Protein Engineering vol 5 no 1 pp.43–51, 1992

A variable gap penalty function and feature weights for protein 3-D structure comparisons

Zhan-Yang Zhu, Andrej Šali and Tom L.Blundell[1]

ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

[1]To whom correspondence should be sent

We have developed a variable gap penalty function for use in the comparison program COMPARER which aligns protein sequences on the basis of their 3-D structures. For deletions and insertions, components are a function of structural features of individual amino acid residues (e.g. secondary structure and accessibility). We have also obtained relative weights for different features used in the comparison by examining the equivalent residues in weight matrices and in alignments for pairs of 3-D structures where the equivalences are relatively unambiguous. We have used the new parameters and the varible gap penalty function in COMPARER to align protein structures in the Brookhaven Data Bank. The variable gap penalty function is useful especially in avoiding gaps in secondary structure elements and the new feature weights give improved alignments. The alignments for both azurins and plastocyanins and N- and C-terminal lobes for aspartic proteinases are discussed.

Key words: Needleman and Wunsch algorithm/relative weights/structure comparison/variable gap penalty function

Introduction

The comparison of protein 3-D structures has become one of the most important ways of studying protein evolution and understanding the relationships between protein sequences and their folds. It is also a basic step in knowledge-based or comparative modelling.

Several approaches to structure comparison have already been developed. In the method of Rossmann and coworkers (Rao and Rossmann, 1973; Eventoff and Rossmann, 1975; Rossmann and Argos, 1977), an initial set of equivalent residues is updated according to both distances between C^{α} atoms and local main-chain orientations. In an alternative method described by Remington and Matthews (1978 , 1980), each possible backbone segment of length n residues from the first protein is compared in turn with each possible backbone segment of the same length from the second protein by a least squares procedure.

Since the differences rarely occur in the elements of secondary structures, Murthy (1984) proposed a method to compare secondary structures represented by vectors. A similar procedure (Richards and Kundrot, 1988), in which the secondary structure elements are represented by a distance matrix, has been used to compare local relationships between secondary structural elements in database searches for given secondary structure patterns.

More recently two groups (Taylor and Orengo, 1989a,b; Šali and Blundell, 1990) have proposed the use of the dynamic programming technique (Needleman and Wunsch, 1970) which was traditionally applied to sequence comparisons. In the work of Taylor and Orengo (1989a,b) interatomic vectors between residues are compared. In the computer program COMPARER of Šali and Blundell (1990), several features of protein sequences and structures are simultaneously compared. Because the method considers the most conserved protein features at a number of structural levels, it can be used to align distantly related protein structures. This is difficult to achieve by using least squares superposition alone.

In COMPARER, a residue-by-residue matrix is calculated for each feature listed in Table I. Each matrix represents the differences in this feature between each residue pair for the two related proteins. An overall residue-by-residue matrix is obtained by summing all normalized individual feature matrices with proper weights. The dynamic programming procedure is then used to get the optimal alignment (Needleman and Wunsch, 1970).

Although a dissimilarity matrix can be calculated easily for the feature involving only one residue (e.g. a property such as accessibility etc.), dissimilarity matrices are harder to obtain when the features involve more than one residue (e.g. relationships such as hydrogen bonding, hydrophobic contacts). In COMPARER, simulated annealing is used to align 'sequence of relationships' first. Then dissimilarity matrices are calculated from the alignments obtained by comparison of relationships. A tree-like addition of proteins has also been implemented for multiple alignments in COMPARER (Hogeweg and Hesper, 1984; Feng and Doolittle, 1987).

In most implementations of the Needleman and Wunsch algorithm, a uniform gap penalty function $g(l) = u \times l + v$ is used, where l is the length of a gap and u and v are constants. This implies that residues at any position of a protein have had the same chance of deletion or insertion during evolution. However, many studies have shown that insertions or deletions occur more often on the protein surface than in the core. Any deletion/insertion within a secondary structure element, especially

Table I. Features used in COMPARER currently

Index	Brief description
1	Residue local fold
2	Residue type properties
3	Residue distance from MGC*
4	Side-chain orientation relative to MGC*
5	Side-chain orientation relative to main-chain
6	Main-chain orientation relative to MGC*
7	Side-chain solvent accessibility
10	Main-chain solvent accessibility
14	Hydrogen bonding and hydrophobic contact relationship
15	Residue identity
17	Residue position in space
18	phi dihedral angle
19	psi dihedral angle
20	Main-chain direction
21	Hydrogen bonding relationship

MGC*, Molecular Gravity Centre.

in a helix, would change the packing of residues. Such information relating to the mechanism of protein evolution was applied to the alignment of distantly related sequences by Lesk *et al.* (1986) and Barton *et al.* (1987). They used a secondary structure dependent gap penalty function, defined by the 3-D structure of one member of the protein family. In the method of Lesk and coworkers, the gap penalty for residues not in a helix or a β-strand was set to 1, for the last two residues of a helix or a β-strand it was increased to 2 and for residues in a helix or a β-strand it was further increased to 8. In a similar way, Barton *et al.* (1987) used the following function $G(l) = Q^*(u^*l + v)$ where $Q = 1.0$ in regions of clear secondary structure and $Q = 0.25$ elsewhere.

Here, we describe a flexible gap penalty function for protein structure alignments and its incorporation into the Needleman and Wunsch algorithm. Because each gap involves a deletion relative to one sequence and an insertion relative to another, we use three components in the gap penalty function. The first and second components depend on structural features of the two proteins in the regions of sequences compared. The third component is a length-independent constant. The approach has been implemented in the structural comparison program COMPARER (Šali and Blundell, 1990).

In this paper, we also discuss relative weights for different features used in COMPARER. We derive the weights from known alignments. We define a similar residue pair as the residue pair whose dissimilarity value is less than a given cutoff. We calculate the probability of the similar residue pairs occurring in all possible residue pairs and the probability of the similar residue pairs occurring in alignments. The difference of the two probabilities is used to represent the relative weight of the feature. The normalized weights were used for further structure comparisons.

