[25] Protein Multiple Sequence Alignment and Flexible Pattern Matching

By GEOFFREY J. BARTON

Introduction

The alignment of two or more protein sequences can provide a wealth of information to guide further experimentation, particularly if one of the aligned proteins has been biochemically or crystallographically well characterized. However, any inference from the alignment is crucially dependent on its accuracy. Thus, in this chapter, alignments obtained from comparison of the protein three-dimensional structures are used as a standard against which to test an automatic method for the pairwise alignment of protein sequences. The accuracy of the resulting alignments is shown to improve when additional nonsequence information is incorporated into the algorithm.

Rigorous methods for the alignment of two protein sequences have long been known;[1] however, the calculation of an optimal alignment of four or more sequences is beyond the capabilities of even the most powerful computers. Alignment of multiple sequences by eye is at best a tedious and time-consuming operation; at worst it is unsystematic and leads to an alignment about which no degree of confidence may be expressed. In this chapter, a practical strategy for the rapid multiple alignment of protein sequences is described. Although not guaranteed to give the mathematically optimal alignment, the algorithm is able to cope with large numbers of sequences. It is also a fast procedure that gives alignments generally as good or better than those obtained by pairwise methods.

When sequence similarity is weak, conventional alignment procedures can fail to identify biologically significant relationships against the background of all known sequences. The sensitivity and selectivity of alignment methods that exploit information from single or multiple sequences with and without additional nonsequence information are also evaluated. Furthermore, a technique that relies on the systematic derivation of *flexible patterns* is shown to be superior to all these methods when applied to the globin family of proteins.

[1] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48,** 433 (1970).

Protein Sequence Comparison

Broadly, there are three categories of methods for the comparison of protein sequences. Segment methods compare all overlapping segments of a predetermined length (e.g., 10 amino acids) from one protein with all segments from the other. The distribution of scores obtained for all segment pairs can be used directly to infer homology.[2] Alternatively, the segment scores may be plotted graphically as a "comparison matrix."[3,4] Segment methods have the advantage of simplicity; however, they do not cater explicitly for insertions and deletions (gaps).

Optimal global alignment methods allow the best overall score for the comparison of the two sequences to be obtained including a consideration of gaps. The Needleman and Wunsch algorithm[1] was the first description of a global alignment method applied to protein sequences, but variants of the basic dynamic programming algorithm have been independently developed and applied in many fields (see Sankoff and Kruskal[5] for review). The advantage of these techniques is that they are guaranteed to find the best overall score for the comparison of the sequences including a consideration of gaps. Furthermore, these methods can also produce one or more alignments consistent with this best score. As a consequence, computer programs based on this method (e.g., NBRF, ALIGN) have been widely used for biological sequence comparison. For these reasons, the algorithm of Needleman and Wunsch[1] forms the nucleus of the techniques for pairwise and multiple sequence alignment described in the following sections.

Finally, *Optimal local alignment* algorithms seek to identify the best *local* similarities between two sequences but, unlike segment methods, include explicit consideration of gaps. The methods are based on modified Needleman and Wunsch-style algorithms (e.g., see Refs. 6 and 7) and represent an important class of comparison algorithm, particularly for the location of significantly similar regions between long sequences.

All the methods require a *scoring scheme* for the matching of each of the 210 possible pairs of amino acids (i.e., 190 pairs of different amino acids plus 20 pairs of identical amino acids). For example, the simple identity scoring scheme gives a score of 1 to identical pairs and 0 to all others. More sophisticated schemes can incorporate knowledge about the

[2] W. M. Fitch, *J. Mol. Biol.* **16,** 9 (1966).

[3] A. D. McLachlan, *J. Mol. Biol.* **61,** 409 (1971).

[4] P. Argos, *J. Mol. Biol.* **193,** 385 (1987).

[5] D. Sankoff and J. B. Kruskal (eds.) "Time Warps, String Edits, and Macromelecules: The Theory and Practice of Sequence Comparison. Addison Wesley, Reading, Massachusetts, 1983.

[6] T. F. Smith and M. S. Waterman, *J. Mol. Biol.* **147,** 195 (1981).

[7] D. R. Bowsell and A. D. McLachlan, *Nucleic Acids Res.* **12,** 457 (1984).

physical properties of the amino acids,[3] minimum allowed base changes,[2] or observed substitutions[8] to give a symmetrical 20 × 20 matrix of scores.

Gap Penalties and Needleman–Wunsch Algorithm

The Needleman–Wunsch algorithm is an elegant procedure that allows the best alignment of two sequences of length N to be calculated in N^2 steps. When comparing two sequences the algorithm seeks to model the real (possibly evolutionary) processes involved in converting one sequence to the other. The scoring scheme that dictates the weight for aligning one type of amino acid with another is part of this model. For any chosen scoring scheme, the Needleman–Wunsch algorithm will find the maximum possible score for the comparison of the two sequences. However, this optimal alignment may require the insertion of an unrealistically large number of gaps (residues aligned with blanks).

In order to overcome this problem and limit the total number of gaps created, an additional factor is introduced into the model. This takes the form of a gap penalty which is subtracted during the process of calculating the best alignment whenever a gap is allowed. One of the most commonly used gap-penalty functions takes the form:

$$P = G_1\, L + G_2 \tag{1}$$

where L is the length of the gap while G_1 and G_2 are user-defined constants. This form of penalty has both length-independent (G_2) and length-dependent (G_1L) terms that are sometimes known as penalties for creation of a gap and extension of a gap, respectively.

Criteria for Assessing Quality of Alignment

Given the Needleman–Wunsch algorithm, a scoring scheme, and gap-penalty function, we have a system that can optimally align any two protein sequences. It is important to bear in mind, however, that this alignment is optimal only with regard to the chosen model; changing the model, either by using a different scoring scheme or a modified gap penalty, can lead to completely different alignments. While all of these alignments will be mathematically optimal it is possible that none of them illustrate genuine, biologically significant equivalences. Thus, when applying and interpreting automatically obtained alignments three questions need to be answered: What is a good protein sequence alignment? How

[8] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *in "Atlas of Protein Sequence and Structure"* (M. O. Dayhoff, ed.), Vol. 5, p. 345. National Biomedical Research Foundation, Washington, D.C., 1978.

closely can automatic alignment procedures reproduce a good alignment? Can we estimate the likely quality of an alignment from sequence information alone?

What Is Good Protein Sequence Alignment?

The protein three-dimensional structure determines its biological activity. It is therefore of crucial importance when two or more protein sequences are aligned that those residues defining a common tertiary fold, together with any common catalytic and binding residues, are correctly equivalenced. It follows that a *good* alignment of two globular proteins is one which faithfully reflects any similarities in three-dimensional structure.

However, although the overall fold may be conserved within a protein family, there is usually considerable variation in the details of structure for individual family members. In particular, although core secondary structures (α helices and β strands) may exhibit similar conformations and relative positions in three dimensions, the loop regions linking these structures may not. Thus, even when the proteins to be aligned have crystallographically determined structures known to high resolution, it can be difficult or impossible to obtain a consistent alignment over the whole length of *all* the family members.

In order to be useful, a protein sequence alignment method must at least align those residues that are performing equivalent structural roles in the two proteins. For this reason when assessing an alignment algorithm for two or more protein sequences, *test zones*[9] are selected that correspond to equivalent secondary structural regions in the proteins. The accuracy of any alignment can then be expressed in terms of the number of residues within the test zones that are equivalenced in the same way as expected from three dimensional structure comparison.

