J. Mol. Biol. (1987) **198**, 327–337

A Strategy for the Rapid Multiple Alignment of Protein Sequences

Confidence Levels from Tertiary Structure Comparisons

Geoffrey J. Barton† and Michael J. E. Sternberg

Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College, Malet Street, London WC1E 7HX, U.K.

(Received 28 April 1987, and in revised form 18 July 1987)

An algorithm is presented for the multiple alignment of protein sequences that is both accurate and rapid computationally. The approach is based on the conventional dynamic-programming method of pairwise alignment. Initially, two sequences are aligned, then the third sequence is aligned against the alignment of both sequences one and two. Similarly, the fourth sequence is aligned against one, two and three. This is repeated until all sequences have been aligned. Iteration is then performed to yield a final alignment.

The accuracy of sequence alignment is evaluated from alignment of the secondary structures in a family of proteins. For the globins, the multiple alignment was on average 99% accurate compared to 90% for pairwise comparison of sequences. For the alignment of immunoglobulin constant and variable domains, the use of many sequences yielded an alignment of 63% average accuracy compared to 41% average for individual variable/constant alignments. The multiple alignment algorithm yields an assignment of disulphide connectivity in mammalian serotransferrin that is consistent with crystallographic data, whereas pairwise alignments give an alternative assignment.

1. Introduction

The advent of fast techniques for DNA sequencing has led to a rapid expansion in the number of known protein sequences (currently $\simeq$4000 in the PIR databank: George *et al.*, 1986). Access to these primary structures leads to the alignment of two or more protein sequences that can identify conserved regions of functional and/or structural importance. Furthermore, if homology can be shown with a biochemically or crystallographically well-characterized protein, many properties or aspects of three-dimensional structure may be predicted (e.g. see Browne *et al.*, 1969).

Since the early work of Fitch (1966) and Needleman & Wunsch (1970), techniques for the comparison and alignment of two protein or DNA sequences have been developed for speed (e.g. see Gotoh, 1982; Taylor, 1984; Fickett, 1984; Wilbur & Lipman, 1983), the identification of local similarities (e.g. see Sellers, 1979; Goad & Kanehisa, 1982; Boswell & McLachlan, 1984) and increased sensitivity (Argos, 1987). Although the alignment of three sequences has been used to confirm weak homology between two sequences (e.g. see Doolittle, 1981), the practical limitations of computer memory and central processing unit (CPU) time restrict the rigorous extension of two sequence methods (e.g. see Needleman & Wunsch, 1970) to short sequences (Murata *et al.*, 1985).

The multiple alignment of four or more sequences cannot in practice be solved by a rigorous method, since the number of segment comparisons that must be carried out is of the order of the product of the sequence lengths (many more if gaps are explicitly considered). Thus, multiple alignment algorithms in common with fast pairwise methods (e.g. see Wilbur & Lipman) seek to identify an optimum alignment by considering only a small number of the total possible residue or segment comparisons.

Sankoff and co-workers (Sankoff & Cedergren, 1976) provided a workable multiple alignment algorithm applicable to nucleic acid sequences, which requires the sequences to be linked by a predetermined evolutionary tree. Sobel & Martinez (1986) described an algorithm that bases the alignment of DNA sequences on the identification of common subsequences of a specified minimum length. Waterman (1986) elaborated a similar

† Author to whom reprint requests should be addressed.

0022-2836/87/220327-11 $03.00/0

technique but also allowed for mismatches, whilst Bains (1986) described a related algorithm that works well for some families of closely similar nucleic acid sequences. None of these methods has been applied directly to protein sequences, although Waterman (1986) described how his algorithm could be so applied.

The algorithm of Taylor (1986) allows large numbers of protein sequences to be aligned but for maximum effect requires that three-dimensional structures are known for some of the sequences in order to provide a "seed" alignment. Johnson & Doolittle (1986) described a more general multiple alignment algorithm for protein sequences whereby a small subset of all possible segment comparisons is considered. Although their algorithm can cope with the three-way alignment of long sequences, alignment of more than four sequences is restricted to short proteins, due to excessive CPU demands. Bacon & Anderson (1986) reduced the number of segment comparisons performed by considering the sequences in an arbitrary order and maintain only the best-scoring segments as each new sequence is added. Their algorithm does not explicitly cater for gaps, nor does it produce a complete alignment of the sequences; however, it presents a sensitive technique for the identification of significant short homologies.

This paper reports an algorithm that can generate a multiple alignment including the consideration of gaps for a large number of protein sequences without the need to introduce additional non-sequence information. Performing all pairwise comparisons for the sequences suggests confidence levels for the multiple alignment of particular sequence groups.

2. Procedures

(a) *Needleman & Wunsch algorithm for two sequences*

(1) A matrix of amino acid pair scores, D, is chosen. Throughout this study the MDM_{78} matrix was used (Dayhoff, 1978) with a constant of 8 added to remove all negative terms.

(2) The protein sequences are defined as $A1_m$, $A2_n$, where m and n are the number of residues in sequence A1, A2, respectively.

(3) A matrix $R_{m,n}$ is generated with reference to D, where each element $R_{i,j}$ represents the score for $A1_i$ *versus* $A2_j$.

(4) $R_{m,n}$ is acted on to generate $S_{m,n}$, where each element $S_{i,j}$ holds the maximum score for a comparison of $A1_{i,m}$ with $A2_{j,n}$.

(5) Either suitable pointers are recorded in (4), or a traceback procedure through $R_{m,n}$ is performed to enable an alignment with the maximum score for $A1_m$ *versus* $A2_n$ to be generated.

In order to limit the total number of residues aligned with blanks, a gap-penalty, G, is subtracted during the process of generating $S_{m,n}$ whenever a gap is introduced. In our earlier work (Barton & Sternberg, 1987) we studied the effect upon the accuracy of pairwise alignment of varying both length-dependent and length-independent gap penalties. The results indicated that a length-dependent penalty is unnecessary. Further unpublished results suggest that for the given D matrix, a length-independent penalty in the range 6 to 10 often yields a reasonable alignment. Ideally a range of penalties should be investigated. However, on the basis of these findings and to provide a consistent benchmark we chose the penalty of 8 (which is not necessarily optimal) for use throughout the current study.

(b) *Multiple alignment*

Let the sequences to be aligned by A1 . . . AN, then:

(1) Align A2 with A1 using Needleman & Wunsch algorithm. Let the length of the aligned sequences be denoted L1, 2.

(2) Align A3 with the alignment of A2 and A1 obtained in step (1).

(3) Align sequence A4 with the alignment of A1, A2 and A3 length L1, 3.

(4) Similarly align the sequences A5 to AN.

In general, for the kth sequence align with the previously obtained alignment for A1 through A($k-1$).

When generating the matrix R in order to align the kth sequence a scoring scheme is adopted that includes a contribution from *all* previously aligned sequences, thus highlighting conserved regions in the alignment.