Materials and methods

Variable gap penalty function

Gap penalty function. If a length of gap between residues j and $(j + 1)$ of structure B is k and it is aligned with residues from i to $i + k - 1$ in structure A, then the penalty for the gap is

$$\sum_{l=0}^{k-1} (A_{1,i+l} + B_{2,j}) + v$$

where $A_{1,i+l}$ is the first component of the gap penalty for residue $(i + l)$ in the first structure A, $B_{2,j}$ is the second component of the gap penalty for residue j in the second structure B. v is a constant, the third component.

Similarly, if the gap is between residues i and $(i + 1)$ in the first structure and it is aligned with residues from j to $j + k - 1$ in structure B, then the penalty for the gap is

$$\sum_{l=0}^{k-1} (A_{2,i} + B_{1,j+l}) + v$$

where $A_{2,i}$ is the second component of the gap penalty for residue i in the first structure A, $B_{1,j+l}$ is the first component of the gap penalty for residue $(j + l)$ in the second structure B.

Calculation of the first and second components of gap penalty. By analysis of the alignments of known 3-D structures we have investigated the correlation between residue fractional side-chain accessibilities defined in Hubbard and Blundell (1987) and probabilities of residues to be deleted in loop regions and in secondary structure elements. We count the total number of residues whose average side-chain accessibilities with their

Table II. Structures used in the analysis

PDB code	Name
1 Globin	
2hhb	haemoglobin (human) α-chain
2hhb	haemoglobin (human) β-chain
3mbn	myoglobin (sperm whale)
1ecd	erythrocuorin (*Chironomous thummi thummi*)
2lhb	haemoglobin (sea lamprey)
1lhl	leghaemoglobin (*Lupinus luteum*)
2 Phospholipase A	
1bp2	phospholipase A2 (bovine)
1p2p	phospholipase A2 (porcine)
1pp2	phospholipase A2 (rattlesnake)
3 Serine proteinase	
1ton	tonin (rat)
2pka	kallikrien A (porcine)
2ptn	trypsin (bovine)
3est	pancreatic elastase (porcine)
3rp2	mast cell proteinase (rat)
1sgt	trypsin (*Streptomyces griseus*)
2sga	proteinase A (*S.griseus*)
2alp	α-lytic proteinase (*Lysobacter enzymogenes*)
3sgb	proteinase B (*S.griseus*)
4cha	α-chymotrypsin (cow)
4 Aspartic proteinase	
2app	penicillopepsin (*Penicillium janthinellum*)
2apr	rhizopuspepsin (*Rhizopus chinensis*)
2cms	chymosin (bovine)
2pep	pepsin (porcine)
4ape	endothiapepsin (*Endothia parasitica*)
5 Retroviral proteinase	
2rsp	Rous sarcoma proteinase
5hvp	HIV-1 proteinase
6 Immunoglobulin variable domain	
1fb4	immunoglobulin Fab KOL HV
3fab	immunoglobulin λ-Fab HV
3fab	immunoglobulin λ-Fab LV
1rei	immunoglobulin Bence-Jones LV (human)
2hfl	immunoglobulin HY-HEL Fab LV (mouse)
1rhe	immunoglobulin Bence-Jones LV (human)
7 Immunoglobulin constant domain	
3fab	immunoglobulin λ-Fab LC
1fbj	immunoglobulin Fab LC (mouse)
1fcl	immunoglobulin Fc HC2
1fb4	immunoglobulin Fab KOL HC
1fbj	immunoglobulin Fab HC (mouse)
8 Azurin	
1AZU	azurin (*Pseudomonas aeruginosa*)
2AZA	azurin (*Alcaligenes faecalis*)
9 Cysteine proteinase	
2act	actinidin (*Actinida chinensis*)
9pap	papain (*Carica papaya*)
10 Cytochrome *c*	
3cyt	cytochrome *c* (albacore)
1ccr	cytochrome *c* (rice)
2c2c	cytochrome *c*2 (*Rhodospirillum rubrum*)
1cyc	ferrocytochrome *c*
155c	cytochrome c550 (*Paracoccus denitrificans*)
11 Cytochrome c5	
351c	cytochrome c551 (*Pseudomonas aeruginosa*)
1cc5	cytochrome c5 (*Azotobacter vinelandii*)

Table II. *Continued*

PDB code	Name
12 Dehydrofolate reductase	
3dfr	dehydrofolate reductase (*Lactobacillus casei*)
4dfr	dehydrofolate reductase (*Escherichia coli*)
13 Ferredoxin	
1fdx	ferredoxin (*Peptococcus aerogenes*)
4fd1	ferredoxin (*Azotobacter vinelandii*)
14 Flavodoxin	
1fx1	flavodoxin (*Desulfovibrio vulgaris*)
3fxn	flavodoxin (*Clostridium sp*.)
15 Lysozyme	
1lzt	lysozyme (chicken)
1lz1	lysozyme (human)

neighbouring residues (one on each side of the residue) are in the same range: 0−5%, 5−10% and so on. We also count the total number of those residues aligned with gaps. The ratio of the two numbers is taken as the probability of a deletion. Based on this analysis, we define the first component of the gap penalty $A_{1,i}$ or $B_{1,i}$, as a function of side-chain accessibilities and secondary structures. For residue i:

$$A_{1,i} \text{ or } B_{1,i} = s_s - s_1 * \text{accessibility}(i)$$

where accessibility(i) is the average side-chain accessibility of residue i and its neighbouring residues (one on each side of residue i is used). s_s and s_1 are parameters. s_s is the secondary structure-dependent parameter. s_1 is a scale factor.

The second component of the gap penalty $A_{2,i}$ or $B_{2,i}$ for an insertion is calculated similarly to the first component, but the side-chain accessibility is the average side-chain accessibility of two neighbouring residues between which the new residue is inserted.

