Protein Engineering vol.1 no.2 pp.89–94, 1987

Evaluation and improvements in the automatic alignment of protein sequences

Geoffrey J.Barton[1] and Michael J.E.Sternberg

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

[1]To whom reprint requests should be sent

The accuracy of protein sequence alignment obtained by applying a commonly used global sequence comparison algorithm is assessed. Alignments based on the superposition of the three-dimensional structures are used as a standard for testing the automatic, sequence-based methods. Alignments obtained from the global comparison of five pairs of homologous protein sequences studied gave 54% agreement overall for residues in secondary structures. The inclusion of information about the secondary structure of one of the proteins in order to limit the number of gaps inserted in regions of secondary structure, improved this figure to 68%. A similarity score of greater than six standard deviation units suggests that an alignment which is greater than 75% correct within secondary structural regions can be obtained automatically for the pair of sequences.

Key words: alignment/protein/sequence/secondary/structure

Introduction

Automatic sequence comparisons (e.g. Needleman and Wunsch, 1970; Sellers, 1974) are used to search for homology between a newly determined sequence and one in the data bank. Recent developments have aimed at identifying local homologies (e.g. Sellers, 1979; Goad and Kanehisa, 1982; Boswell and McLachlan, 1984) or at increasing speed and reducing memory requirements (e.g. Gotoh, 1982; Taylor, 1986; Fickett, 1984; Wilbur and Lipman, 1983). This paper concentrates on the accuracy of the alignment.

From a sequence alignment of two (or more) proteins, conserved regions are identified which may provide pointers to functionally and/or structurally important regions of the molecules. This sequence information alone can be the basis for studies such as site-directed mutagenesis (Winter and Fersht, 1984) or the selection of peptides against which antibodies are raised (Sutcliffe *et al.*, 1983). Furthermore, a newly determined sequence may show sequence homology with a protein whose structure has been obtained crystallographically and this can lead to a three-dimensional atomic model by model-building techniques (e.g. Browne *et al.*, 1969; Blundell *et al.*, 1983; Travers *et al.*, 1984).

Most sequence comparison methods are based on dynamic programming algorithms such as those originally applied in molecular biology by Needleman and Wunsch (1970) and formalized by Sellers (1974) and Waterman *et al.* (1976). These methods carry out a global comparison of the sequences and the result is an optimal score for the comparison and an alignment with that score. Although the score may be valuable for identifying homology against a background of randomized sequences, the actual sequence alignment may not be correct. Even for two very short sequences (<20 amino acids when translated) taken from chicken alpha and beta haemoglobin, varying the user-supplied gap-penalty parameters will produce a number of different optimal alignments (Fitch and Smith, 1983).

Several workers have noted that alignments obtained by the superposition of homologous high-resolution protein structures can be quite different to those produced by automatic sequence based methods (e.g. Dickerson *et al.*, 1976; Chothia and Lesk, 1982; Delbaere *et al.*, 1979). In this paper, the quality of an automatically obtained sequence alignment is systematically assessed by reference to alignments based on structure superposition. Furthermore, in order to model more effectively the observed evolutionary preference for insertions/deletions to occur in the loop regions which join secondary structures (e.g. Perutz *et al.*, 1965), we have introduced a secondary structure dependent function (Q) which has the effect of reducing the penalty for a gap in non-secondary structural regions. A similar approach has been independently applied by A.M.Lesk, M.Levitt and C.Chothia (Lesk *et al.*, 1986) to align proteins within the globin and serine proteinase families.

In this paper five pairs of structurally homologous proteins were considered and the sequence alignments tested over residues which are clearly in homologous secondary structures. The inclusion of Q derived from a knowledge of the three-dimensional structure of one protein in each pair resulted in an overall improvement in alignment accuracy.

Materials and methods

Proteins aligned

Five pairs of proteins and domains, for which alignments based on three-dimensional structures are available, were taken as test data. For each pair of proteins a set of unambiguously assigned equivalent residues were selected from central portions of homologous secondary structures (Table I). The percentage accuracy of each automatic sequence alignment was calculated from the number of amino acid pairs within the defined zones which were aligned as in the structural alignment.

Needleman and Wunsch algorithm

The computer program written for this study implemented a variant of the Needleman and Wunsch (1970) algorithm.

