CABIOS

Vol 10 no.1. 1994
Pages 19 – 29

Improved sensitivity of profile searches through the use of sequence weights and gap excision

Julie D.Thompson, Desmond G.Higgins and Toby J.Gibson

Abstract

Position-specific substitution matrices, known as profiles, derived from multiple sequence alignments are currently used to search sequence databases for distantly related members of protein families. The performance of the database searches is enhanced by using (i) a sequence weighting scheme which assigns higher weights to more distantly related sequences based on branch lengths derived from phylogenetic trees, (ii) exclusion of positions with mainly padding characters at sites of insertions or deletions and (iii) the BLOSUM62 residue comparison matrix. A natural consequence of these modifications is an improvement in the alignment of new sequences to the profiles. However, the accuracy of the alignments can be further increased by employing a similarity residue comparison matrix. These developments are implemented in a program called PROFILEWEIGHT which runs on Unix and Vax computers. The only input required by the program is the multiple sequence alignment. The output from PROFILEWEIGHT is a profile designed to be used by existing searching and alignment programs. Test results from database searches with four different families of proteins show the improved sensitivity of the weighted profiles.

Introduction

Various methods now exist to search sequence databases for distantly related members of protein families. These methods fall into two classes: those based on single sequences and those based on aligned sequence families. Pairwise sequence comparisons with programs such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson and Lipman, 1988) can detect most, but not all, members of divergent families. Multiple sequence alignments can form the basis of searches either by pattern matching (Taylor, 1986; Barton, 1990; Sibbald and Argos, 1990a; Staden, 1990) or they may be used to provide substitution matrices, known as PROFILES (Gribskov *et al.*, 1987) for use by dynamic programming algorithms. There has been some debate as to whether pattern matching or profile searching is the more sensitive. However, the methods are perhaps best viewed as complementary in that some protein families show tightly conserved features that are amenable to description by simple patterns while other families lack strongly conserved features and may be better described in terms of substitution matrices. An example of the latter are the cyclins (Evans *et al.*, 1983), which drop below 10% sequence identity and lack absolutely conserved residues at any position.

In the original implementation of profile searches, the profiles are prepared in a simple and straightforward manner. The PAM250 amino acid substitution matrix (Dayhoff *et al.*, 1978) is used to provide probabilities for amino acid replacement. By summing the probabilities according to the observed amino acids, a replacement matrix is prepared for each column of the alignment. Additionally, where gaps in aligned sequences occur, locally reduced gap penalties are introduced.

We have been considering ways in which the performance of profiles may be enhanced. In this, we have been aided by the separation of the method of construction of the profiles from the actual search itself (Gribskov *et al.*, 1987). We have found a number of modifications to the ways in which profiles can be prepared that significantly improve discrimination of true and false hits when the profile is used with the standard PROFILESEARCH program supplied in the GCG package (Genetics Computer Group, 1991).

The success of the profile search is naturally dependent on the original choice of sequences to be included in the probe alignment. Several highly related sequences can bias the search result more than one distant sequence. There is general agreement that the sequences in an alignment need to be weighted (Altschul *et al.*, 1989; Vingron and Argos, 1989; Sibbald and Argos, 1990b). However, none of the published weighting schemes seemed ideal for the purpose of weighting profiles. We have developed a new method to weight the sequences, where more distant sequences are assigned higher weights than closely related ones according to branch length values of neighbour-joining trees (Saitou and Nei, 1987) prepared from the aligned sequences.

The Gribskov profile uses an adaptation of the Dayhoff PAM250 residue comparison matrix for scoring pairs of residues. An alternative matrix, BLOSUM62 (Henikoff and Henikoff, 1992), using data from a much larger set of sequence alignments, provides better comparison scores in single sequence searches. We find this to be generally true of profile searches too.

As new sequences are added to an alignment, the proportion of indels rises inexorably. These positions usually carry limited information and can be harmful to the searches. Improved sequence alignments have been achieved where some secondary

European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, 69012 Heidelberg, Germany

structure information is available, by reducing the penalty for inserting a gap in loop regions (Barton, 1990). Analogous treatments for insertions and deletions are described here, which both improve search results and facilitate the alignment of new sequences to the profile.

Database searches with profiles follow the Smith–Waterman (1981) approach, which detects the best local similarities and does not attempt a global match. It is accepted that, on average, this method is the most sensitive dynamic programming strategy. For output alignments with profiles, however, global matches are often desired, for which the Smith–Waterman method is inappropriate.

The improvements to the searches correlate with improved alignment of new sequences to the profiles. However, due to an imbalance in profile matrix scores of positions that allow gaps, versus positions that do not, profiles based on dissimilarity (i.e. matrices including negative numbers) cannot routinely provide correct alignments of sequences against the profiles. This problem has been resolved by specifying a positive value known as the matrix bias, which is added to every element in the comparison matrix (Dayhoff *et al.*, 1983). Thus, residue versus residue matrix scores are generally higher than residue versus gapped positions. These positive profiles routinely yield completely correct alignments with the standard PROFILEGAP program (Genetics Computer Group, 1991) except for the most divergent sequences.

The tree weights and other developments have been incorporated into one program, PROFILEWEIGHT, that will produce a profile (for use with GCG programs) given only a multiple alignment. In addition to our branch-proportional weighting method, we have included an implementation of the Altschul sequence weighting algorithm as an option. In this paper, the development and use of PROFILEWEIGHT are outlined. The performance of profiles prepared with multiple alignments of four different sequence families are compared for each of the enhancements and in combination.

The utility of the sequence-weighted profile has already been demonstrated for two sequence families. Firstly, profiles of the ~50 residue KH domain family revealed many new members in RNA-associated proteins (Gibson *et al.*, 1993b). Secondly, the improved profile method revealed sequence similarities between the two previously unconnected families of cyclins and TFIIB proteins, which show conserved sequence similarity in the range 10–20% (Gibson *et al.*, 1994).

System and methods

The alignments we used as test cases to develop our algorithm were prepared in the GDE (Genetic Data Environment) multiple sequence editor supplied by S.Smith, Harvard University. The GDE has a flat file output format for the aligned sequences which can be read directly by PROFILEWEIGHT. The phylogenetic trees produced by PROFILEWEIGHT are constructed using code taken from the CLUSTALV program (Higgins *et al.*, 1992) and incorporated into our program. The tree file format produced by CLUSTALV has been retained in order to maintain compatibility and to provide the option to use a pre-constructed tree in the sequence weighting algorithm. For example, if the biological root of the tree is known, this can be specified in the PHYLIP RETREE program (Felsenstein, 1989). We used the PHYLIP programs DRAWTREE and DRAWGRAM (Felsenstein, 1989) to display the unrooted and rooted phylogenetic trees output by PROFILEWEIGHT (Figures 1 and 2).

