies can be further enhanced using the expected patterns for residue variance (M.S. Johnson & J.P. Overington, unpubl. results). A related procedure has been proposed by Eisenberg and coworkers (Bowie et al., 1991; Lüthy et al., 1991).

These procedures represent an approach to identifying the sequences that are compatible with a fold (Ponder & Richards, 1987), generally termed the "inverse protein folding problem." The tertiary templates can be used to search the data base of sequences for other sequences that

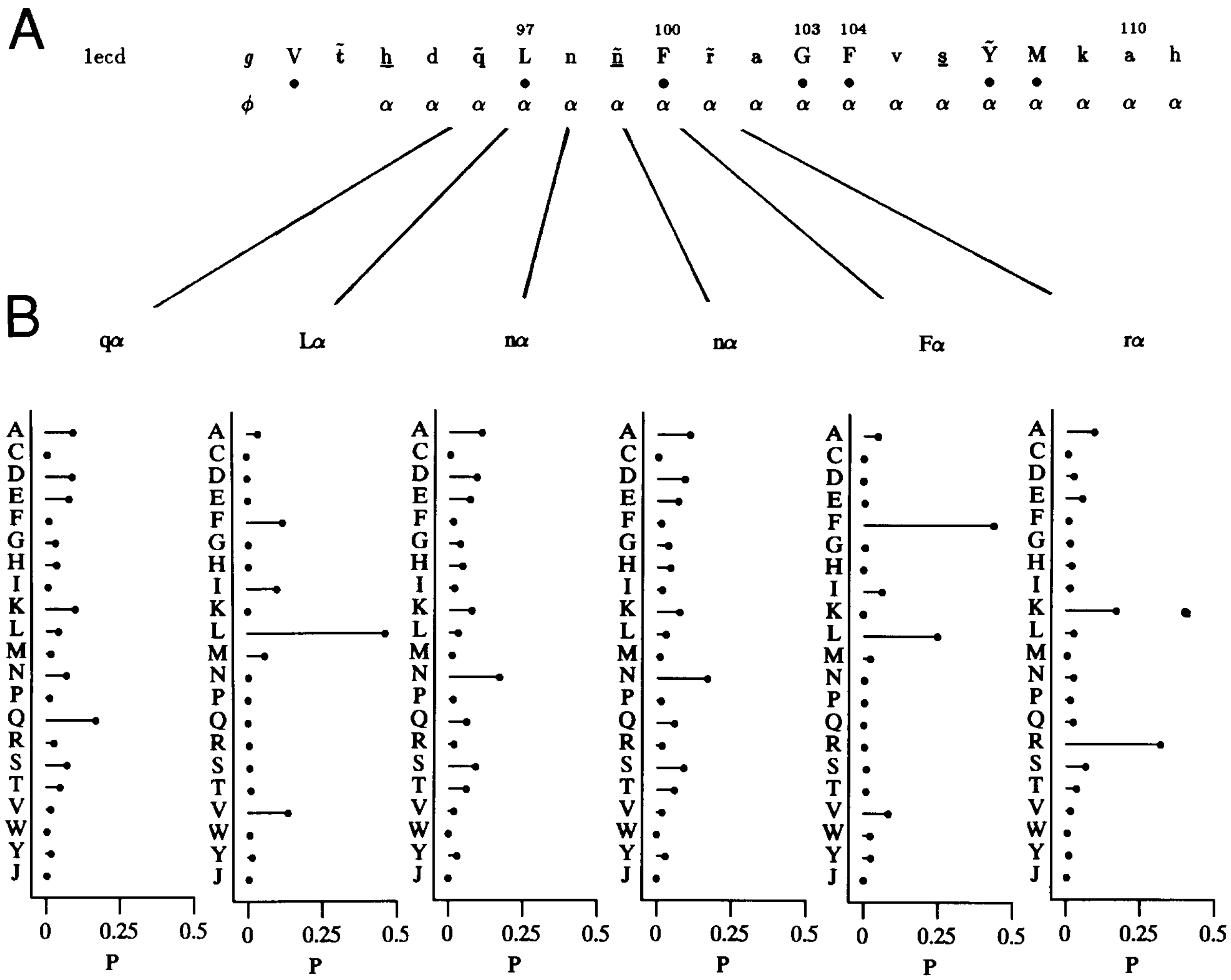


Fig. 3. Example of environment-specific scores used in template generation. A: The sequence and local structural restraints for part of the *Chironomus thummi thummi* globin structure (PDB data set 1ECD) (Steigemann & Weber, 1979). Buried positions are emphasized by bullets underneath the sequence. The extent of the α-helix (defined using the DSSP program of Kabsch & Sander [1983]) is also shown. Numbering is as in the PCB data set 1ECD. B: The predicted substitution patterns for residues in the observed environmental class.

are likely to adopt the same fold. This approach should allow a large proportion of new sequences to be associated with known folds, even though overall sequence similarities for pairs of sequences may not be statistically significant indicators of homology.

