Flexible Protein Sequence Patterns

A Sensitive Method to Detect Weak Structural Similarities

Geoffrey J. Barton[1,3]† and Michael J. E. Sternberg[2,3]

[1]*Biomedical Computing Unit and* [2]*Biomolecular Modelling Laboratory*
Imperial Cancer Research Fund Laboratories
P.O. Box No. 123
Lincoln's Inn Fields
London WC2A 3PX, U.K.

[3]*Laboratory of Molecular Biology*
Department of Crystallography
Birkbeck College, London WC1E 7HX, U.K.

(Received 10 August 1989; accepted 28 November 1989)

The concept of a flexible protein sequence pattern is defined. In contrast to conventional pattern matching, template or sequence alignment methods, flexible patterns allow residue patterns typical of a complete protein fold to be developed in terms of residue positions (elements), separated by gaps of defined range. An efficient dynamic programming algorithm is presented to enable the best alignment(s) of a pattern with a sequence to be identified. The flexible pattern method is evaluated in detail by reference to the globin protein family, and by comparison to alignment techniques that exploit single sequence, multiple sequence and secondary structural information. A flexible pattern derived from seven globins aligned on structural criteria successfully discriminates all 345 globins from non-globins in the Protein Identification Resource database. Furthermore, a pattern that uses helical regions from just human α-haemoglobin identified 337 globins compared to 318 for the best non-pattern global alignment method. Patterns derived from successively fewer, yet more highly conserved positions in a structural alignment of seven globins show that as few as 38 residue positions (25 buried hydrophobic, 4 exposed and 9 others) may be used to uniquely identify the globin fold. The study suggests that flexible patterns gain discriminating power both by discarding regions known to vary within the protein family, and by defining gaps within specific ranges. Flexible patterns therefore provide a convenient and powerful bridge between regular expression pattern matching techniques and more conventional local and global sequence comparison algorithms.

1. Introduction

The most successful approach to prediction of the structure and function of a protein is to identify similarities between the sequence of the molecule and other well-characterized proteins. When a strong sequence similarity exists with a protein of known three-dimensional structure, then model-building techniques may be applied with some success (e.g. see Browne *et al.*, 1969; Blundell *et al.*, 1987). Even when there is incomplete similarity to a protein of known three-dimensional structure, the observation of conserved motifs (e.g. the E-F hand calcium-binding loop (Tufty & Kretsinger, 1975; Argos *et al.*, 1977), or β-α-β dinucleotide binding fold (Wierenga *et al.*, 1986)) can provide important clues to the identification of the functional domains of the protein, and give guidance for the design of site-specific mutagenesis studies. When the structure of at least one member of a protein family is known, multiple alignments of the sequences can provide insights into the tolerance of substitutions within the protein core and on the surface regions (Bashford *et al.*, 1987).

In the absence of a crystal structure, observations of residue conservation, and the location of insertions/deletions in multiple alignments of sequences, can suggest the location of loop regions and core secondary structures (Zvelebil *et al.*, 1987; Sternberg *et al.*, 1987; Crawford *et al.*, 1987).

† Author to whom all correspondence should be addressed at: Laboratory of Molecular Biophysics, The Rex Richards Building, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

0022-2836/90/060389-14 $03.00/0

Conventional global sequence comparison methods give good alignments when the best score for the comparison of the sequences is greater than 6·0 standard deviations from the mean of scores for shuffled sequences (Barton & Sternberg, 1987*a*,*b*). However, when the similarities are weaker, or confined to short stretches separated by variable regions, global alignment methods can fail to give alignments with scores significantly higher than for randomized or unrelated sequences. Alignments obtained under these circumstances are, at best, unpredictable in their quality. Several authors have considered this problem, and suggested the use of multiple sequence alignments, in some cases together with secondary or tertiary structural information, to encapsulate the principal features of a protein or domain fold. Taylor (1986*a*) describes the use of "consensus templates" in his analysis of the immunoglobulins, and in the identification of an alignment of the weakly similar retroviral proteases and non-viral aspartyl proteases (Pearl & Taylor, 1987). Bashford *et al.* (1987) define "templates" that use information gleaned from a detailed study of the globins, whilst Gribskov *et al.* (1987, 1988) derive "profiles" for globin and immunoglobulin variable domain families. Patthy (1987) also discusses a similar procedure and its use in the superfamily of complement related repeats, while Staden (1988) describes a programme that combines many features of these methods.

The method of Gribskov *et al.*, in common with our earlier work (Barton & Sternberg, 1987*a*) and that of Lesk *et al.* (1986), permits lower gap penalties to be applied within regions known to be variable within the family. However, none of these pattern or template definitions allows the explicit encoding of the observed variability in length between conserved regions in protein families. Regular expression pattern-matching techniques can encode flexible length gaps (e.g. see Abarbanel *et al.*, 1984; Lathrop *et al.*, 1987) but, since they do not permit arbitrary weights to be assigned to the alignment of a pattern element and a character, they are of limited use for protein sequence comparison.

In this paper we define the flexible pattern, a concept that allows any chosen position-specific scoring or weighting scheme to be applied, and it permits regions between weighted positions to be explicitly defined to have a range of lengths. An efficient algorithm is presented for the location of the best scoring alignment between a flexible pattern and another sequence (the target). This method also permits repeats and sub-optimal matches to be explored. The method is systematically evaluated for the globin family of proteins by comparison to conventional pairwise alignment techniques and techniques that exploit structural and multiple sequence information. The limits in sensitivity of flexible patterns are explored, and we suggest that less than 30% of the protein sequence may be used to identify uniquely all other family members.

2. Methods

(a) *Flexible pattern definition*

Fig. 1 provides an overview of a hypothetical flexible pattern and illustrates its general features. The pattern is made up of alternating elements and gaps. The elements may be derived from a multiple alignment, structural, biochemical or other information, whilst the gaps signify the exclusion of variable regions that have predetermined length ranges. An important feature of the pattern is that not every residue in the protein used to derive the pattern is represented.

Each element represents a single residue position, whilst gaps define the allowed range of residues between the elements (including zero). Thus, a fixed length template of 5 amino acids length, e.g. as used by Taylor (1986*b*), may be defined by 5 elements, separated by 4 gaps of length zero. Such a series of 5 elements is illustrated in Table 1. A lookup table (Table 1) defines the score obtained for aligning each element with each type of

Table 1
A simple flexible pattern

Pattern segments (E_i)			Gap lengths (F_i) Min	Max	Lookup table ($T_{N,23}$) (partial)								
					A	R	G	L	P	E	D	W	F. . .
1	A	A			1·0	0	0	0	0	0	0	0	0
			0	2									
2	R	G			0	1·0	1·0	0	0	0	0	0	0
			0	1									
3	P	A			1·0	0	0	0	1·0	0	0	0	0
			0	5									
4	D	E			0	0	0	0	0	1·0	1·0	0	0
			0	0									
5	W	F			0	0	0	0	0	0	0	1·0	1·0

A simple flexible pattern consisting of 5 elements (E_i) derived from an alignment of 2 protein sequences. Four gaps (F_i) are defined to have minimum and maximum values, whilst each row of the lookup table identifies the score for aligning an element with each amino acid type. In this example, the simple identity scoring scheme is used (see the text).