Let i be the position of an aligned residue in sequences A1 . . . A($k-1$) such that $1 \leq i \leq L1$, ($k-1$). Let j be the position of a residue in sequence Ak then:

$$R_{i,j} = \frac{1}{k-1} \sum_{p=1}^{p=k-1} D_{Ap_i, Ak_j}. \qquad (1)$$

For example, if 3 sequences have been aligned so far and we are considering the comparison of the ith position in the alignment (Ala-Val-Leu) with the jth amino acid in the 4th sequence (Ala), then the score ($R_{i,j}$) is given by the score for (Ala *versus* Ala) + (Ala *versus* Val) + (Ala *versus* Leu) $\times$ 1/3 $= (10+8+6)/3 = 8$. The value of D_{Ap_i, Ak_j} when Ap_i is a gap is set to the minimum value for any residue to residue score (0 in this work).

The multiple alignment obtained in (4) may be refined by realigning each sequence with the completed alignment less that sequence. Accordingly, sequence A1 is aligned with the alignment of sequences A2 . . . AN (having first removed any gaps that are common to A2 . . . AN). A2 is then realigned with the alignment of A1, A3 . . . AN. This process is repeated until AN has been realigned with A1 . . . A(N-1). The complete cycle may then be repeated.

(c) *Criteria for assessing the quality of alignment*

The unique conformation that a globular protein adopts and the resultant disposition of key catalytic or binding residues ultimately determines its biological activity. Therefore, when 2 or more protein sequences are aligned it is of crucial importance that residues defining a common tertiary fold are correctly equivalenced. Although the general fold may be conserved, there can be considerable variation in 3-dimensional structure within a protein family. In particular, the presence of insertions or deletions makes it impossible to assign structurally equivalent residues over the full length and common to *all* members of a family. Even when insertions and deletions are absent, it can be difficult to justify a particular structural alignment especially in the loop regions that connect elements of secondary structure. In the light of these observations we use those regions that are common and also found in the

core β-strands or α-helices of the proteins as test zones. An automatic alignment method should at least be able to align these zones correctly, although sequence alignments based on structure may also be justified outside these regions.

We define the *accuracy* of an alignment of 2 sequences as the percentage of residues within these zones that are aligned in the same way as in the reference alignment.

(d) *Order of alignment*

The order in which the sequences are added may be expected to have an effect on the final multiple alignment. However, for N sequences there are $N!$ alternative orders, so it is important to have a systematic approach to selecting the single order. Our previous findings suggested that a pairwise sequence comparison that gives a significance score of $>6{\cdot}0$ S.D. may be aligned to $>75\%$ accuracy within regions of secondary structure (Barton & Sternberg, 1987). Further results presented in this paper for 49 pairwise comparisons of immunoglobulin and globin sequences (Fig. 1) support this observation and indicate that the alignment accuracy is correlated with the significance score. Accordingly, our strategy for determining the alignment order first identifies the pair of sequences that have the highest pairwise significance score. Having established A1 and A2, A3 is identified as the sequence having the highest significance score with either A1 or A2. Similarly, A4 is the sequence that exhibits the highest significance score with A1, A2 or A3. This procedure is continued until all sequences in the group have been entered in the order.

(e) *Reducing calculation time*

Before applying the ordering algorithm to a group of sequences it is necessary to perform all pairwise comparisons for the group. The calculation of significance scores for all pairwise alignments of N sequences is an expensive procedure since if M randomizations per pair are performed then $N \times (N-1) \times M/2$ comparisons must be carried out. If no pairwise comparisons have previously been made then it is necessary to determine the cheapest (in CPU time) approach to determining an order.

Feng *et al.* (1985) considered how many randomizations need to be performed on a pair of sequences before consistent results are obtained and suggested on the basis of 4 pairs of sequences that as few as 25 could produce a genuinely reflective score. We have repeated this analysis with a larger data set by carrying out all pairwise comparisons for 9 members of the immunoglobulin superfamily and 6 serine proteinase sequences (47 pairs in all), using from 10 to 100 randomizations in steps of 10. The results indicate that instabilities in significance score do not damp out until at least 60 randomizations have been performed. It would seem impractical therefore to use this approach routinely for establishing an alignment order when more than a small number of sequences are involved.

Doolittle (1981) has demonstrated the usefulness of scoring systems based upon a single alignment score and not involving randomization of the sequences. Fig. 2 illustrates the relationship between one such scheme, the match score divided by the length of the shortest sequence (NASs) and the significance score for 2 groups of

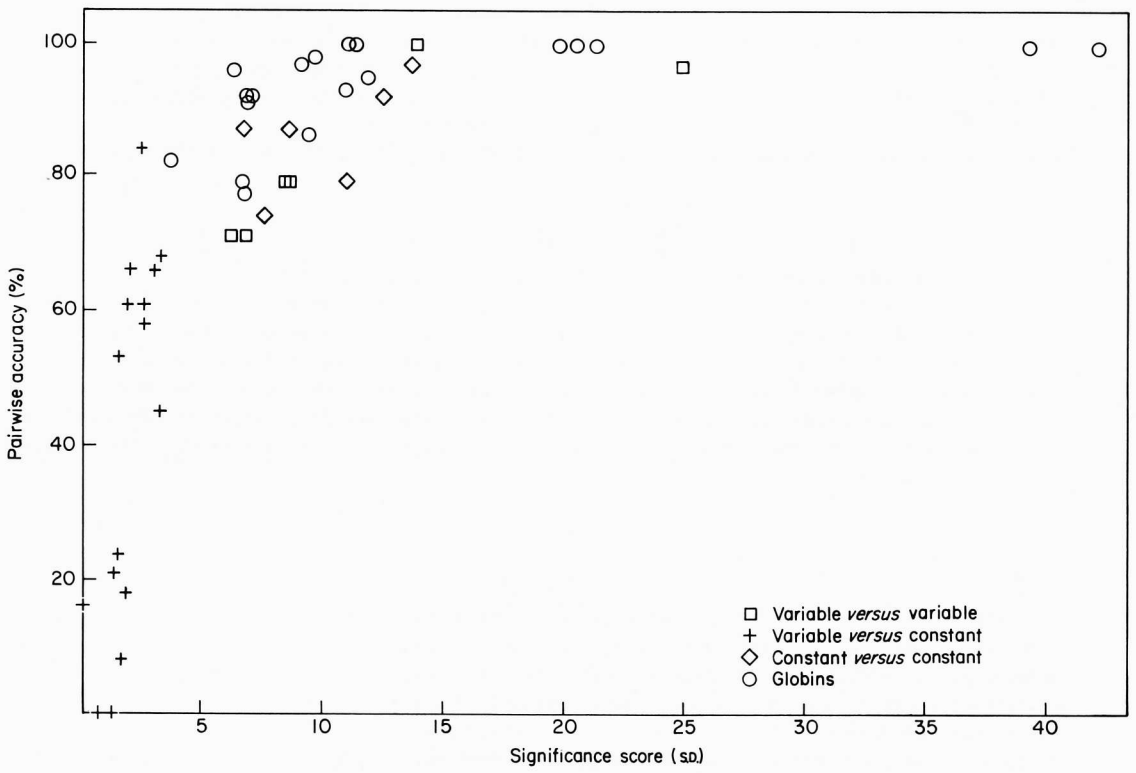


Figure 1. The relationship between alignment accuracy as measured by reference to alignments obtained from 3-dimensional structure superposition and the significance score for 21 pairwise alignments of 7 globin sequences and 28 alignments of 8 immunoglobulin domains (Variable refers to immunoglobulin variable domains, Constant to immunoglobulin constant domains).