How Well Do Automatic Alignment Methods Perform?

The option of using different scoring scheme and gap-penalty combinations complicates the evaluation of alignment methods. Studies performed by Dayhoff *et al.*[8] and more recently by Feng *et al.*[10] indicate that, on average, the Dayhoff mutation data matrix (MDM) is more effective than the simple identify matrix, genetic code, or physical property scoring schemes at detecting homology between distantly related proteins. Using Dayhoff's matrix, Barton and Sternberg[9] considered a range of gap-penalty constants ($G_1 = 0-10, G_2 = 0-10$ in integer steps) for five polypeptide

[9] G. J. Barton and M. J. E. Sternberg, *Proteın Eng.* **1,** 89 (1987).

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

pairs for which alignments based on three-dimensional structure were known. For each pair of sequences, test zones were selected from the common core secondary structural regions and mean, maximum, and minimum accuracies of alignment obtained over the 121 comparisons *(Fig. 1). Clearly, some protein pairs align very well (1, 4 and 5), whereas* others give poor alignments (2, 3). Furthermore, the best alignment obtained for each pair did not require a length-dependent gap penalty (i.e, $G_1 = 0$).

Use of Significance Scores to Estimate Likely Quality of Alignment

With the increase in the number of protein sequences derived directly from cDNA one frequently wishes to align two sequences for which there is little or no additional nonsequence information available to guide the alignment (e.g., three-dimensional structure, catalytic residues). Under

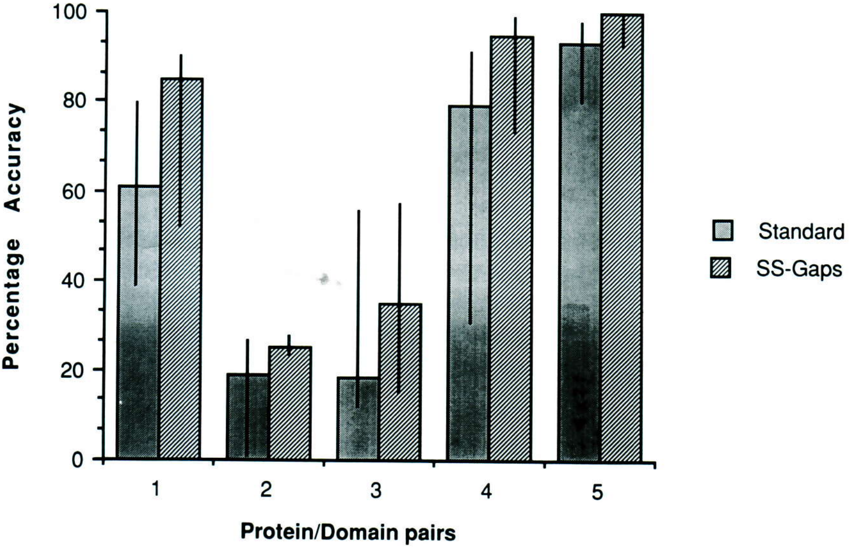

FIG. 1. Accuracy of pairwise sequence alignments by the Needleman–Wunsch method by comparison with tertiary structural alignments. Values of G_1 and G_2 from 0 to 10 in integer steps were used with MDM (+8 to remove negative elements). 1, Immunoglobulin light chain variable region (FABVL) versus heavy chain variable region (FABVH); 2, FABVH versus light chain constant region (FABCL); 3, plastocyanin versus azurin; 4, human α-hemoglobin versus leghemoglobin; 5, trypsin versus elastase. Standard: Mean values for 121 comparisons using the conventional Needleman–Wunsch algorithm. SS-Gaps: Effect of including secondary structure-dependent gap penalties. Upper and lower extremities of vertical bars indicate the best and worst alignments obtained.

these circumstances it is invaluable to have an indication of the likely accuracy of any automatic alignment performed.

A commonly used method of assessing the similarity between two sequences proceeds as follows. First, the best score V for the comparison of the native sequences is obtained. The sequences are then randomized a number of times (typically 100) to give artificial sequences with the same length and composition as the native. The best score for aligning each pair of randomized sequences is then obtained, and the mean m and standard deviation (S.D.) of this distribution of random scores are calculated. The similarity of the native sequences is then expressed in terms of the number of standard deviation units away from the mean of the random distribution [i.e., Score = $(V - m/\text{S.D.}$].

A study was performed[11] which considered all pairwise comparisons within seven globin sequences and eight immunoglobulin domains (49 unique pairs in all). The results shown in Fig. 2 illustrate that alignments scoring above 15.0 S.D. (seven 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, whereas scores below 5.0 SD (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* really poor alignments; however, in the lower standard deviation range small changes in observed significance score can indicate a considerable difference in alignment accuracy.

These studies provide guidelines for the quality of a protein sequence alignment. Clearly, a near ideal alignment is indicated by significance scores above 15.0 S.D. Scores above 5.0 S.D. suggest a "good" alignment, whereas an alignment giving a score below 5.0 S.D., although possibly good, must be regarded with greater caution.

Improving Sequence Alignments by Using Secondary Structure-Dependent Gap Penalties

The alignment model consisting of a scoring scheme and gap-penalty function can produce good alignments, however, the gap-penalty function [Eq. (1)] acts equally over the entire sequence length. This feature does not reflect observations on families of known protein three-dimensional structures where there is a clear preference for insertions/deletions to occur in loop regions linking the core secondary structures (e.g., see Refs. 12 and

[11] G. J. Barton and M. J. E. Sternberg, *J. Mol. Biol.* **198,** 327 (1987).

[12] M. F. Perutz, J. C. Kendrew, and H. C. Watson, *J. Mol. Biol.* **104,** 59 (1965).

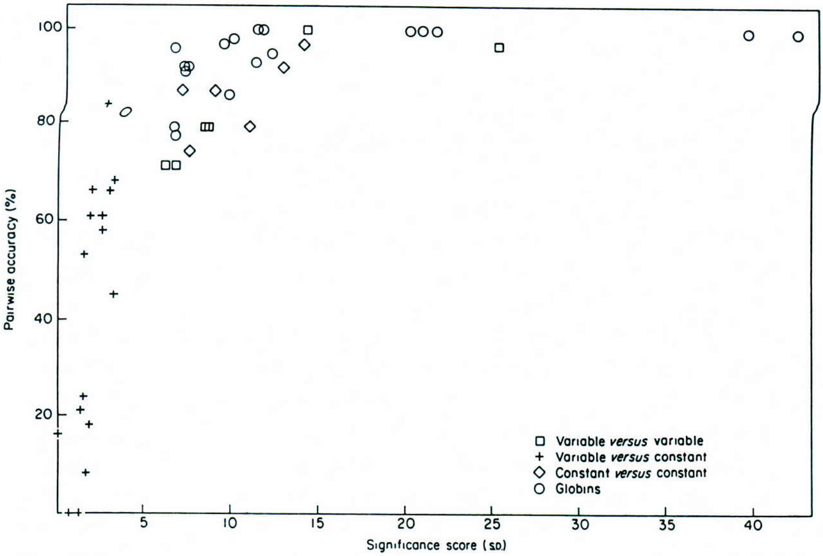

FIG. 2. Pairwise alignment accuracy versus significance score (100 randomizations) for seven globins [human α-hemoglobin (HAHU), human β-hemoglobin (HBHU), horse α-hemoglobin (HAHO), horse β-hemoglobin (HBHO), myoglobin (MYWHP), lamprey globin (P1LHB), leghemoglobin (LGHB)] and eight immunoglobulin domains [consisting of one light chain constant domain (FABCL), three heavy chain constant domains (FABCH, FCCH2, FCCH3), two light chain variable domains (FABVL, FB4VL), and two heavy chain variable domains (FABVH, FB4VH)]. Dayhoff's matrix was used: (250 PAM) + 8, $G_2 = 8$, $G_1 = 0$.