To perform the database searches we used the GCG program PROFILESEARCH. The default options to average sequence composition and normalise comparison scores for length were found to affect adversely the searches and were always turned off. The alignments of sequences to profiles were produced by a second GCG program, PROFILEGAP. Complete lists of known members of the example protein families were extracted from SWISS-PROT (Bairoch and Boeckmann, 1991) by the SRS (Sequence Retrieval System) program (Etzold and Argos, 1993).

PROFILEWEIGHT is written in the C programming language, and conforms to ANSI/ISO standards. Although the program was developed on a SGI Indigo I12 computer running the IRIX v. 4.0.5F operating system, the program should compile under any Unix operating system. A version is also available for Vax computers, and has been tested on a Vax 9000.

Test alignments

Four aligned sequence families have been chosen on the basis of variable conservation patterns, secondary and tertiary structures and sequence length. In each case the alignments were originally prepared due to an interest in the biology of the families and their use is published elsewhere (Musacchio *et al.*, 1992a; Gibson *et al.*, 1993a,b; Higgins *et al.*, 1994).

Algorithm

The profile is built using the linear position-weighting method described by Gribskov *et al.* (1987), from a sequence alignment and a residue comparison matrix. The algorithm is described here for completion. For a group of N aligned sequences of length L, let a_{ij} be the amino acid residue in sequence i at position j. The rows of the profile correspond to positions in the aligned sequences. The profile has a column for each character that occurs in the aligned sequences. The score in the profile at row r and column c is given by

$$\text{Profile}(r,c) = \sum_{d=1}^{20} W_d \, \text{Comp}(\text{residue}_d, \text{residue}_c)$$

where Comp is the value in the comparison matrix for two residues: each possible residue (residue_d) and the residue represented by the column of the profile. The weight, W_d, is given by

$$W_d = \frac{\sum_{i=1}^{N} w_i \delta_d}{\sum_{i=1}^{N} w_i} \qquad \delta_d = \begin{cases} 1 \text{ if } a_{i,r} = \text{residue}_d \\ 0 \text{ if } a_{i,r} \neq \text{residue}_d \end{cases}$$

The weights w_i are the weights for each sequence, and by default are set to 1.0.

We describe here a number of modifications to the Gribskov profile algorithm.

Branch-proportional sequence weighting

The algorithm described by Gribskov *et al.* anticipates the use of weights assigned to each sequence in the alignment, although in their implementation these are all unit weights. PROFILEWEIGHT calculates a profile using sequence weights based on the information provided by a phylogenetic tree. The branch lengths of a phylogenetic tree built by the neighbour-joining method (Saitou and Nei, 1987) are proportional to the divergence of the sequences. In principle therefore, they may be used directly to provide sequence weights.

Figure 1(a) shows a simple tree constructed to illustrate the method. The phylogenetic tree thus generated is unrooted. In order to be able to calculate sequence weights, an arbitrary root is selected such that the means of the branch lengths on either side of the root are equal. This root can be specified by the following algorithm. Let P be any point on the tree. Then P divides the tree into two parts, called the left and right subtrees. Let the n_l sequences on the left of point P be denoted by L_1, L_2, . . ., L_{nl}. Similarly, let the n_r sequences on the right of point P be denoted by R_1, R_2, ..., R_{nr}. For a sequence S_i, let the distance from the sequence to the point P be denoted by $d_{p,Si}$. Then the difference, Δp, between the mean branch lengths on the left and on the right at point P is defined by:

$$\Delta_p = \frac{\sum_{i=1}^{n_l} d_{p,L_i}}{n_l} - \frac{\sum_{i=1}^{n_r} d_{p,R_i}}{n_r}$$

A root of the tree is defined as any point where Δp is equal to zero. To find all possible roots of the tree, the program first finds all nodes that have a positive Δp and whose parent node has a negative Δp, implying that a point along the branch joining the node to its parent has $\Delta p = 0$. From the list of all possible roots, the root that forms the shallowest tree (i.e. the root that minimizes the longest distance from any sequence to the root) is selected.

Figure 1(b) shows the tree depicted in Figure 1(a) after rooting. The resulting tree is an ordered tree in which each leaf (sequence) has a unique path to the root. The branches along the path from a sequence to the root are said to be owned by that sequence. The order of a branch is defined as the number of sequences that own the branch. Let the N sequences in the alignment be denoted by S_1, S_2, S_3, . . ., S_N. Let the path from any sequence S to the root consisting of n branches be denoted by b_1, b_2, b_3, . . ., b_n. Then, let l_i be the length of b_i, and let o_i be the order of branch n. The weight W of sequence S is defined as:

$$W = \sum_{i=1}^{n} \frac{l_i}{o_i}$$

For example, in the tree of Figure 1(b) the sequence labeled seq1 has weight W given by

$$W = \frac{3.0}{1} + \frac{2.0}{2} + \frac{0.7}{3} = 4.2$$

Finally the weights are normalized, with the maximum weight equal to 100. The weights are listed in the header of the profile output file.

Gap treatment

PROFILEWEIGHT includes an option to exclude from the profile any positions in the alignment that have insertions or

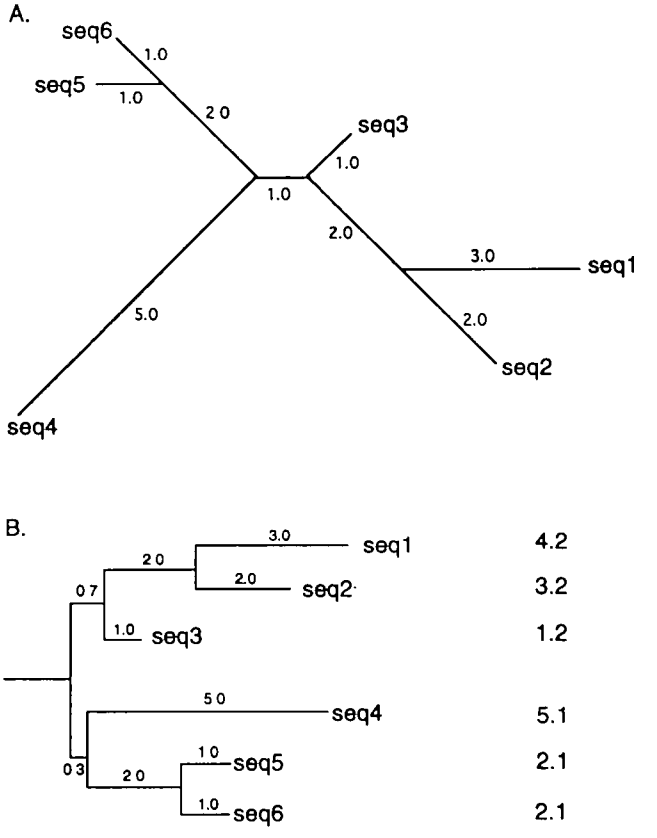

Fig. 1. (A) An unrooted tree such as can be built by the neighbour-joining method. The numbers along the branches are the percentage divergence of the sequences. (B) The same tree rooted by PROFILEWEIGHT. The weights assigned to each sequence are shown on the right-hand side.

