Evaluation and improvements in the automatic alignment of protein sequences

Geoffrey J. Barton¹ and Michael J. E. Sternberg

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

¹To 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 MDM256 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 MDM256 matrix (Feng et al., 1985).

(ii) The protein sequences are defined as A1,m, B1,n where m and n are the number of residues in sequence A, B, respectively.

(iii) A matrix Rm,n is generated by reference to D where each
penalty function to the form:

\[ Q = Q_1 \times L + G_2 \]

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:

\[ P_{ss} = Q \times (G_1 \times L + G_2) \]

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 define \( Q \) 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 \( Q = 1.0 \) (\( Q_i \)) in regions of clear secondary structure, and \( Q = 0.25 \) elsewhere (\( Q_j \)) (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 \( Q_j \) from 0 to 0.75 in steps of 0.25. Note that a value of \( Q_j = Q_L = 1.0 \) sets \( P_{ss} = P \). \( Q \) therefore has the effect of making the formation of a gap more likely in the regions linking secondary structures.

### Table I. Protein pairs used to test alignment methods

<table>
<thead>
<tr>
<th>Protein/domain A (abbreviation) (number of residues)</th>
<th>Protein/domain B (abbreviation) (number of residues)</th>
<th>Source of structural alignment</th>
<th>Number of residues taken for test</th>
<th>Number of zones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunoglobulin light chain variable region (FABVL) (103)</td>
<td>Immunoglobulin heavy chain variable region (FABVH) (117)</td>
<td>Cohen et al. (1981)</td>
<td>41</td>
<td>7</td>
</tr>
<tr>
<td>Immunoglobulin heavy chain variable region (FABVH) (117)</td>
<td>Immunoglobulin light chain constant region (PLABCL) (105)</td>
<td>Cohen et al. (1981)</td>
<td>38</td>
<td>7</td>
</tr>
<tr>
<td>Human alpha haemoglobin (HAHU) (141)</td>
<td>Root nodule leghaemoglobin (LEGHE) (156)</td>
<td>Lesk and Chothia (1980)</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>Tryptsin (P2PTN) (223)</td>
<td>Tosyl-elastase (PIEST) (240)</td>
<td>M.Zvelebil (personal communication)</td>
<td>63</td>
<td>11</td>
</tr>
</tbody>
</table>
Alignments carried out

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

<table>
<thead>
<tr>
<th>Run</th>
<th>End gaps weighted?</th>
<th>$Q_s$</th>
<th>$Q_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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 additional information was gained by exploring the $G_1/G_2$ 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 $Q_s = Q_t = 1.0$ 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 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 $Q_t = 1.0$, 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 ($Q_1 = Q_2 = 1.0$), plain areas the result of including secondary structural information ($Q_1 = 1.0$, $Q_2 = 0.25$). $W =$ end-gaps penalized, UW = end-gaps not penalized.

Table II. Result of similarity tests

<table>
<thead>
<tr>
<th>Comparison</th>
<th>$G_1$</th>
<th>$G_2$</th>
<th>Percent correctly aligned</th>
<th>Significance score (SD units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2PTN versus PIEST</td>
<td>0</td>
<td>5</td>
<td>98</td>
<td>17.5</td>
</tr>
<tr>
<td>HAHU versus LEGH</td>
<td>0</td>
<td>9</td>
<td>92</td>
<td>6.5</td>
</tr>
<tr>
<td>FABVL versus FABVH</td>
<td>0</td>
<td>6</td>
<td>78</td>
<td>6.1</td>
</tr>
<tr>
<td>PLSTO versus AZURIN</td>
<td>0</td>
<td>2</td>
<td>50</td>
<td>5.1</td>
</tr>
<tr>
<td>FABVH versus FABCL</td>
<td>0</td>
<td>2</td>
<td>26</td>
<td>1.7</td>
</tr>
</tbody>
</table>

"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 $(V - M)/SD$.

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
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 FABVL versus FABCBL 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., 1984). 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 structural features 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. immunoglobulin 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, Marketa 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**


G.J. Barton and M.J.E. Sternberg


Received on August 27, 1986