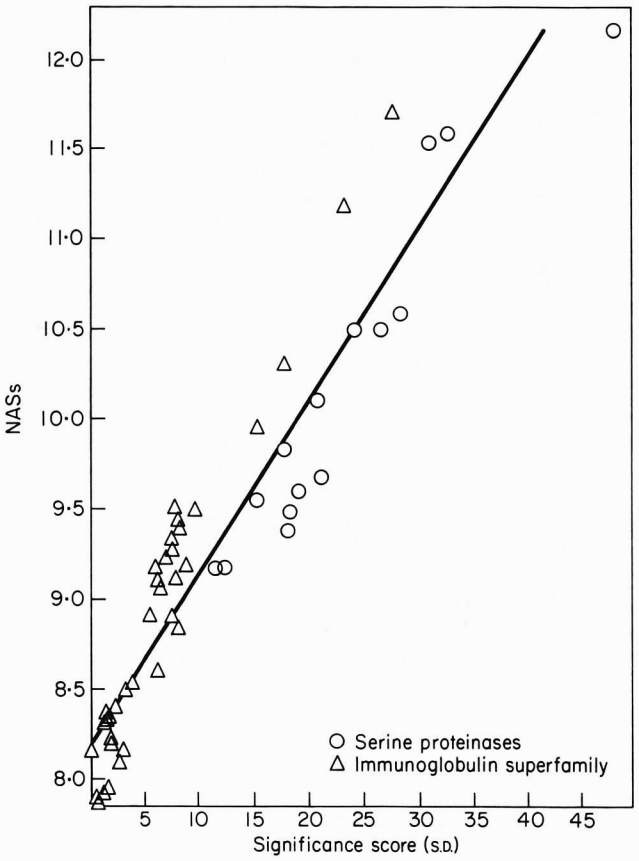


Figure 2. Relationship between the normalized alignment score (NASs) calculated from the match score divided by the length of the shortest sequence and the significance score for each of 47 pairwise comparisons within the serine proteinases and immunoglobulin superfamily. The data are correlated at 0·957. NASa (match score divided by the number of residues not aligned with a gap) for the same sequences gave a correlation value of 0·958.

sequences that contain some very closely related members and many with tenuous similarities. Although these data are not sufficiently representative of proteins *in general* to allow a conclusion to be drawn on the quantitative relationship between NASs and significance, the qualitative relationship is clear. For groups of proteins where randomization procedures would be too expensive, NASs values or the slightly more expensive NASa (match score divided by the number of residues not aligned with gaps), may be used to generate systematically a rational order for multiple alignment.

(f) *Test sequences and reference alignments*

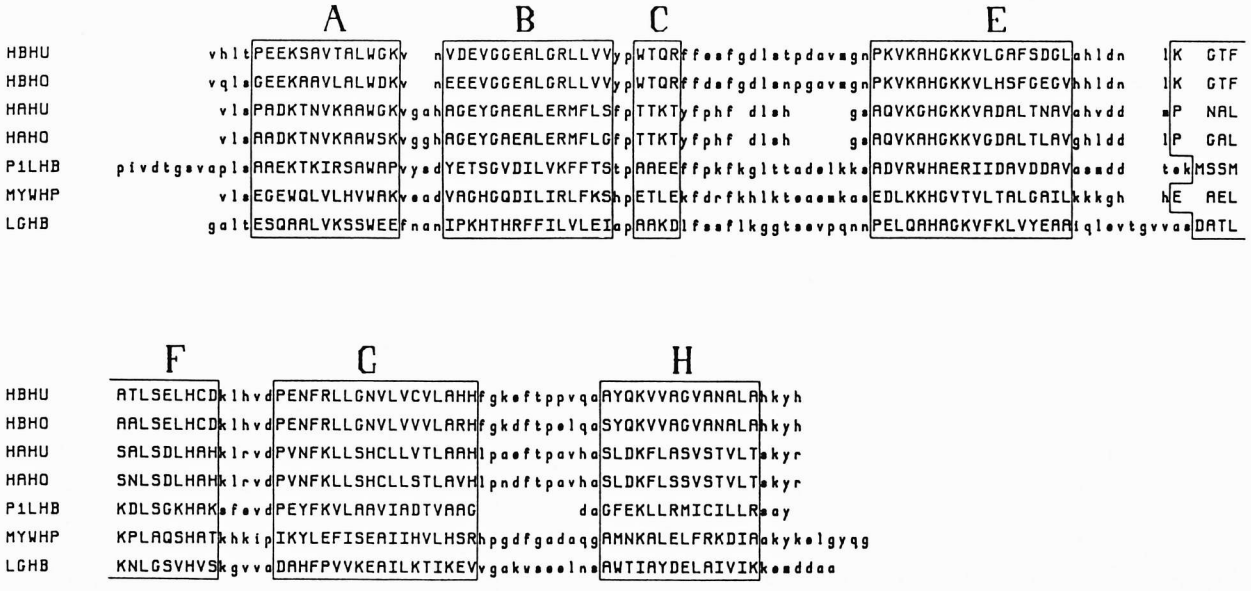
(i) *Globins*

Globins belong to the α/α class of proteins and have a common fold that is highly conserved in proteins from organisms distantly related in evolution. The sequences, however, show considerable variation and provide an interesting test for the alignment method. Seven globin sequences and their reference alignment were taken from Lesk & Chothia (1980) without modification (human haemoglobin α-strand (HAHU), human haemoglobin β-strand (HBHU), horse haemoglobin α-strand (HAHO), horse haemoglobin β-strand (HBHO), sperm whale myoglobin (MYWHP), sea lamprey cyanohaemoglobin (P1LHB) and root nodule leghaemoglobin (LGHB)). Seven zones totalling 95 residues and corresponding to the A, B, C, E, F, G and H α-helices were defined for each sequence as illustrated in Fig. 3(a).

(ii) *Immunoglobulin domains*

The immunoglobulin domains belong to the β/β class of proteins and have been studied in detail in terms of both sequence and tertiary structure (e.g. see Amzel & Poljak, 1976; Beale & Feinstein, 1976; Lesk & Chothia, 1982). Although the overall fold of the domains is conserved, there is considerable sequence variation, particularly between the variable and constant domains. These sequences thus provide a particularly stringent test for an alignment method.

Eight domains were selected (Brookhaven data bank



(a)

Fig. 3.