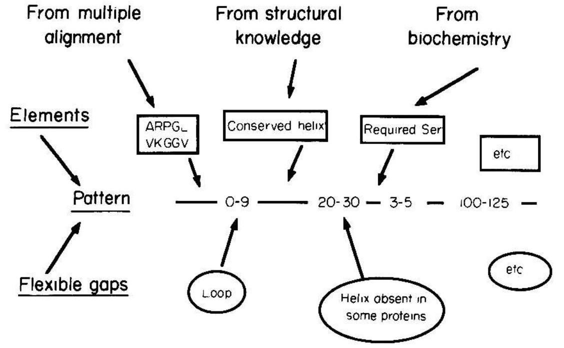

Figure 1. Hypothetical flexible pattern showing the potential alternative sources of information used in defining the elements, and flexible gaps.

amino acid. Values for the lookup table may be obtained automatically by application of a scoring scheme (see below) to each element, or by manually defining weights to highlight particular structural or functional features.

A simple flexible pattern is illustrated in Table 1. The pattern consists of 5 elements (E_i); in this example, each element is represented by pairs of amino acids from the alignment of 2 protein sequences. The minimum and maximum values allowed for each flexible gap are shown, and part of a lookup table ($T_{N,23}$) that defines the score for matching each amino acid type with each element of the pattern. In this pattern, the Table was derived from the alignment by applying identity IDE scoring (see below).

More formally; a pattern P is defined in terms of a series of N elements E_i and $N-1$ gaps, F_i 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). This definition allows all conventional scoring systems to be accommodated. Thus, an element might represent a specific amino acid (e.g. glycine), group of amino acids derived from a multiple alignment (e.g. VALAVLLG) or more general properties (e.g. hydrophobic). 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. For example, element E_1 is in the first position of a pattern, and might be defined to give scores of 20·5, −10·0, 15·0 when matched with Trp, Glu, Phe, respectively, and −50·0 when matched with all other amino acids.

Gaps are defined to have a specific length range $\geqslant 0$. For example, the 1st gap (F_1) might be set to 0, the 2nd to have a value of $5 \leqslant F_2 \leqslant 12$ and the 3rd (F_3) a value of $110 \leqslant F_3 \leqslant 110$. 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.

(b) *Scoring schemes*

Five scoring schemes that may be applied directly to an alignment of protein sequences were considered:

(1) Dayhoff scoring (DAY) uses the Dayhoff MDM_{78} pairscore matrix (Dayhoff, 1978). The score for aligning a particular amino acid with an element is given by the mean score for aligning that amino acid with each amino acid in the element (after Barton & Sternberg, 1987*b*).

(2) Conservation scoring (CON) is based upon the quantification by Zvelebil *et al.* (1987) of Taylor's Venn diagram of amino acid properties (Taylor, 1986*a*), as represented by conservation numbers ranging from 0 (totally unconserved properties) to 1 (totally conserved properties).

(3) Identity scoring (IDE) gives a score of 1 if the amino acid at the current position is also in the pattern element, or 0 if it is not (e.g. see Table 1).

(4) In frequency scoring (FRE), the number of each type of amino acid in a pattern element is counted, and this value is used when matching with an amino acid of that type. Thus, if an element contains 30 Ala, 1 Glu and 2 Gly residues, it will score 30 for matching Ala, 1 for matching Glu, 2 for Gly and 0 for all other amino acid types.

(5) Weight scoring (WGT) uses the protocol of Dodd & Egan (1987), which normalizes frequency scores by the relative abundance of the amino acids in the database. In this study, the abundancies were taken from the Doolittle (1981) NEWAT database.

The exact manner in which the elements and gaps of a flexible pattern (FP) are derived, is dependent upon the information available. A single sequence may be sufficient if the protein is of known three-dimensional structure. For example, residues defining the core secondary structures of the protein might be selected as the pattern elements, together with known catalytic or other key residues. Alternatively, if 2 or more homologous sequences are available, but no structure, then a pattern may be derived by selecting the conserved positions identified in a multiple alignment of the sequences. Examples of these derivation methods are evaluated below, and compared to conventional sequence alignment techniques.

(c) *Algorithm to locate the best alignment(s) of the pattern with a sequence*

Given the flexible pattern, a method is required to identify the sequences that most closely match the pattern, and to generate an alignment of the pattern with each sequence.

If relatively few flexible gaps are defined, then the most straightforward approach is to generate all possible fixed length patterns, then use a simple protocol to compare each pattern in turn to the database. However, since the number of patterns that must be defined is given by the product of the flexible gap ranges, even the simple flexible pattern shown in Table 1 would require 36 fixed-length patterns to be stated explicitly (i.e. 3 × 2 × 6 × 1). A pattern defining the Kringle domain found in plasminogen and related molecules, and containing 14 flexible gaps, would need $\sim 10^{10}$ fixed-length patterns to represent it (Barton, 1987). A less cumbersome method based upon an extension of the Needleman & Wunsch (1970) dynamic programming algorithm was, therefore, developed.

The operation of the algorithm is illustrated in Table 2, for a comparison between the pattern and lookup table shown in Table 1, and a sequence of 10 amino acids (A_{10}).

(1) A matrix $R_{M,N}$ is derived by reference to the lookup table such that each element $R_{i,j}$ holds the score for the comparison of A_i with E_j. For example, residue 3 (A) scores 1 when aligned to element 1 (E_1).

(2) The matrix $R_{M,N}$ is converted to $S_{M,N}$ column by column using a procedure similar to the Needleman & Wunsch (1970) algorithm, but constrained by the specified gap lengths (F), and the restriction which disallows deletions within the pattern. This process is shown partially completed in Table 2A. The matrix element currently being processed ($S_{4,2}$) takes the value for a comparison of element E_2 (R, G) with residue A_4 (G), plus

the maximum value from the previously processed column of the matrix that is within the specified gap range ($0 \leqslant F_2 \leqslant 1$).

(3) Once the matrix S is complete (Table 2B), the elements in the 1st column contain the best score for an alignment of the pattern P, or the beginning of P with A starting at A_i. The elements of the 1st row (for $j > 1$) contain the best score for an alignment of the tail end of P starting at E_j. As with the Needleman & Wunsch (1970) algorithm, the best overall score is given by the maximum value in the 1st row or column of S. In addition, a traceback through the matrix as illustrated in Table 2B allows the best scoring alignment of P and A to be generated.

In the example illustrated, there is only 1 alignment with score 5. In general, however, it is possible for there to be several alignments with scores the same or similar to the best, but starting in different elements of the 1st column of S. These may represent possible repeats of P in A. In addition, as with a Needleman & Wunsch (1970) comparison, there is the possibility of more than 1 alignment starting in a given element of S. When scanning a database, it is sufficient to know initially just the best match score between a pattern and each sequence. Having identified a high scoring match, possible alternative alignments may then be investigated using the same algorithm.