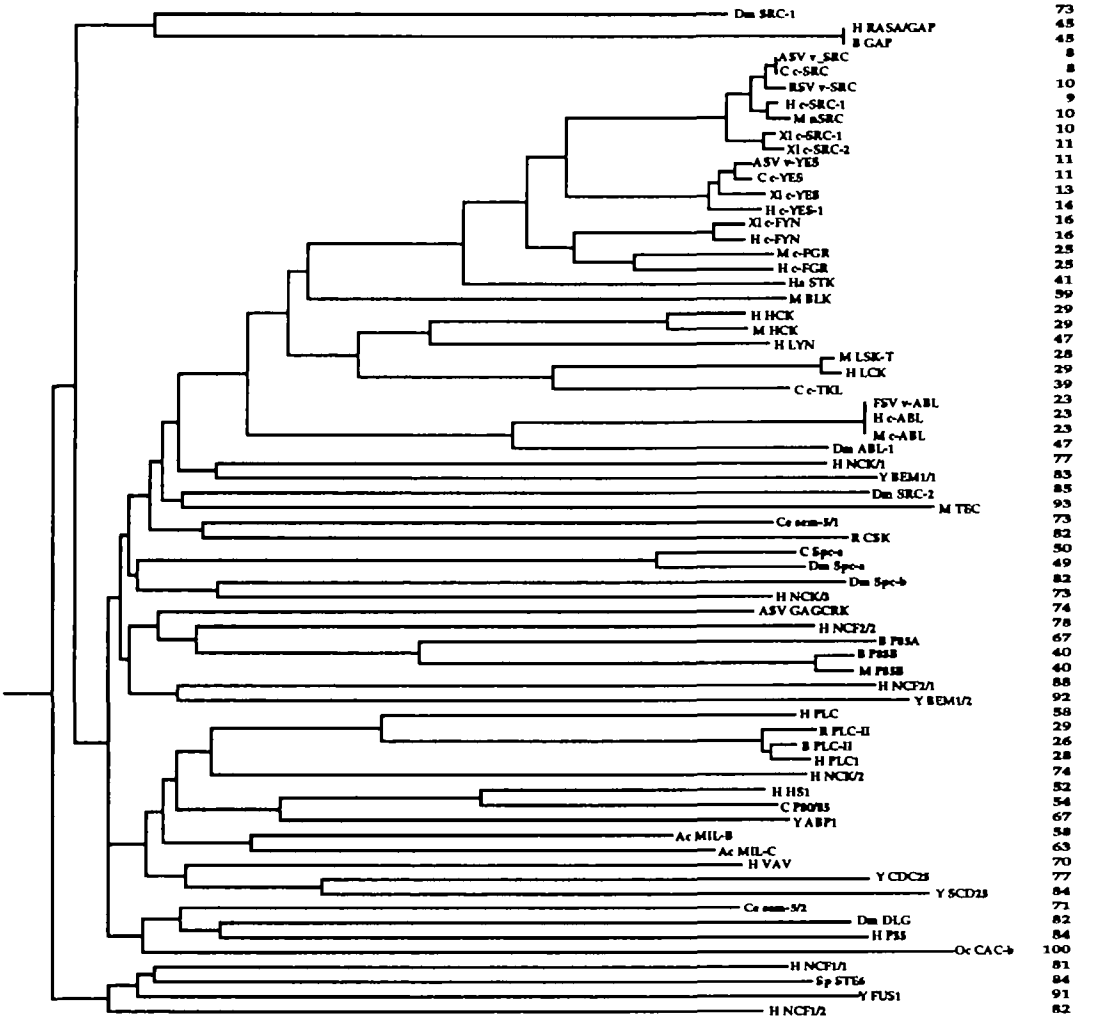

Fig. 2. The rooted tree constructed by PROFILEWEIGHT for the family of proteins containing the SH3 domain. The figures on the right-hand side are the weights assigned to each sequence.

deletions. The positions to be excluded are defined by specifying the percentage of amino acids required at included positions. The penalties for gap opening and extension for the position immediately following the gap are then modified according to user-specified values, or given the default values of 10, 100. The penalties for gap opening and gap extension at non-gap positions are specified as 100, 100.

Sequence to profile alignment

The profiles described above can be used as input for PROFILEGAP and PROFILESEGMENTS of the GCG software package (Genetics Computer Group, 1991), which use a Smith–Waterman (1981) dynamic programming algorithm for alignment. The profiles specify large negative values for aligning incompatible residues. However, the matrix value for aligning a residue on an existing gap position is usually close to zero. This causes homologous but divergent sequences to be misaligned on gaps in the profile. An option is provided in PROFILEWEIGHT to build a profile that has only positive values (with penalties for aligning residues on gaps remaining near zero). A positive residue comparison matrix is constructed by adding an offset such that the minimum score for any pair of residues is equal to zero. Using the positive version of the BLOSUM62 matrix in the Gribskov algorithm results in the required positive profile.

Visual comparison of profiles

An important visual tool in sequence comparison is the inspection of diagonal similarity plots. As well as indicating the presence or absence of significant similarity matches, the plots allow an independent check of automatically generated alignments. However, diagonal plots involving profiles have not so far been implemented.

