Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation

András Fiser[a,b], István Simon[b], Geoffrey J. Barton[a,*]

[a] *University of Oxford, Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK*
[b] *Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, PO Box 7, Budapest H-1518, Hungary*

Received 18 July 1996; revised version received 25 September 1996

Abstract Analysis of the relationship between surface accessibility and amino acid conservation in multiple sequence alignments of homologous proteins confirms expected trends for hydrophobic amino acids, but reveals an unexpected difference between the conservation of Asp, Glu and Gln. Even when not in an active site, Asp is more highly conserved than Glu. There is a clear preference for conserved and buried Asp to be present in coil, but there is no tendency for Asp to conserve ϕ/ψ in the ++ region of the Ramachandran map. Glu does not show any preference to be conserved in a particular secondary structure. Analysis of recently derived substitution matrices (e.g. BLOSUM) confirms that Glu tends to substitute more frequently with other amino acids than does Asp. Analysis of relative accessibility versus relative conservation for individual amino acid positions in alignments shows a negative correlation for all amino acid types. With the exception of Arg, Lys, Gly, Glu, Asp and Tyr, a relative conservation of >2 suggests the amino acid will have a relative accessibility of $<50\%$. Observation of conserved Cys, Gly or Asp in a reliable multiple alignment suggests a position important for the structure of the protein. Furthermore, the Asp is likely to be involved in polar interactions through its side chain oxygen atoms. In contrast, Gln is the least conserved amino acid overall.

Key words: Conservation analysis; Multiple sequence alignment; Protein structure prediction

1. Introduction

Knowledge of protein sequences is growing much faster than knowledge of either three-dimensional structure or function. Accordingly, the interpretation of sequence data to identify structurally or functionally important residues is essential if the data are to be effective in furthering understanding of biological systems. Multiple sequence alignments of families of protein sequences are now used routinely to indicate residues of key importance to the function of the protein. A position in an alignment that has identical residues in all members of a protein family may have a key catalytic role. A position where similar physico-chemical properties (e.g. hydrophobicity) are shared may suggest importance in stabilising the native conformation of the protein [1,2]. Identification of such conserved features in multiple alignments has been used to good effect to improve the accuracy of prediction of secondary structure and buried residues (α-helix and β-strand) (e.g. [3–7]).

Here we report a systematic study of residue conservation in multiple alignments where at least one protein is of known tertiary structure. Our analysis complements that of Overington et al. [8] who considered only pairwise substitution frequencies for amino acids in structurally aligned families.

2. Materials and methods

2.1. Data base

A non-redundant set of 81 proteins was generated from the April 1993 release of the Brookhaven Protein Data Bank (PDB) [9]. The set was chosen in a two-step procedure. First, all pairs of chains (over 50 residues and resolution better than 2.5 Å) in the data bank were compared by calculating correlation coefficients between the dipeptide frequencies in each protein. A set of 101 protein chains was selected such that all pairs had a correlation of <0.4. All pairs in this set were then compared by a rigorous sequence comparison method [10,11] followed by cluster analysis. This reduced the set to 81 protein chains that show no obvious sequence similarity (PDB code and chain identifiers: 155C 1ACX 1ALC 1BBP_A 1CC5 1ECA 1FKF 1FNR 1GCR 1GP1_A 1HDSB 1HIP 1HOE 1LRD_4 1PAZ 1PCY 1PHH 1PRC_C 1RBP 1RHD 1RNH 1SN3 1TGS 1TPK_A 1WSY_B 256B_A 2ALP 2AZA_A 2CAB 2CD4 2CDV 2CPP 2FXB 2GN5 2LH7 2LIV 2LTN_A 2OR1_L 2PAB_A 2RNT 2RSP_A 2SEC_I 2SNL_E 2SNS 2SOD_B 2SSI 2STV 2TS1 2UTG_A 3ADK 3B5C 3CLA 3FXC 3GAP_B 3LZM 3SGB_I 451C 4BP2 4FD1 4FXN 4HHB_A 4PEP 4PFK 4PTP 4TNC 5CTS 5CYT_R 5EBX 5RUB_A 5RXN 6LDH 6TMN_E 7PTI 8ADH 8ATC_B 8CAT_A 8DFR 9PAP 9RSA_A 9WGA_A).