Dynamic programming procedure that includes the variable gap penalty function. If the numbers of residues in the first and second structures compared are M and N respectively, W is the residue by residue weight matrix, and D is the distance matrix for back-tracking the alignment, then the modified dynamic programming formulae that give matrix D *are:*

$$D_{i,j} = \min \begin{cases} P_{i,j} \\ D_{i-1,j-1} + W_{i,j} \\ Q_{i,j} \end{cases}$$

$$P_{i,j} = \min \begin{cases} D_{i-1,j} + A_{1,i} + B_{2,j} + v \\ P_{i-1,j} + A_{1,i} + B_{2,j} \end{cases}$$

$$Q_{i,k} = \min \begin{cases} D_{i,j-1} + A_{2,i} + B_{1,j} + v \\ Q_{i-1,j} + A_{2,i} + B_{1,j} \end{cases}$$

The arrays D,P and Q are initialized as follows:

$$D_{i,0} = \begin{cases} 0 & i \le e_1 \\ \Sigma_{k=e_1+1}^{i} A_{1,k} + v & e_1 < i \le M \end{cases}$$

$$D_{0,j} = \begin{cases} 0 & j \le e_2 \\ \Sigma_{k=e_2+1}^{j} B_{1,k} + v & e_2 < j \le N \end{cases}$$

$$P_{i,0} = Q_{i,0} = +\infty, i = 1,2, \ldots, M$$

$$P_{0,j} = Q_{0,j} = +\infty, i = 1,2, \ldots, N$$

where parameters e_1 and e_2 are the maximum numbers of elements at sequence termini that are not penalized with a gap penalty if not equivalenced. The minimal score $d_{M,N}$ is obtained from

$$d_{M,N} = \min(D_{i,N}, D_{M,j})$$

where $i = M, M-1, \ldots, M - e_1$ and $j = N, N-1, \ldots, N - e_2$ to allow for the overhangs. Alignment is obtained by backtracking in matrix D from $D_{i,j} = d_{M,N}$.

Multiple alignment. Using tree-like addition of proteins for multiple alignment (Hogeweg and Hesper, 1984; Feng and Doolittle, 1987; Šali and Blundell, 1990), different sets of gap penalty parameters can be used simultaneously when homologous structures covering a wide range of similarities are aligned.

The first and the second components of the gap penalty at each position of a sub-alignment are defined as the averages of the corresponding components of gap penalties of residues at the same position. If a gap occurs in a sub-alignment, the first component of the gap penalty of the 'residue' is 0; and its second component is equal to the second component of a residue i if the gap occurs between residues i and $i + 1$ of a structure.

Relative weights of features

If the sequence of protein A is $a_1, a_2, \ldots, a_{n_1}$ (n_1 is length of the sequence A) and the sequence of protein B is $b_1, b_2, \ldots, b_{n_2}$ (n_2 is length of the sequence B), we define the set of all residue pairs between A and B as S:

$$S = \{(a_i, b_j) | i = 1,2, \ldots, n_1; j = 1,2, \ldots, n_2\}$$

The dissimilarity value is assigned to each residue pair in S. For the alignment of A and B, we also define a set of residue pairs S':

$S' = \{(a_i, b_j)$ if a_i, b_j are aligned in the alignment$\}$

Obviously, S' is a subset of S: $S' \subset S$.

In S' the pairs between gaps and residues are not included. But we can still derive a unique alignment from S' as follows:

(i) sort all pairs in S' according to the indices of the first elements (sequence A) or the second elements (sequence B) of these pairs;
(ii) add missing residues in sequences A and B; these residues are aligned with gaps;

So, there is a one to one relationship between an alignment and S'; We simplify the dynamic programming procedure as a transformation called DP:

$$S \overset{DP}{\Rightarrow} S'$$

The transformation is used to select elements from S according to the dissimilarity values and gap penalty functions. The selected elements constitute S'. We define a similar residue pair as the residue pair whose dissimilarity value is less than a given cutoff. For each feature f, we calculate the ratio p_r between the total number of similar residue pairs and total number of residue pairs in the set S and the same ratio p_a^f in S'. The parameter p_r^f is the probability of DP that selects similar pairs from S randomly. p_a^f is the probability of similar pairs occurring in the reference alignment if S is obtained from the reference alignment. If the difference between p_a^f and p_r^f is large, it means similar residue