(i) A matrix of amino acid pair scores D is chosen: in its simplest form this may indicate a score of 1 for identity and 0 for all other states, while more sophisticated systems incorporate information about conservative substitutions by using a weighting scheme. In this study the MDM_{250} matrix was used (Dayhoff, 1972, 1978) with a constant of 8 added to remove all negative terms. Preliminary studies showed this pair score matrix to be superior to either identity or genetic code types. Alternative matrices (e.g. McLachlan, 1972; Feng *et al.*, 1985) although not included in this study may be expected to perform at least as well as the MDM_{250} matrix (Feng *et al.*, 1985).

(ii) The protein sequences are defined as $A_{1,m}$, $B_{1,n}$ where m and n are the number of residues in sequence A, B, respectively.

(iii) A matrix $R_{m,n}$ is generated by reference to D where each

Table I. Protein pairs used to test alignment methods

Protein/domain A (abbreviation) (number of residues)	Protein/domain B (abbreviation) (number of residues)	Source of structural alignment	Number of residues taken for test	Number of zones
Immunoglobulin light chain variable region (FABVL) (103)	Immunoglobulin heavy chain variable region (FABVH) (117)	Cohen *et al.* (1981)	41	7
Immunoglobulin heavy chain variable region (FABVH) (117)	Immunoglobulin light chain constant region (FABCL) (105)	Cohen *et al.* (1981)	38	7
Plastocyanin (PLASTO) (99)	Azurin (AZURIN) (128)	Chothia and Lesk (1982)	48	7
Human alpha haemoglobin (HAHU) (141)	Root nodule leghaemoglobin (LEGHE) (156)	Lesk and Chothia (1980)	100	7
Trypsin (P2PTN) (223)	Tosyl-elastase (P1EST) (240)	M.Zvelebil (personal communication)	63	11

For each protein pair, a number of zones were identified from the central regions of homologous beta-strands and alpha-helices (FABVL:FABVH, Leu4-Pro8:Leu4-Gly8, Thr19-Gly24:Ser19-Val24; His34-Gln39:Tyr33-Arg38; Ser58-Ser62:Thr68-Asn72; Ser64-Ile70:Asn76-Leu82; Tyr81-Tyr86:Tyr93-Asn98; Val92-Thr97:Val106-Ser111. FABVH:FABCL, Leu4-Gly8:Val8-Pro12; Ser19-Val24:Thr24-Ile29; Tyr33-Val37:Thr38-Lys43; Thr68-Asn72:Gly51-Thr55; Asn76-Leu82:Ala67-Leu73; Tyr93-Asn98:Tyr84-Thr89; Val106-Ser111:Val95-Val99. PLASTO:AZURIN, Asp2-Gly6:Ser4-Gln8; Ala13-Ile21:Gln14-Val22; Glu25-Asn32:Lys27-Ser34; Val40-Asp42:Val49-Ser51; Gly67-Leu74:Lys92-Val99; Gly78-Cys84:Glu106-Cys112; Met92-Asn99:Met121-Lys128. HAHU:LEGH, Pro4-Lys16:Glu5-Glu17; Ala21-Ser35:Ile22-Ile36; Thr38-Thr41:Ala39-Asp42; Ala53-Val70:Pro60-Ala77; Pro77-His89:Asp89-Ser101; Pro95-His112:Asp107-Val124; Pro119-Thr137:Glu131-Lys149. P2PTN:P1EST, Gln15-Asn19:Gln15-Gln19; His23-Ile30:His28-Ile35; Trp34-Ser37:Trp39-Thr42; Gln63-Val72:Gln70-Val79; Met86-Leu90:Ala95-Leu99; Cys116-Gly120:Cys127-Gly131; Lys136-Ala140:Gln146-Leu150; Met160-Ala163:Met172-Ala175; Pro180-Cys183:Pro191-Cys194; Lys186-Trp193:Ala201-Phe208; Gly204-Lys208:Thr221-Arg225). Automatic alignments were tested by considering how many residues within these zones were equivalenced as expected from the structure based alignment. The secondary structure dependent function Q was derived by reference to the first sequence in each pair such that $Q = Q_s$ within each zone and $Q = Q_L$ outside the zone.

element $R_{i,j}$ represents the score for A_i versus B_j.