To aid comparisons of profiles we have developed a program called PROPLOT which produces two-dimensional plots of

Table I. Example sequence searches

	Number of hits in top 300	Highest positive score	Lowest positive score	Highest negative score	Number of false positives[a]	Hits below top negative score
SH3						
Profile 1	69	27.87	11.61	13.87	191	4
Profile 2	73	24.43	7.12	8.64	47	4
Profile 3	73	18.14	7.15	7.68	2	4
Profile 4	73	19.87	6.53	6.42	0	0
Profile 5	73	20.74	6.34	6.13	0	0
SH3 subset						
Profile 1	68	33.61	11.47	13.75	117	5
Profile 2	69	32.44	6.85	9.45	129	6
Profile 3	69	17.87	5.62	7.07	72	11
Profile 4	73	19.81	5.10	6.20	5	3
Profile 5	73	20.68	4.98	6.45	9	4
HLH						
Profile 1	92	21.88	9.68	12.35	6	4
Profile 2	92	17.71	6.13	8.78	1	3
Profile 3	92	15.97	6.82	8.98	1	3
Profile 4	92	16.26	6.68	8.59	1	3
Profile 5	92	15.60	6.89	8.66	1	3
HLH subset						
Profile 1	92	26.45	10.41	12.22	1	7
Profile 2	92	24.22	7.13	8.60	1	1
Profile 3	92	18.15	6.24	8.57	1	11
Profile 4	92	18.83	5.43	8.60	8	13
Profile 5	92	15.67	4.86	8.58	16	14

	cyclin	cyclin TFIIB	cyclin	cyclin TFIIB	cyclin	cyclin TFIIB	cyclin	cyclin TFIIB	cyclin	cyclin TFIIB	cyclin	cyclin TFIIB
Cyclins												
Profile 1	48	48	88.22	88.22	16.88	16.88	20.84	20.84	0	0	0	0
Profile 2	49	53	74.03	74.03	11.29	9.14	11.72	10.94	3	60	1	1
Profile 3	49	53	64.02	64.02	10.51	8.26	10.49	9.30	0	6	0	1
Profile 4	49	53	71.75	71.75	11.14	9.26	11.84	10.46	3	13	1	1
Profile 5	49	54	50.99	50.99	9.01	4.33	6.02	6.02	1	145	1	2
Cyclin subset												
Profile 1	48	48	97.20	97.20	20.44	20.44	22.74	22.74	10	10	2	2
Profile 2	49	52	83.85	83.85	10.79	10.79	10.97	10.97	9	174	1	4
Profile 3	49	52	62.29	62.29	10.24	7.95	10.36	10.36	1	97	1	4
Profile 4	49	53	70.15	70.15	9.82	6.89	9.31	9 31	0	190	0	4
Profile 5	49	52	61.29	61.29	9.34	8.24	8.79	8.79	0	3	0	3

Results of test sequence searches for three families of proteins. The sensitivities of different profiles are compared. Profile 1 is the default profile described by Gribskov et al. Profile 2 uses the BLOSUM62 matrix for pairwise residue comparisons. Profile 3 adds branch-proportional sequence weighting to the BLOSUM62 comparison matrix. Profile 4 also excludes positions where 80% of sequences have an insertion. Profile 5 is based on the Altschul sequence weighting scheme. The first family contains the SH3 domain. The second group is the HLH family of proteins. The third set is the family of cyclins: the second column gives the results when TFIIB sequences are counted as positive hits.

[a]A false positive is defined as a sequence that scores higher than the lowest positive score in the top 300.

either (i) a pair of sequences, (ii) a sequence versus a profile or (iii) a pair of profiles. The output from the program is a Postscript file which can be printed on any printer with a Postscript capability. In the first case, the algorithms used by the DIAGON graphical program (Staden, 1982) are used. Each point (i,j) on the plot corresponds to the similarity of the two residues at position i in the first sequence and position j in the second. Patterns in the plots are detected by eye. To plot a sequence against a profile, we take the residue at each position in the sequence, then simply look up its score in the corresponding column of the profile. The scoring method for plotting a pair of profiles is less intuitive. Let $P1$ and $P2$ be two profiles. Let the ith column of profile P be denoted by $P[i]$. The score at any point (i,j) on the plot is taken to be the normalized RMSD (root mean squared deviation) of the two columns $P1[i]$ and $P2[j]$. This allows all positions in the profile to be sampled, including the highest and lowest scores. The user can control two filters which affect which points are actually plotted: (i) a variable window centred on residues i and j and (ii) a cutoff value for the significance of the scores.

Table II. Titin immunoglobulin searches

	PAM250 comparison matrix			BLOSUM62 comparison matrix		
	No. of hits in top 1000	No. of hits in top 500	No. of hits in top 300	No. of hits in top 1000	No of hits in top 500	No. of hits in top 300
Titin IG						
Profile 1	442	306	219	309	236	192
Profile 2	441	313	225	331	249	198
Profile 3	413	312	215	478	339	252
Profile 4	671	434	290	622	441	287
Titin IG subset						
Profile 1	337	249	188	265	220	186
Profile 2	370	278	200	293	227	201
Profile 3	378	278	195	385	286	209
Profile 4	415	303	218	420	320	228

Results of test searches for the immunoglobin family using an alignment of titin class II-type sequences. Profile 1 is a default profile using either the PAM250 or the BLOSUM62 comparison matrix. Profile 2 adds sequence weighting to the profile. Profile 3 also excludes gap positions.Profile 4 is based on the Altschul weighting method, and excludes gap positions.

Implementation

PROFILEWEIGHT is run interactively and can be given a number of options on the command line. Help is available by invoking PROFILEWEIGHT with no command line options. The sequence alignment can be input in one of two formats: either the PIR format (Sideman *et al.*, 1988) with gaps specified by '−' characters, or the GDE editor flat file format (S.Smith, Harvard University). The phylogenetic tree used for sequence weighting may be specified by the user, otherwise a tree is constructed by the neighbour-joining method as implemented in CLUSTALV. The unrooted and rooted trees used for the weighting may be output for informational purposes. The output file format is compatible with the PHYLIP package (Felsenstein, 1989). The profile output file produced by PROFILEWEIGHT can be used as input to the GCG programs PROFILESEARCH, PROFILESEGMENTS and PROFILEGAP.

Test results

Four families of real protein sequences are used as examples to illustrate the degree of improvement that may be achieved in database searching and in sequence to profile alignments. Five types of profile were constructed for each of the four example families, and the results of searching the SWISS-PROT release 25 sequence database (Bairoch and Boeckmann, 1991) with the PROFILESEARCH program are shown in Tables I and II. Profile type 1 is the default profile constructed by the Gribskov method. The comparison matrix employed is the PAM250 matrix. No sequence weighting is included and gap penalties are specified according to the Gribskov algorithm. Profile type 2 is the same as type 1 except that the BLOSUM62 matrix is used for residue comparisons. The third profile, type 3 adds branch-proportional sequence weighting to the profile based on the BLOSUM62 matrix. In the type 4 profile, positions where $>80\%$ of the sequences in the alignment have insertions are excluded. Finally, the type 5 profile uses the Altschul sequence weighting scheme, in addition to the exclusion of gap positions. A second set of searches were performed using only a subset of the sequences in the alignment. The five most divergent and the five least divergent sequences provide a controlled subset.