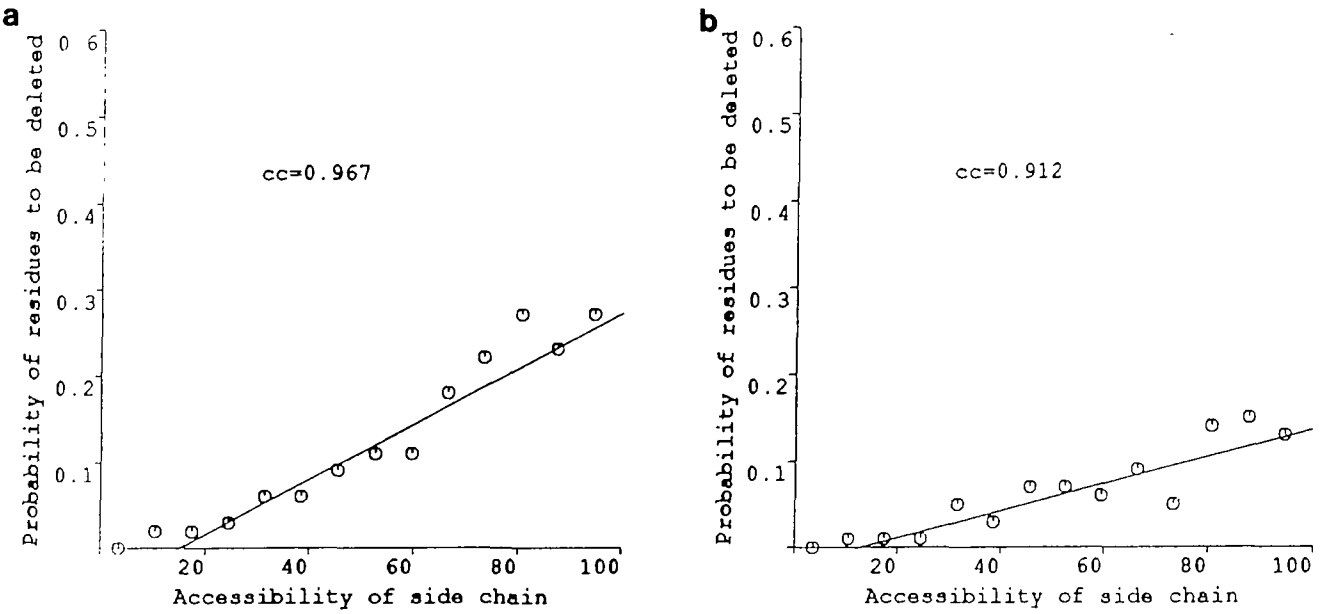


Fig. 1. The correlations between residue side-chain accessibilities and probabilities of residues to be deleted for alignments of proteins with sequence identities (a) 20–30% and (b) 40–50%. Accessibilities are calculated as described in the text. The protein structures used in the analysis are listed in Table II The circles indicate the observed values. Solid lines are least square fits to the observed values, cc is the correlation coefficient.

Table III. Normalized relative weights

Feature no.	Cutoff	Weights for different sequence identities						
		≤20	20–30	30–40	40–50	50–60	60–70	70–100
1	3.0	1 035	1.076	1 081	1 114	1 068	1.095	1.024
2	2.3	0.522	0.571	0 623	0 716	0.780	0.893	1.105
3	10.0	1.393	1.357	1.227	1.197	1 030	1.049	0.958
4	180.0	0.090	0.117	0.057	0.083	0.061	0.049	0 039
5	180.0	0.872	0.762	0.660	0.623	0 542	0.619	0.567
6	180.0	0.932	0.908	0 791	0.795	0.701	0.775	0.749
7	100.0	0 838	0.761	0.691	0.721	0 680	0.735	0.755
10	100.0	0 547	0.438	0.465	0.412	0 448	0.454	0.487
14	1.0	0 515	0.787	0.836	0 953	1 083	0.915	0 884
17	7.0	1.206	1 177	1.279	1.196	1.271	1 266	1 242
18	180.0	0.595	0.547	0.589	0.563	0.577	0.475	0.456
19	180.0	0.807	0.717	0.784	0.740	0.723	0 643	0 672
20	10 0	0.647	0.783	0.916	0.887	1.038	1 030	1 063
21	1 0	0.538	0.802	0.917	0.918	1 019	0 737	0.581

Note: see Table I for feature numbers.

pairs occur in alignments with higher probability than the probability that similar residue pairs occur randomly. This reflects the fact that this feature is important for the reference alignment. Hence it should have a great contribution to the overall matrix. We use the normalized difference as a relative weight of the feature.

In our analysis, we group all known pairwise alignments (these may be inferred from multiple alignments) (Table II) according to the sequence identity of each pair of proteins compared. For example, all alignments whose sequence identities are <20%, 20–30%, etc. are in different groups. Then for each group of the alignments, p_r^f and p_a^f were calculated for several different cutoffs for similar pairs. The cutoff is selected so that most residue pairs in the alignment are similar pairs.

The difference between p_r^f and p_a^f for each feature is normalized by the sum of all the differences.

Results and discussion

Figure 1 shows a linear correlation between amino acid accessibilities and the probabilities of deletions. In our analysis we have used the average accessibility of three residues: the residue to be deleted and those on either side. For an insertion we have used the average accessibilities for the two residues between which the insertion is to be made. In all cases the accessibilities are those of the side-chain. The correlation between the probability of a deletion and the accessibility is better for the three residues than for the single residue deleted, indicating that the probability of a deletion or an insertion is dependent on the general accessibility of a region of the polypeptide rather than that of the individual amino acid. This is consistent with the idea that some movement is required in order to accommodate the insertion or deletion. The analysis shows that the increase in the probability of a deletion with increased accessibility is greater

Table IV. Correlation coefficient of relative weights

Groups	≤20	20–30	30–40	40–50	50–60	60–70	70–100
≤20	1 000	0.886	0 799	0.734	0 560	0.683	0.599
20–30	0.886	1 000	0.960	0.945	0.824	0.865	0 742
30–40	0.799	0.960	1.000	0.984	0.934	0.926	0.915
40–50	0.734	0.945	0.948	1.000	0.959	0 948	0.848
50–60	0.560	0.824	0.934	0.959	1.000	0.944	0.866
60–70	0.638	0.865	0.926	0.948	0.944	1.000	0.966
70–100	0 599	0.742	0.915	0.848	0.866	0.966	1.000

Table V. Structures used in the comparisons

PDB code	Name	*N* residues	Resolution	R factor
(a) Azurin/plastocyanin				
1AZU	azurin (*Pseudomonas aeruginosa*)	126	2.7	35.0
1PAZ	pseudoazurin (*Alcaligenes faecalis*)	120	1.5	15.9
1PCY	Plastocyanin (poplar)	99	1 6	17.0
2AZA	azurin (*Alcaligenes faecalis*)	129	2 5	15.7
(b) Aspartic proteinase				
2APP	pencillopepsin (*Penicillinium janthinellum*)	323	1.8	13.6
2APP	rhizopuspepsin (*Rhizopus chinensis*)	325	1.8	14.3
2CMS	chymosin (bovine)	323	2.2	
2PEP	pepsin (porcine)	326	2.0	18.0
4APE	endothiapepsin (*Endothia parasitica*)	330	2.1	17.8