(iv) $R_{m,n}$ is acted on to generate $S_{m,n}$ where each element $S_{i,j}$ holds the maximum score for a comparison of $A_{i,m}$ with $B_{j,n}$.

(v) Suitable pointers are recorded in (iv) to enable an alignment with the maximum score for $A_{1,m}$ versus $B_{1,n}$ to be generated.

In order to limit the total numbr of gaps introduced (residues in one sequence aligned with blanks), a gap-penalty is subtracted during the process of generating $S_{m,n}$ whenever a gap is introduced. We followed the recommendations of Fitch and Smith (1983) by using a gap-penalty function having both length-dependent and length-independent terms of the form:

$$P = G_1 \times L + G_2$$

where L is the length of gap and G_1 and G_2 are user-defined constants. The program also allows gaps at the ends of the sequences (terminal gaps) to be optionally weighted, a feature not inherent to the standard Needleman and Wunsch (1970) algorithm.

Extension of gap-penalty function to include secondary structural information

The secondary structure dependent function Q modifies the gap penalty function to the form:

$$P_{\mathrm{ss}} = Q \times (G_1 \times L + G_2)$$

where $0 \leq Q \leq 1$ and the suffix ss denotes the inclusion of secondary structural information. In its most general form, Q may be derived from a property of the sequence which exhibits a maximum for regions likely to be involved in secondary structures or other conserved regions, and a minimum for regions likely to be subject to greater variability. One might therefore derive Q from a secondary structure prediction profile (e.g. Garnier *et al.*, 1978; Chou and Fasman, 1977), a smoothed profile based on hydrophobicity (e.g. Levitt, 1976), or a profile of likely buried residues (e.g. Janin, 1979). For the purposes of this study, however, a step function was derived through a knowledge of the tertiary structure of one of the pair of proteins such that $Q = 1.0$ (Q_s) in regions of clear secondary structure, and $Q =$ 0.25 elsewhere (Q_L) (see legend to Table I). The value of 0.25 was found in a preliminary study to be optimal for all five pairs of proteins, by varying Q_L from 0 to 0.75 in steps of 0.25. Note that a value of $Q_s = Q_L = 1.0$ sets $P_{\mathrm{ss}} = P$. Q therefore has the effect of making the formation of a gap more likely in the regions linking secondary structures.

		G_2											
		0	1	2	3	4	5	6	7	8	9	10	
	0	30	29	31	32	32	32	32	29	29	29	29	
	1	30	29	32	32	32	32	29	29	29	29	29	
	2	30	30	32	32	32	29	29	29	29	29	29	
	3	31	30	32	32	32	29	29	29	29	29	29	
	4	31	30	32	32	29	29	29	29	29	29	29	
G_1	5	31	30	32	32	29	29	29	29	29	29	29	[a]
	6	31	30	32	18	18	18	18	18	18	18	18	
	7	31	16	18	18	18	18	18	18	18	18	18	
	8	16	16	18	18	18	18	18	18	18	18	18	
	9	16	16	18	18	18	18	18	18	18	18	18	
	10	16	16	18	18	18	18	18	18	18	18	18	

		G_2											
		0	1	2	3	4	5	6	7	8	9	10	
	0	30	29	29	29	30	36	36	36	36	36	36	
	1	31	31	31	32	33	36	36	36	36	36	36	
	2	31	31	32	33	36	36	36	36	36	36	36	
	3	31	32	33	36	36	36	36	36	36	36	36	
	4	32	33	36	36	36	36	36	36	36	36	36	
G_1	5	30	33	36	36	36	36	36	36	36	36	36	[b]
	6	33	36	36	36	36	36	36	36	36	36	36	
	7	36	36	36	36	36	36	36	36	36	36	36	
	8	36	36	36	36	36	36	36	36	36	36	36	
	9	36	36	36	36	36	36	36	36	36	36	36	
	10	36	36	36	36	36	36	36	36	36	36	36	