13). Indeed, errors in sequence alignment can frequently be attributed to the misplacing of a gap in a core secondary structural region.

The alignment model may be improved to better match the observed pattern of insertions by using a modified gap-penalty function:

$$P_{ss} = Q\,(G_1L + \mathrm{G}_2) \tag{2}$$

where $0 \leq Q \leq 1$ and the subscript *ss* denotes the inclusion of secondary structural information. This change has the effect of reducing the penalty for a gap in loop regions relative to secondary structural regions.

In its simplest form, Q takes a value of 1.0 for regions of secondary structure and a value of less than 1.0 for loop regions. The effect of applying this type of penalty is illustrated in Fig. 1. All five protein pairs show improvements in mean accuracy and improvements in the worst alignment obtained, and, with the exception of FABVH *versus* FABCL, the best alignment obtained also gives a higher accuracy.

[13] A. M. Lesk and C. Chothia, *J. Mol. Biol.* **136,** 225 (1980).

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. Q might therefore be derived from a secondary structure prediction profile,[14,15] a smoothed profile based on hydrophobicity,[16] or a profile of likely buried residues.[17] Unfortunately, none of these methods predict the location of secondary structural elements with sufficient accuracy to improve the alignment quality. Indeed, in studies using several alternative predictive schemes to derive values of Q over the sequences, the overall accuracy of alignment actually decreased.

This observation clearly limits the applicability of the modified gap-penalty function to systems where one of the proteins has a known X-ray structure. For such systems the improvement in accuracy obtained justifies the inclusion of secondary structural information into the alignment and is of particular use when the alignment is to be used for subsequent building of a three-dimensional model by homology (e.g., see Ref. 18). Lesk and co-workers[19] described a similar technique and showed that it improves the alignment of sequences within the globin and serine proteinase families.

Simultaneous Alignment of More than Two Sequences (Multiple Alignment)

Needleman and Wunsch[1] suggested that their dynamic programming algorithm could be extended to the simultaneous comparison of many sequences. Waterman *et al.*[20] also described how dynamic programming could be used to align more than two sequences. In practice, however, the need to store an N-dimensional array (where N is the number of sequences) limits these extensions to three-sequence applications (e.g., see Ref. 21). In addition, the time required to perform the comparison of even three sequences is proportional to N^5. Murata *et al.*[22] described a modification of the Needleman–Wunsch procedure for three sequences which ran in time proportional to N^3; unfortunately, this approach required an additional

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

[15] P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47,** 45 (1978).

[16] M. Levitt, *J. Mol. Biol.* **104,** 59 (1976).

[17] J. Janin, *Nature (London)* **277,** 491 (1979).

[18] T. L. Blundell, B. L. Sibanda, and L. Pearl, *Nature (London)* **304,** 273 (1983).

[19] A. M. Lesk, M. Levitt, and C. Chothia, *Protein Eng.* **1,** 77 (1986).

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

[21] R. A. Jue, N. W. Woodbury, and R. F. Doolittle, *J. Mol. Evol.* **15,** 129 (1980).

[22] M. Murata, J. S. Richardson, and J. L. Sussman, *Proc. Natl Acad. Sci. U.S.A.* **82,** 3073 (1985).

three-dimensional array, thus further limiting its application to short sequences.

The multiple alignment of four or more sequences cannot in practice be performed by a rigorous method since even when gaps are not explicitly considered, the number of segment comparisons that must be made is of the order of the product of the sequence lengths. Algorithms for multiple sequence alignment therefore seek to identify an optimum alignment by considering only a small number of the total possible residue or segment comparisons. Several authors have described multiple alignment algorithms; however, they either do not give an overall alignment,[23] are restricted to relatively few sequences,[24] or are specifically intended for aligning nucleic acid sequences and do not allow the flexibility in scoring scheme that is useful for protein sequence comparison.[25-26] In the following section a method that permits large numbers of protein sequences to be aligned quickly is described and tested by comparison with alignments obtained from the comparison of protein three-dimensional structures.

Effective and Rapid Strategy for Multiple Protein Sequence Alignment[11]

The alignment algorithm described here reduces the multiple alignment of N sequences to a set of $N - 1$ pairwise alignments and is summarized in Fig. 3. (1) Sequences A and B are optimally aligned by the Needleman–Wunsch algorithm. (2) The third sequence is optimally aligned with the *alignment* resulting from Step 1. Average scores are used when comparing a residue in sequence C to an aligned position in the result of Step 1. For example, the score for matching the alignment of Ala and Val with Ala would be given by the score for (Ala versus Ala) plus (Val versus Ala) divided by 2. Gaps that are already present in the alignment from Step 1 are maintained, and a low score is assigned to matching an amino acid in sequence C with any such gap. This score is used when calculating the average score at the aligned position. For example, if the aligned position is AlaGap then the score for matching Ala would be given by (Ala versus Ala) plus (Gap versus Ala) dived by 2. (3) The multiple alignment of sequences A, B, and C obtained in Step 2 is now optimally aligned with sequence D using a procedure similar to Step 2. (4) Step 3 is repeated until all sequences have been added to the alignment. (5) The alignment from Step 4 may optionally be refined by reoptimizing the alignment of each sequence with the completed alignment less that sequence.

[23] D. J. Bacon and W. F. Anderson, *J. Mol. Biol.* **191,** 153 (1986).

[24] M. S. Johnson and R. F. Doolittle, *J. Mol. Evol.* **23,** 267 (1986).

[25] E. Sobel and H. M. Martinez, *Nucleic Acids Res.* **14,** 363 (1986).

[26] W. Bains, *Nucleic Acids Res.* **14,** 159 (1986).

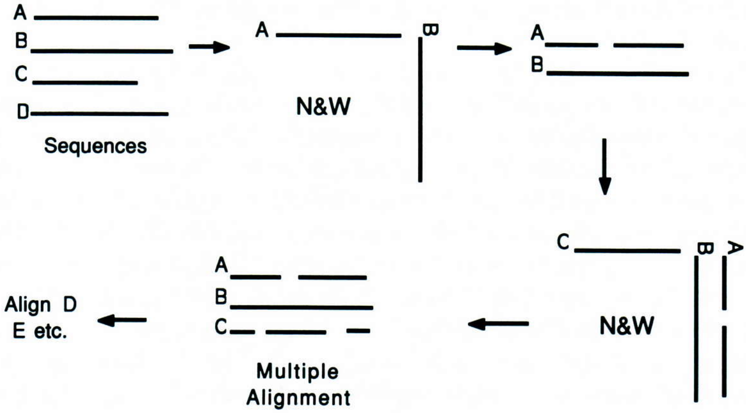

FIG. 3. Summary of the multiple alignment process.