Table 2

Illustration of the algorithm to locate the best score for aligning the flexible pattern from Table 1 with a ten amino acid sequence

A. *Partially completed matrix*

Min : max 0.2 0:1 0.5 0.0 Gaps (F_j)

	A A	R G	P A	D E	W F	Segments (E_i)
V			2			1
I			2			2
A	1		3			3
G		4	2			4
T		3	2			5
P		1	3			6
E		1	1	2		7
F		1	1		1	8
S			1			9
D				1		10
	1	2	3	4	5	

B *Traceback to give best alignment*

V	3	3	2		
I	3	3	2		
A	5	2	3		
G	3	4	2		
T	1	3	2		
P	1	1	3		
E	1	1	1	2	
F		1	1		1
S			1		
D				1	

In practice, steps 1 and 3 of the algorithm are combined. Furthermore, if only 1 alignment for each element of S is required (usually that with the shortest gaps), then the entire S matrix need not be stored, but merely 2 arrays of length m and a pointer array that can be of a compact data type (e.g. 2 byte integer). This considerably reduces the computer time and memory required to trace out the best path for alignment, and was the procedure adopted for this study. Further memory savings may be made if no alignment is required, since no pointer array is then needed.

The computer program that implements the flexible pattern method is written in Fortran-77 and integrated with the AMPS (Alignment of Multiple Protein Sequences: Barton, 1990) package for multiple protein sequence alignment. Thus, a multiple alignment of related sequences may be generated rapidly (Barton & Sternberg, 1987*b*), then the alignment edited to define the elements and flexible gaps of a pattern. This pattern may then be used to scan the database to identify weaker family relationships.

(d) *Efficiency considerations*

Regular expression pattern-matching algorithms can run very efficiently, even when flexible gaps are allowed (e.g. see Abarbanel *et al.*, 1984). However, these methods rely for their speed on the ability to abandon a comparison when an element of the pattern does not exactly match the target sequence. When weighted matching is required, these speeding devices cannot be applied.

As discussed, the approach of generating all patterns, then testing each against the target sequence, is one solution. However, the number of steps required by this approach is approximately proportional to the product of the length of flexibility in each gap. In contrast, the algorithm presented in this paper requires:

$$\sum_{i=1}^{N-1} U_i(M - U_i)$$

steps, where N is the number of elements in the pattern, M is the number of residues in the target sequence, and U_i is the range of flexibility in the ith gap. For example, to find the best match of a pattern containing 10 flexible gaps, each with a range of 5, to a sequence of 1000 residues would require $\sim 5 \times 10^5$ steps, compared to $\sim 10^{10}$ steps for the "generate all patterns and test" method. The flexible pattern scans of PIR 14·0 presented in this study required from 8 min (pattern 8, Table 3) to 36 min (pattern 1, Table 3) VAX 8700 central processing unit (c.p.u.) time, using code, not fully optimized for speed.

(e) *Alternative methods for the identification of proteins with similar folds*

Given the sequence of one protein, there are several sequence comparison techniques that may be applied in order to identify other proteins in the database that may have similar complete folds.

(1) The Needleman & Wunsch (1970) algorithm (NW) gives a score guaranteed to be the best possible for the global alignment of 2 sequences. When calculating the score, gaps inserted within the alignment are penalized, but gaps at the ends are not. If the gap penalty is set sufficiently high, this feature allows the algorithm to be used to search for a common domain within a much longer sequence, with little risk of generating an unrealistically large number of gaps.

(2) FASTP (Lipman & Pearson, 1985) is a widely used program that, in order to gain speed and permit implementation on small computers, does not perform the rigorous search of the Needleman & Wunsch (1970) algorithm. Speed is obtained by performing an initial screen for identical amino acids, followed by a restricted optimization scoring with the MDM_{78} matrix.

(3) The Needleman & Wunsch (1970) algorithm with secondary structure dependent gap penalties (NW-SS), allows the probability of inserting a gap in a core secondary structure to be reduced. This gives a better model of the observed preferences for gaps in non-core loops in protein families, and can yield useful improvements in alignment accuracy (Barton & Sternberg, 1987*a*; Lesk *et al.*, 1986), providing that the 3-dimensional structure of the query sequence is known.

If 2 or more clearly related sequences are available, then techniques based upon the more sensitive multiple alignment procedures are possible.

(4) Barton & Sternberg's (1987*b*) method (BS) utilizes the additional information from an alignment of 2 or more sequences when optimizing the alignment with each sequence in the database. The method applies an adapted Needleman & Wunsch (1970) algorithm for each alignment-query *versus* database-entry comparison.

(5) If the 3-dimensional structure of at least 1 of the aligned sequences used as a query is known, then it is possible to incorporate secondary structure dependent gap penalities in ther BS method, giving the BS-SS method.

(f) *Evaluation of alignment methods by database scanning: the globins*

An alignment procedure may be assessed using a well-characterized protein family, by considering the methods' ability to identify members of the family against the background of all known protein sequences. The evaluation procedure consists of optimally aligning the query pattern, sequence or alignment against every sequence in the database, then rank ordering the scores. The specificity of the method is then estimated by counting how many of the known family members have higher scores than the first non-family protein, whilst the sensitivity of the procedure is shown by the overall profile of scores given for the family members.

A sensitive procedure may yield consistently high scores for the family members, yet give poor specificity, by giving equivalent scores for non-family proteins. The ideal query has perfect specificity, where all family members give scores greater than any other sequence.

Before any comparison of methods can be made, it is important to know which proteins belong in the family and which do not. For this reason, the well-characterized globin family was used as a test system. The globins have the advantage of being well represented in the PIR database (George *et al.*, 1986: 345 complete sequences as well as 17 fragments in PIR release 14·0), with sequences from varied biological sources (including representatives from mammals, plants, annelids and bacteria). Furthermore, several members of the family have been characterized by X-ray crystallography at high resolution indicating that, despite considerable sequence divergence, all members possess very similar 3-dimensional structures.

Every scan of the PIR 14·0 database generates 6418 scores, each of which represents the optimal score for aligning the query with a sequence. A query with perfect specificity will yield a score distribution where all the globins give higher scores than non-globins. A poorer query will yield a distribution where some globins give scores lower than known non-globins. Rather than present the entire score distribution for each scan performed, in this study, the results are represented by 3 values.

Value 1. The number of globins giving higher scores than the 1st non-globin.

Value 2. The number of globins not in value 1 but still in the top 500 scoring sequences.

Value 3. The number of globins not found in the top 500 scoring sequences.

Value 1 illustrates the specificity of the method, whilst values 2 and 3 show the gross features of the score distribution. For example, query *A* might give values of 300, 10 and 35; whilst query *B* gave 300, 35 and 10. Although both queries score the same number of globins before non-globins (300), *B* is a more sensitive method, since the distribution for globins is more skewed towards higher scores.

Since the aim is to identify proteins that contain the complete globin fold, the 17 database entries listed as fragments were excluded from the evaluation. Methods requiring a single query sequence were tested using human α-hemoglobin. Those methods requiring secondary structural information and/or more than 1 query sequence, drew upon the information in the 3-dimensional structure-based alignment of 7 globins shown by Bashford *et al.* (1987).

3. Results

(a) *Comparison of alignment methods*

Table 3 summarizes the result of scans using the six methods considered. Scans 1 to 3 utilized human α-hemoglobin as the query. The scan using FASTP (scan 1) reported only 297 globins before the first non-globin, with 41 globins not in the top 500 scores. Scanning with the Needleman & Wunsch (1970) method (scan 2) yielded a small improvement, but still 31 globins were not found in the top 500 scores. A further improvement was obtained by the inclusion of secondary structure-dependent gap penalties (scan 3) with 311 globins scoring above the first non-globin, and 25 not found in the top 500 scores.