Fig. 1. Result of varying constants G_1 and G_2 for a comparison of FABVL with FABVH. **(a)** Standard algorithm (no penalty for end gaps). **(b)** As **(a)** but including secondary structural information ($Q_L = 0.25$, $Q_s = 1.0$). Each score represents the total number of correctly equivalenced residues within the seven specified zones. All numbers are therefore out of a maximum of 41 (see Table I and Figure 2).

```
            |-------A-------|                                           |---------B---------|
Z   S   V   L   T   Q   P   P   S   V   S   G   A   P   G   Q   R   V   T   I   S   C   T   Q   S   S   S   N   I   Q
Z   V   Q   L   E   Q   S   Q   P   G   L   V   R   P   S   Q   T   L   S   L   T   C   T   V   S   Q   S   T   F   S
            |-------A-------|                                           |---------B---------|

                                                    |-------------C-------------|
-   -   -   -   -   -   -   -   -   -   A   Q   N   H   V   K   W   -   -   Y   Q   Q   L   P   Q   T   A   P   K   L
N   D   Y   Y   T   W   V   R   Q   P   P   G   R   Q   L   E   W   I   Q   Y   V   F   Y   H   Q   T   S   D   T   D
        |---------C---------|                                                                                        [a]

                                    |-------D-------|       |-----------E-----------|
L   I   F   H   N   N   A   R   F   S   V   S   K   S   G   S   S   A   T   L   A   I   T   Q   L   Q   A   E   D   E
T   P   L   R   S   R   V   T   M   L   V   N   T   S   K   N   Q   F   S   L   R   L   S   S   V   T   A   A   D   T
                            |-------D-------|               |-----------E-----------|

        |-------------F-------------|                       |---------G---------|
A   D   Y   Y   C   Q   S   -   -   Y   D   R   S   L   R   V   F   G   G   G   T   K   L   T   V   L   R
A   V   Y   Y   C   A   R   N   L   I   A   G   C   I   D   V   W   G   Q   G   S   L   V   T   V   S   S
        |---------F---------|                               |---------G---------|


            |-------A-------|                                           |---------B---------|
Z   S   V   L   T   Q   P   P   S   V   S   G   A   P   G   Q   R   V   T   I   S   C   T   Q   S   S   S   N   I   Q
Z   V   Q   L   E   Q   S   Q   P   G   L   V   R   P   S   Q   T   L   S   L   T   C   T   V   S   Q   S   T   F   S
            |-------A-------|                                           |---------B---------|

            |---------C---------|
A   Q   N   H   V   K   W   Y   Q   Q   L   P   Q   T   A   P   K   L   L   -   -   I   F   -   H   -   -   N   N   A
N   -   D   Y   Y   T   W   V   R   Q   P   P   G   R   G   L   E   W   I   G   Y   V   F   Y   H   Q   T   S   D   T
            |---------C---------|                                                                                    [b]

                                        |-------D-------|       |-----------E-----------|
-   -   -   -   R   -   -   -   -   F   S   V   S   K   S   Q   S   S   A   T   L   A   I   T   Q   L   Q   A   E   D
D   T   P   L   R   S   R   V   T   M   L   V   N   T   S   K   N   Q   F   S   L   R   L   S   S   V   T   A   A   D
                                |-------D-------|               |-----------E-----------|

            |---------F---------|                               |---------G---------|
E   A   D   Y   Y   C   Q   S   Y   D   R   S   -   -   L   R   V   F   G   G   G   T   K   L   T   V   L   R
T   A   V   Y   Y   C   A   R   N   L   I   A   G   C   I   D   V   W   G   Q   G   S   L   V   T   V   S   S
            |---------F---------|                               |---------G---------|
```

Fig. 2. The effect of including secondary structural information into the alignment of FABVL (top sequence) with FABVH (bottom sequence). The seven zones A–G correspond to beta-strands in the immunoglobulin domains (Cohen *et al.*, 1981). **(a)** Standard algorithm (no penalty for end-gaps, $G_1 = 3$, $G_2 = 5$). The alignments of the C and D and F strands are incorrect. On including secondary structural information ($Q_s = 1.0$, $Q_L = 0.25$) **(b)**, the C and F strand alignments are corrected but D is still displaced by two residues.

Alignments carried out

For each pair of protein sequences, runs using the following values for end gap weighting, Q_s and Q_L were performed:

Run	End gaps weighted?	Q_s	Q_L
1	Yes	1	1
2	Yes	1	0.25
3	No	1	1
4	No	1	0.25

For each run, parameters G_1 and G_2 were varied from 0 to 10 in integer steps providing 121 comparisons for analysis. A preliminary study indicated that little more information was gained by exploring G_1/G_2 space to $G_1 = G_2 = 20$. In addition, conventional homology tests for each pair of sequences were carried out by randomizing the sequences 100 times, obtaining an alignment score for each random pair of sequences and expressing the score obtained from the original sequences in standard deviation units from the mean of the randomized comparisons.