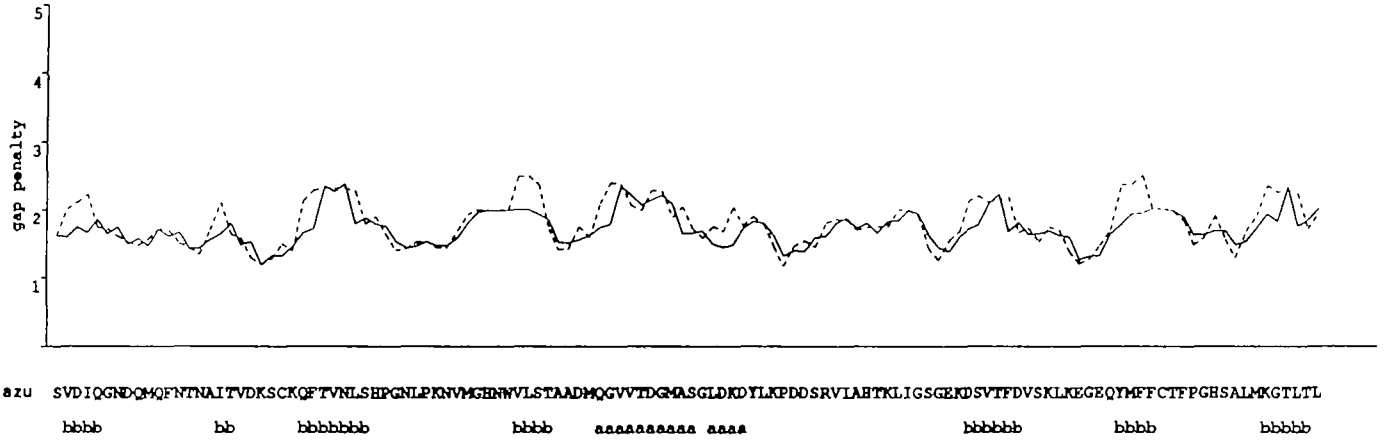


Fig. 2. The gap penalties for each of the residues in protein P1AZU that were used to obtain the sub-alignment of P1AZU and P2AZAA by COMPARER. The solid line represents the first components of gap penalties The dashed line represents their second components. The first line below the *x*-axis is the sequence of P1AZU, in which the standard one-letter code for amino acids is used. The second line indicates the secondary structures of residues which are defined by the program DSSP (Kabsch and Sander, 1983): 'a', α-helix, 'b', β-strand.

for comparison of sequences with lower percentage sequence identity.

Our analysis of the probability of finding insertions or deletions in secondary structure elements confirms the generally held view that they rarely occur in α-helices and β-strands, although their existence in β-strands can be accommodated with a β-bulge and in a helix by a kink. We have defined three values of the parameters s_s which characterize residues in α-helices, β-strands and loop regions. From an analysis of alignments using different values of these parameters we have demonstrated that values for s_s of 2.5, 2.5 and 2.00 are appropriate. We have also introduced a further parameter s_1 which relates the variation of gap penalties for the different secondary structure elements to the variation of side-chain accessibilities. Again, using various values in trial alignments, we found that an optimal value was 1. The analysis also shows that deletions may occur at the termini of secondary structure elements. Therefore, we reduce the first component gap penalties for the first and last two residues at each end of the secondary structural elements by 20%.

Table III shows seven groups of normalized weights for each feature. Table IV, which gives their correlation coefficients, shows that the seven groups of parameters have no significant differences, although they were derived from alignments which have different similarities. This may be a consequence of the fact that these parameters are relative weights.

Features 1, 3 and 17, which are related to the residue positions in space, have relatively high weights. However, although the residue positions are important in 3-D structure comparisons, they do not dominate the parameters in Table III. Table III also shows that side-chain accessibility (feature 7) has higher weight than

main-chain accessibility (feature 10) and that the psi dihedral angles (feature 19) have higher weight than phi angles (feature 18). The weight of side-chain orientation relative to molecular gravity centre (feature 4) is the lowest.

We have used the modified COMPARER with the new weights and the variable gap penalty function to align the proteins in Table V and other proteins or motifs from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). The parameters for the relevant percentage sequence identities (Table III) are used without manual intervention.

Alignment of azurin and plastocyanin proteins

Azurin and plastocyanin are blue copper proteins. The four proteins, for which high resolution 3-D structures have been determined, are listed in Table V. Each of the proteins contains two β-sheets.

We first compared P1AZU and P2AZAA and P1PAZ with P1PCY. The two resulting sub-alignments were then compared to get the family alignment. Figure 2 shows the gap penalty used in the alignment of P1AZU. The two sub-alignments and their gap penalities in each position are shown in Figure 3. Figure 4 shows the COMPARER alignment of the four proteins.

The alignments of azurin and poplar plastocyanin (corresponding to P1AZU and P1PCY in our data set) have been previously considered by Chothia and Lesk (1982), Adman (1985) and Taylor and Orengo (1989b). For sheet S-2 indicated in Figure 4, our alignment agrees with theirs. However, for sheet S-1, our