The first example is an alignment of 68 SH3 domain sequences (modified from Musacchio *et al.*, 1992a). The SH3 domain alignment consists of five conserved blocks and four indels which correspond to loops in the β-barrel tertiary structure (Musacchio *et al.*, 1992b). The alignment consists of 109 residue columns, of which 57 contain indels. Figure 2 shows the rooted phylogenetic tree produced by PROFILEWEIGHT. The default type 1 profile found 69 of the 73 SH3 domain proteins in SWISS-PROT, while the remaining four modified profiles found all 73 proteins (Table I). The type 4 and type 5 profiles based on the BLOSUM62 comparison matrix, excluding gap positions and either branch-proportional or Altschul sequence weighting, found no false positives. In Figure 3(A), the scores for the first 300 sequences found in each search are plotted for comparison purposes. Using the subset of 10 SH3 sequences, the type 1 profile performance is the worst. Only 68 of the 73 SH3 domains are found in the top 300 sequences. The type 2 profile is marginally better with 69 detections in the top 300. In comparison, the type 4 profile still succeeds in correctly identifying the 73 sequences containing the motif, with just five false positives. The type 5 profile using Altschul weights also finds the 73 SH3 domains, but has nine false positives.

In the second example an alignment of 40 helix−loop−helix (HLH) domains is used (modified from Gibson *et al.*, 1993a). This alignment consists of 70 residues containing a single indel region of length 21. The conserved blocks are predicted to be all α-helix. Here the default type 1 profile finds six false positives, but the type 2−5 profiles result in only one false

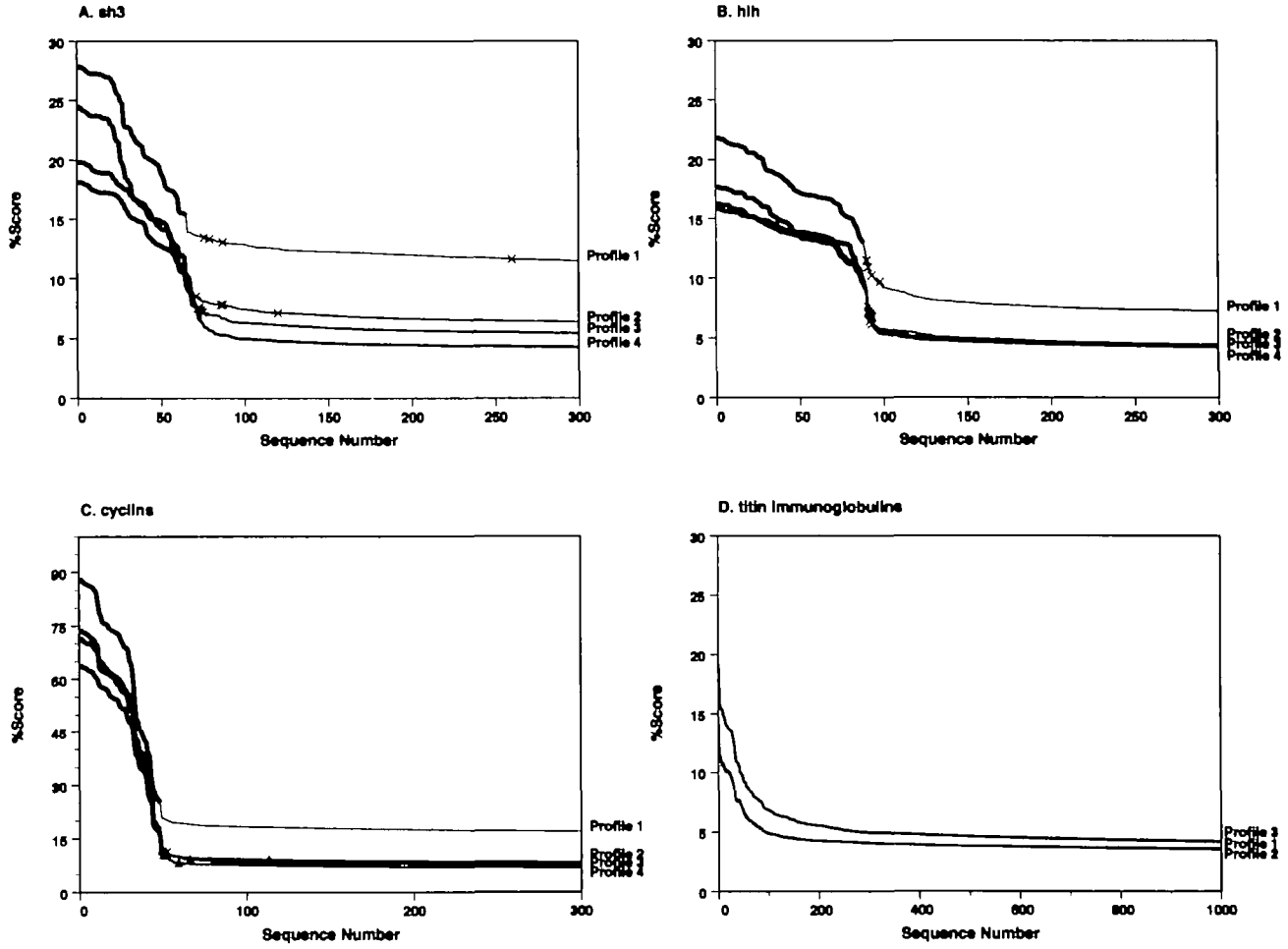

Fig. 3. Plots of the PROFILESEARCH scores (on the *y*-axis) for the highest scoring sequences (on the *x*-axis) found by searching the SWISS-PROT database with the profiles. The profile types are described in the text. (**A**) SH3 domain searches; (**B**) HLH searches; (**C**) cyclin searches; (**D**) titin class II immunoglobulin searches. The thicker lines at the start of each plot indicate the area which consists of all positive hits. The crosses mark the positions of family members which score less than the first false positive. Triangles mark the TFIIB hits in the cyclin profiles.

positive (Table I). Figure 3(B) plots the scores of these searches. Using the subset of 10 sequences described above, all profiles find all 92 HLH proteins. In this example profiles with gap excision perform slightly worse than the others. The type 5 profile based on the Altschul sequence weights finds the same number of HLH domains but with 16 false positives, compared to the type 4 profile which finds eight false positives.