Scans 4 to 6 all utilize the additional information from the seven-sequence structural alignment given by Bashford *et al.* (1987). Scanning with the alignment, but no explicit secondary structural information in the gap penalty identified 309 globins before the first non-globin (scan 4). This is slightly worse than the best single-sequence scan (scan 3); however, the sensitivity is better, since only 17 globins failed to score in the top 500 sequences (cf.

Table 3
Database scans using queries derived from globin sequences

Scan number	Source of query	Method (gap penalty)	Additional structural information?	Globins before first non-globin	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	Seven 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	Seven globins (automatic multiple alignment)	FP	No	327	18	0

The result of database scans against the PIR 14·0 sequence database (6418 sequences, 345 globins) using queries derived from globins, but with different comparison methods. Scan number, index used in the text. Source of query, the sequence, or alignment that was used to derive the query. Method, comparison technique, as described in text, figures in parentheses refer to the length-independent gap-penalty employed. Additional structural information, Yes if the method includes the explicit definition of secondary structure positions, otherwise No. See the text for explanation of Globins before first non-globin, Globins remaining in top 500 scores and Globins not in top 500 scores. Scans 3 and 5 include secondary structure-dependent gap penalties. The penalty shown was multiplied by a factor of 4 within the helical regions, scans using a factor of 10 gave identical results. 3D, 3-dimensional.

25). The inclusion of secondary structural information in the gap penalty gave a further improvement in the specificity for globin sequences (scan 5), with 318 globins before the first non-globin and only 15 globins not in the top 500 scores.

The flexible pattern utilized in scan 6 was derived directly from the seven-globin alignment. The pattern elements consisted of positions for which each of the seven sequences had a residue in an observed helix. Flexible gaps were then defined within the observed ranges of allowed loop connection (±4), between the conserved helices. For example, the longest loop connection between helix A and B is eight residues and the shortest is two. The flexible gap at this position was therefore given the range 0 to 12. This pattern gave the perfect result by scoring all 345 globin sequences in the database, before a non-globin. Scans using alternative scoring schemes based upon the frequency of occurrence of amino acids in the pattern, either normalized (WGT scoring: Dodd & Egan, 1987) or not, gave equivalent results. Scans using scoring schemes based upon amino acid identity, or physical properties performed slightly less well. Identity scoring (IDE) gave 339 globins before the first non-globin with one not scoring in the top 500 sequences, whilst physical property scoring based upon conservation numbers (CON) identified 337 globins before the first non-globin, with all globins in the top 500 scoring sequences. Although not perfect, both these scoring systems performed better than the best non-pattern method (scan 5).

(b) *Flexible pattern: why is it more effective?*

The flexible pattern used in scan 6 incorporates the sequence and secondary structural information from seven well-characterized proteins. It might justifiably be expected to out-perform techniques that utilize only part of this information. However, the BS-SS method (scan 5) also makes use of similar secondary structural and aligned sequence information yet does not perform as well. From where then is the benefit coming?

There are two principal differences between the pattern and the multiple alignment; the incorporation in the pattern of only the structurally conserved regions, and the specification of gaps to be permissible only over a specific range. Together, these have the effect of removing the background "noise" associated with matching to the more variable loop regions, and reducing the chance of a spurious good match with a long sequence. These factors are illustrated by the results of scans 7 and 8 (Table 3). Scan 7 makes use of a flexible pattern that shares the same elements and flexible gaps as that used in scan 6. However, instead of deriving scores from all seven aligned sequences, only the residues present in human α-hemoglobin (HAHU†) were used. The pattern performs almost as well as the scan 6 pattern, with only one globin sequence not identified in the top 500.

Given the encouraging results of scan 7, can the single-sequence pattern be improved by incorporating the additional information from one more protein? Pairwise comparison of the seven globin sequences indicates that overall, GGICE3 is the least similar to HAHU. The residues from this sequence when aligned with HAHU will, therefore, give an indication of the range of variability permitted at each position. A pattern that makes use of the same elements, and flexible gaps as scans 6 and 7, yet uses residues from both GGICE3 and HAHU when scoring, gave near-perfect results (scan 8). All but one globin was identified before the first non-globin, and all globins scored in the top 500 sequences.

Scans 7 and 8 demonstrate the utility of using a flexible pattern, even when the information from only one or two proteins of known tertiary structure is available.

(c) *Sensitivity to number of pattern elements and flexible gap lengths*

The successful pattern applied in scan 6 consists of 107 pattern elements, or 79% of the shortest sequence (GGICE3, 136 amino acids) in the set. In order to discover whether this high percentage of the alignment was actually required to give total discrimination for the globins, seven further patterns were developed containing successively fewer elements.

In each example, the pattern elements were defined by making use of the concept of "conservation numbers" at each aligned position. The derivation of these numbers was described by Zvelebil *et al.* (1987) and the numbers range from zero, for poor conservation, to one, for total identity at an aligned position. For example, a score of 0·9 would mean that all physico-chemical properties are conserved (by Taylor's (1986*a*) definitions), yet the amino acids are not all identical. Conservation numbers provide a convenient numerical scale to classify positions in a multiple alignment. Accordingly, patterns 2 to 8 (Table 4) were derived by imposing successively higher conservation number cutoffs to the original pattern. Thus, pattern 4 consists of only those elements giving conservation scores ⩾ 0·4 and pattern 8 only those scoring ⩾ 0·8. For every pattern, the length of each allowed flexible gap was increased to accommodate the removal of pattern elements from the ends of helical regions. Where pattern elements are deleted from within helical regions, a fixed length gap of equivalent length was inserted. For example, with reference to Table 4, elements A2 and A3 are not present in pattern 4, so a fixed gap of length 2 is allowed between elements A1 and A4.

Table 5 illustrates the effect of reducing the number of defined elements in the pattern. Clearly, there is a decrease in sensitivity and specificity as the number of defined positions is reduced. However, even pattern 6 with only 28 elements identifies 335 globins before the first non-globin; a distinct improvement over the best non-pattern method in scan 5 (318 globins).

In addition to patterns with flexible gaps defined within specific ranges, eight patterns were developed having totally flexible gaps (i.e. allowed gap ranges from zero to the total length of the target sequence). The result of scanning these patterns is illustrated in parentheses in Table 5. As expected, patterns without constrained gap lengths give consistently poorer specificity than the constrained patterns. The specificity decays faster as the number of pattern elements is reduced.

(d) *Structural features of the pattern elements*

Patterns 2 to 8 were all derived purely from application of conservation values to pattern 1, without drawing on knowledge of the protein three-dimensional structure. But how does the choice of pattern elements relate to the protein tertiary structure? Bashford *et al.* (1987) classified 32 common hydrophobic sites where residues are buried, and highly conserved throughout all globins (including the absolutely conserved Pro at C2). They also identified 32 conserved sites where residues are exposed (Table 4). Pattern 1 includes all these positions plus 45 other elements. As the conservation cutoff is made more severe, from pattern 2 through to pattern 5, the total number of pattern elements is reduced from 107 to 38 (Table 4). However, of the 67 elements discarded, 42% are exposed positions, 51% "others" and only 7% buried hydrophobics. Thus, pattern 5 consists of 38 elements, where 25 are buried, four are exposed, and there are nine others. The observation that pattern 5 gives good discrimination for globins (343 before 1st non-globin, 1 not in top 500), suggests that it is principally the conserved hydrophobic elements that confer the pattern specificity. However, some sites additional to the 32 conserved hydrophobic positions identified by Bashford *et al.* (1987) are important, since a pattern using just the 32 elements found only 327 globins before the first non-globin (pattern not shown).