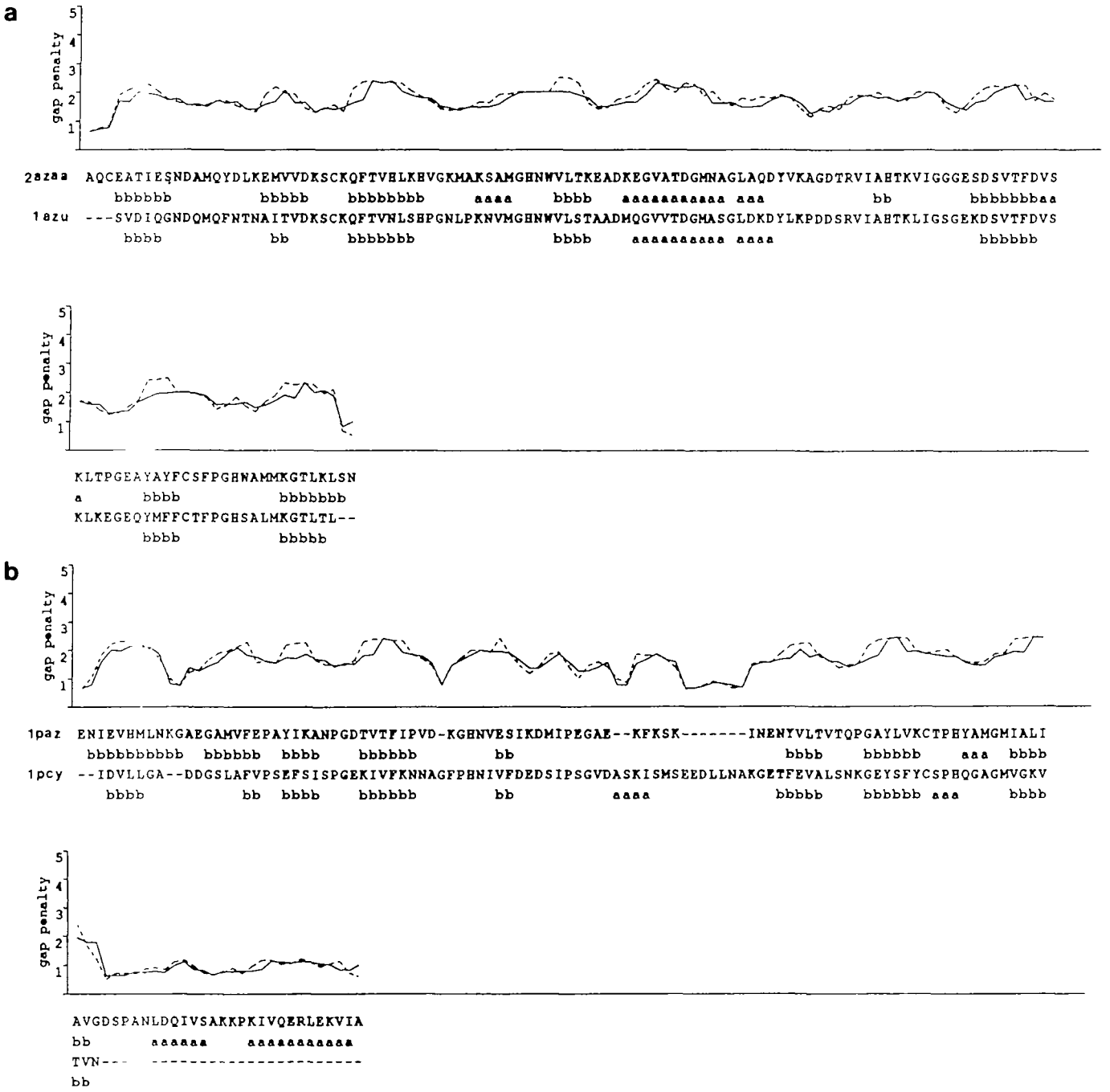


Fig. 3. The gap penalties for each of the 'residues' (a) in the sub-alignment of P1AZU and P1AZAA and (b) in the sub-alignment of P1PAZ and P1PCY which were used to obtain the family alignment (Figure 4). The same convention as in Figure 2 is used for gap penalties and protein sequences.