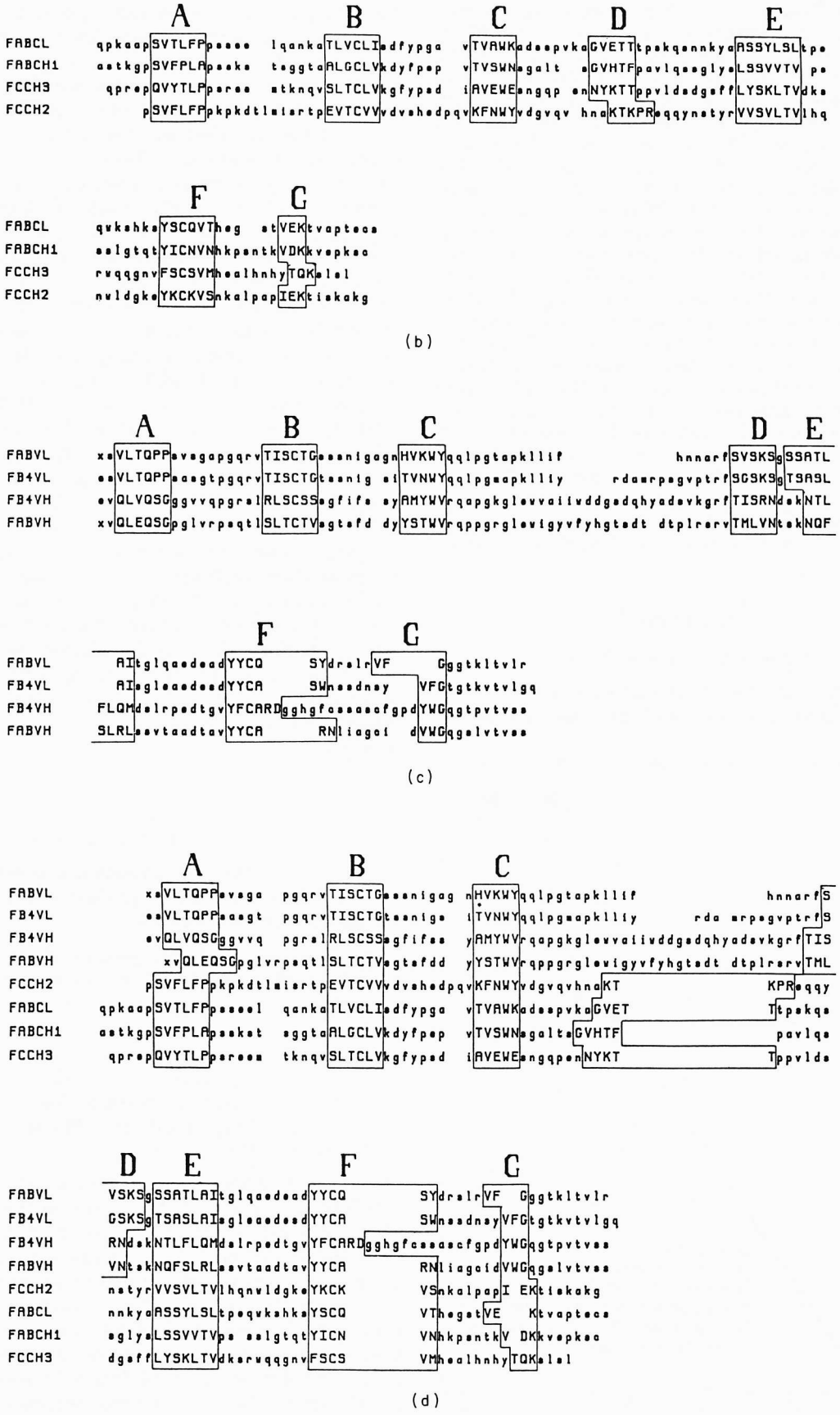


Figure 3. Test multiple alignments. Regions of the sequences written in capitals and boxed correspond to test zones selected from homologous secondary structures. (a) Alignment (1) 7 globin sequences: A, B, C, E, F, G, H are α-helixes; (b) alignment (2), 4 immunoglobulin constant domains; (c) alignment (3), 4 immunoglobulin variable domains; (d) alignment (4), 8 immunoglobulin domains including variable and constant. In (b) to (d), A, B, C, D, E, F and G are β-strands.

codes). Four from 3FAB: (1) light chain constant region Cλ (FABCL); (2) light chain variable region Vλ (FABVL); (3) heavy chain constant region 1 Cγ1 (FABCH1); (4) heavy chain variable region Vγ (FABVH). Three from 1FC1: (1) heavy chain constant region 2 Cγ2 (FCCH2); (2) heavy chain constant region 3 Cγ3 (FCCH3). Two from 1FB4: (1) light chain variable region Vλ (FB4VL); (2) heavy chain variable region Vγ (FB4VH). The reference alignment for the 8 domains was taken principally from Cohen *et al.* (1981) and Lesk & Chothia (1982). However, the most recent version of the co-ordinates deposited in the Brookhaven data bank (Bernstein *et al.*, 1977) for 3FAB shows a modified sequence in the B–C loop and C strand of FABVL. From a consideration of hydrogen bonding, and least-squares fitting of the domains, the alignment was revised in this area to take account of these changes. In addition it should be noted that the sequence of FB4VL used here (taken from the Brookhaven data bank structure 1FB4) differs slightly from the earlier version used by Lesk & Chothia: threonine (T) has been substituted for serine at residue numbers 23 and 33, whilst alanine (A) substitutes for glycine (G) at 75. Seven test zones comprising 38 residue positions in total and corresponding to the 7 homologous β-strands A to G were defined as illustrated in Fig. 3(b) to (d)).

3. Results

(a) *Pairwise comparisons*

For each of the 28 unique pairwise comparisons for the immunoglobulins and 21 comparisons for the globins, the percentage agreement with the reference alignment was calculated. In addition, a conventional test for significance was carried out by randomizing each pair of sequences 100 times and calculating the mean (m) and standard deviation (S.D.) of the distribution. The significance score is quoted as $(V-m)$/S.D., where V is the alignment score for the two native sequences.

Figure 1 illustrates the result of these comparisons. Alignments that score >15·0 S.D. (7 examples) give at or near 100% agreement with the reference alignment. Those scoring between 5·0 and 15·0 S.D. (25 examples) give better than 70% agreement with the reference alignment, whilst scores below 5·0 S.D. (17 examples) show a sharp rise in alignment accuracy correlated with significance score and ranging from 0% (0·57 S.D.; FABCH1 *versus* FB4VH) to 84% (2·4 S.D.; FABVL *versus* FABCL). Above 5·0 S.D. there are no poor alignments; however, in the lower S.D. range small changes in observed significance score can indicate a considerable difference in alignment accuracy.

When aligning two sequences it is useful to bear in mind these findings since they can suggest the likely quality of the alignment obtained. As an approximate guide we consider an S.D. score above 5·0 to indicate a "good" alignment, with the confidence in alignment increasing with alignment score. An alignment giving a score below 5·0 S.D. we regard with a caution that becomes more stringent as the score decreases.