Each protein in the set was compared by the Smith-Waterman algorithm [11,12] to the NBRF-PIR sequence data bank (Release 38) and all sequences that gave probability values of $<10^{-6}$ by a length-dependent scoring scheme (program SCANPS ftp://geoff.biop.ox.ac.uk/programs/scanps) were multiply aligned with the query sequence by the algorithm of Barton and Sternberg [13]. This gave 81 alignments with between 3 and 499 sequences in each (median of 28 sequences).

2.2. Calculation of conservation and accessibility

Conservation scores based upon the physico-chemical properties of the amino acids were calculated for each position in each alignment according to Livingstone and Barton [2]. Conservation scores range from 0 to 10 and represent the number of the properties: Hydrophobic, Positive, Negative, Polar, Charged, Small, Tiny, Aliphatic, Aromatic, Proline and their negations (e.g. *not* positive) that are shared at a position. The program AMAS [2], which calculates conservation values from a multiple alignment, may be run over the World Wide Web (http://geoff.biop.ox.ac.uk/servers/amas-server.html).

Although conservation scores are absolute, the relative importance of a conservation score is dependent on the overall similarity between the sequences in the multiple alignment. For example, in an alignment of 20 sequences that all share $>90\%$ pairwise identity conservation scores above 8 may be interesting. In contrast, if the pairwise identity is below 30% then lower conservation scores will be informative. Accordingly, in this study we normalised conservation scores by the average conservation for each alignment to give relative conservation scores C_r. We refer to a position as 'conserved' if $C_r > 1$.

Accessible surface areas were calculated by the program DSSP [14] and converted to relative accessibilities by dividing by the accessibility of the residue in a Gly-X-Gly tripeptide [15]. Two relative accessibility classes were considered $0 \leq A \leq 0.25$ (buried) and $0.25 < A \leq max(A)$ (exposed).

*Corresponding author. Fax: (44) (1865) 510454.
E-mail: gjb@bioch.ox.ac.uk

PII S0014-5793(96)01181-7

2.3. Exclusion of active site residues

It was anticipated that residues involved in active sites will be more highly conserved than residues in the bulk of the protein and that this might bias any analysis. Accordingly, active site and binding residues were identified in the data set and the data were examined with and without these residues. From the 81 proteins, 21 have a site record in their PDB file. For the remainder, the original literature on the structures was consulted. This added a further 35 proteins with sites. Unfortunately, the description of active sites varies from author to author. A precise definition is often difficult because either the active site pocket does not make covalent connections with the substrate, e.g. bilin binding protein [16], or it does not take part in the enzymatic action, e.g. rhodanese [17]. We considered those residues active site residues that either are attached to a prosthetic group (e.g. haem, Fe-S cluster) or take part in the enzymatic reaction, or if they were considered crucial by the authors of the structures even if they make only second-order interactions with the substrate (van der Waals interactions or hydrogen bonds). Among the 81 proteins 56 have active site residues giving a total of 331 residues. The most frequent active site residues are Cys(52/341), His(39/327), Tyr(21/520), Trp(7/184), Met(11/294) and Asp(28/818). The most conserved are His (mean C_r=1.6), Cys (1.5), Asp (1.4) and Gly (1.4).

3. Results and discussion

3.1. Distribution of amino acids

Fig. 1 shows the distribution of residues in buried and exposed positions. There are no surprises in this distribution with the amino acids that are predominantly hydrophobic (W, M, F, I, V, L, A) seen to be more frequently buried than exposed, and polar amino acids (T, S, N, Q, R, D, E, K) seen to be more frequently exposed than buried. Glycine and histidine are seen equally exposed and buried while proline is predominantly on the surface, presumably due to its frequent location in turns [18]. Cys is the most highly conserved residue in this data set and the rarest on the surface probably because it has the most reactive side chain [19]. The distribution of half-cystines and cysteines among the buried and exposed residues is approximately equal (79% and 80% inside, respectively). The average relative conservation score between the two covalent forms of Cys is so wide, that even the standard deviations (σ) are comparable with the difference: 1.56 (σ = 0.53) and 1.13 (σ = 0.44) for cystines and cysteines, respectively. The data set excluding the active site residues shows no appreciable differences (data not shown).

