CABIOS

Vol.9. no.2. 1993
Pages 183 – 189

A fast, sensitive pattern-matching approach for protein sequences

K.Rohde[1] and P.Bork[1,2]

Abstract

Pattern-matching algorithms are a powerful tool for finding similarities and relationships among the steadily growing amount of known protein sequences. We present a fast, sensitive pattern-matching algorithm that describes a pattern by its physico-chemical properties rather than by occurrence of amino acids, using a fast, dynamic programming algorithm. Selected examples will demonstrate applications and advantages of our approach.

Introduction

The steadily increasing number of known protein sequences is a permanent challenge to assemble similarity families, to develop evolutionary trees and to relate similar sequence regions or modules to common structures or functions of the proteins. Of particular interest is the similarity of local regions, which may coincide with certain protein domain structures.

Such a common feature of a protein group is not defined as a single rigid sequence of amino acids but rather as a (flexible) pattern, characterizing each position of the region by a set of amino acids allowed or forbidden at that position or by possible insertions/deletions. The process of constructing such (flexible) patterns from similar sequence regions and then searching protein sequence databases for further sequences, matching those patterns, has been studied by Barton (1990). A number of useful pattern-matching algorithms have been published (Bashford *et al.*, 1987; Gribskov *et al.*, 1987; Boswell, 1988; Staden, 1988; Bork and Grunwald, 1989, 1990; Barton and Sternberg, 1990; Sali and Blundell, 1990; Sibbald and Argos, 1990; Smith and Smith, 1990; Smith *et al.*, 1990), but nevertheless the problem of pattern matching still leads into problems of discriminating, within reasonable computing time, weak similarities (twilight zone) from random matches.

To detect weak similarity we describe a protein pattern not in the usual way, by its amino acid frequency, but rather by the physico-chemical properties at pattern positions. This method leads to a better description of the pattern, as it is not simply a group of similar sequence positions but rather a structure with special properties and functions. In order to find a 'sound' compromise between sensitivity and computing time, we confine ourselves to a simple set of physico-chemical properties (Bork and Grunwald 1989, 1990), avoiding a sophisticated one such as used by Sali and Blundell (1990). With the database search of patterns so defined, we look for deviations (instead of similarities) of the patterns in sequence windows, using an accelerated Sankoff-type algorithm (Sankoff, 1972). This is a fast and sensitive pattern-matching routine.

[1]*Max-Delbrueck-Centre for Molecular Medicine, Robert-Roessle-Strasse 10, O-1115 Berlin-Buch,* [2]*European Molecular Biology Laboratory, Biocomputing Group, Meyerhofstrasse 1, W-6900 Heidelberg, Germany*

System and methods

The programs described in this paper are written in C and were developed on a VAX 11/785. The code is portable and has been run as well on a SUN Sparcstation-2 and will also run on personal computers if there is sufficient hard disk capacity to store the necessary sequence database. In our search we used MIPSX (1991) as well as SwissProt (Bairoch and Boeckmann, 1991) as the data resource.

Algorithm

Pattern construction (program MAKPAT)

Our starting point is a reasonable multiple alignment—from any of the available routines and/or edited by hand—so that each position in the pattern, together with its amino acid occupancy, is preliminarily fixed.

Such a 'reasonable' multiple alignment should reflect the structural and functional constraints of the pattern as expressed by conservation of amino acid properties at sequence level rather than by the conservation of the amino acids (of 'letters') themselves. To switch from amino acid letters to amino acid properties offers the great advantage that any amino acid not contained in the multiple alignment (learning set) nevertheless may be represented by a property pattern. Therefore we describe every amino acid by a vector of 11 physico-chemical properties (e.g. hydrophobicity, polarity, size and so on—see Figure 1a) according to Taylor (1986), Zvelebil *et al.* (1987) and Bork and Grunwald (1990). For the sake of simplicity each component of the vector (property) is a binary variable of value: 0 (property denied) or 1 (property shown). Any gap in the alignment or an unknown amino acid is represented by a vector of zeros.