(e) *Globins that score lower than non-globins*

Table 6 itemizes those globins that did not score above a non-globin when optimally aligned with patterns 3 to 7. The globins that are missed by patterns 3, 4 and 5 are from the marine worm

† Abbreviation used: HAHU, human α-hemoglobin.

Table 4
Structural alignment of seven globins

									Flexible patterns							
Acc	Structure	Alignment							1	2	3	4	5	6	7	8
							P									
							I									
							V									
							D									
							T									
							G									
							S									
							V									
			V			A	G									
		V	H	V		P	A	G								
		L	L	L	L	L	L	L								
	A1	S	T	S	S	S	T	S	|	|	|	|	|	|	|	
	A2	P	P	E	A	A	E	A	|	|						
	A3	A	E	G	D	A	S	A	|	|	|					
e	A4	D	E	E	Q	E	Q	Q	|	|	|	|	|	|		
	A5	K	K	W	I	K	A	R	|							
e	A6	T	S	Q	S	T	A	Q	|	|	|	|	|			
	A7	N	A	L	T	K	L	V	|	|						
b	A8	V	V	V	V	I	V	I	|	|	|	|	|	|	|	|
	A9	K	T	L	Q	R	K	A	|	|						
e	A10	A	A	H	A	S	S	A	|	|						
b	A11	A	L	V	S	A	S	T	|	|	|	|				
b	A12	W	W	W	F	W	W	W	|	|	|	|	|	|	|	|
e	A13	G	G	A	D	A	E	K	|	|						
e	A14	K	K	K	K	P	E	D	|	|						
b	A15	V	V	V	V	V	F	I	|	|	|	|	|	|		
	A16	G		E	K	Y	N	A								
		A		A	G	S	A	G								
								N								
								D				Flexible gap				
	B1	H	N	D		T	N	N								
	B2	A	V	V		Y	I	G								
	B3	G	D	A		E	P	A								
	B4	E	E	G		T	K	G								
	B5	Y	V	H	D	S	H	V	|							
	B6	G	G	G	P	G	T	G	|	|	|	|	|			
	B7	A	G	Q	V	V	H	K	|							
	B8	E	E	D	G	D	R	D	|	|						
b	B9	A	A	I	I	I	F	C	|	|	|	|	|			
b	B10	L	L	L	L	L	F	L	|	|	|	|	|	|	|	
	B11	E	G	I	Y	V	I	I	|							
e	B12	R	R	R	A	K	L	K	|	|						
b	B13	M	L	L	V	F	V	H	|	|	|					
b	B14	F	L	F	F	F	L	L	|	|	|	|	|	|	|	
	B15	L	V	K	K	T	E	S	|							
	B16	S	V	S	A	S	I	A	|	|	|	|				
												Flexible gap				
	C1	F	Y	H	D	T	A	H	|							
s	C2	P	P	P	P	P	P	P	|	|	|	|	|	|	|	|
	C3	T	W	E	S	A	A	Q	|	|						
b	C4	T	T	T	I	A	A	M	|	|	|	|	|			
	C5	K	Q	L	M	Q	K	A	|	|						
e	C6	T	R	E	A	E	D	A	|	|						
	C7	Y	F	K	K	F	L	V	|	|	|					
b	CD1	F	F	F	F	F	F	F	|	|	|	|	|	|	|	|
e	CD2	P	E	D	T	P	S	G	|	|						
	CD3		H	S	R	Q	K									
	CD4	F	F	F	F	F	F	F								
			G	K	A	K	L	S								
		D	D	H	G	G	K	G								
		L	L	L		L										
		S	S	K	K	T	G									
	D1		T	T	D	T	T									
	D2		P	E	L	A	S									
	D3		D	A	E	D	E	A				Flexible gap				
	D4		A	E	S	Q	V	S								
	D5		V	M	I	L	P									
	D6		H	M	K	K	K	Q								
	D7		G	G	A	G	K	N								

Table 4 *(continued)*

												Flexible patterns				
Acc	Structure	Alignment							1	2	3	4	5	6	7	8
e	E1	S	N	S	T	S	N	D	│	│	│	│	│			
e	E2	A	P	E	A	A	P	P	│	│						
e	E3	Q	K	D	P	D	E	A	│							
b	E4	V	V	L	F	V	L	V	│	│	│	│	│	│		
e	E5	K	K	K	E	R	Q	A	│	│						
e	E6	G	A	K	T	W	A	D	│							
	E7	H	H	H	H	H	H	L	│	│	│	│				
b	E8	G	G	G	A	A	A	G	│	│	│	│	│	│	│	│
e	E9	K	K	V	N	E	G	A	│							
e	E10	K	K	T	R	R	K	K	│	│	│	│	│			
b	E11	V	V	V	I	I	V	V	│	│	│	│	│	│	│	│
b	E12	A	L	L	V	I	F	L	│	│	│	│	│			
e	E13	D	G	T	G	N	K	A	│	│						
b	E14	A	A	A	F	A	L	Z	│	│	│					
b	E15	L	F	L	F	V	V	I	│	│	│	│	│	│		
	E16	T	S	G	S	N	Y	G	│	│	│	│				
e	E17	N	D	A	K	D	E	V	│	│						
b	E18	A	G	I	I	A	A	A	│	│	│	│	│	│		
b	E19	V	L	L	I	V	A	V	│	│	│	│	│	│		
e	E20	A	A	K	G	A	I	S	│	│						
	EF1	H	H	K	E	S	Q	H	│	│						
		V	L		L	M	L	L								
					E											
					V											
			P	D	T	G										
		D	D	K		D	G	D								
						T	V	Z				Flexible gap				
				G		E	V	G								
		D	N	H	N	K	V	K								
		M	L	H	I	M	S	M								
		P	K	E	E	S	D	V								
		N	G	A	A	M	A	A								
		A	T	E	D	K	T	Q								
	F1	L	F	L	V	L	L	M	│	│	│	│	│	│		
e	F2	S	A	K	N	R	K	K	│	│	│					
e	F3	A	T	P	T	D	N	A	│	│	│					
b	F4	L	L	L	F	L	L	V	│	│	│	│	│	│		
b	F5	S	S	A	V	S	G	G	│	│	│	│	│			
e	F6	D	E	Q	A	G	S	V	│	│						
	F7	L	L	S	S	K	V	R	│	│						
	F8	H	H	H	H	H	H	H	│	│	│	│	│	│	│	│
	F9	A	C	A	K	A	V	K	│	│	│					
	F10	H	D	T	P	K	S	G	│	│						
e	FG1	K	K	K	R	S	K	Y	│	│	│					
	FG2	L	L	H	G	F	G	G	│	│						
								N								
								K								
	FG3	R	H	K		Q		H				Flexible gap				
b	FG4	V	V	I	V	V	V	I								
e	G1	D	D	P	T	D	A	K	│							
	G2	P	P	I	H	P	D	G	│							
	G3	V	E	K	D	Q	A	Q	│							
	G4	N	N	Y	Q	Y	H	Y	│	│	│	│				
b	G5	F	F	L	L	F	F	F	│	│	│	│	│	│	│	
e	G6	K	R	E	N	K	P	E	│	│						
	G7	L	L	F	N	V	V	P	│	│	│					
b	G8	L	L	I	F	L	V	L	│	│	│	│	│	│		
e	G9	S	G	S	R	A	K	G	│	│	│					
e	G10	H	N	E	A	A	E	A	│							
	G11	C	V	A	G	V	A	S	│	│	│	│	│			
b	G12	L	L	I	F	I	I	L	│	│	│	│	│	│	│	
	G13	L	V	I	V	A	L	L	│	│	│	│	│	│		
	G14	V	C	H	S	D	K	S	│							
b	G15	T	V	V	Y	T	T	A	│	│	│	│				
b	G16	L	L	L	M	V	I	M	│	│	│	│	│	│	│	
e	G17	A	A	H	K	A	K	E	│							
	G18	A	H	S	A	A	E	H	│							
	G19	H	H	R	H	G	V	R	│							
		L	F	H	T		V	I								
		P	G	P			G	G								