3.2. Relationship between accessibility and conservation

Fig. 2 illustrates average relative accessibility versus average relative conservation of each amino acid. Uppercase letters show data for all amino acids in the set, lowercase letters

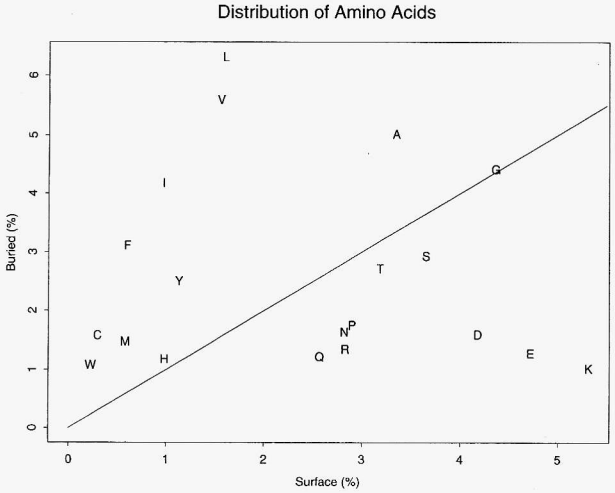

Fig. 1. Distribution of amino acids in the data base analysed by relative accessibility. Buried: percentage of amino acids <25% exposed to solvent. Surface: percentage of amino acids ≥25% exposed to solvent. The line indicates where buried = exposed. The axes refer to percentages of the total number of amino acids in the sample.

show the data set with active site residues removed. As expected, accessibility is negatively correlated with conservation. For example, hydrophobic residues which are often conserved are usually buried (M, V, I, L, F) while hydrophilic residues are less conserved and usually exposed (E, K, N). However, there are five interesting outliers. The three outliers tryptophan, cysteine and glycine show high conservation for their mean accessibility values. The simple explanation for this is that tryptophan is nearly always buried and mutation of the large residue to any other amino acid is likely to disrupt the protein core. Similarly, Cys when participating in a disulphide bridge will not favour mutation to another residue as this would leave a single free sulphydryl group. The unique properties of glycine, which can adopt φ/ψ angles unfavoured by other residues, allowing tight packing of the polypeptide chain, lead to its conservation.

The most surprising observations are the positions of Asp and Gln. Asp has a slightly smaller relative accessibility than Glu, but is significantly more conserved. Gln is significantly less conserved than Glu. The differing interactions and environments of Asp and Glu are examined in more detail in the following sections. Exclusion of the active site residues from the data set has the greatest effect on His and Cys, which

Table 1
Comparison of mean conservation for Asp and Glu in different secondary structures

	Asp (buried)		Asp (exposed)		Glu (buried)		Glu (exposed)	
SS	N	C_r $\bar{x}$ (σ)	N	C_r $\bar{x}$ (σ)	N	C_r $\bar{x}$ (σ)	N	C_r $\bar{x}$ (σ)
Helix (H)	58	1.11 (0.249)	199	0.92 (0.162)	76	1.12 (0.224)	303	0.80 (0.141)
Strand (S)	49	1.13 (0.293)	51	0.84 (0.163)	44	0.96 (0.287)	84	0.81 (0.156)
Coil (C)	121	1.25 (0.216)	340	0.95 (0.175)	51	1.01 (0.186)	275	0.81 (0.162)
H+S	107	1.13 (0.277)	250	0.90 (0.163)	120	1.05 (0.188)	387	0.80 (0.144)
Total (H+S+C)	228	1.19	590	0.93	171	1.04	662	0.81

N = number of residues in sample, $\bar{x}$ = mean relative conservation, σ = standard deviation. Secondary structure was defined by DSSP [14], then reduced to 3 states as follows: helix (H) = α, 3_{10} and π helix (DSSP H, G and I). Strand (S) = β sheet and bridge (DSSP E and B). Coil (C) = DSSP bend (S) and turn (T).

appear less conserved, but it does not affect the relative positions of the amino acids.