A table of properties ascribed every amino acid in this way is given by Bork and Grunwald (1990). In Figure 1(b) you may find property vectors of single amino acids if you follow the

Pattern search in databases (program PROPATS)

The use of a flexible pattern, including insertions/deletions, excludes fast algorithms as in FASTA (Pearson and Lipman, 1988) or BLASTP (Altschul *et al.*, 1990) for the screening of databases, and requires dynamic programming as introduced by Needleman and Wunsch (1970). In view of the steadily increasing size of databases—the current version of the MIPSX (1991) contains ~36 000 sequences—some modifications have been made to gain efficiency.

First we are looking for deviations instead of similarities as proposed by Sankoff (1972) and Sellers (1974). That means a complete match of the pattern by a test sequence as a score of zero, any mismatch or gap adds non-negative integers to the score. If for any reason the deviation should surpass a certain level, the calculation may stop for all paths through the score matrix that have already reached that level.

This 'certain' level or threshold is a crucial point in database search, since too low a threshold would reduce the power of the search and lead to many similar patterns being missed; too high a threshold would increase the number of false positive patterns. In the absence of any reliable expectation score for similar patterns, we start with a relatively low threshold (having in mind the length of the pattern and the number of property deviations we may accept) and increase it until the number of unrelated patterns (noise) found markedly increases too.

A second feature is the use of length-independent gap penalties, since inside a flexible pattern (with the exception of non-penalized spacers) we expect only short gaps, whose single position gap penalties simply add up. In calculating the score for a score matrix position, this reduces the time-consuming inspection of all preceding row and column positions to the inspection of the direct precursor positions.

In general, the pattern length is much shorter than that of the test sequences. In the case of pattern positions that are obligatorily occupied by certain amino acids, we choose the amino acid with the lowest frequency in the database and look for its occurrence in the test sequence. Instead of carrying out the time-consuming dynamic minimization routine for the whole test sequence, we may confine our search to the neighbourhood of the mandatory amino acid in the test sequence, which allows ways through the score matrix up to the given threshold. Repeated occurrence of that amino acid leads to an extension of that region or to a completely new search region within the test sequence.

In the process of database searching, all optimal pattern matches are immediately printed out. To avoid clustering of successive similar alignments within a database test sequence, a pattern match below the threshold is kept in mind for a given number of sequence positions. If this pattern match proves to be the best in this region, it is given out as optimal alignment; if a better one appears in this region, we take this as optimal alignment and keep it in mind for the same number of sequence positions until it is printed out or replaced by a still better pattern match. To be as simple as possible the number of sequence positions kept in mind is set equal to the value of the threshold. This works well for minor thresholds; in case of major ones we recommend an extra parameter for this 'dead time' to avoid loss of sensitivity.

Implementation

Two programs have been developed: MAKPAT for creating the property pattern matrix from a given multiple alignment; and PROPATS, which uses this property pattern for a fast database search. The programs run in interactive mode as well as in batch mode (VAX 11/785).

A database search of a property pattern of a normal length of 20 amino acids (e.g. the middle conservative block in the following example) using a threshold level of 30, running over the MIPSX database (1991)—currently the most voluminous protein database, containing ~36 000 sequences—with PROPATS runs on a Sun Sparcstation-2 in only ~7 min; a Vax 11/785 takes ~10 times longer. The speed may be increased if certain amino acids are declared imperative. Program documentation is available on request.

Discussion

Applications

There are several situations where pattern search methods are preferable to global homology search. Sometimes only a few regions remain above the 'twilight zone' due to functional or structural constraints (e.g. active centres) which then become marker regions. The pattern library PROSITE (Bairoch, 1991) is an example for powerful use of such specific regions. The same holds for mosaic proteins containing many structurally distinct units dispersed by exon shuffling. Each of them has to be considered as separate pattern. Often the most conserved regions of a protein family may no longer be described as a sequence of amino acids, because even in these conserved regions the protein family shows considerable letter variability. Nevertheless there are functional and structural constraints, which can be described by combination of required amino acid properties. If the connecting segments between these relatively conserved regions vary both in length and amino acid composition, the preference of the property pattern approach described here seems to be contracting.