Results and discussion

An assessment of the standard Needleman and Wunsch alignment method

The sensitivity of the standard Needleman and Wunsch algorithm to changes in G_1 and G_2 is illustrated by Figure 1a, whilst Figure 2a illustrates one possible alignment. Figure 1a shows the result of 121 alignments carried out on two immunoglobulin variable domains (FABVL, FABVH) with $Q_s = Q_L = 1.0$ and no penalty for end gaps. Even the best alignment has only 32 out of the 41 selected residues correctly equivalenced whilst the worst scores 16/41. Although the values of G_1 and G_2 corresponding to these alignments are distant, the scores of 16 border on regions of 31, 30, 32 and 18/41. A change of one unit in G_1 and/or G_2 can therefore lead to a great reduction in alignment quality. Furthermore, this pattern of scores is not common to all five protein pairs examined indicating that a universal recommendation for gap-penalty values is not possible. The choice of G_1 and G_2 available may be, however, considerably reduced since for every pair of protein sequences studied here, at least one example of the best alignment obtainable for the pair may result with a value of $G_1 = 0$ (e.g. see Figure 1a, values of 32/41). This finding differs from that of Fitch and Smith (1983) who showed that the expected alignment of two short sequences from chicken alpha and beta haemoglobin could only be obtained by including a length dependent gap-term (G_1).

Figure 3 includes the result of runs for $Q_L = 1.0$, with and without end-gap weighting. The standard Needleman and Wunsch algorithm does not explicitly include gap weights for terminal gaps. However, it is natural to weight terminal gaps when aligning homologous protein sequences since generally there are equivalent residues near the ends. Weighting the end gaps leads to an im-