The third example is an alignment of 37 cyclins (Gibson *et al.*, 1994), consisting of 326 residues (with indels at 157 positions). The results of the cyclin searches are presented for cyclin hits alone and for cyclin plus TFIIB hits. Counting just the cyclins, the type 2 profile based on the BLOSUM62 matrix performs better with 49 positives, compared to the PAM250-based type 1 profile, which finds only 48 positives (Table I). The exclusion of gap positions causes an increase in the number of false positives from zero (with the type 3 profile) to three with the type 4 profile. In this example the Altschul weighted type 5 profile successfully matches five TFIIB sequences, one more than the type 4 profile. The profiles derived from the subset of 10 cyclins behave similarly to the full set, except that excluding gaps from the Type 4 profile does not produce a larger number of false positives. There is a straightforward correlation between improved cyclin detection and improved TFIIB detection, in agreement with the proposal that they are related. The top scores of the searches are plotted in Figure 3(C).

Finally, an alignment of 76 class II domains from titin and other muscle proteins (Higgins *et al.*, 1994), which are homologous to immunoglobulins, was used. Because of the diversity of the immunoglobin family (which includes numerous extracellular proteins and membrane-bound receptors) with as low as 5% residue identities, and the highly biased subset chosen, this profile provides a test system with a large number of false positives. Results are recorded in Table II and Figure 4(D) shows the scores for the 1000 highest scoring sequences. In the case of the type 1 profile based on the PAM250 matrix,

the number of hits in the top 1000 sequences matched is 442. The profile based on the BLOSUM62 matrix performs less well here with only 309 hits in the first 1000 sequences. Employing branch-proportional sequence weighting (profile 2) and excluding gap positions (profile 3) improves the sensitivity of the search when used in conjunction with the BLOSUM62 matrices. The best score of 671 is achieved by the profile with Altschul sequence weighting and the PAM250 comparison matrix. The profile produced from a subset of titin sequences resulted in similar comparative scores, from lower absolute values, although in this case the profile with Altschul sequence weighting and the BLOSUM62 comparison matrix performs marginally better than the same profile based on the PAM250 matrix.

Sequence to profile alignments

PROFILEWEIGHT includes a BLOSUM62 similarity matrix in which all values are positive. This is constructed by adding an offset to the default matrix, which ensures that the minimum residue−residue score is zero. Use of this positive comparison matrix leads to a profile that is also positive. The matrix scores for aligning residues on gaps in this case remain close to zero depending on the number of sequences that have gaps. When the default dissimilarity matrix (in which the scores are distributed about zero) is used by the PROFILEGAP program, divergent sequences tend to be aligned on gaps in the profile. A second problem is that the Smith−Waterman algorithm will often output only a partial alignment, losing the ends as in Figure 4(A,B). The accuracy of the alignment can be improved by using the positive profile described above. Positive profiles are always completely aligned to the sequence. As an example consider the alignments shown in Figure 4(B,C). Two alignments are shown of SH3 profiles with the strongly divergent mouse KTEC protein, which also contains the largest insertion of all SH3 motifs. Figure 4(B) shows the sequence aligned to the negative profile, while Figure 4(C) shows the same sequence aligned to the positive matrix. The latter is in fact a perfect alignment of the sequence against the profile.

Discussion

We have found several ways to enhance the utility of profiles in the comparative analysis of protein sequences. For the convenience of the user, modifications to the profiles have been implemented in a single program PROFILEWEIGHT which needs only the multiple alignment to produce the profile. PROFILEWEIGHT is effectively a substitute for the GCG program PROFILEMAKE. The output profiles are designed to be read and used by the GCG searching and alignment programs PROFILESEARCH, PROFILEGAP and PROFILESEGMENTS.

We have also developed an accessory program, PROPLOT, which enables profiles to be used in diagonal similarity plots which are important tools in visual identification and assessment of matching regions of proteins. In addition to discussing relevant points concerning the usage of these programs, it is appropriate to review here certain features of sequence weighting as well as the significance of the tree rooting methodology.

Weighting sequences

A number of methods are already available for weighting sequences in an alignment. In a recent comparison (Vingron and Sibbald, 1993) two methods were found to perform the best, depending on the nature of the problem to be solved. For different reasons, neither of the available weighting methods in the literature is ideal in the case of profile searches. The Voronoi method (Sibbald and Argos, 1990b) does not depend on a pre-constructed phylogenetic tree, and provides weights in proportion to the divergence of the sequences in an otherwise unbiased manner. However, the algorithm has not been implemented to deal with insertions and deletions. A more serious drawback for larger domains (>100 residues) is that the number of iterations required in the Monte Carlo algorithm to achieve consistent weights is increased. The time required to calculate the profile becomes impractical.

The Altschul *et al.* (1989) algorithm weights sequences based on a phylogenetic tree for which the root is known. The method upweights sequences with long branch lengths, but also those that are closest to the tree root. In this scheme the most divergent sequences furthest from the root are usually downweighted, which is not generally desirable for profile searches. In order to search a sequence database for distant members of a family, the information about the most divergent family members is of great significance. If the biological root of the tree is not known, the same rooting method as for the branch-proportional system may be successfully applied as shown here. Nevertheless, a biologically incorrect root could radically alter the sequence weights in an unrealistic manner. In tests where deliberate misalignments were introduced, the Altschul method performed poorly (data not shown). In real datasets, misalignments are an almost ever-present feature. However, given a high-quality alignment the Altschul method has a particular advantage. When the sequences forming the alignment comprise a closely related subset of the protein family, then the sequences form a subtree of the family tree. A profile based on Altschul weights will be closer to the biological root of the tree than a profile employing branch-proportional weighting. Proteins that diverged earlier than the alignment subset will be detected more easily by the Altschul system, as shown by results for the titin−immunoglobulin and cyclin/TFIIB searches, while the branch-proportional method may be more sensitive to other members of the subset. The branch-proportional tree-weighting scheme introduced here is less sensitive to either misalignments or root placement than the Altschul method and will produce quite similar weights to the Voronoi scheme but is much more

computationally efficient. In practice, it is recommended to use branch-proportional weighting routinely, since robustness to error is a major consideration, and limit the Altschul weights to highly refined alignments.

The four test alignments used here have all shown that weighting the sequences improves the sensitivity of the profiles. The test families were chosen because we have previously invested effort in their alignment due to an interest in the biology (Musacchio *et al.*, 1992a,b; Gibson *et al.*, 1993a, 1994; Higgins *et al.*, 1993). In each case the alignments were visually examined and hand edited: for the divergent members, errors are certainly lower than can be achieved by current multiple alignment algorithms. Thus the test scores reflect accurate data sets. Comparable results cannot be expected for profiles made from poor multiple alignments and it is critical to invest proper effort in minimizing error in the alignments. The dictum 'rubbish in—rubbish out' applies very much to profile searching.