Order of Alignment

Since the multiple algorithm shares the Needleman–Wunsch procedure with pairwise methods, the alignment will be dependent on both the scoring scheme and gap penalty. In addition, there are *N*! alternative orders in which the sequences could be aligned. A systematic procedure for determining the alignment order must therefore be applied.

The pairwise comparison tests shown in Fig. 2 demonstrate that the accuracy of alignment is correlated with the significance score. The single alignment order may therefore be determined by first calculating significance scores for all unique *pairwise* comparisons within the sequence set. Then, when generating the multiple alignment, the pair of sequences that gives the highest significance score is aligned first. Of the remaining sequences, the one which gives the highest score when compared to A or B is then aligned. The process is repeated for all remaining sequences, where every *ith* sequence being added to the alignment is the one that gives the highest pairwise significance score with the $i-1$ sequences already aligned.

Cluster Analysis

A useful method of visualizing the pairwise comparison data is to apply the technique of single linkage cluster analysis. This provides a convenient representation in the form of a dendrogram that can illustrate some of the interrelationships between the members of a sequence group.

The dendrograms illustrated in Fig. 4a,b for the seven globins and eight immunoglobulins used to evaluate pairwise methods clearly show the

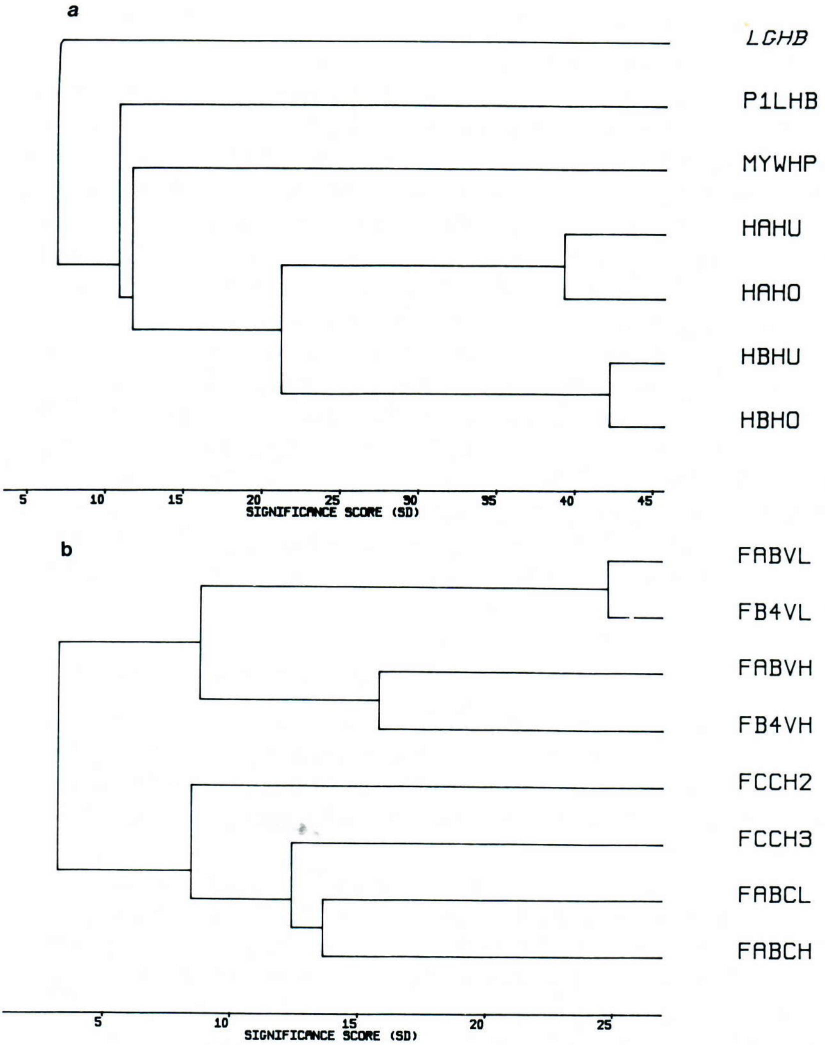

FIG. 4. Single-linkage dendrograms for (a) globins and (b) immunoglobulin domains.

sequences grouped by similarity. The maximum level of similarity between the groups is also readily apparent. Given that the relationship between significance score and alignment accuracy is known, the dendrogram in conjunction with pairwise scores can help to identify quickly which pairs of sequences may align to high accuracy. Furthermore, the high scoring clusters may indicate groups of sequences that will also align well by the multiple alignment algorithm.

Reduction of Calculation Time

Before a group of sequences can be ordered, or cluster analysis performed, it is necessary to calculate scores for all sequence pairs. This is an expensive procedure since if M randomizations are performed, $N(N-1)M/2$ alignments must be generated. Feng *et al.*[10] considered how many randomizations need be performed on a pair of sequences before consistent significance values are obtained. On the basis of 4 pairs of sequences they suggested that as few as 25 could produce a genuinely reflective score. A related study using a larger dataset (47 protein pairs)[11] indicated that instabilities in significance score do not damp out until at least 60 randomizations are performed. It is therefore impractical to use a randomization procedure to establish the order when large numbers of long sequences are to be aligned. However, it is possible to derive a normalized alignment score (NAS) directly from the match score V without the need for randomization.[10,11] Scores of this type correlate well with the significance score, suggesting that when central processing unit (CPU) time would otherwise by prohibitive NAS values can be used to establish an alignment order and reduce the number of comparisons that need be made by a factor of at least 60.

Evaluation of Multiple Alignment Algorithm: Comparison with Pairwise

The alignment procedure described above is able to produce a multiple alignment for *any* set of sequences. However, as with pairwise methods, it is of vital importance that the properties and limitations of the method are well understood so that its best features can be exploited when it is applied to new systems.

The seven globin and eight immunoglobulin sequences used to evaluate pairwise methods also provide a good test system for the multiple algorithm. In common with the pairwise method, the gap penalty and scoring scheme may be varied; however, there are three additional factors to be considered: (1) Which sequences should be included in a multiple alignment? (2) Does the order of alignment have a serious effect? (3) Can alignments be improved by iteration (Step 5 above)?

Point (1) was considered by multiply aligning four groups of sequences derived from the globins and immunoglobulins: Alignment 1, the seven globin sequences HBHU, HBHO, HAHU, HAHO, MYWHP, P1LHB, and LGHB (see Fig. 5); Alignment 2, the four constant immunoglobulin domains FABCL, FABCH, FCCH3, AND FCCH2; Alignment 3, the four variable domains FABVL, FB4VL, FB4VH, AND FABVH; and Alignment 4, the eight immunoglobulin domains used in Alignments 2 and 3.