Another problem, which involves relating a sequence to a three-dimensional fold, is to test if a protein is correctly folded. Such tests will be very important for protein structure prediction in the future. It is well documented that potential energy-based methods are poor discriminators between correctly folded and misfolded protein structures (Novotny et al., 1984). More successful attempts to identify correctly folded proteins have included analysis and comparison of structural features such as the ratio of buried polar and surface nonpolar atoms (Baumann et al., 1989, Novotny et al., 1988). In our approach we test whether the pattern of residue variation observed in homologous sequences is consistent with the fold predicted in the modeling (Overington et al., 1990; C. Topham, A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson, & T.L. Blundell, unpubl. results).

Prediction

A single sequence is often used to predict the secondary and/or the tertiary structure of a protein of interest. Because the general fold of a protein family is conserved, the use of aligned homologous sequences can result in more accurate consensus predictions. We have attempted to use our substitution data in an alternative way to predict secondary structure from aligned homologous proteins.

The patterns of sequence variation are clearly dependent on the physical environment of a residue in the

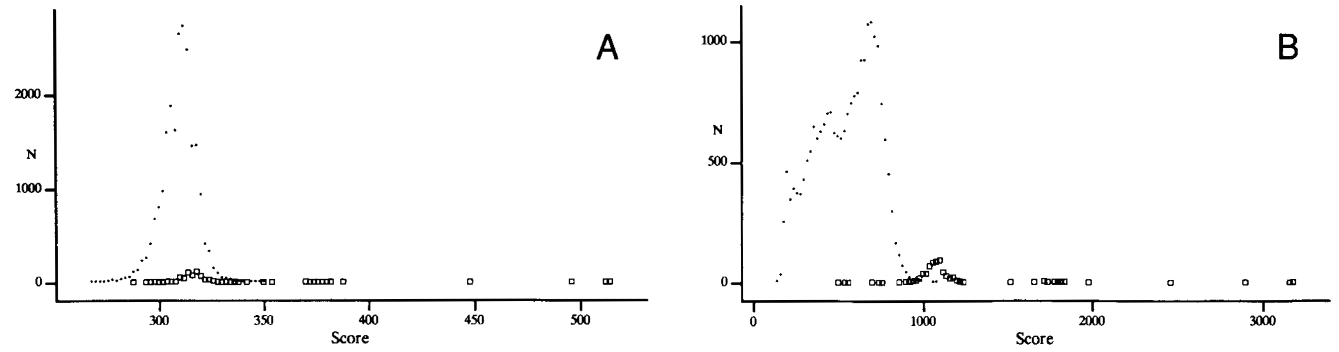
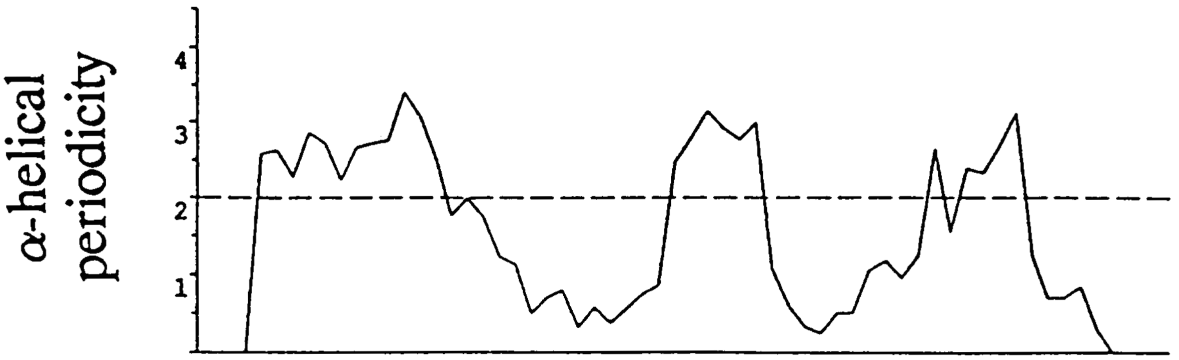