Analysis of relative accessibility versus relative conservation for individual amino acid positions shows a negative correlation for all amino acid types (data not shown). With the exception of Arg, Lys, Gly, Glu, Asp and Tyr a relative conservation of >2 suggests the amino acid will have a relative accessibility of $<50\%$ (data not shown).

3.2.1. Why do Asp and Glu show different conservation? Asp might be unusually conserved due to backbone conformation preferences, secondary structure preferences, or specific side-chain interactions. To decide which is responsible, we first examined the proportion of conserved Asp in the ++ ϕ/ψ conformation. 20/411 (4.55%) conserved Asp residues are in the ++ conformation, while 40/818 (4.86%) of all Asp are in this conformation. These data do not suggest a preference to conserve Asp due to maintenance of unusual backbone information.

The secondary structure distribution for Asp and Glu is summarised in Table 1. The highest mean C_r is seen for buried Asp in coil (1.25). This is significantly higher than the mean C_r for Asp in strand and helix (*t*-test gives probability of 95.5% for difference). In contrast, Glu shows no such preference for coil in either buried or exposed states. Thus, there appears to be a preference for buried Asp to be conserved in coil.

Since our data do not suggest a significant preference for ++ ϕ/ψ, the preferred conservation of Asp is likely to be due to differing side-chain interactions. The most obvious hypothesis is that since Glu has a higher proportion of non-polar atoms than Asp it can make more non-specific interactions and so there are fewer constraints on its environment. In order to test this idea, we examined the residue types that interact with Asp and Glu.

Residue pairs were considered to be interacting if the distance between any of their heavy atoms was smaller than the sum of the van der Waals radii plus 1 Å. The occurrence of Asp and Glu in our set of proteins was almost equal (818 and 833, respectively). Despite its smaller size, Asp has the same number of interactions as Glu on average (10 276/818 = 12.56 and 10 314/833=12.38). This may be due to the observed

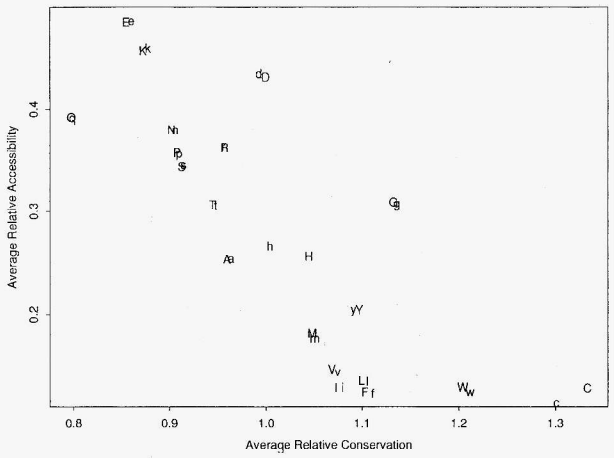

Fig. 2. Average relative accessibility versus average relative conservation for amino acids including active site residues (uppercase letters) and without active site residues (lowercase letters). The outliers are Gln (Q, q), Asp (D, d), Gly (G, g), Trp (W, w) and Cys (C, c); see text.