In this respect, the tree-based approach has a special usage that 'black box' methods such as the Voronoi cannot provide. If the rooted tree produced by PROFILEWEIGHT is printed, misaligned sequences will be on particularly long branches that protrude from the tree. This is a direct consequence of the proportionality of branch length to sequence dissimilarity. For example, the sequence of FUS1 was manually added to the SH3 alignment, initially incorrectly. Suspicions were aroused because the branch protruded much further than any other sequence. FUS1 was re-examined, including by alignment to the positive profile, which suggested the correct alignment. It is one of the most divergent and difficult to detect of the SH3 domains but now in the revised alignment tree it only just protrudes from the nearby sequences (Figure 2).

Most methods for building trees, including the neighbour-joining method (Saitou and Nei, 1987) produce unrooted trees. Unfortunately, it is not always possible to place the root on purely biological grounds. The use of an outgroup to place the root is not generally appropriate in the case of profile searches because it is most desirable to include the divergent homologs. Therefore, a rooting method has been implemented which, while it produces roots that lack biological relevance, is safe with respect to the derived sequence weights. The root is placed at the point where the tree is shallowest given the condition that the mean branch length on either side of the root is identical. For the method of choosing weights from branch segments according to their ownership number, a change in this deep root will have trivial effects on the weight values for almost all real cases. It must be emphasized, however, that it is unsound to use these roots to provide biological inference.

Aligning sequences to profiles

Previously, alignment to profiles has been poor for divergent sequences, limiting their use as tools for multiple alignment. The reason is that the profiles are designed for Smith–Waterman (1981) searches, which look for best local similarity

```
A. Ktec_Mouse x profile based on PAM250 matrix

S     99 EEIVVAMYDFQAT.......EAHDLRLER.GQEYIILEKNDL........ 132
P      1 vTLVVALYDYEAkidddiprteDDLSFKKkGERLQILNKTeqalgfsdgn 50
S    133 .......HWWRARDKYGWYCRNTNRSKAEQLLRTEDKEGGFMVRDSSQPG 175
P     51 eacpedGDWWEARSLygwycrntnrskaeqlkgredaegddddn...... 94

S    326 GVCTQQKPIYIVTEFME 342
P     95 ...ttGKTGYIPSNYVA 108

B. Ktec_Mouse x profile based on BLOSUM62 dissimilarity matrix

S     99 EEIVVAMYDFQA.......TEAHDLRLER.GQEYIILEKNDL........ 132
P      1 vTLVVALYDYEAkidddiprteDDLSFKKkGERLQILNKTeqalgfsdgn 50
S    133 .......HWWRARDKYGWYCRNTNRSKAEQLLRTEDKEGGFMVRDSSQPG 175
P     51 eacpedGDWWEARSLygwycrntnrskaeqlkgredaegd.......... 90
S    176 LYTVSLYTKPGGEGSSGFRHYHIKETATSPKKYYLAEKHAP........G 217
P     91 .....................................dddnttGKTG 100
S    218 SIP 220
P    101 YIP 103

C. Ktec_Mouse x profile based on BLOSUM62 similarity matrix

S     99 EEIVVAMYDFQA.......TEAHDLRLER.GQEYIILEKND......... 131
P      1 vTLVVALYDYEAkidddiprteDDLSFKKkGERLQILNKTeqalgfsdgn 50
S    132 ......LHWWRARDKYGWYCRNTNRSKAEQLLRTEDKEGGFMVRDSSQPG 175
P     51 eacpedGDWWEARSLygwycrntnrskaeqlkgredaegddddnttGKTG 100
S    176 LYTVSLYTK 184
P    101 YIPSNYVAP 109
```

Fig. 4. Three PROFILEGAP alignments of the mouse KTEC protein against profiles based on the SH3 domain sequences. (A) The first profile is based on the PAM250 matrix with no sequence weighting. (B) The second profile is based on a BLOSUM62 dissimilarity matrix and includes sequence weighting. (C) The third profile is based on a BLOSUM62 similarity matrix and also includes branch-proportional sequence weighting. In each case, S indicates the sequence and P the profile. Lowercase characters indicate positions in the profiles which have gaps in one or more of the aligned sequences. Correctly matched blocks are marked by |.

matches, giving up when the score becomes negative. As well as producing incomplete alignments, residues and even whole blocks tend to be placed into gaps in the profile because the matrix score for placing a gap versus residue is much more favourable than matching two diverged residues. Thus they are inappropriate tools for global alignment. This is remedied straightforwardly by making the profile from a similarity matrix, one with positive amino acid replacement scores. When this is done, it is always more favourable to align residues to residues than residues to gaps. The choice of residue match now depends on the highest score, with new gaps being inserted preferentially opposite existing gaps, which is the correct general strategy.