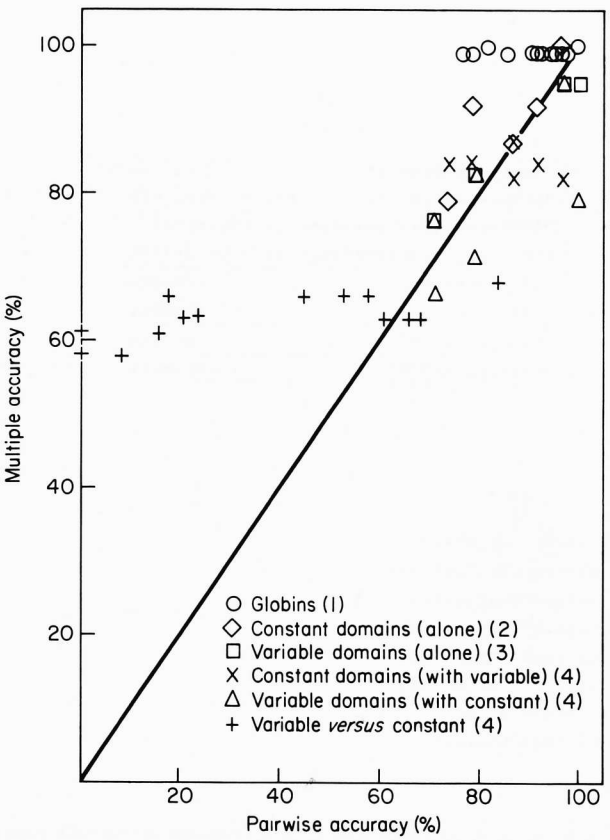


Figure 4. Comparison of accuracy for multiple alignments (1) to (4) with the same sequences aligned pairwise. Points above the diagonal line indicate an improvement in accuracy on multiple alignment.

(b) *Test multiple alignments: comparison with pairwise*

In each of four test alignments performed, the sequences were ordered by similarity on the basis of 100 randomizations as shown.

(1) The seven globin sequences HBHU, HBHO, HAHU, HAHO, MYWHP, P1LHB, LGHB.

(2) The four constant domains FABCL, FABCH1, FCCH3, FCCH2.

(3) The four variable immunoglobulin domains FABVL, FB4VL, FB4VH, FABVH.

(4) The eight immunoglobulin domains FABVL, FB4VL, FB4VH, FABVH, FCCH2, FABCL, FABCH1, FCCH3.

Figure 4 shows the accuracy of alignment obtained for pairs of sequences within multiple alignments (1) to (4) (Fig. 3, (a) to (d)) compared with the same sequences aligned pairwise. Points above the diagonal line represent an improvement in alignment when the multiple alignment algorithm is applied. Multiple alignment of the globins (1) results in an improvement from 90% to 99% overall with the most dramatic improvements for the more distantly related sequences P1LHB, MYWHB and LGHB and the largest change occurring for the comparison of LGHB with HBHO (77 to 99%). The final alignment shown in Figure 3(a) has 94 out of 95 defined residues correctly aligned for all seven sequences. Furthermore, the D

α-helix, which is only present in HBHU, HBHO, MYWHP and P1LHB, is also correctly aligned. The single error occurs at the beginning of the F helix and may in part be caused by the choice of score used for a residue *versus* a pre-introduced gap (discussed further below). Analysis of the globin significance scores by the technique of single linkage cluster analysis (Dayhoff *et al.*, 1972; Sokol & Sneath, 1973) in the light of these results suggests that sequences that cluster above 5·0 S.D. align at least as well by the multiple algorithm as they do pairwise. The immunoglobulin constant domains (2) that cluster at 8·5 S.D. and align to 86% accuracy pairwise and 90% by the multiple algorithm with 30/38 positions correctly equivalenced across the complete four-sequence alignment (Fig. 3(b)) and the variable domains, which also cluster at 8·5 S.D. with mean accuracies of 83% (pairwise), 84% (multiple) and 29/38 positions correctly aligned across all four sequences (Fig. 3(c)), lend further support to this hypothesis.

When all eight immunoglobulins are aligned (4), the accuracies of variable *versus* variable and constant *versus* constant domain comparisons marginally deteriorate ($84 \pm 10\%$ to $81 \pm 7\%$); however, there is a striking improvement in the alignment accuracy for variable *versus* constant domains from a mean value of $41 \pm 28\%$ to $63 \pm 3\%$. This is most noticeable for the alignment of FB4VH *versus* FABCH1 and FB4VH *versus* FCCH3, which were completely misaligned when compared pairwise. FABVH *versus* FABCL, which could only be aligned pairwise to $<30\%$ accuracy even after optimizing the gap penalty (Barton & Sternberg, 1987), are aligned by the multiple algorithm to an accuracy of 63%. Thus a considerable improvement in the alignment of distantly related sequences is obtained at a slight loss in accuracy for more closely related sequences.

The eight-sequence alignment (4) has 22 out of 38 residues (58%) correctly aligned in all sequences (Fig. 3(d)). This may seem to be a rather poor score; however, it must be noted that it is difficult to show any significant homology between immunoglobulin constant and variable domains (Edelman, 1970; Moore & Goodman, 1977), and pairwise methods can fail even to align the two cystine residues correctly. Closer examination of the alignment shows the B, C and E strands to be totally correct along with the first four out of six residues in the F strand, the last two residues of this strand are only misaligned in relation to FB4VH, principally as a result of a long insertion in that sequence. The A and G strands are misaligned due to convincing sequence similarities that do not coincide with the structural alignment. Whilst the alignment of the D strand, which shows considerable sequence variation, is confused by the presence of a long insertion in the variable domains between C and D (the C′ strand).

Although the overall trend indicated by Figure 4 is one of improvement on multiple alignment, 13 out of the 61 points shown indicate a deterioration in accuracy. This drop is not surprising for seven of the 13 examples, since these are for variable *versus* variable and constant *versus* constant domain alignments *within* the complete constant and variable domain alignment (4). The inclusion of information from less-similar sequences can clearly have a detrimental effect on the alignment of closely similar sequences. This effect is also noticed for FABVL *versus* FB4VL (97% pairwise, 94% multiple (3)) and FABVH *versus* FB4VH (100% pairwise, 95% multiple (3)) where the errors in the F and G strands (Fig. 3(c)) are the result of convincing alternative alignments that would not be available to a pairwise comparison. The remaining reductions in accuracy are for four comparisons of variable and constant domains, and are difficult to rationalize; however, these are small errors to accept in the light of the considerable improvements in overall alignment accuracy for these domains.

In general, the multiple alignments reflect the type of alignment that might be produced by hand if only sequence information was available. The most obvious errors in alignment occur within the severe test presented by alignment (4) (Fig. 3(d)), in particular the segment QLEQSGpg in strand A of FABVH would certainly be shifted two residues to the left by a human expert at the expense of introducing an additional gap. In other parts of the alignment alternative arrangements may be proposed; however, these become more difficult to justify when the positions of the secondary structures are unknown.