Table 4 *(continued)*

Acc	Structure	Alignment							Flexible patterns							
									1	2	3	4	5	6	7	8
		A	K	G			A	G								
		E	E	D	D		K	K								
		F	F	F	F		W	M				Flexible gap				
	H1	T	T	G	A		S	N								
	H2	P	P	A			E	A								
	H3	A	P	D	G		E	A								
	H4	V	V	A	A		L	A								
e	H5	H	Q	Q	E	D	N	K	│	│	│					
e	H6	A	A	G	A	A	S	D	│	│	│	│				
	H7	S	A	A	A	G	A	A	│	│	│	│	│	│	│	
b	H8	L	Y	M	W	F	W	W	│	│	│	│	│	│		
e	H9	D	Q	N	G	E	T	A	│	│	│					
e	H10	K	K	K	A	K	I	A	│	│	│					
b	H11	F	V	A	T	L	A	A	│	│	│	│				
b	H12	L	V	L	L	M	Y	Y	│	│	│	│	│			
e	H13	A	A	E	D	S	D	A	│	│	│					
	H14	S	G	L	T	M	E	D	│	│						
b	H15	V	V	F	F	I	L	I	│	│	│	│	│	│		
	H16	S	A	R	F	C	A	S	│	│						
	H17	T	N	K	G	I	I	G	│	│						
	H18	V	A	D	M	L	V	A	│	│						
b	H19	L	L	I	I	L	I	L	│	│	│	│	│	│	│	│
	H20	T	A	A	F	R	K	I	│							
	H21	S	H	A	S	S	K	S	│	│						
	H22	K	K	K	K	A	E	G	│	│						
	H23	Y	Y	Y	M	Y	M	L	│	│	│	│	│	│		
	H24	R	H	K			D	Q								
	H25			E			D	S								
	H26			L			A									
				G			A									
				Y												
				Q												
				G												
		Number of exposed segments							32	25	13	5	4	1	0	0
		Number of buried segments							30	30	30	28	25	20	11	6
		Number of other segments							45	32	17	13	9	7	4	2

Structural alignment of 7 globins as described by Bashford *et al.* (1987) and derivation of flexible patterns with reduced number of elements. The proteins are (left to right (PIR code)); human hemoglobin α-chain (HAHU), human hemoglobin β-chain (HBHU), sperm-whale metmyoglobin (MYWHP), larval *Chironomous thummi* globin (GGICE3), sea lamprey cyanohemoglobin (GGLMS), *Lupinus luteus* leghemoglobin (GPYL2) and annelid worm *Glycera dibranchata* hemoglobin (GGNW1B). Acc, conserved buried hydrophobic (b), and exposed positions (e) as identified by Bashford *et al.* (1987). Flexible patterns, the regions of the alignment used to derive the elements for flexible patterns 1 to 8 are shown by vertical bars whilst the position of gaps allowed to range between specified minimum and maximum values are shown also. The number of exposed, buried and other elements as defined by Bashford *et al.* (1987) in each pattern is summarized at the end.

Tylorrhynchus heterochaetus (GGWNS), and from the genus of filamentous aerobic bacterium *Vitreoscilla* (GGZLB). Bashford *et al.* (1987), in their detailed study of the globin family, identified implausible residues in GGWNS, at G16 and H15, which are both conserved positions utilized by patterns 3 to 5. Bashford *et al.* (1987) also identified a deletion in GGWNS at C4. Even with the flexibility allowed between helical regions, this deletion would prevent the highly conserved elements at C2 and CD1 from simultaneously aligning with their correct counterparts and hence lead to a lower score for matching the patterns to this sequence. The bacterial globin was also observed by Bashford *et al.* (1987) to give relatively low scores with their templates. Our patterns 3, 4 and 5 effectively discriminate against the GGZLB, with pattern 5 (38 elements) giving the greatest specificity for non-bacterial globins, whilst patterns 1 and 2 give full generality by successfully discriminating all globins from non-globins.

Pattern 6 is the first to give low scores to more than two globins. However, of the eight further globins given low scores, four were identified by Bashford *et al.* (1987) to contain implausible residues at key positions (GGGACR, HBFGRE, GPVF and GPSYC2), whilst the alignments obtained in this study for the remaining four proteins (GPPMI, GPSYC1, GPSYS and MYRKJ) also suggest implausible residues (GPPMI: K at A15, I at B14, N at F1; GPSYC1: I at B14, D at F1; GPSYS: I at B14, D at F1; MYRKJ: D at A15).

Table 5

Scans using patterns derived from seven globins: alignment at increasing conservation value cutoffs

Pattern number	Conservation number cutoff	Number of pattern elements	Percentage of shortest sequence	Globins before first non-globin	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	60	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)	17 (64)
8	0·8	8	6	1 (0)	298 (281)	46 (64)

Result of scans using patterns from Fig. 5 against the PIR version 14·0 database with DAY scoring. Percentage of shortest sequence is determined by dividing the number of pattern elements by the length of GGICE3 (136 amino acids). Numbers in parentheses are the result of scans using similar patterns, but with the flexible gaps of unconstrained length.

It is only with the reduction of the pattern to 15 elements (pattern 7, 11% of GGICE3) that sensitivity and specificity is seriously affected.

(f) *Development of a flexible pattern when no three-dimensional structure is known*

In general, the sequences of a protein family may be known, but no details of three-dimensional structure may be available to guide the derivation of a pattern. However, an effective flexible pattern can be developed in the absence of such compelling structural information.

The seven globin sequences were aligned without reference to the secondary or tertiary structures, by an automatic multiple alignment method (Barton & Sternberg, 1987*b*). A pattern was then derived by discarding all positions within the alignment at which gaps occurred, and of the remaining positions, only those that gave conservation scores above 0·4 were maintained. Gaps were made flexible between pattern elements where insertions and deletions had been included by the automatic algorithm, but were kept to fixed lengths where no insertions/deletions were observed. The resulting flexible pattern consisted of 39 elements. When scanned against the PIR database (scan 9, Table 3), all globins scored in the top 500, with 327 giving scores above non-globins. This result is superior to the best non-pattern method considered (scan 5), where 15 globins did not score in the top 500 sequences.