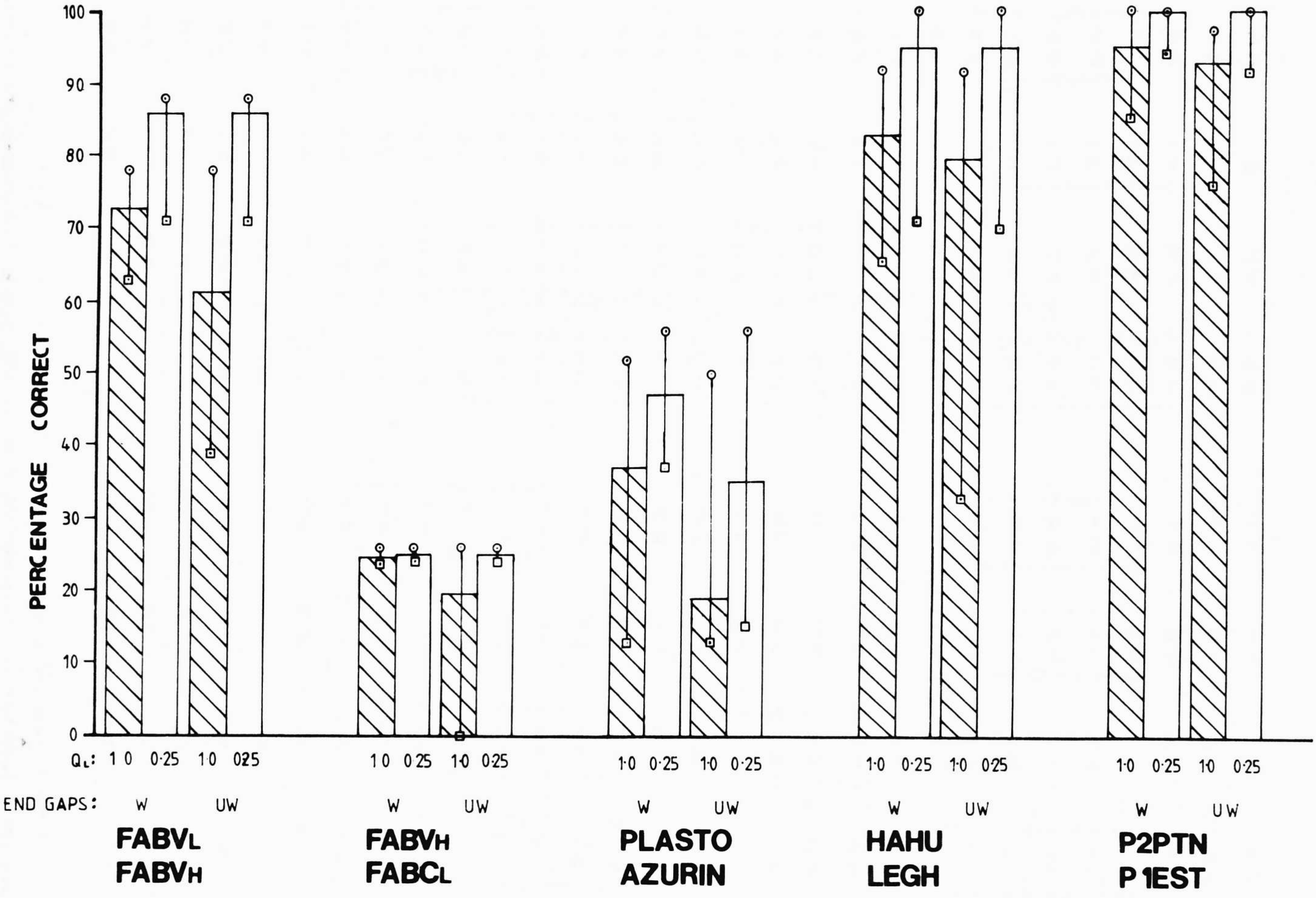

Fig. 3. Summary of alignment accuracy for five pairs of proteins/domains. Blocks represent the mean of 121 alignments for $G_1 = 0-10$, $G_2 = 0-10$ in integer steps. Small circles and small squares indicate the best and worst alignment obtained in each 121 comparisons, respectively. Hatched areas are the results of the standard algorithm ($Q_s = Q_L = 1.0$), plain areas the result of including secondary structural information ($Q_s = 1.0$, $Q_L = 0.25$). W = end-gaps penalized, UW = end-gaps not penalized.

provement in mean alignment quality for all five pairs of sequences (54–62%). In addition, the best alignment obtained for each pair improved for P2PTN versus P1EST and PLASTO versus AZURIN, whilst it stayed the same for FABVL versus FABVH, HAHU versus LEGH and FABVH versus FABCL. In all five pairs, the worst alignment obtained improved (32–51% overall). Furthermore for FABVH versus FABCL alignments in which no residues were correctly equivalenced were removed.

The protein pairs FABVL versus FABVH, HAHU and LEGH and P2PTN versus P1EST score much better overall (90 as against 39% with weighted end gaps) than FABVH versus FABCL and PLASTO versus AZURIN. In order to align the latter two pairs correctly, a long gap must be introduced and/or the algorithm must ignore regions of poor similarity. A long gap will only be allowed by a Needleman and Wunsch type algorithm if the sequence similarity between inserted residues, and the residues bordering the potential gap is very weak. Normally, several shorter gaps will be introduced instead in order to match regions of the insertion with segments of the other sequence. This smearing of the alignment is inherent to the global alignment method, although adoption of an algorithm which weights strings of insertions or deletions lower than multiple isolated single insertions and deletions may overcome this deficiency (Krushkal and Sankoff, 1983).

Given any two sequences to align for which the three-dimensional structures are unknown, one would like to discover how successful the alignment is likely to be. The scores obtained by a conventional test for relatedness (as described in Materials and methods) should follow the same trends as the structure-based test used here including any deficiencies in the treatment of gaps as discussed above. Table II illustrates the results of such homology tests and shows a correlation between the percentage of residues within secondary structures correctly aligned, and the similarity score (which is based on the whole sequence). This

Table II. Result of similarity tests

Comparison	G_1	G_2	Percent correctly aligned	Significance score (SD units)
P2PTN versus P1EST	0	5	98	17.5
HAHU versus LEGHE	0	9	92	6.5
FABVL versus FABVH	0	6	78	6.1
PLSTO versus AZURIN	0	2	50	5.1
FABVH versus FABCL	0	2	26	1.7