(c) *Effect of iteration and order of alignment*

Table 1 illustrates the effect on alignment accuracy of performing up to four iterations. With the exception of alignment (3) there is little change in alignment accuracy for iterations beyond two. The multiple alignments described above are therefore the result of adding the sequences in order of similarity, then performing two iterations. The improvements for the variable domain alignment for three to four iterations (3) (Table 1) are due to rearrangements of residues at the end of the F and beginning of the G strands. However, although the residues are similarly arranged in this region of alignment (4) after two iterations, further iteration does not improve the alignment. This difference in behaviour is due to the score of zero given to the matching of a residue in the A*kth* sequence with a previously generated gap. Application of equation (1) results in the weighting down of gaps that occur simultaneously in more than one sequence (desirable to highlight conservation). However, once a gap has been established in the initial alignment, the score for matching with that gap may be too low to ever allow a residue to align at that position. This can lead to the formation of columns of aligned residues (e.g. the beginning of the F helix in alignment (1)), which iteration cannot overcome. Alternative scoring schemes where

Table 1

Effect of iteration on the mean alignment accuracy of pairs of sequences within the multiple alignments

	Iteration (mean accuracy ± 1 S.D.)				
Alignment†	0	1	2	3	4
(1)	98·9±1	99·5±0·5	99·5±0·5	99·5±0·5	99·5±0·5
(2)	88·2±6	88·2±6	89·5±7	89·5±7	89·5±7
(3)	83·3±9	83·3±9	84·2±9	85·5±9	87·3±7
(4)	62·6±14	70·5±12	70·8±10	70·5±10	71·6±10

† (1) 7 globins; (2) 4 immunoglobulin constant domains; (3) 4 immunoglobulin variable domains; (4) 8 immunoglobulin domains (4 constant, 4 variable).

residue *versus* pre-existing gap is given a score >0 during iteration can help to alleviate this problem but may also introduce errors at other points in the alignment (results not shown). Averaging also has an effect on the non-gap regions of the alignment, so that the effect of iteration becomes less apparent as the number of sequences is increased. This property can be an advantage for very large alignments, since a single alignment pass with no iterations may be sufficient to yield a final alignment.

The alignment of sequences in one specific order is the main route by which this algorithm reduces the number of comparisons to manageable proportions. To investigate the importance of order on the final result, ten unique alternative orders were generated for alignments (1) and (4). The mean alignment accuracy for the ten globin orders was little different, at 98·7%, from that obtained for alignments ordered by S.D. score (99·5%) or NASa (98·9%). However, the ten immunoglobulin orders gave a mean value of 57·6% compared to 70·8% for the alignment ordered by S.D. score. This result is hardly surprising, since many of the orders start with the poor alignment of a variable and constant domain that is not subsequently corrected. Indeed, the order that performed least well of the ten was one in which variable and constant domains alternated (FABVH, FABCH1, FB4VH, FCCH3, FABVL, FABCL, FB4VL, FCCH2). This gave only 38% accuracy before iteration with no correctly aligned positions across all eight sequences. After two iterations, the accuracy had improved to 46% and 4/38 residues (the first four of the F-strand were in complete alignment). The order defined by NASa scores (FABVL, FB4VL, FB4VH, FABVH, FABCH1, FABCL, FCCH3, FCCH2) is very similar to the S.D. score order and gave an alignment accuracy of 67·2%.

Although it might appear from the variable and constant domain example that a bad initial alignment will always lead to a generally poor overall alignment, this is not necessarily true. For example, it is possible for a multiple alignment of 20 sequences to be poor for the first ten, yet good for the second ten provided that the second ten are closely related. This feature is another consequence of the scoring scheme shown by equation (1), since one good comparison can be identified against a background of average scores.

4. Applications and Conclusions

One advantage of the algorithm described here is its speed. For example, the complete seven-sequence globin alignment (Fig. 3(a)) (2 iterations) required 65 seconds, whilst the same operation for the eight-sequence immunoglobulin alignment (Fig. 3(d)) took 50 seconds CPU time on a VAX 11/750. Pairwise comparisons to establish an order without randomization required 44 seconds for the globins and 34 seconds for the immunoglobulins. For comparison, the Johnson & Doolittle (1986) algorithm requires 60 minutes CPU time to align five sequences of less than 50 residues in length.

The determination of an alignment order *via* pairwise comparisons is the most time-consuming part of the procedure, particularly if randomizations are performed. However, such an analysis is often carried out as part of the characterization of a newly determined sequence and would not need to be repeated to permit multiple alignment. If the time required to perform all pairwise comparisons is prohibitive then the seven-globin alignment suggests that an arbitrary order may perform almost as well for an alignment of similar sequences. Each iterative pass of our algorithm requires time approximately proportional to NM^2, where N is the number of sequences and M is the length of the sequences when aligned. Although it is expensive to align long sequences, the task is not impossible: however, the longest alignment that may be produced is limited by the need to store one array of dimensions $M \times M$.

Aligning large numbers of medium-length protein sequences (150 to 300 residues) is therefore a matter of routine. For example, the alignment of 128 globin sequences, including haemoglobin-α and β, myoglobin and leghaemoglobin from a wide range of species, required 25 minutes of CPU time including two iterations (Fig. 5). The sequences used in the previous section to test the method are indicated on the Figure together with the test