4. Discussion and Conclusions

The techniques and analyses described here were all performed on PIR version 14·0. However, pattern 1, when scanned against PIR 19·0 (10,527 sequences) still totally discriminates the globins from non-globins, whilst the reduced patterns (2 to 8) give results similar to those from the smaller database. This suggests that the patterns truly represent the important features of the globin fold.

Bashford *et al.* (1987) developed general pattern-like templates for the globins. However, their general template (template II) did not completely discriminate globins from non-globins. The lack of any constraints on the gaps between defined helices contributed to this deficiency. However, a program that allows the inter-helical gap lengths to be constrained still scored three globins below non-globins (Boswell, 1988). For comparison, a flexible pattern based upon the aligned positions used in template II was derived. Like pattern 1, this pattern gives perfect specificity when used in conjunction with DAY scoring. This finding suggests that the use of a general scoring scheme such as Dayhoff's matrix can be superior to methods that require detailed analysis of complete families.

With the exception of the FASTP program, the techniques that are compared to the flexible pattern method here, all seek to identify the best match between the complete query and the target sequence. This approach is justified, bearing in mind our aim of finding proteins in the database with the same overall fold as the query. However, if the goal is the more general problem of locating common sub-structures between two longer sequences, then these global comparison methods seem inappropriate. Collins & Coulson (1988) have applied to database scanning the Smith & Waterman (1981) dynamic programming algorithm for the location of the best local similarity between sequences. When this method is used to scan human α-hemoglobin against PIR version 14·0, 319 globins are identified before the first non-globin, with 13 globins not scoring in the top 500 sequences. Although this is a slightly better result than the best non-pattern method previously considered (scan 5; 318 globins before the 1st non-globin, 15 not found), it does not approach the specificity and sensitivity of the single-sequence flexible pattern scan (scan 7; 337 globins before the 1st non-globin).

Gribskov *et al.* (1987, 1988) also utilized the Smith & Waterman (1981) technique to locate the best sub-sequence score for aligning a query profile to

Table 6

Details of globin sequences not identified by patterns 1 to 7 from Table 3

Pattern	Comment	Rank	ID	Title	Score
1				NO GLOBINS MISSED	
2				NO GLOBINS MISSED	
3	Last globin	346	gpvf	Leghemoglobin i-broad bean	93·86
	Globins after first non-globin	351	ggwns	Globin extracellular small chain-*Tylorrhynchus*	71·71
		352	ggzlb	Bacterial hemoglobin-*Vitreoscilla* sp.	71·14
4	Last globin	347	ggwns	Globin extracellular small chain-*Tylorrhynchus*	77·29
	Globins after first non-globin	429	ggzlb	Bacterial hemoglobin-*Vitreoscilla* sp.	64·14
5	Last globin	346	gpsyc2	Leghemoglobin c2-soybean	83·86
	Globins after first non-globin	350	ggwns	Globin extracellular small chain-*Tylorrhynchus*	75·43
		> 500	ggzlb	Bacterial hemoglobin-*Vitreoscilla* sp.	—
6	Last globin	337	a05133	Hypothetical leghemoglobin-soybean	81·14
	Globins after first non-globin	345	gggacr	Globin-water snail	78·86
		350	gppmi	Leghemoglobin i-garden pea	78·14
		351	gpsyc1	Leghemoglobin c1-soybean	78·14
		352	gpsys	Leghemoglobin a-soybean	78·14
		357	myrkj	Myoglobin-Port Jackson shark	77·43
		364	hbfgre	Hemoglobin β-chain-edible frog	75·57
		366	gpvf	Leghemoglobin i-broad bean	75·0
		389	ggwns	Globin extracellular small chain-*Tylorrhynchus*	72·29
		408	gpsyc2	Leghemoglobin c2-soybean	71·0
		> 500	ggzlb	Bacterial hemoglobin-*Vitreoscilla* sp.	—
7	Last globin	295	gpfba	Leghemoglobin a-kidney bean	69·0
	Globins after first non-globin	298	ggewa3	Globin aiii-common earthworm	68·0
		299	ggwn2c	Globin iic-extracellular-*Tylorrhynchus*	66·86
		303	hagsm	Hemoglobin α-a-chain-magpie goose	66·0
		304	hall	Hemoglobin α-chain-llama Arabian camel	65·71
		305	hagda	Hemoglobin α-a-chain-American flamingo	65·71
		310	hbbof	Hemoglobin β fetal chain-bovine	64·86
		312	haos	Hemoglobin α-chain-ostrich	64·29
		313	haeh	Hemoglobin α-a-chain-greater rhea	64·29
		314	hakoaw	Hemoglobin α-a-chain-white stork	64·29
		315	hadk	Hemoglobin α-a-chain-ducks	64·29
		316	hags	Hemoglobin α-a-chain-western greylag goose	64·29
		317	hagsi	Hemoglobin α-a-chain-bar-headed goose	64·29
		318	hagsc	Hemoglobin α-a-chain-Canada goose	64·29
		319	haws	Hemoglobin α-a-chain-mute swan	64·29
		320	haqc	Hemoglobin α-a-chain-golden eagle	64·29
		321	hach2	Hemoglobin α-a-chain-chicken	64·29
		322	hafea	Hemoglobin α-a-chain-ring-necked pheasant	64·29
		323	hadla	Hemoglobin α-a-chain-blue-and-yellow macaw	64·29
		324	hajsa	Hemoglobin α-a-chain-starling	64·29
		325	b26429	Hemoglobin α-a-chain-black vulture	64·29
		326	a24692	Hemoglobin α-a-chain-Andean condor	64·29
		327	hbgtf	Hemoglobin β fetal chain-goat and sheep	63·86
		328	a24625	Hemoglobin α-a-chain-tree sparrow	63·86
		358	ggwns	Globin extracellular small chain-*Tylorrhynchus*	62·0
		368	hblua	Hemoglobin β-chain-South American lungfish	61·57
		371	a24653	Hemoglobin α-chain-spiny dogfish	61·43
		380	haxl1	Hemoglobin α major chain-African clawed frog	61·0
		383	hafg3t	Hemoglobin α-chain-bullfrog tadpole	60·86
		389	hasnv	Hemoglobin α-chain-aspic viper	60·57
		406	ggice3	Globin ctt-iii-midge larva	59·86
		421	ggice4	Globin ctt-iv-midge larva	59·57
		424	hapn	Hemoglobin α-chain-emperor penguin	59·43

Table 6 *(continued)*

Pattern	Comment	Rank	ID	Title	Score
		489	gpyl	Leghemoglobin i-yellow lupin	57·86
		> 500	a05133	Hypothetical leghemoglobin-soybean	—
		> 500	ggzlb	Bacterial hemoglobin-*Vitreoscilla* sp.	—
		> 500	gppmi	Leghemoglobin i-garden pea	—
		> 500	gpsycl	Leghemoglobin cl-soybean	—
		> 500	gpsyc2	Leghemoglobin c2-soybean	—
		> 500	gpsyc3	Leghemoglobin c3-soybean	—
		> 500	gpsys	Leghemoglobin a-soybean	—
		> 500	gpvf	Leghemoglobin i-broad bean	—
		> 500	harkj	Hemoglobin α-chain-Port Jackson shark	
		> 500	haxl2	Hemoglobin α minor chain-African clawed frog	—
		> 500	hbfgc	Hemoglobin β-chain-bullfrog	—
		> 500	hbfgre	Hemoglobin β-chain-edible frog	—
		> 500	hbgtc	Hemoglobin β-c-chain-goat sheep and fragments	—
		> 500	hbshbc	Hemoglobin β-c(na) chain-Barbary sheep	—
		> 500	myca	Myoglobin-carp	—
		> 500	myrkj	Myoglobin-Port Jackson shark	—
		> 500	mytuy	Myoglobin-yellowfin tuna	—