'Percent correctly aligned' refers to the percentage of residues aligned within the selected zones as expected from the structure based alignment. Values for G_1 and G_2 were taken to give the best alignment possible (see Figure 3). The significance score was calculated as follows: a value (V) was calculated for the comparison of the two sequences; values for the comparison of 100 randomized sequences of the same composition and length were then obtained and the mean (M) and standard deviation (SD) calculated; the significance score quoted is equal to $(V - M)$/SD.

suggests that an alignment is likely to be good if a score greater than 6 SD can be obtained. However, preliminary trials (data not included), have shown that the significance score obtained for a comparison can vary by ± 1 SD depending on the particular values of gap-penalties used. We suggest that a significance score greater than 7 SD for a comparison means that most of the residues within secondary structures will be aligned correctly. A score of below 5 SD indicates that the alignment is poor.

Inclusion of secondary structural information into the alignment when one X-ray structure is known

The effect of introducing secondary structural information from one sequence of known X-ray structure is illustrated in Figure 1b. Not only has the maximum alignment score increased to 36/41, but the method has converged on this value with increasing G_1 and G_2. This pattern is observed for FABVL versus FABVH, HAHU versus AZURIN and P2PTN versus P1EST, with and without end-gap weighting and indicates that the observed dependence upon G_1 and G_2 is reduced when reliable secondary structural information is included from one sequence. Figure 2b illustrates one alignment for FABVL versus FABVH with 36/41 residues correctly equivalenced, this corresponds to the correct alignment of A, B, C, E, F and G strands. However, the standard method using the same values of G_1 and G_2 misaligned the C-, D- and F-strands giving a score of only 29/41 (Figure 2a).

The result of including secondary structural information into the alignment of all five protein pairs is presented in Figure 3. There is an improvement in mean alignment accuracy on including Q_L = 0.25 from 54 to 68% overall (unweighted end gaps) and 62 to 70% (weighted end gaps). The best alignment obtained for each pair also increases, except for FABVH versus FABCL where the values stay the same. No alignments having zero residues correctly equivalenced were obtained when secondary structural information was included.

The improvements in alignment within secondary structural regions are of value in providing a more reliable automatic starting point for protein model-building studies. They are also valuable when aligning whole families of proteins since less manual intervention is required. Such multiple alignments help to increase understanding of the structural and functional importance of particular residue positions. However, Needleman and Wunsch type algorithms cannot easily be extended to the simultaneous alignment of more than three sequences due to prohibitive memory and computer-time requirements (Murata *et al.*, 1985). Taylor (1986) has developed an algorithm which aligns on the basis of pre-defined 'templates' corresponding to conserved regions (e.g. beta-strands, alpha-helices) in the proteins. This method although not as inherently flexible as the Needleman and Wunsch approach, allows multiple alignments which conserve structural motifs to be carried out in a reasonable time.

Conclusions

In this study we have used sequence alignments based on the superposition of known three-dimensional protein structures as a standard against which to test: (i) The Needleman and Wunsch (1970) global sequence comparison method, and (ii) an extended Needleman and Wunsch algorithm which includes information based on known secondary structures. The following are the main findings.

(i) If the user defined length-dependent and length-independent gap-penalties are varied, a number of different alignments can be obtained, many of which are only partly correct, or sometimes completely different to the alignment expected from the superposition of three-dimensional structures.

(ii) The best alignment using the standard Needleman and Wunsch alignment for each pair of proteins studied may be obtained without including a length dependent gap-penalty.

(iii) Protein sequence comparisons where one sequence has a long insertion, or non-structurally homologous region are not well aligned by this method (38% overall). Treating the sequences in sections bordered by known key residues may help the overall alignment for these protein pairs (Smith *et al.*, 1981).

(iv) The standard deviation used in Needleman and Wunsch homology tests correlates very well with the quality of alignment as indicated by reference to alignments based on structure superposition. For those proteins tested, a score greater than 6 SD suggests that the alignment(s) obtained for the two sequences will be good (>75% correct within secondary structures).

(v) The modification of the gap-penalty function to include information about the positions of secondary structures from one of the proteins being aligned, and thus limit the number of gaps introduced in helix/sheet regions, improves the overall alignment quality and reduces the sensitivity to changes in length-dependent and length-independent gap-penalty constants.