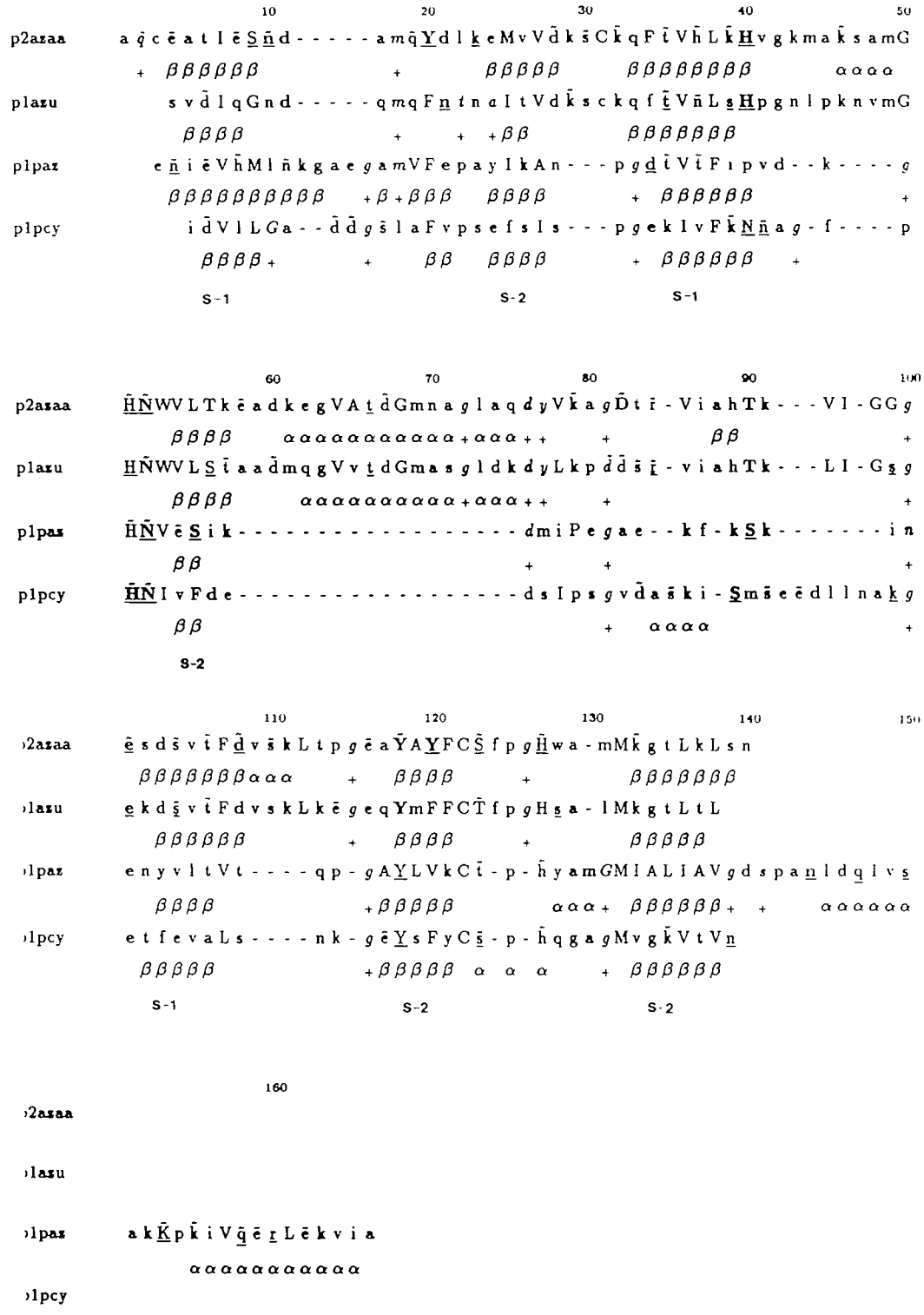
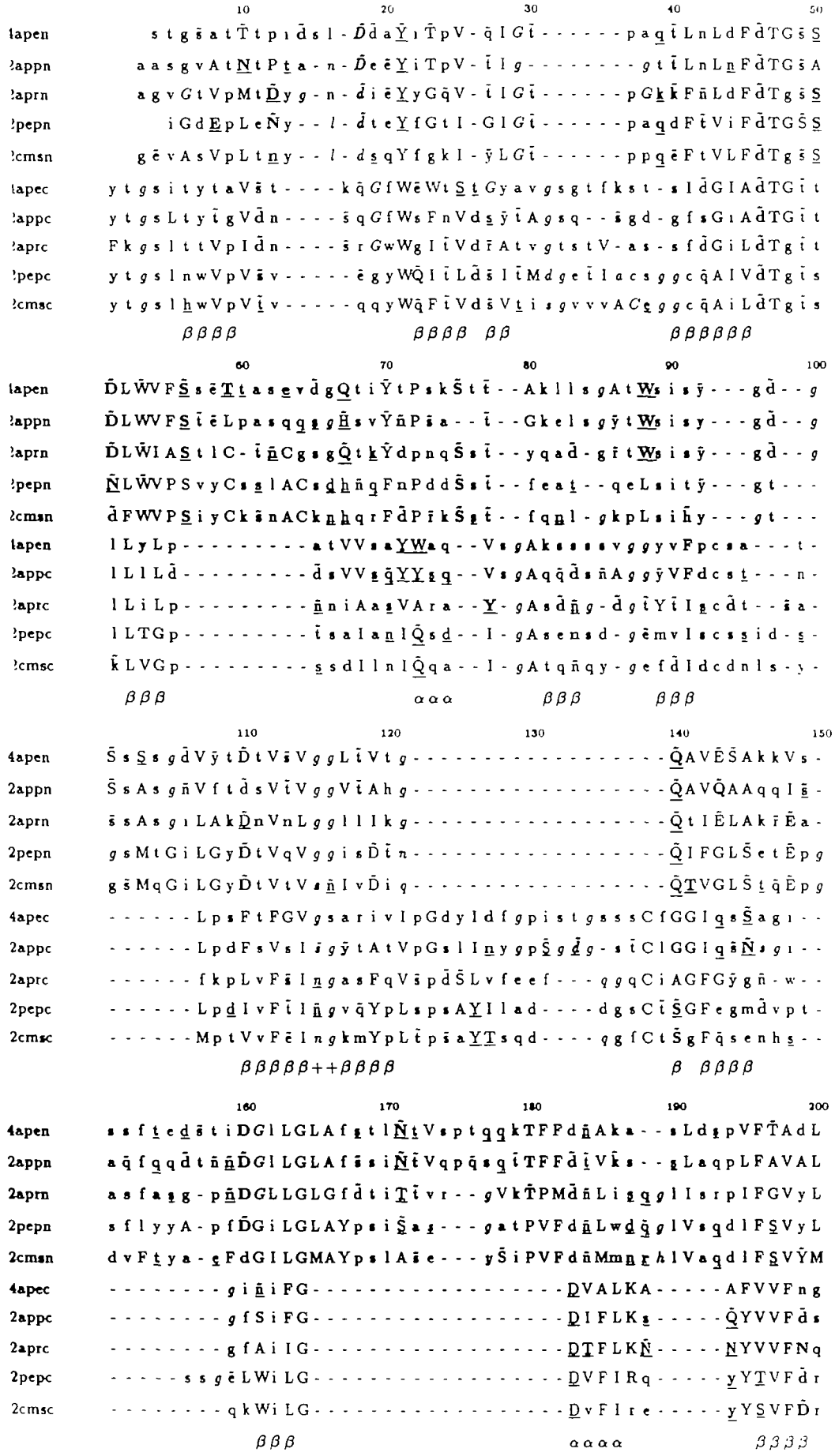


Fig. 4. The COMPARER alignment of azurin and plastocyanin proteins See Table V for protein codes. The amino acid code is the standard one-letter code formatted using the following convention (Overington *et al.*, 1990): *Italic* for positive phi; UPPER CASE for solvent inaccessible residues (less 7%); 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

alignment agrees with those of both Adman *et al.* (1985) and Taylor and Orengo (1989b) but is shifted two residues over the whole sheet compared with that of Chothia and Lesk (1982). Comparisons of sequence and structure features of these two proteins together with those of two further related proteins as shown in Figure 4 indicate that our alignment is preferable. In the second strand of sheet 2, we have two conserved residues: an exposed and side-chain hydrogen bonded threonine and a buried valine (position numbers 35 and 36 in Figure 4) in the first three proteins. This conservation, together with the similarity between P1PCY and P1PAZ, supports the alignments of P1PCY with P1AZU in the strand shown in Figure 4. In the third strand