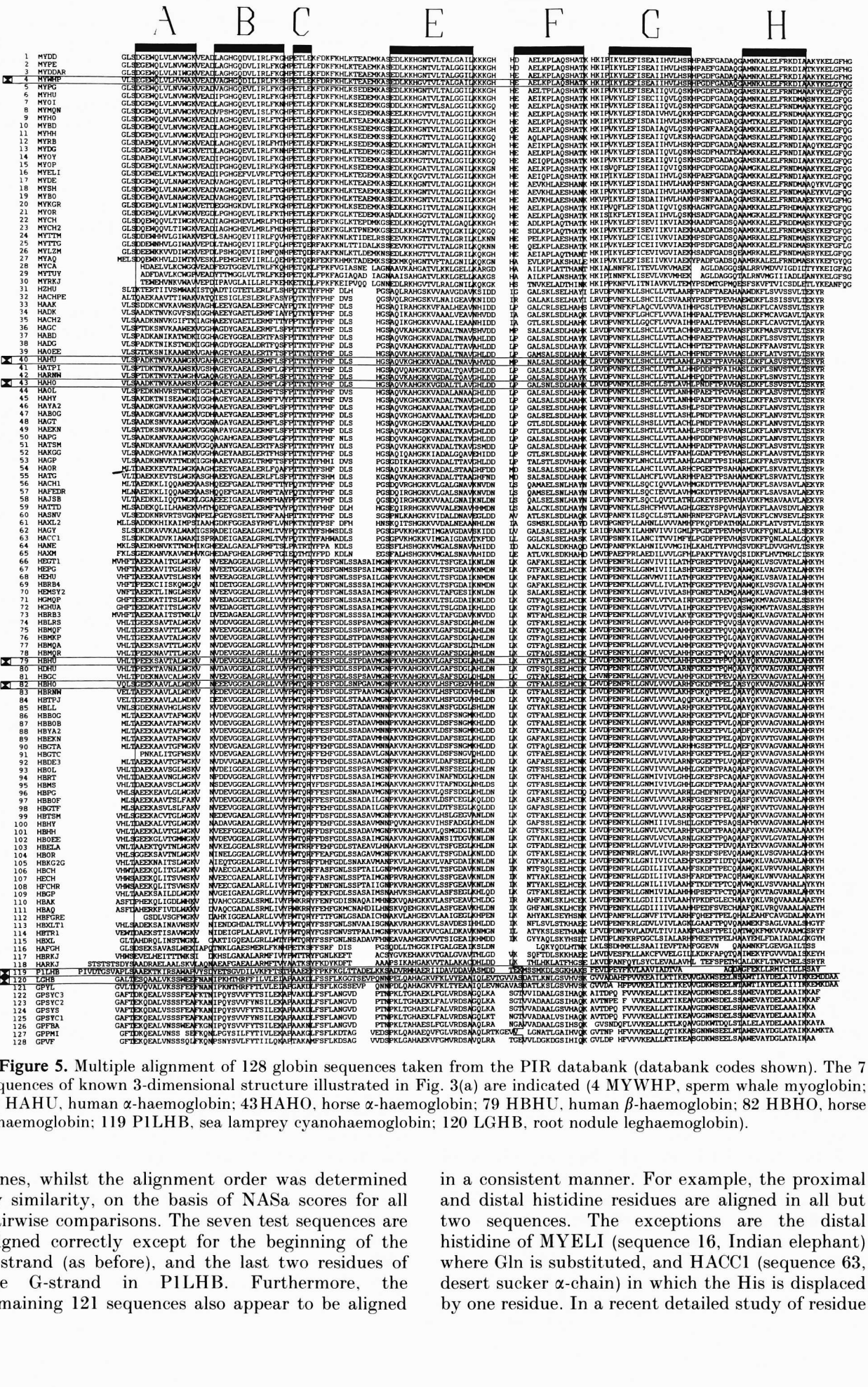


Figure 5. Multiple alignment of 128 globin sequences taken from the PIR databank (databank codes shown). The 7 sequences of known 3-dimensional structure illustrated in Fig. 3(a) are indicated (4 MYWHP, sperm whale myoglobin; 40 HAHU, human α-haemoglobin; 43HAHO, horse α-haemoglobin; 79 HBHU, human β-haemoglobin; 82 HBHO, horse β-haemoglobin; 119 P1LHB, sea lamprey cyanohaemoglobin; 120 LGHB, root nodule leghaemoglobin).

zones, whilst the alignment order was determined by similarity, on the basis of NASa scores for all pairwise comparisons. The seven test sequences are aligned correctly except for the beginning of the F-strand (as before), and the last two residues of the G-strand in P1LHB. Furthermore, the remaining 121 sequences also appear to be aligned in a consistent manner. For example, the proximal and distal histidine residues are aligned in all but two sequences. The exceptions are the distal histidine of MYELI (sequence 16, Indian elephant) where Gln is substituted, and HACC1 (sequence 63, desert sucker α-chain) in which the His is displaced by one residue. In a recent detailed study of residue

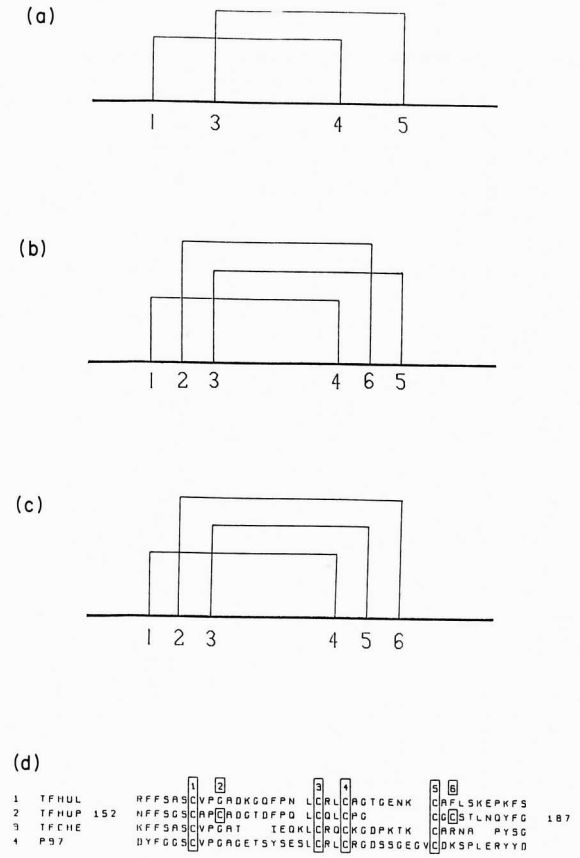


Figure 6. Connectivity of disulphide bridges in the transferrins. Bridges 1 to 4, 3 to 5 and 2 to 6 correspond to bridges 4, 5 and 11 in the nomenclature of Williams (1982). TFHUL, human lactotransferrin; TFHUP, human serotransferrin; TFCHE, chicken ovotransferrin; P97, human melanoma antigen. (a) Connectivity for TFCHE; (b) connectivity suggested by Metz-Boutigue *et al.* (1984), and Rose *et al.* (1986); (c) connectivity suggested by the rabbit serotransferrin crystal structure; and (d) multiple alignment that supports the connectivity in (c).

conservation in the globins based upon pairwise alignment with manual corrections (Bashford, Chothia & Lesk, personal communication), the second example was identified as a sequencing error, since the order of residues as shown (H G K K) would lead to a shift in the E helix by a quarter turn; the sequence should be (K H G K).

We stress that the alignment shown in Figure 5 was produced entirely automatically, without any manual intervention or pre-alignment of key regions. To our knowledge there is no other algorithm that will permit an objective global alignment of so many protein sequences and to such a high level of accuracy.

Application of the multiple alignment algorithm has proved valuable during the crystallographic determination of a mammalian serotransferrin (rabbit) in this laboratory (Gorinsky *et al.*, 1979; P. Lindley *et al.*, personal communication). Human transferrin (TFHUP) shows strong sequence homology with chicken transferrin (TFCHE), human lactotransferrin (TFHUL) and human melanoma antigen (P97). TFCHE has been studied biochemically and the connectivity of the disulphide links determined (Williams *et al.*, 1982). Figure 6(a) illustrates the topology of disulphides 1 to 4 and 3 to 5, which are unambiguous in the alignment of TFCHE, TFHUL and P97. However, TFHUP in common with rabbit serotransferrin has an additional disulphide in this region, and the topology as indicated by the published alignments for TFHUP with TFCHE and TFHUL (Metz-Boutigue *et al.*, 1984) as well as TFHUP with P97 (Rose *et al.*, 1986) is shown in Figure 6(b). However, inspection of the electron density map for serotransferrin at 3·3 Å (1 Å = 0·1 nm) resolution suggested that this topology could not be accommodated, but that the arrangement shown in Figure 6(c) was more likely. To provide independent evidence, the multiple alignment algorithm was applied to the four sequences. Pairwise comparisons showed that the sequences clustered at 41 S.D., suggesting a high level of confidence in the alignment, whilst multiple alignment of the complete sequences ($\simeq$800 residues) performed with two iterations lead to the alignment partly shown in Figure 6(d). This alignment supports the crystallographic interpretation of topology shown in Figure 6(c).