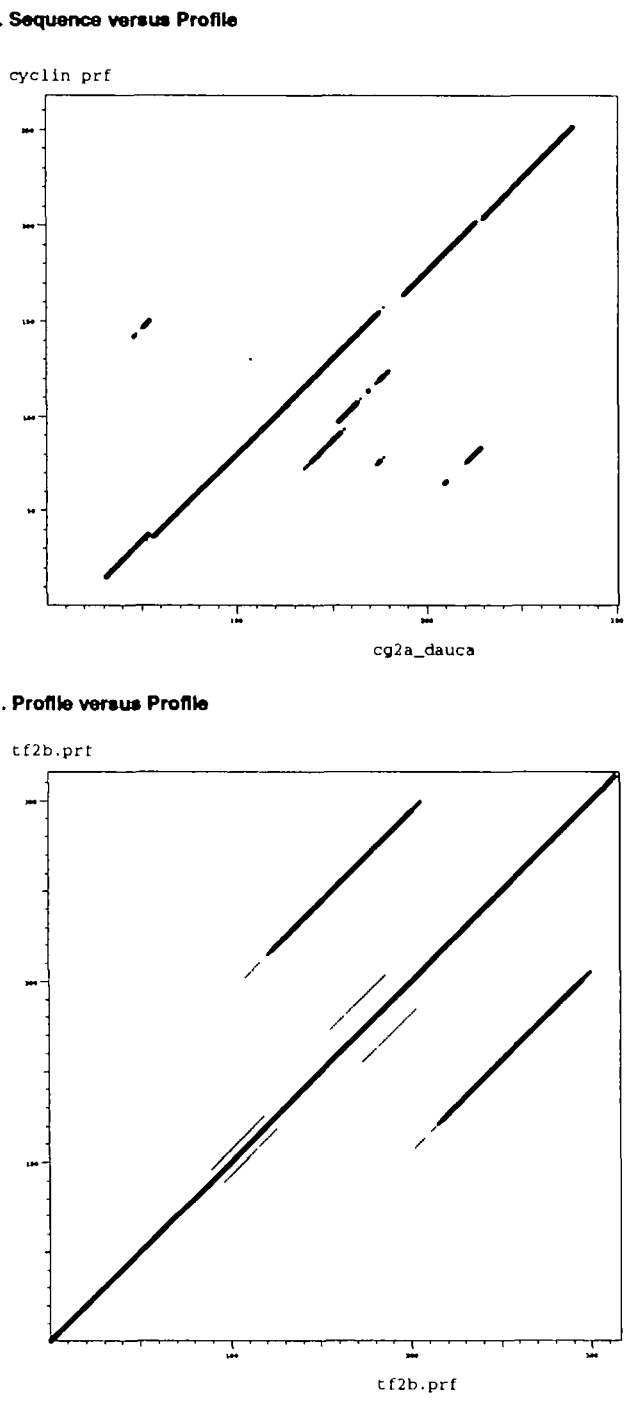

Fig. 5. Dot plots produced by the PROPLOT program. (A) The cyclin cg2a_dauca is plotted against a profile based on the alignment of 37 cyclins. Dots are plotted at the centre of high scoring windows. (B) A profile based on an alignment of TFIIB sequences is plotted against itself. Dots are plotted along the length of a window centred on the highest scoring positions.

This can lead to completely correct alignment of very divergent sequences against the profile, as is seen for the KTEC example (Figure 4). Partial sequences may, however, align better to the standard profiles since local alignment is then appropriate e.g. the fragment entry KABL_Caeel fails to align to the SH3 positive profile.

The positive profiles are comparably effective for database searches—effectively they operate according to the Needleman–Wunsch (1970) algorithm—except that they penalise partial sequences, e.g. the KABL_Caeel fragment entry is lost.

Graphical representation of profile matches

A further deficiency in the use of profiles in sequence alignment has been the absence of a visual method to assess the best matches. Dynamic programming alignment methods find a best path through a matrix. They frequently fail to take recognizably correct segments, particularly when large indels need to be crossed. Dotplots are the most unbiased visual method of allowing a user to assess whether an alignment output by a dynamic programming method has taken the best matches. Therefore, to complement the automatic profile alignments we have implemented in PROPLOT a dotplot method for profile alignment. Dotplots of a sequence to a profile are straightforward since a residue looks up its own value for each position in the profile. For profile versus profile, the method we adopted samples the fit of every amino acid at each position in the profiles. The result is sensitive, easily detecting the TFIIB repeats which share $<20\%$ identity (Figure 5B).

A further use of the profiles in graphical detection of repeated sequences has recently been introduced by Heringa and Argos (1993). Scores are plotted for sliding the profile past a sequence. This is done by sequentially feeding PROFILEGAP with fragments of the sequence that are the same length as the profile. High scoring peaks reliably identify repeats in the sequence.

The source code for PROFILEWEIGHT is available from the EMBL file server (Stoehr and Ormond, 1989; Fuchs, 1990).

Acknowledgements

We are grateful to Thure Etzold, Dmitri Frishman and Peter Rice for helpful discussions and code. We would like to thank Simon Hubbard for the use of his program to produce two-dimensional plots. We also wish to thank Peter Sibbald for suggesting the node scoring idea used to place the root of phylogenetic trees.

References

Altschul,S.F., Carroll,R.J. and Lipman,D.J. (1989) Weights for data related by a tree. *J. Mol. Biol.*, **207**, 647–653.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Boeckmann,B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19**, 2247–2249.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.*, **183**, 403–428.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524–545.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model for evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. Volume 5, Suppl. 3, pp. 345–358.

Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput Applic. Biosci.*, **9**, 49–57.

Evans,T., Rosenthal,E.T., Joungblom,J., Distel,D. and Hunt,T (1983) Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell*, **33**, 389–396.

Felsenstein,J. (1989) PHYLIP—phylogeny inference package. *Cladistics*, **5**, 164–166.

Fuchs,R. (1990) Free molecular biological software available from the EMBL file server. *Comput. Applic. Biosci.*, **6**, 120–121.

Genetics Computer Group (1991) Program manual for the GCG package, Version 7. University of Wisconsin, Madison, WI.

Gibson,T.J., Thompson,J.D. and Abagyan,R.A. (1993a) Proposed structure for the DNA-binding domain of the HLH family of eukaryotic gene regulation proteins. *Prot. Engng.*, **6**, 41–50.

Gibson,T.J., Thompson,J.D. and Heringa,J. (1993b) The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. *FEBS Lett.*, **324**, 361–366.

Gibson,T.J., Thompson,J.D., Blocker,A. and Kouzarides,T. (1994) Evidence for a protein domain superfamily linking the cyclins, TFIIB and RB/p107. *Nucleic Acids Res.*, submitted.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Heringa,J. and Argos,P. (1993) A method to recognise distant repeats in protein sequences. *Proteins*, **17**, 391–411.

Higgins,D.G., Labeit,S., Gautel,M. and Gibson,T.J. (1994) The evolution of titin and related giant muscle proteins. *J. Mol. Evol.*, in press.

Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Applic. Biosci.*, **8**, 189–191.

Musacchio,A., Gibson,T., Lehto,V.-P. and Saraste,M. (1992a) SH3—an abundant protein domain in search of a function. *FEBS Lett.*, **307**, 55–61

Musacchio,A., Noble,M., Pauptit,R., Wierenga,R. and Saraste,M. (1992B) Crystal structure of a Src-homology 3 (SH3) domain. *Nature*, **359**, 851–855.

Needleman,S.B. and Wunsch,C. (1970) A general method to search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 444–453.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Saitou,N. and Nei,M. (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sibbald,P.R. and Argos,P. (1990A) SCRUTINEER: a computer program which flexibly seeks and describes motifs and profiles in protein sequence databases *Comput. Applic. Biosci.*, **6**, 279–288.

Sibbald,P.R. and Argos,P. (1990B) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.*, **216**, 813–818.

Sideman,K.E., George,D.G., Barker,W.C. and Hunt,L.T (1988). The protein identification resource (PIR). *Nucleic Acids Res.*, **16**, 1869–1870.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Staden,R. (1982) An interactive graphics program for comparing nucleic acid and amino acid sequences. *Nucleic Acids Res.*, **10**, 2951–2961.

Staden,R. (1990) Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol.*, **183**, 193–211.

Stoehr,P. and Omond,R. (1989) The EMBL file server. *Nucleic Acids Res.*, **17**, 6763–6764.

Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignments. *J. Mol. Biol.*, **188**, 233–258.

Vingron,M. and Sibbald,P.R. (1993) Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. USA*, in press.

Received on April 30, 1993; accepted on September 30, 1993