The effect of order was addressed by considering alternative alignment

```
                                   A                    B                                        E
HBHU              vhltPEEKSAVTALWGKv   nVDEVGGEALGRLLVVypWTQRff••fgdl•tpdav■gnPKVKAHGKKVLGAFSDGLahldn      lK  GTF
HBHO              vql•GEEKAAVLALWDKv   nEEEVGGEALGRLLVVypWTQRffd•fgdl•npgav■gnPKVKAHGKKVLHSFGEGVhhldn      lK  GTF
HAHU               vl•PADKTNVKAAWGKvgahAGEYGAEALERMFLSfpTTKTyfphf dl•h     g•AQVKGHGKKVADALTNAVahvdd      ■P  NAL
HAHO               vl•AADKTNVKAAWSKvgghAGEYGAEALERMFLGfpTTKTyfphf dl•h     g•AQVKAHGKKVGDALTLAVghldd      lP  GAL
P1LHB     pivdtg•vapl•AAEKTKIRSAWAPvy•dYETSGVDILVKFFTStpAAEEffpkfkgltta d•lkk•ADVRWHAERIIDAVDDAVa•■dd     t•kMSSM
MYWHP              vl•EGEWQLVLHVWAKv•adVAGHGQDILIRLFKShpETLEkfdrfkhlkt•a•■ka•EDLKKHGVTVLTALGAILkkkgh      hE  AEL
LGHB              galtESQAALVKSSWEEfnanIPKHTHRFFILVLEIapAAKDlf••flkggt••vpqnnPELQAHAGKVFKLVYEAAiql•vtgvva•DATL

                  F                 G                               H
HBHU       ATLSELHCDklhvdPENFRLLGNVLVCVLAHHfgk•ftppvqaAYQKVVAGVANALAhkyh
HBHO       AALSELHCDklhvdPENFRLLGNVLVVVLARHfgkdftp•lqaSYQKVVAGVANALAhkyh
HAHU       SALSDLHAHklrvdPVNFKLLSHCLLVTLAAHlpa•ftpavhaSLDKFLASVSTVLT•kyr
HAHO       SNLSDLHAHklrvdPVNFKLLSHCLLSTLAVHlpndftpavhaSLDKFLSSVSTVLT•kyr
P1LHB      KDLSGKHAK•f•vdPEYFKVLAAVIADTVAAG         daGFEKLLRMICILLR•ay
MYWHP      KPLAQSHATkhkipIKYLEFISEAIIHVLHSRhpgdfgadaqgAMNKALELFRKDIAakyk•lgyqg
LGHB       KNLGSVHVSkgvvaDAHFPVVKEAILKTIKEVvgakv•••ln•AWTIAYDELAIVIKk••ddaa
```

FIG. 5. Multiple alignment of seven globin sequences. Boxed regions with capital letters refer to test zones. Regions A, B, C, E, F, G, and H are all α helical in the known protein structures.

orders for Alignments 1 and 4, and the effect of applying up to four iterations was investigated for all four alignments. Figure 6 shows the accuracy of alignment obtained for pairs of sequences within the four multiple alignments compared to the accuracy obtained when the sequences are aligned pairwise. Points above the diagonal represent an improvement in alignment when the multiple algorithm is applied. The globin multiple alignment (1) gives an overall improvement from 90 to 99% accuracy, with the largest improvement for the comparison of leghemoglobin with human β-hemoglobin (77 to 99%). Alignments 2 and 3 also show an overall improvement in accuracy; for the constant domains this is from 86 to 90%, while the variable domains improve from 83 to 84%. The

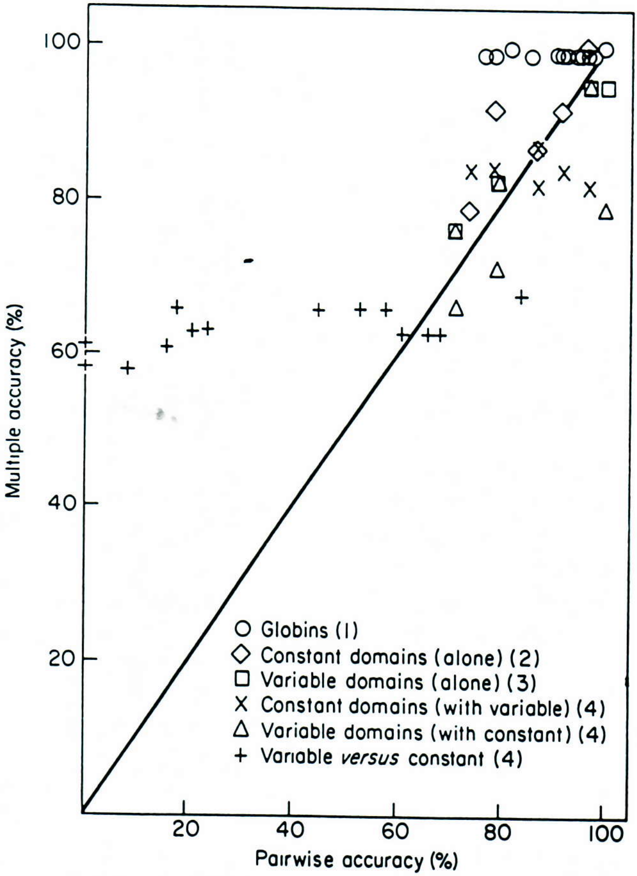

FIG. 6. Accuracy of alignment obtained by the multiple method versus the conventional pairwise method.

most striking improvement is for variable versus constant domains within Alignment 4. Some sequence pairs which were completely misaligned by the pairwise procedure gave around 60% accuracy when multiple aligned (e.g., FB4VH versus FABCH). However, this improvement was obtained at the expense of a slight degradation of variable *versus* variable and constant *versus* constant alignments.

Although Alignment 4 shows a large improvement in accuracy for low scoring sequence pairs, further studies (results not shown) suggest that improvements in accuracy can be very variable for this type of sequence. Furthermore, when ten alternative alignment orders were generated for the seven globin and eight immunoglobulin alignments (1 and 4), the alternative orders had very little effect on the globin accuracy (< 1%), but for the immunoglobulin most alternative orders gave poorer alignments (mean of 57.6% compared to 70.8% for the order based on S.D. score). These findings are consistent with the observed variation in alignment accuracy below significance scores of 5.0 S.D. for pairwise comparisons (Fig. 2).

The use of up to four iterations to refine the initial multiple alignment showed that in general there was no benefit in performing more than two iterations. Alignments 1–3 improved by approximately 1% over an alignment with no iterations, although, once again, Alignment 4 proved to be atypical with an improvement of around 9%.

In summary, this evaluation suggests that for sequence groups that on pairwise comparison cluster above 5.0 S.D. (e.g., Alignments 1, 2, and 3) the resulting multiple alignment is likely to be as good or better than corresponding pairwise alignments. In common with the findings for pairwise methods, the multiple alignment obtained for sequences that cluster below 5.0 S.D. (e.g., Alignment 4) is likely to be unpredictable in quality. Furthermore, time can be saved when large numbers of sequences are to be aligned by using the minimum calculation of normalized alignment scores to establish the alignment order rather than significance scores. The observation that alignment order has little effect on the result for some groups of sequences (e.g., the seven globins) suggests that an arbitrary alignment order may often be acceptable, thus removing the time-consuming need to perform all pairwise comparisons prior to multiple alignment.

Tree-Based Multiple Alignment

The algorithm evaluated in the previous section considers the sequences in a single linear order. However, the dendrogram representation of the pairwise comparison data (Fig. 4a,b) suggests an alternative way in which to order the multiple alignment process. Rather than starting with the most similar pair and adding successively to that alignment, the dendrogram or tree is followed exactly from its branches to the root. For example, with the seven-globin alignment the following series of alignments are performed: (1) Align the most similar pair of sequences, HBHU and HBHO, to give *alignment* HBHU:HBHO. (2) Align the next most similar pair, HAHU and HAHO, to give alignment HAHU:HAHO. (3) Now align the two *alignments* HBHU:HBHO and HAHU:HAHO to give the four-sequence alignment HBHU:HBHO:HAHU:HAHO. (4) Align MYWHP to the four-sequence alignment obtained in Step 3, then P1LHB to the resulting five-sequence alignment, and finally LGHB to the six-sequence alignment to give the final seven-sequence alignment.