Fig. 4. Comparison of sequence-based and template-based data-base searching. A: The results of a search of the PIR data base using the sequence of *Chironomus thummi thummi* erythrocrourin (an atypical globin) as a probe and the Dayhoff 250PAM matrix to score residue comparisons. The dots represent nonglobin sequences, and the squares represent globin sequences. The vertical axis is the number of observed comparisons with a given alignment score (shown on the horizontal axis). As is apparent, many globins are within the region of nonglobin sequences. B: The same search but with environment-specific substitution patterns used for the residue scoring. The environment at each position was classified into one of eight classes (four for secondary structure and two for accessibility). Position-dependent substitution profiles were then used to scan the data base. As can be seen, the globins are more clearly differentiated from the nonglobins using the position-dependent scoring scheme. Most of the globin sequences that are within the cluster of nonglobins are partial sequences, and so scoring and assessment are more difficult.



```
         1        10        20        30        40        50        60
en       DEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWF QNKRAKIKKST
enl      EDKRPRTAFTAEQLQRLKAEFQANRYITEQRRQTLAQELSLNESQIKIWF QNKRAKIKKAT
Antp     ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWF QNRRMKWKKEN
Fts      DSKRTRQTYTRYQTLELEKEFHFNRYITRRRRIDIANALCNTERQIKIWF QNRRMKWKKDS
Ubx      LRRRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALCLTERQIKIWF QNRRMKLKKEI
Dfd      EPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWF QNRRMKWKKDN
Xhox-1A  EAKRSRTAYTRQQVLELEKEFHYNRYLTRRRRVEIAHTLRLSERQIKIWF QNRRMKWKKDN
Hox 1.6  QPNAVRTNFTTKQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWF QNRRMKQKKRE
eve      SVRRYRTAFTRDQLGRLEKEFYKENYVSRPRRCELAAQLNLPESTIKVWF QNRRMKDKQQR
prd      KQRRCRTTFSASQLDELERAFERTQYPDIYTREELAQRTNLTEARIQVWF SNRRARLRKQH
Hox 1.5  SSKRGRTAYTRPQLVELEKEFHFNRYLMRPRRVEMANLLNLTERQIKIWF QNRRMKYKKDQ
bcd      RPRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWF KNRRRRHKIQS
cad      TKDKYRVVYTDFQRLELEKEYCTSRYITIRRKSELAQTLSLSERQVKIWF QNRRAKERTSN
zen 2    KSKRSRTAFSSLQLIELEREFHLNKYLARTRRIEISQRLALTERQVKIWF QNRRMKLKKST
MAT a2   YRGHRFTKENVRILESWFAKKNIEPYLDTKGLENLMKNTSLSRIQIKNWV SNRRRKEKTITQ
PHO2      QRPKRTRAKGEALDVLKRKFEINPTPSLVERKKISDLIGMPEKNVRIWF QNRRAKLRKKQ
                  ααααααααααααα     αααααααααα    αααααααα ααααααααα
```