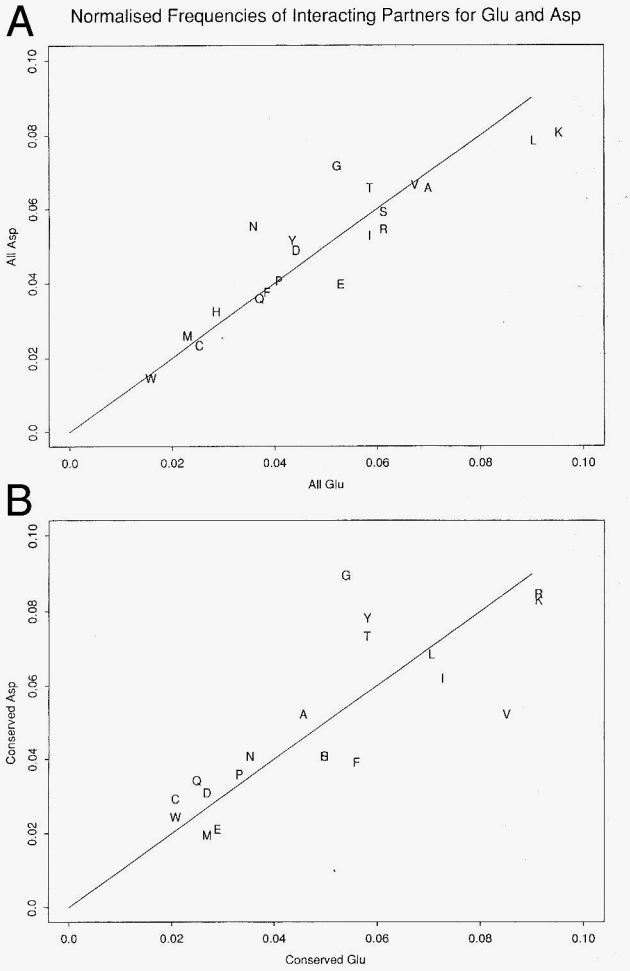

Fig. 3. Normalised frequencies of interacting partners for Glu and Asp. A: All alignment positions. B: Only conserved positions.

greater relative accessibility of Glu when compared to Asp (Fig. 2).

We calculate the normalised frequency of interaction for Asp, P_{Asp}, with each of the 20 amino acids as follows:

$$P_{\mathrm{Asp}_i} = \frac{N_{\mathrm{Asp}_{A_i}}}{\sum_{i=1}^{20} N_{\mathrm{Asp}_{A_i}}}$$

where $N_{\mathrm{Asp}_{A_i}}$ is the number of interactions between an Asp residue and amino acid type A_i. Similar frequencies were calculated for Glu.

If we consider interactions between Asp/Glu and other residues at least 5 amino acids distant in the chain, then some interesting trends emerge (Fig. 3A,B). A cutoff of 5 amino acids was chosen to exclude local secondary structure interactions. Fig. 3A shows data for all Asp and Glu residues, while Fig. 3B shows data only for Asp/Glu that are conserved. If there was no difference in the interactions of Asp and Glu, all points would lie on the line in Fig. 3A,B. In Fig. 3A, most amino acids cluster close to the line indicating equivalent interactions for Asp and Glu, but Gly and Asn appear to favour Asp. The most common interacting residues are Lys and Leu, both with a slight preference for Glu.

When only conserved Asp/Glu are considered as shown in Fig. 3B, greater scatter from the line is observed. Arg moves

from a position close to the line, to become equally favoured with Lys. Val and Phe move from being equally favoured to being preferred by Glu, while the preference of Asp for Gly is accentuated. These differences may be due to conserved Asp tending to occur in coil, where Gly is common. The longer aliphatic side chain of Glu can participate in more hydrophobic interactions than Asp and so conserved Glu residues tend to interact more frequently with hydrophobic amino acids than do conserved Asp residues. Singh and Thornton [20] reported frequencies for interacting pairs for all amino acid combinations. Their data show similar trends to ours for all data, but they did not gather statistics for conserved versus unconserved positions.

3.3. *Analysis of substitution matrices*

In this study, we consider conservation of amino acid residues across complete families. This shows that Asp is significantly more conserved than Glu. However, one might expect that this trend would also be seen in substitution matrices derived from pairwise comparisons of aligned sequences. Accordingly, we examined a number of commonly used substitution matrices to see if a preference for Asp-Asp substitutions when compared to Glu-Glu was observed.

We considered the more recently published matrices in the following articles: [8,21–34].