The only new operation involved in this process is the alignment of two *alignments* shown in Step 3. This step is essentially the same as adding a single sequence to an alignment, only now it is necessary to calculate mean scores over all unique pairs of residues at each position. As before, gaps that already exist in either alignment are maintained, and the score for matching two such gaps is given a low score.

The tree-based approach is intuitively better than a single order alignment method. However, for sequences that all cluster at high scores, the differences in alignment are only slight, and for the seven globins the end result is identical. Where there are two or more distinct high scoring clusters that do not form a single high scoring cluster (as for the eight immunoglobulins), then a tree-based alignment will give better results within the high scoring clusters. However, the problem of variable alignment quality when low scoring sequence pairs are compared still remains. Thus the eight-immunoglobulin alignment when performed tree-wise gives better accuracy for constant *versus* constant and variable *versus* variable domains but equally unpredictable results for variable *versus* constant domains. The tree-based, or *progressive,* multiple alignment method has been described by Feng and Doolittle[27] (also this volume,[23]), who demonstrated that better phylogenetic trees could be obtained from its application.

Speed of Multiple Alignment and Applications

An advantage of the multiple alignment algorithm described here is its speed. For example, the complete seven-sequence globin alignment required only 65 CPU sec (2 iterations) on a VAX 11/750. Pairwise comparisons to establish the order without randomization required 44 sec, giving a total time of 109 sec. If an arbitrary order with no iterations had been used, the total time required would be approximately 20 sec. This compares

[27] D. F. Feng and R. F. Doolittle, *J Mol. Evol.* **25,** 351 (1987).

favorably with the algorithm of Johnson and Doolittle[24] which requires 60 *min* of CPU time to align five sequences of less than 50 residues in length and cannot easily be extended to cope with large numbers of sequences.

Aligning large numbers of medium length sequences (150–300 residues) by single order or tree methods is therefore a matter of routine. For example, the alignment of 128 globin sequences including α and β hemoglobin, myoglobin, and leghemoglobin from a wide range of species required only 8.5 min of CPU time to produce an alignment prior to refinement by iteration.[11] The alignment of longer sequences is also practical. This is shown in one application of the algorithm to the prediction of potential T and B lymphocyte-defined epitopes on the *env, gag,* and *pol* viral polyproteins of the human immunodeficiency virus (HIV). Four viral isolates were aligned (500–1000 amino acids in length), and analysis of residue conservation in combination with structure prediction methods allowed potential epitopes to be identified.[28,29]

The speed and accuracy of the alignment method have also permitted an improved secondary structure prediction method to be developed.[30] The prediction algorithm combines Robson prediction values[14] averaged over all aligned sequences with a measure of the residue conservation at each aligned position. The use of a conservation value has the effect of reducing the likelihood of predicting secondary structure (α helix or β strand) in regions where gaps have been inserted in the alignment. The overall improvement in accuracy over the standard Robson method was 8.5% obtained for 11 protein families representative of the most common structural classes (α/α, β/β, α/β, and $\alpha + \beta$).

Refinements to Improve Speed and Sensitivity

Since the multiple algorithm is built from successive applications of a pairwise technique, any refinements available to pairwise methods may also be incorporated into the multiple alignment procedure. For example, the time required to perform a Needleman–Wunsch alignment can be reduced by "cutting corners" during calculation of the best alignment.[31] Fast but approximate pairwise methods (e.g., the algorithm of Lipman and

[28] A. R. M. Coates, J. Cookson, G. J. Barton, M. J. Zvelebil, and M. J. E. Sternberg, *Nature (London)* **326,** 549 (1987).

[29] M. J. E. Sternberg, G. J. Barton, M. J. J. Zvelebil, J. Cookson, and A. R. M. Coates, *FEBS Lett.* **281,** 231 (1987).

[30] M. J. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. E. Sternberg, *J. Mol. Biol.* **195,** 957 (1987).

[31] J. B. Kruskal and D. Sankoff, *in "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence"* (D. Sankoff and J. Kruskal, eds). p 265. Addison-Wesley, Reading, Massachusetts, 1983.

Pearson[32] may also be used in place of the rigorous Needleman–Wunsch algorithm. Both these methods are most useful when long sequences that have relatively few differences are to be aligned.

Nonsequence information may also be incorporated to guide multiple alignments. Secondary structure-dependent gap penalties as described earlier may be incorporated when at least one sequence has a known three-dimensional structure. Furthermore, position-specific weights may be assigned to residues of known importance (e.g., catalytic amino acids) to increase their likelihood of aligning with similar amino acids.

Guidelines for Performing Multiple Protein Sequence Alignments and Assessing Accuracy

Given a group of sequences to multiply align the following steps may be followed: (1) Ideally all pairwise comparisons for the sequences should be performed using at least 60 randomizations to establish significance scores. (2) Cluster analysis may then be applied to the pairwise data resulting from Step 1 and a dendrogram drawn to represent the results. (3) The dendrogram should be inspected to locate any high scoring clusters of sequences. The sequences that cluster above 5.0 S.D. can be multiply aligned with a high degree of confidence. Outlying sequences (those that do not belong to high scoring clusters) should be removed for possible incorporation in Step 5. (4) The sequences within the high scoring clusters identified in Step 3 should be multiply aligned following the order suggested by the pairwise significance scores. (5) Steps 1–4 will produce one or more "core" alignments that are largely correct. The next step is to align the remaining weakly similar sequences to one or more cores, making use of additional nonsequence information where possible, for example, the location of known catalytic or structural regions common to both sequence groups. A flexible pattern derived from the core alignments may be used to assist in this procedure (see following section). (6) The final alignment(s) must *always* be discussed in the light of the likely error rates implied by the pairwise significance scores and any assumptions made in combining core alignments.

Flexible Patterns: Sensitive Method to Detect Weak Structural Similarities

The score obtained when two sequences are optimally aligned by the Needleman–Wunsch algorithm tells us how similar the sequences are

[32] D. J. Lipman and W. R. Pearson, *Science* **227,** 1435 (1985).

according to the model of evolutionary change implied by the scoring scheme and gap penalty. However, this scheme can give lower scores for protein pairs that are known to have similar tertiary structures than for either random sequences of the same length and composition or an arbitrary pair of unrelated protein sequences. In other words, the similarity between the proteins may be hidden in the noise generated by chance high scoring alignments.

The overall improvement in alignment accuracy observed when multiple rather than pairwise alignments are used suggests one route by which the sensitivity of an alignment procedure may be improved. Reliably aligned protein families clearly contain information that is not available from a single sequence, for example, the importance of conservation at particular residue positions and the disposition of gaps. Multiple alignment information of this type has been exploited to improve the sensitivity of comparison between families of aligned proteins (e.g., see Refs. 4 and 33), and, as suggested in the previous section, they can also be used in conjunction with additional nonsequence information (e.g., additional weights for important residues, secondary structure-dependent gap penalties).