Fig. 5. Prediction of α-helical segments in an alignment of homeodomain sequences using Fourier analysis of substitution patterns. Underneath the alignment is shown the extents of the helical segments found by X-ray crystallography (Kissinger et al., 1990).

folded protein; therefore, the residue variation in an alignment can be used to infer the local environment of a residue (see Kinemage 3). The analysis of the relative strengths of structural features as constraints has shown that solvent accessibility is the major single factor. Amphipathic α-helices and β-sheets have distinctive periodicity in the pattern of residue accessibility, and so this can be used to identify such regions from alignments.

We have calculated substitution patterns for surface and buried residues in α-helices and combined the use of these with Fourier analysis (Eisenberg et al., 1984) to identify periodicity indicative of amphipathic secondary structural elements (D. Donnelly, J.P. Overington, & T.L. Blundell, unpubl. results). Although, in principle, the Fourier conservation method is equally applicable to β-strands, it is more useful for α-helices. This is due to the higher propensity for α-helices to be amphipathic.

Figure 5 shows the α-helical regions predicted using substitution patterns for several homeodomain sequences and compares the prediction with the experimentally observed helices. The prediction correctly identifies all three α-helical regions, although their precise limits are not well defined. An improvement in the prediction accuracy can be achieved by analysis of other features within the alignment, for example, the conservation of helix-capping residues at the termini of helices (D. Donnelly, J.P. Overington, & T.L. Blundell, unpubl. results).

Conclusion

The study of residue substitutions as a function of local environment has highlighted the clear differences observed for various local structural constraints. In each position of a protein fold, the local environment constrains the accepted mutations to those that will not disrupt function and structure. The most conserved positions are in the solvent-inaccessible core of the protein and especially the buried polar residues within the core. It is the conservation at these positions that will provide the most powerful recognition of homology between two sequences.

Acknowledgments

We thank the Imperial Cancer Research Fund, the Science and Engineering Research Council, Pfizer, and Merck Sharp and Dohme for financial support. We are grateful to Janet Thornton, Alasdair MacLeod, Chris Topham, Pamela Thomas, and many other colleagues for comments.

References

Baumann, G., Frömmel, C., & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng. 2*, 329-334.

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, Y. (1977). The protein data bank: A computer based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535-542.

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

Dayhoff, M.O., Barker, W.C., & Hunt, L.T. (1983). Establishing homologies in proteins. *Methods Enzymol. 91*, 524-545.

Eisenberg, D., Weiss, R.M., & Terwilliger, T.D. (1984). The helical hydrophobic moment: A measure of the amphiphilicity of a helix. *Nature 299*, 199-203.

Hubbard, T.J.P. & Blundell, C. (1987). Comparison of the solvent-inaccessible cores of homologous proteins: Definitions useful for protein modeling. *Protein Eng. 1*, 159-171.

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

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., & Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell 63*, 579-590.

Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol. 55*, 379-400.

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

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

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

Novotny, J., Rashin, A.A., & Bruccoleri, R. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins 4*, 19-30.

Overington, J.P., Johnson, M.S., Šali, A., & Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity. *Proc. R. Soc. Lond. Ser. B. 241*, 132-145.

Ponder, J. & Richards, F.M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol. 193*, 775-791.

Ramachandran, G.N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem. 23*, 283-438.

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. Mol. Biol. 212*, 403-442.

Steigemann, W. & Weber, E. (1979). Structure of erythrocruorin in different ligand states refined at 1.4 Å resolution. *J. Mol. Biol. 127*, 309-338.