In order to assess the relative mutability of Asp and Glu when compared to each other a cumulative index was calculated for each mutation matrix as follows:

$$I_x = \sum_{i=1}^{20} \frac{\mathrm{Asp}(A_i) - \mathrm{Glu}(A_i)}{\mathrm{Asp}(A_i) + \mathrm{Glu}(A_i)}$$

where $\mathrm{Asp}(A_i)$ and $\mathrm{Glu}(A_i)$ are the mutation scores between Asp and Glu, and all 20 amino acids. This cumulative index results a positive score if the overall mutability of Asp is greater than that of Glu, zero if they mutate equally and negative if Glu mutates more frequently. In Table 2 we list the analysed mutational matrices with the calculated I_x. The matrices were converted into all positive values before calculating the index according to Johnson and Overington [29].

Table 2
Amino acid pair substitution matrices examined for preference to conserve Asp over Glu

Mutation matrix	I_x (Asp, Glu)
Risler et al. [31]	−5.21
Henikoff et al. [24]	−2.77
Pongor [28]	−2.66
Gonnet et al. [22]	−2.30
Miyata et al. [27]	−2.12
Johnson et al. [29]	−1.76
Henikoff et al. [23]	−1.71
Tusnady et al. [34]	−1.44
Rao [30]	−0.84
Altschul [26]	−0.42
Dayhoff et al. [21]	−0.29
Overington et al. [8]	0.00
Tudos et al. [33]	0.55
Levin et al. [32]	0.67
Tusnady et al. [34]	0.83
Jones et al. [25]	0.95

Henikoff et al. [24] refers to the BLOSUM62 matrix. Dayhoff et al. [21] refers to the PAM250 matrix, Altschul [26] refers to PAM120, Henikoff et al. [23] refers to a matrix derived from the structural alignments of Overington et al. [8].

Although the majority of the matrices show negative values of I_x there is no consistent explanation for the values. For example, the Risler et al. matrix [31] is derived from structural alignments and shows $I_x = -5.21$ while the BLOSUM62 matrix [24] ($I_x = -2.77$) is derived purely from sequence alignment. It is difficult to make direct comparisons between all the matrices shown in Table 2 since they are calculated for alignments of differing similarity. For example, BLOSUM62 represents sequences at a shorter evolutionary distance than PAM250 or PET92 [25]. The families we have analysed in the present study only include sequences that are readily alignable by sequence comparison methods. Accordingly, our results are more likely to be consistent with a matrix such as BLOSUM62 than one at a greater evolutionary distance, e.g. PET92.

4. Conclusions

In this study we have analysed multiple alignments for 81 non-homologous protein families each of which has at least one member of known three-dimensional structure. We have examined the relationship between the conservation of physico-chemical properties at a position and the relative accessibility. The principal new observations are that Asp is more highly conserved for its accessibility than Glu (Fig. 2), and that conserved Asp is most often found in coil (Table 1). The differences in interacting partners for Asp and Glu show Glu to favour non-polar partners more than Asp (Fig. 3). This may be explained simply by the higher proportion of non-polar atoms in the Glu side chain. Although carboxylate-amino interactions in proteins have been studied in some detail [35,20,36], these studies did not discriminate between conserved and variable positions and so do not help explain our current observations.

Why, then, is Asp most highly conserved when buried in coil? The short Asp side chain is restricted in mobility yet able to make strong polar interactions. It is possible that Asp may form a 'pin' that stabilises non-regular structures in loops. Further work will be required to dissect the precise role of conserved Asp in specific coil structures.

This study has elucidated the structural reasons for the greater conservation of Asp over Glu, but it is intriguing to speculate why this situation may have arisen during evolution. Indeed, why is Gln found in this study to be the least conserved of all amino acids? The differences we see here may be due to the relative lability of Asp/Asn and Glu/Gln. Asp may cyclise into a succinimidyl ring, then hydrolyse back to Asp in both D and L isomers, causing the death of the protein. For Asn the half-life is 1.4 days for cyclisation, for Asp 53 days, but Gln will only cyclise at the N-terminus [19]. Thus, Asp and Asn residues could be regarded as time bombs in proteins whereas Gln is a useful and safe 'filler'. These chemical pressures may contribute to the observed greater conservation of Asp over Glu and Gln.