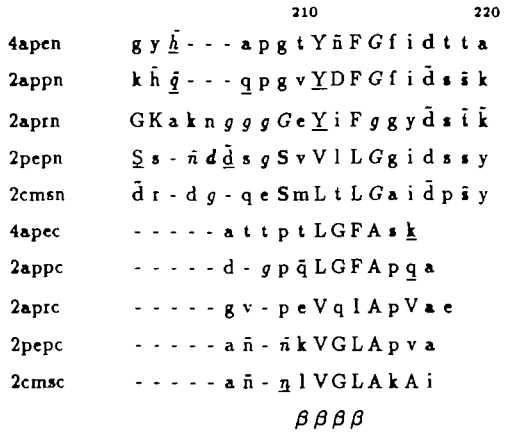


Fig. 5. The COMPARER alignment of 10 aspartic proteinase lobes obtained using the variable gap penalty function. See Table V for protein codes. The same convention as in Figure 4 is used for the amino acid codes.

of sheet 1, a conserved glutamate (position number 101 in Figure 4) is a solvent accessible residue but hydrogen bonded to a main-chain amide nitrogen in all the four proteins. We also align the only buried residue of strand 3 of each structure (position number 107 in Figure 4). This analysis shows the importance of using several protein features at the same time in the structure comparison. Figure 4 also shows that gaps align with residues in loop regions and do not break secondary structure elements with the exception of a very short helix in protein P1PCY (position numbers 123–127 in Figure 4).

Alignment of aspartic proteinases

The aspartic proteinases are a family of mainly β-sheet proteins. In the eukaryotic pepsins there are two topologically similar lobes (Tang *et al.*, 1978): the N-terminal lobe (N-lobe) and the C-terminal lobe (C-lobe). The proteins whose 3-D structures are known are listed in Table V. Because sequences of equivalent lobes are quite similar, it is not too difficult to align them correctly either by hand or by COMPARER. However, the differences in lengths of secondary structures, together with insertions and deletions of secondary structure elements in the two lobes, makes alignment of the N- and C-lobes together a challenge.

Figure 5 shows the COMPARER alignment of 10 aspartic proteinase lobes obtained using the variable gap penalty function. Compared with the previous COMPARER alignment (Šali and Blundell, 1990) obtained using a uniform gap penalty function, there are two major differences between the two alignments. First, the gaps in the helix (position numbers 56–74 in Figure 5) of C-lobes in the previous alignment are moved to the loop region between a β-strand and the helix, so that the helices are not interrupted. Secondly, the new alignment of the helix (position numbers 183–188 in Figure 5) is displaced by four residues over the whole helix compared with the previous alignment. The superpositions of N-lobes with C-lobes do not give any obvious equivalences in this region. However, the new alignment gives four common helical residues (there are two in the previous alignment) and one completely conserved aspartate in the 10 lobes. The conserved aspartate residue still occurs if we compare the lobes without weight on sequence identities. The other main differences correspond to shifts of some residues that are at termini of secondary structure elements.

Our definition of the variable gap penalty function is more consistent with models of protein evolution than a uniform gap penalty function. Furthermore our procedure has proved to be useful especially in avoiding gaps in secondary structure elements in either of the compared structures.

In our approach, we used variable gap penalties not only for different residues but also for different structure comparisons in order to obtain multiple alignments according to their evolutionary trees. This makes it much easier to obtain alignments of a number of different homologous protein structures correctly.

Acknowledgements

We are grateful to our colleagues Mark Johnson, John Overington, Pam Thomas, Dan Donnelly, Chris Topham, Alasdair McLeod, Lynn Sibanda and David Smith for many stimulating discussions. We thank the Imperial Cancer Research Fund, the Royal Society, Slovene Research Council and The J.Stefan Institute for financial support.

References

Adman,E.T. (1985) In Harrison,P.M. (ed.), *Metalloproteins*. Verlag Chemie, Weinheim, pp. 1–142.
Barton,J.G. and Sternberg,J.E.M. (1987) *Protein Engng*, **1**, 89–94.
Berstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanovichi,T. and Tasumi,M. (1977) *J. Mol Biol.*, **122**, 535–542.
Chothia,C. and Lesk,A M. (1982) *J. Mol. Biol.*, **160**, 309–323.
Eventoff,W. and Rossmann,M.G. (1975) *CRC Crit. Rev. Biochem.*, **3**, 111–140
Feng,D.-F. and Doolittle,R.F. (1987) *J. Mol. Evol.*, **25**, 351–360.
Hogeweg,P. and Hesper,B. (1984) *J. Mol. Evol.*, **20**, 174–186.
Hubbard,T.J.P. and Blundell,T.L. (1987) *Protein Engng*, **1**, 159–171
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Lesk,A.M., Levitt,M. and Chothia,C. (1986) *Protein Engng.*, **1**, 77–78.
Murthy,M.R.N. (1984) *FEBS Lett.*, **168**, 97–102.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453
Overington,J., Johnson,M., Šali,A. and Blundell,T.L. (1990) *Proc. R. Soc. Lond. B*, **241**, 132–145
Rao,S.T. and Rossmann,M.G (1973) *J. Mol. Biol.*, **153**, 1027–1042.
Remington,S.J. and Matthews,B.W. (1978) *Proc. Natl Acad. Sci. USA*, **75**, 2180–2184.
Remington,S.J. and Matthews,B W. (1980) *J. Mol. Biol.*, **140**, 77–99.
Richards,F.M and Kundrot,C.E. (1988) *Protein*, **3**, 71–84.
Rossmann,M.G. and Argos,P (1977) *J. Mol. Biol.*, **109**, 99–129
Šali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.
Tang,J , James,M., Sieleck,A., Jenkins,J.A. and Blundell,T.L. (1978) *Nature*, **271**, 618–621
Taylor,W.R. and Orengo,C.A. (1989a) *Protein Engng*, **2**, 505–519.
Taylor,W.R. and Orengo,C.A (1989b) *J. Mol. Biol.*, **208**, 1–22.

Received on August 7, 1991; accepted on November 21, 1991