For each pattern, Last globin identifies the lowest scoring full-length (i.e. non-fragment) globin that scores higher than all non-globin sequences.

each sequence in the database. Their profile is derived in a similar manner to our lookup table, with position-specific weights assigned to matching each amino acid to each aligned position. In common with the BS-SS method, the gap penalty is also made dependent upon observed secondary structures, or other key residues. Gribskov *et al.* (1987) describe a profile derived from five globins aligned on structural criteria and consisting of 124 positions. When compared to the protein sequence database, this profile shares the flexible pattern's success in discriminating all globins from non-globins. As a further comparison, we scanned the alignment shown in Table 4 against the database, using the local similarity algorithm and DAY scoring, but without structure-dependent gap penalties. This scan scored 341 globins above non-globins, with only one not in the top 500 sequences (GGWN2C), thus confirming the effectiveness of the local similarity algorithm when used in conjunction with a multiple alignment. Although successful when using a complete alignment as a query, the Smith & Waterman (1981) algorithm, even when combined with variable gap penalities, does not allow the same degree of control over gap lengths that is possible with the flexible pattern method. Such control is essential when specifying sparse patterns like patterns 2 to 8, where a few carefully chosen elements are widely separated in the amino acid sequence. This feature of the flexible pattern method is highly advantageous when attempting to locate the best alignment between one or more aligned proteins of known three-dimensional structure, and a weakly related homologue, prior to model-building the structure by protein extension techniques. Unlike conventional sequence comparison methods, or profiles, a flexible pattern allows the alignment to be concentrated only on the most structurally conserved regions, a common starting point for modelling (e.g. see Greer, 1981; Sutcliffe *et al.*, 1987*a*,*b*).

In summary, we have defined the concept of a flexible protein sequence pattern, and evaluated this concept with reference to the globin family of proteins. The general conclusions are:

(1) When scanning a single protein sequence (HAHU) of known three-dimensional structure against the sequence database, a flexible pattern derived from the core secondary structural regions gives substantially better discrimination for proteins of the same family than conventional global or local sequence comparison methods (Table 3).

(2) Including just one further sequence in the derivation of the pattern is sufficient to give near-perfect specificity for the globin family (scan 8, Table 3).

(3) A general-purpose scoring system (Dayhoff's MDM_{78} matrix), can be as successful as schemes tailored specifically to the protein family, yet still allow effective patterns to be derived from a single sequence.

(4) A pattern with only 38 elements (28% of the sequence) can be objectively derived from a structural alignment, yet give near-perfect discrimination for globins (Tables 4 to 6).

(5) A flexible pattern derived from an automatic multiple alignment of seven globins, performed in the absence of secondary or tertiary structural information, is more sensitive than the best conventional global comparison method (scan 7, Table 3).

(6) Flexible patterns gain their improved discrimination power, from both the capability to remove poorly conserved regions from the query and the ability to define gaps within specific ranges (scans 7 and 8; Tables 3 and 5).

The flexible patterns discussed here have concentrated on one protein family and shown that clear improvements in discriminating power may be obtained over conventional alignment methods. Improvements of this nature are seen also when the procedure is applied to other well-characterized protein families (e.g. the immunoglobulin superfamily, study in progress). However, in common with other pattern-based comparison methods, flexible patterns require assumptions to be made about the relative importance of individual residues in a protein, or positions in an alignment. Patterns can be difficult to define unambiguously when there is little residue-specific information known other than the protein sequence. Despite this, the globin test system illustrates that a simple procedure based upon screening out the less highly conserved positions in an alignment can be a systematic and effective pattern derivation method. However, a general solution to the problem of deriving discriminating patterns remains a research goal closely allied to the development of effective techniques for the prediction of protein structure and function.

Flexible patterns allow the specific definition of gap-length ranges, yet permit the application of any chosen weighting scheme for matching a pattern element with each amino acid type. Due to these features, flexible patterns provide a convenient and powerful bridge between regular expression pattern-matching techniques and more conventional local and global sequence comparison algorithms.

We thank Professor T. L. Blundell, and Drs C. J. Rawlings and J. Fox for their encouragement. This study was supported by the Science and Engineering Research Council, and the Imperial Cancer Research Fund.

References

Abarbanel, R. M., Wieneke, P. R., Mansfield, E., Jaffe, D. A. & Brutlag, D. L. (1984). *Nucl. Acids Res.* **12**, 263–280.

Argos, P., Rossmann, M. G. & Johnson, J. E. (1977). *Biochim. Biophys. Acta*, **439**, 261–273.

Barton, G. J. (1987). Ph.D. thesis, University of London.

Barton, G. J. (1990). *Methods Enzymol.* **183**, 403–428.

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

Barton, G. J. & Sternberg, M. J. E. (1987*b*). *J. Mol. Biol.* **198**, 327–337.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). *J. Mol. Biol.* **196**, 199–216.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). *Nature (London)*, **326**, 326–347.

Boswell, D. R. (1988). *Comp. Appl. Biol. Sci.* **4**, 345–350.

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.

Collins, J. F., Coulson, A. F. W. & Lyall, A. (1988). *Comp. Appl. Biol. Sci.* **4**, 67–71.

Crawford, I. P., Niermann, T. & Kirchner, K. (1987). *Proteins*, **2**, 118–129.

Dayhoff, M. O. (1978). In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 1–8, National Biomedical Research Foundation, Washington, DC.

Dodd, I. B. & Egan, J. B. (1987). *J. Mol. Biol.* **194**, 557–564.

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

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

Greer, J. (1981). *J. Mol. Biol.* **153**, 1027–1042.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Gribskov, M., Homyak, M., Edenfield, J. & Eisenberg, D. (1988). *Comp. Appl. Biol. Sci.* **4**, 61–66.

Lathrop, R. H., Webster, T. A. & Smith, T. F. (1987). *Commun. ACM* **30**, 909–921.

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

Lipman, D. J. & Pearson, W. R. (1985). *Science*, **227**, 1435–1441.

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

Patthy, L. (1987). *J. Mol. Biol.* **198**, 567–577.

Pearl, L. & Taylor, W. R. (1987). *Nature (London)*, **329**, 351–354.

Smith, T. F. & Waterman, M. S. (1981). *J. Mol. Biol.* **147**, 195–197.

Staden, R. (1988). *Comp. Appl. Biol. Sci.* **4**, 53–60.

Sternberg, M. J. E., Barton, G. J., Zvelebil, M. J. J., Cookson, J. & Coates, A. R. M. (1987). *FEBS Letters*, **281**, 231–237.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987*a*). *Protein Eng.* **1**, 377–384.

Sutcliffe, M. J., Hayes, F. R. F. & Blundell, T. L. (1987*b*). *Protein Eng.* **1**, 385–392.

Taylor, W. R. (1986*a*). *J. Theoret. Biol.* **119**, 205–218.

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

Tufty, R. M. & Kretsinger, R. M. (1975). *Science*, **187**, 167–169.

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

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

Edited by A. R. Fersht