Acknowledgements: We thank Professor L.N. Johnson for providing a stimulating working environment, Drs. R.B. Russell and C.D. Livingstone for advice and guidance. This work was carried out during 1993 when A.F. was supported by a 9 month Soros Fellowship. G.J.B. thanks the British Council and Royal Society for support.

References

[1] Casari, G., Sander, C. and Valencia, A. (1995) Nature Struct. Biol. 2, 171–178.
[2] Livingstone, C.D. and Barton, G.J. (1993) Comp. Appl. Biosci. 9, 745–756.
[3] Barton, G.J. (1995) Curr. Opin. Struct. Biol. 5, 372–376.
[4] Livingstone, C.D. and Barton, G.J. (1996) Methods Enzymol. 266, 497–512.
[5] Benner, S.A. and Gerloff, D. (1990) Adv. Enzyme Reg. 31, 121–181.
[6] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[7] Russell, R.B. and Sternberg, M.J.E. (1995) Curr. Biol. 5, 488–490.
[8] Overington, J., Donnelly, D., Johnson, M.S., Sali, A. and Blundell, T.L. (1992) Protein Sci. 1, 216–226.
[9] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[10] Barton, G.J. (1993) Protein Eng. 6, 37–40.
[11] Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195–197.
[12] Barton, G.J. (1993) Comp. Appl. Biosci. 9, 729–734.
[13] Barton, G.J. and Sternberg, M.J.E. (1987) J. Mol. Biol. 198, 327–337.
[14] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[15] Rose, G., Geselowitz, A., Lesser, G., Lee, R. and Zehfus, M. (1985) Science 229, 834–838.
[16] Huber, R., Schneider, M., Mayr, I., Mueller, R., Deutzmann, R., Suter, F., Zuber, H., Falk, H. and Kayser, H. (1987) J. Mol. Biol. 198, 499–513.
[17] Ploegman, J.H., Drent, G., Kalk, K.H. and Hol, W.G.J. (1979) J. Mol. Biol. 127, 149–162.
[18] Wilmot, A.C.M. and Thornton, J.M. (1988) J. Mol. Biol. 203, 221–232.
[19] Creighton, T.E. (1993) Proteins: Structures and Molecular Properties, 2nd Edn. W.H. Freeman, New York.
[20] Singh, J. and Thornton, J.M. (1990) J. Mol. Biol. 211, 595–615.
[21] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in: Atlas of Protein Sequence and Structure (Dayhoff, M.O., Ed.), Vol. 5, pp. 345–358. National Biomedical Research Foundation, Washington, DC.
[22] Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) Science 256, 1443–1444.
[23] Henikoff, S. and Henikoff, J.G. (1993) Proteins 17, 49–61.
[24] Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad. Sci. USA 89, 10915–10919.
[25] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) Comp. Appl. Biosci. 8, 275–282.
[26] Altschul, S.F. (1991) J. Mol. Biol. 219, 555–565.
[27] Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) J. Mol. Evol. 12, 219–236.
[28] Pongor, S. (1987) Methods Enzymol. 154, 450–473.
[29] Johnson, M.S. and Overington, J.P. (1993) J. Mol. Biol. 233, 716–738.
[30] Rao, J.K.M. (1987) Int. J. Peptide Protein Res. 29, 276–281.
[31] Risler, J.L., Delorme, M.O., Delacroix, H. and Henaut, A. (1988) J. Mol. Biol. 204, 1019–1029.
[32] Levin, J.M., Robson, B. and Garnier, J. (1986) FEBS Lett. 205, 303–308.
[33] Tudos, E., Cserzo, M. and Simon, I. (1990) Int. J. Peptide Protein Res. 36, 236–239.
[34] Tusnady, G.E., Tusnady, G. and Simon, I. (1995) Protein Eng. 8, 417–423.
[35] Singh, J., Thornton, J.M., Snarey, M. and Campbell, S.F. (1990) FEBS Lett. 224, 161–171.
[36] Mitchell, J.B.O., Thornton, J.M., Singh, J. and Price, S.L. (1992) J. Mol. Biol. 226, 251–262.