Another route by which greater alignment sensitivity has been achieved is to abstract a pattern of allowed residues that represents a particular protein fold then use this pattern rather than the complete sequence to identify the fold in another protein (e.g., ADP-binding proteins[34]). The flexible pattern method[35] allows patterns of this type to be readily expressed and compared to any number of protein sequences. Briefly, a flexible pattern is defined in terms of a series of *n elements,* E_i, and $n - 1$ *gaps,* F_j, where each pattern starts and ends with an element (e.g., E_1, F_1,E_2,F_2,E_3,F_3,E_4). In its most general form an element is a place marker defined in terms of its position and the score obtained when it is aligned with each amino acid type. This definition allows all conventional scoring systems to be accommodated. *Gaps* are defined with a specific length range ≥ 0. For example F_1 might be set to $0,F_2$ to a value of $5 \leq F_2 \leq 12$. This definition implies that deletions within the pattern are not allowed, although deletions from the ends (where gap lengths are not explicitly stated) may occur. Figure 7 illustrates a hypothetical pattern derived from a number of different information sources. A modified Needleman–Wunsch algorithm was developed to allow the best alignment between a pattern and a sequence to be determined. This algorithm also allows any repeats of the pattern to be located within the sequence.[36]

[33] W. M. Fitch, *J Mol Bıol.* **49,** 1 (1970).

[34] R. K. Wierenga, P. Terpstra, and W. G. J. Hol, *J. Mol. Biol.* **187,** 101 (1986).

[35] G. J. Barton, *Ph D Thesis,* University of London, 1987.

[36] G. J. Barton and M. J. E. Sternberg, *J. Mol. Biol.* submitted (1989).

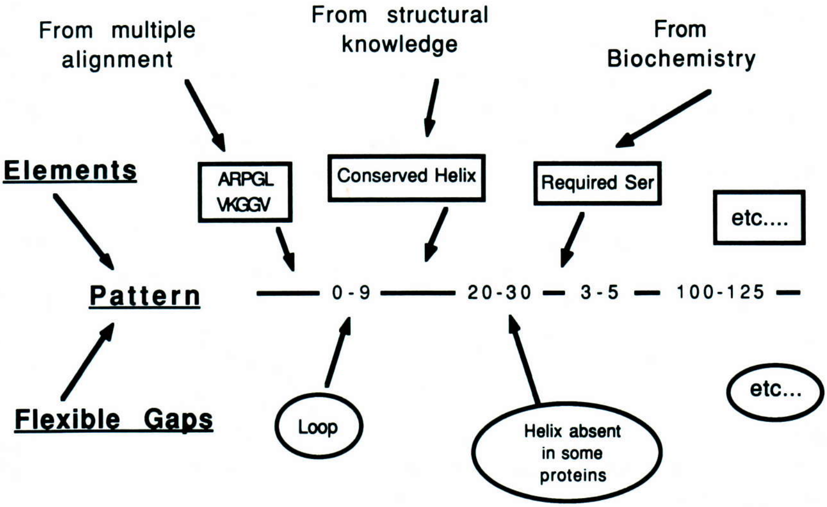

FIG. 7. Generalized flexible pattern consisting of pattern elements, which may be derived from a variety of sources defining alternative scoring schemes, and flexible gaps, which permit an allowed range of insertions but no others.

Alignment Methods Considered

If we accept that using multiple sequence data or patterns is an improvement on using a single sequence to identify similarity, it remains to decide which is the best approach and what are the limits of its sensitivity. The following approaches may be used: FASTP, the database scanning procedure of Lipman and Pearson, which uses a single query sequence and identity scoring scheme and treats gaps uniformly; NW, the Needleman–Wunsch rigorous pairwise alignment procedure, which uses a single query sequence and treats gaps uniformly; NW–SS, the NW method but with secondary structure-dependent gap penalties; BS, the multiple alignment procedure of Barton and Sternberg, in which an *alignment* of two or more sequences is optimally aligned with each entry in the database in turn and which treats gaps uniformly; BS–SS, the BS method but with secondary structure-dependent gap penalties, a procedure similar to that of Gribskov *et al.*;[37] and FP, the flexible pattern approach (with patterns derived from one sequence or a sequence alignment), which may include secondary or tertiary structural information and for which gap ranges between pattern elements are explicitly defined.

[37] M. Gribskov, A. D. McLachlan, and D. Eisenberg, *Proc Natl. Acad. Sci. U.S.A.* **84,** 4355 (1987).

Evaluation of Alignment Methods by Database Scanning

A convenient method to assess the sensitivity and selectivity of an alignment procedure is to test its ability to identify known members of a protein family from the database of all known sequences. The evaluation procedure consists of optimally aligning the query sequence(s) or pattern with each sequence in the database then rank ordering the scores. The selectivity of the method is then estimated by counting how many of the known family members have higher scores than the first nonfamily protein. The sensitivity of the procedure is shown by the overall profile of scores given for family members.

In the studies described here the globin family was used as a test system since there are a large number of entries in the database (345 complete sequences as well as 17 fragments in PIR Release 14), which vary in biological source (representatives from mammals, plants, and bacteria). Furthermore, the globins exhibit very similar protein folds, and several globin structures have been determined to high resolution by X-ray crystallography. For each scan performed, three values were determined: (1) the number of whole globins giving higher scores than the first nonglobin, (2) the number of whole globins not in group 1 but still present in the top 500 scoring sequences, and (3) the number of whole globins not in groups 1 or 2. The three-dimensional structure-based alignment of seven globins performed by Bashford *et al.*[38] was used as the information source for those algorithms that require multiple alignment data and/or secondary structural information (see Fig. 8).

Comparison of Alignment Methods

The result of the scans performed is summarized in Fig. 9. Scans 1, 2, and 3 used the complete human α-hemoglobin (HAHU) sequence to query the database. The NW procedure (scan 2) performed better than the widely used FASTP program (scan 1) both in terms of selectivity (NW placed 306 globins before first the nonglobin, whereas FASTP only 297) and sensitivity (only 31 globins not in the top 500 scores versus 41). The inclusion of secondary structural information in the form of modified gap penalties (scan 3) gave a further improvement; however, there were still 34 globins that gave alignment scores below that of a nonglobin, 25 of which were not in the top 500 sequences.

Scans 4, 5, and 6 applied multiple sequence information from the structural alignment of seven globin sequences. As expected, this additional information makes the BS algorithm (scan 4) more selective and

[38] D. Bashford, D. C. Chothia, and A. M. Lesk, *J Mol Biol* **196,** 199 (1987).