The algorithm presented in this paper represents a practical solution to the problem of automatically aligning more than two protein sequences when only sequence information is available. It appears most valuable when there are weaker similarities (e.g. immunoglobulin constant *versus* variable domains). However, the great sensitivity of alignment accuracy below 5·0 S.D. (Fig. 1) to changes in significance score and the sensitivity to alternative alignment orders make the level of success in an alignment that includes weakly similar sequences difficult to predict.

The test systems presented here suggest that when a group of sequences cluster at >5·0 S.D. multiple alignment by our algorithm will provide a convenient representation, which is likely to be >70% correct within secondary structures and more accurate than individual pairwise alignments.

Our results suggest the overall accuracy of alignment that might be expected for a particular significance score; however, the problem still remains of identifying which *regions* of the alignment are correct. Argos (1987) has described a sensitive procedure for identifying significant local homologies between two sequences or pre-aligned families of sequences and shown that these often correspond to regions of similarity in three-dimensional structure. Thus, when there are two or more distinct clusters of sequences to be aligned, the Argos method may be applied to alignments obtained automatically by our algorithm and provide an indication of which regions are correctly equivalenced.

Although the incorporation of properties in addition to the Dayhoff matrix can improve the sensitivity of sequence comparison methods by reducing background noise (Argos, 1987), without

a more complete understanding of the relationship between sequence and three-dimensional structure it is difficult to envisage a scoring scheme that would, for example, lead to the correct alignment of the A β-strands of immunoglobulin variable and constant domains (Fig. 3(d)). When sequence similarity is weak, sequence alignment becomes an exercise in structure prediction and correct alignment is constrained by the fact that the code relating amino acid sequence to three-dimensional structure is degenerate.

Alignments of clearly similar sequences generated by our algorithm have been used as the basis of an improved secondary structure and active site prediction algorithm (Zvelebil *et al.*, 1987) and also to align four different strains of human immunodeficiency virus (HIV) *env* (800 residues), *gag* (500 residues) and *pol* (1000 residues) polyproteins with the aim of predicting potential T and B-lymphocyte-defined epitopes (Coates *et al.*, 1987; Sternberg *et al.*, 1987).

It has been suggested that a few key residues can be sufficient to define a tertiary fold (e.g. see Wierenga *et al.*, 1986). The multiple alignment algorithm provides a useful tool for identifying such patterns from closely related sequences. We are currently developing and calibrating techniques for identifying these patterns, and rapidly scanning the protein sequence databank to identify proteins of potentially similar tertiary folds.

We thank Professor T. Blundell for his continued support, M. Zvelebil and I. Haneef for helpful discussions, and R. Garrett, B. Gorinsky and P. Lindley for presenting the transferrin problem. This work was funded by the Science and Engineering Research Council.

References

Amzel, L. M. & Poljak, R. (1979). *Annu. Rev. Biochem.* **48**, 961–997.

Argos, P. (1987). *J. Mol. Biol.* **193**, 385–396.

Bacon, D. J. & Anderson, W. F. (1986). *J. Mol. Biol.* **191**, 153–161.

Bains, W. (1986). *Nucl. Acids Res.* **14**, 159–177.

Barton, G. J. & Sternberg, M. J. E. (1987). *Protein Eng.* **1**, 89–94.

Beale, D. & Feinstein, A. (1976). *Quart. Rev. Biophys.* **9**, 135–180.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, D. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

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

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

Coates, A. R. M., Cookson, J., Barton, G. J., Zvelébil, M. J. & Sternberg, M. J. E. (1987). *Nature (London)*, **326**, 549–550.

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

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

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

Dayhoff, M. O., Park, C. M. & McLaughlin, P. J. (1972). In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), p. 12. National Biomedical Research Foundation, Washington, DC.

Doolittle, R. F. (1981). *Science*, **214**, 149–159.

Edelman, G. M. (1970). *Biochemistry*, **9**, 3197–3205.

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

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

Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 9–16.

George, D. G., Barker, W. C. & Hunt, L. T. (1986). *Nucl. Acids Res.* **14**, 11–15.

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

Gorinsky, B., Horsbaugh, C., Lindley, P. F., Moss, D. S., Parkar, M. & Watson, J. L. (1979). *Nature (London)*, **281**, 157–158.

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

Johnson, M. S. & Doolittle, R. F. (1986). *J. Mol. Evol.* **23**, 267–278.

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

Lesk, A. M. & Chothia, C. (1982). *J. Mol. Biol.* **160**, 325–342.

Metz-Boutigue, M. H., Jollès, J., Mazurier, J., Schoentgen, F., Legrand, D., Spik, G., Montreuil, J. & Jollès, P. (1984). *Eur. J. Biochem.* **145**, 659–676.

Moore, G. W. & Goodman, M. (1977). *J. Mol. Evol.* **9**, 121–130.

Murata, M., Richardson, J. S. & Sussman, J. L. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 3073–3077.

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

Rose, T. M., Plowman, G. D., Teplow, D. B., Dreyer, W. J., Hellstrom, K. E. & Brown, J. P. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 1261–1265.

Sankoff, R. J. & Cedergren, G. L. (1976). *J. Mol. Evol.* **7**, 133–149.

Sellers, P. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 3041.

Sobel, E. & Martinez, H. M. (1986). *Nucl. Acids Res.* **14**, 363–374.

Sokol, R. R. & Sneath, P. A. H. (1973). In *Numerical Taxonomy*, p. 201, Freeman, San Francisco.

Sternberg, M. J. E., Barton G. J., Zvelébil, M. J. J., Cookson, J. & Coates, A. R. M. (1987). *FEBS Letters*, **218**, 231–237.

Taylor, P. (1984). *Nucl. Acids Res.* **12**, 447–455.

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

Waterman, M. S. (1986). *Nucl. Acids Res.* **14**, 9095–9102.

Wilbur, W. J. & Lipman, D. J. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 726–730.

Williams, J. (1982). *Trends Biochem. Sci.* **7**, 394–397.

Williams, J., Elleman, T. C., Kingston, I. B., Wilkins, A. G. & Kuhn, K. A. (1982). *Eur. J. Biochem.* **122**, 297–303.

Wierenga, R. K., Terpstra, P. & Hol. W. G. J. (1986). *J. Mol. Biol.* **187**, 101–107.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). *J. Mol. Biol.* **195**, 957–961.

Edited by R. Huber