The problems of aligning sequences where there are regions of poor structural homology, or large insertions (e.g. immunogobulin variable versus constant domains), is unlikely to be overcome by global methods using a simple gap-penalty function as described here. The inclusion of secondary structural information from a knowledge of the three-dimensional structure of one of the proteins to be aligned by using a variable gap-penalty can provide, however, improvements in alignment accuracy. In particular, for protein sequences which though distantly related in evolution do not exhibit large insertions relative to each other (e.g. human alpha-haemoglobin versus root nodule leghaemoglobin) a useful improvement in alignment accuracy can be obtained. The inclusion of secondary structural information into the alignment is therefore to be recommended whenever possible.

Acknowledgements

We would like to thank Dr J.Thornton for her comments on the manuscript, Markéta Zvelebil for her alignment of trypsin with elastase and Professor T.L.Blundell for his continued support. This work was supported by the Science and Engineering Research Council.

References

Blundell,T.L., Sibanda,B.L. and Pearl,L.H. (1983) *Nature*, **304**, 273–275.

Boswell,D.R. and McLachlan,A.D. (1984) *Nucleic Acids Res.*, **12**, 457–464.

Browne,W.J., North,A.C.T., Phillips,D.C., Brew,K., Vanaman,T.C. and Hill,R.C. (1969) *J. Mol. Biol.*, **120**, 97–120.

Chothia,C. and Lesk,A.M. (1982) *J. Mol. Biol.*, **160**, 309–323.

Chou,P.Y. and Fasman,G.D. (1977) *J. Mol. Biol.*, **115**, 135–175.

Cohen,F.E., Novotny,J., Sternberg,M.J.E., Campbell,D.G. and Williams,A.F. (1981) *Biochem. J.*, **195**, 31–40.

Dayhoff,M.O. (1972) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation Washington, DC, Vol. 5, pp. 89–110.

Dayhoff,M.O. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation Washington, DC, Vol. 5, suppl. 3., pp. 345–358.

Delbaere,L.T.J., Brayer,G.D. and James,M.N.G. (1979) *Nature*, **279**, 165–168.

Dickerson,R.E., Timkovich,R. and Almassy,R.J. (1976) *J. Mol. Biol.*, **100**, 473–491.

Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) *J. Mol. Evol.*, **21**, 112–125.

Fickett,J.W. (1984) *Nucleic Acids Res.*, **12**, 175–179.

Fitch,W.M. and Smith,T.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1382–1386.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Goad,W.B. and Kanehisa,M.I. (1982) *Nucleic Acids Res.*, **10**, 247–263.

Gotoh,O. (1982) *J. Mol. Biol.*, **162**, 705–708.

Janin,J. (1979) *Nature*, **277**, 491–492.

Krushkal,J.B. and Sankoff,D. (1983) In Sankoff,D. and Krushkal,J.B. (eds), *Time Warps String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*. Addison-Wesley, pp. 296–299.

Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.

Lesk,A.M., Levitt,M. and Chothia,C. (1986) *Prot. Eng.*, **1**, 77–78.

Levitt,M. (1976) *J. Mol. Biol.*, **104**, 59–107.

McLachlan,A.D. (1972) *J. Mol. Biol.*, **64**, 417–437.

Murata,M., Richardson,J.S. and Sussman,J.L. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 3037–3077.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Perutz,M.F., Kendrew,J.C. and Watson,H.C. (1965) *J. Mol. Biol.*, **104**, 59–107.

Sellers,P. (1974) *SIAM J. Appl. Math.*, **26**, 787–793.

Sellers,P. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 3041.

Smith,T.F., Waterman,M.S. and Fitch,W.M. (1981) *J. Mol. Evol.*, **18**, 38–46.

Sutcliffe,J.G., Shinnick,T.M., Green,N. and Lerner,R.A. (1983) *Science*, **219**, 660–666.

Taylor,W.R. (1986) *J. Mol. Biol.*, **188**, 233–258.

Travers,P., Blundell,T.L., Sternberg,M.J.E. and Bodmer,W.F. (1984) *Nature*, **310**, 235–238.

Waterman,M.S., Smith,T.F. and Beyer,W.A. (1976) *Adv. Math.*, **20**, 367–387.

Wilbur,W.J. and Lipman,D.J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 726–730.

Winter,G. and Fersht,A.R. (1984) *Trends Biotech.*, **2**, 115–119.

Received on August 27, 1986