Secondary Structure	Alignment							Elements of Flexible Patterns							
								1	2	3	4	5	6	7	8
.	.	.	.	.	.	.	.								
.	.	.	.	.	.	.	.								
		V			A	G									
	V	H	V		P	A	G								
	L	L	L	L	L	L	L								
A 1	S	T	S	S	S	T	S	│	│	│	│	│	│	│	
A 2	P	P	E	A	A	E	A	│	│						
A 3	A	E	G	D	A	S	A	│	│	│					
A 4	D	E	E	Q	E	Q	Q	│	│	│	│	│	│		
A 5	K	K	W	I	K	A	R	│							
A 6	T	S	Q	S	T	A	Q	│	│	│	│	│			
A 7	N	A	L	T	K	L	V	│	│						
A 8	V	V	V	V	I	V	I	│	│	│	│	│	│	│	│
A 9	K	T	L	Q	R	K	A	│	│						
A10	A	A	H	A	S	S	A	│	│						
A11	A	L	V	S	A	S	T	│	│	│	│				
A12	W	W	W	F	W	W	W	│	│	│	│	│	│	│	│
A13	G	G	A	D	A	E	K	│	│						
A14	K	K	K	K	P	E	D	│	│						
A15	V	V	V	V	V	F	I	│	│	│	│	│	│	│	
A16	G		E	K	Y	N	A								
	A		A	G	S	A	G	**Flexible Gap**							
							N								
							D	**Explicit gap range**							
B 1	H	N	D		T	N	N	**for Pattern 1**							
B 2	A	V	V		Y	I	G								
B 3	G	D	A		E	P	A	**0 - 12 residues**							
B 4	E	E	G		T	K	G								
B 5	Y	V	H	D	S	H	V	│							
B 6	G	G	G	P	G	T	G	│	│	│	│	│			
B 7	A	G	Q	V	V	H	K	│							
B 8	E	E	D	G	D	R	D	│	│						
B 9	A	A	I	I	I	F	C	│	│	│	│	│			
B10	L	L	L	L	L	F	L	│	│	│	│	│	│	│	
B11	E	G	I	Y	V	I	I	│							
B12	R	R	R	A	K	L	K	│	│						
B13	M	L	L	V	F	V	H	│	│	│					
B14	F	L	F	F	F	L	L	│	│	│	│	│	│	│	
B15	L	V	K	K	T	E	S	│							
B16	S	V	S	A	S	I	A	│	│	│	│				
.	.	.	.	.	.	.	.								
.	.	.	.	.	.	.	.								
etc.															

FIG. 8. Derivation of flexible patterns from multiple alignment and secondary structure assignment for seven globin sequences, showing a partial pattern. The secondary structure column gives the position of α helices A and B. Alignment is based on tertiary structure comparisons. Flexible pattern elements are shown by vertical bars. Patterns 2–8 have progressively fewer elements until only the most highly conserved positions remain (see Fig. 10).

Scan number	Source of query	Method (Gap penalty)	Additional Structural Information?	Globins before first nonglobin	Globins remaining in top 500 scores	Globins not in top 500 scores
1	Single sequence (HAHU)	FASTP	No	297	7	41
2		NW(16)	No	306	8	31
3		NW-SS(16)	Yes	311	9	25
4	7 Globins (3D structure alignment)	BS(16)	No	309	19	17
5		BS-SS(16)	Yes	318	12	15
6		FP	Yes	345	0	0
7	Single sequence (HAHU)	FP	Yes	337	7	1
8	Two sequences (HAHU, GGICE3)	FP	Yes	344	1	0
9	7 globins (automatic multiple alignment)	FP	No	327	18	0

FIG. 9. Comparison of alignment procedures by database scanning with queries derived from globin sequences.

sensitive than the NW single sequence method (scan 2). Similarly, the BS–SS procedure (scan 5) further improved on the results obtained by the NW–SS method (scan 3) by identifying 318 globins before the first nonglobin (cf. 311), with only 15 globin sequences not among the top 500 scores (cf. 25). The most startling improvement in performance, however, was obtained by the flexible pattern method (FP, scan 6) which gave perfect selectivity for globins with no nonglobin sequences scoring higher than the 345 whole globins in the database.

The successful scan 6 used a pattern that consisted of 107 pattern elements and 5 flexible gaps (Pattern 1, Fig. 8). The elements consisted of all aligned positions that had no gaps in any of the sequences and were also within secondary structural regions, while the scoring scheme took mean values from the MDM over all seven aligned sequences. Scan 7 used the same pattern elements as scan 6, but the scores were derived from only the human α-hemoglobin sequence (HAHU). This scan also performed significantly better than the multiple alignment method (scan 5), with only eight sequences not scoring higher than a nonglobin. It confirms that the bulk of the improvement from using the FP method comes from discarding the variable regions of the protein sequence, rather than from the use of multiple sequences.

The small deficiency in scan 7 is virtually eliminated by including one further sequence when describing the pattern elements. Scan 8 illustrates the result of this scan using a pattern derived from HAHU and GGICE3; only the bacterial hemoglobin fails to score higher than a nonglobin.

Pattern 1 has 107 elements or 79% of the shortest sequence. In order to investigate whether this high percentage of the alignment was actually required, a series of seven patterns (see Fig. 8) with successively fewer elements was derived and tested against the database. The result of scanning each pattern is summarized in Fig. 10. As expected, the overall trend in sensitivity and selectivity is downward as fewer elements are included. However, even pattern 6, which contains only 28 elements (21% of the shortest sequence), performs better than the full multiple alignment BS–SS method (335 globins before first nonglobin, cf. 318 for scan 5). Values in parentheses are for patterns in which the flexible gaps are unconstrained. The poorer performance of these patterns demonstrates the importance of defining flexible gaps to model the observed variation in sequence length within a protein family.

Derivation of Flexible Pattern When No Three-Dimensional Structure Is Known

In general, a sequence family may be known but with no details of three-dimensional structure available to guide the alignment or derivation of a pattern. Can an effective pattern be derived from just the sequences? To answer this question the seven globins used for scans 4–6 were multiply aligned by the single order algorithm described above; pairwise scores clustered at 7.9 S.D., suggesting confidence in the alignment. All positions at which gaps occurred were discarded, and, of the remaining positions, only those that had conservation values above 0.4 were maintained. Finally, gaps were made flexible between elements where insertions and

Pattern Number	Conservation Number Cutoff	Number of pattern elements	Percentage of shortest sequence	Globins before first nonglobin (total in PIR = 346)	Globins remaining in top 500 scores	Globins not in top 500 scores
1	0.0	107	79	345 (343)	0 (2)	0 (0)
2	0.2	87	64	345 (343)	0 (0)	0 (2)
3	0.3	61	45	343 (341)	2 (2)	0 (2)
4	0.4	46	34	344 (329)	1 (13)	0 (3)
5	0.5	38	28	343 (318)	1 (22)	1 (5)
6	0.6	28	21	335 (306)	9 (19)	1 (20)
7	0.7	15	11	295 (0)	33 (281)	18 (64)
8	0.8	8	6	1 (0)	298 (281)	46 (67)

FIG. 10. Result of scans using patterns with progressively fewer elements (see Fig. 8) derived from the seven-globin alignment[38] at increasing conservation value cutoffs.

deletions had been included by the automatic alignment algorithm, but were kept to fixed lengths where no insertions/deletions were observed. The resulting pattern consisted of 39 elements and, when scanned against the PIR database (scan 9), scored all globins in the top 500, with 327 globins giving scores above nonglobins. Thus, flexible patterns can be derived purely from sequence information and show a useful improvement in sensitivity and selectivity over the multiple alignment method (scan 4, 17 globins not in top 500 scores, only 309 globins before first nonglobin), or conventional single sequence methods.

Implementation and Availability of Programs

The techniques described in this chapter are all implemented in the AMPS package (alignment of multiple protein sequences), which provides the functions described together with additional features for multiple sequence manipulation and analysis within an easy-to-use environment. The package is available for a nominal fee to academic users. It is implemented on a VAX/VMS or Sun 3 with fp68881 coprocessor, and the current program limits for multiple alignment are 250 sequences of up to 1200 amino acids in length.

Acknowledgments

I would like to thank Dr. Mike Sternberg for support and advice. I would also like to thank Drs. Marketa Zvelebil, Chris Rawlings, and Prof. Tom Blundell. This work was supported by the Science and Engineering Research Council (UK) and The Imperial Cancer Research Fund.