With the following example (P.Bork, C.Sander and A.Valencia, unpublished) we want to demonstrate the usage, sensitivity as well as the speed of our database search. It is already known (Adams *et al.*, 1988) that the galactokinases of different species show no significant global homology but rather a local one, confined to three conserved blocks, each ~20 amino acids long, separated by unrelated sequence regions of sometimes different length.

We thus start with the multiple alignment of the three conser-

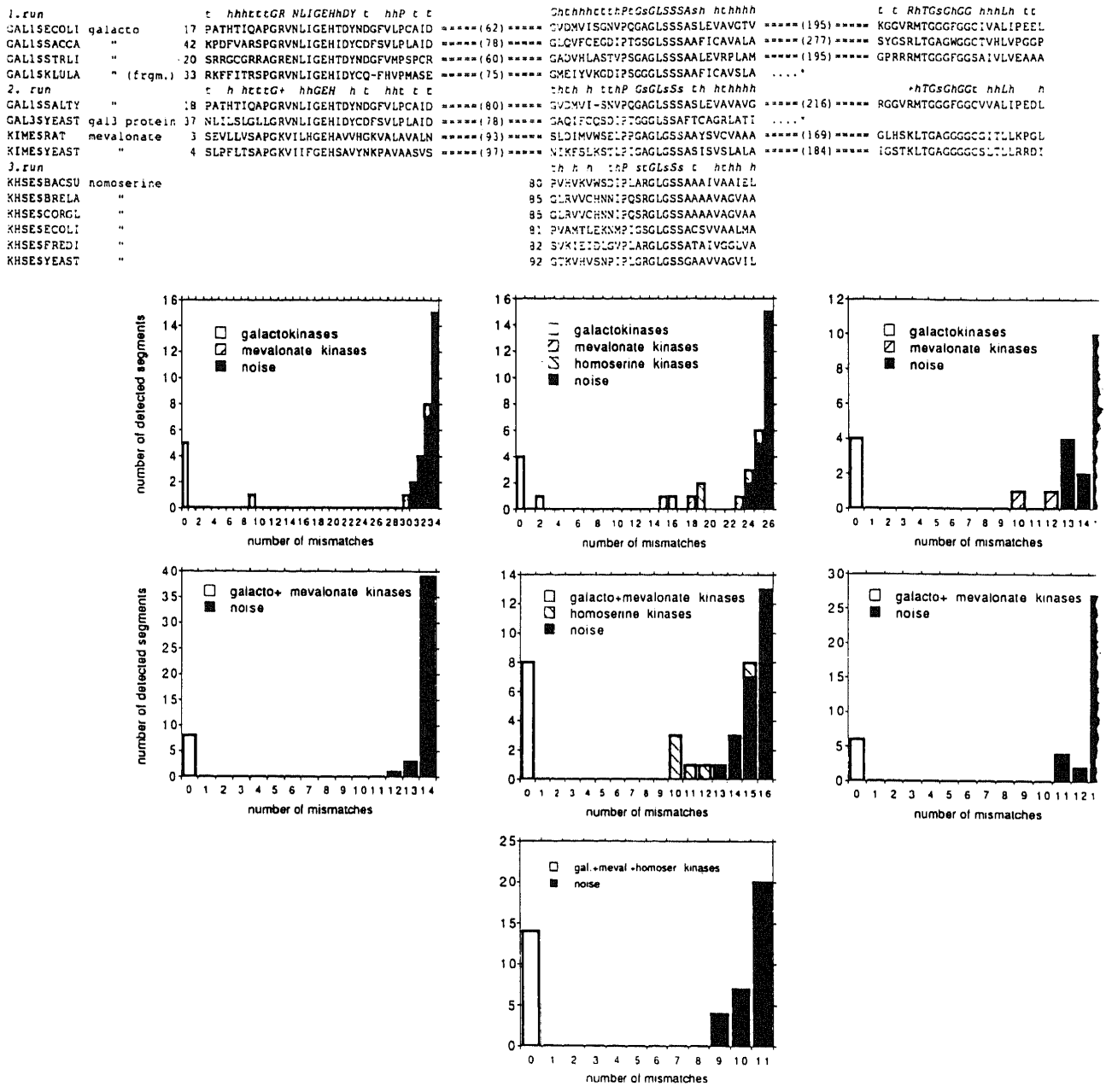

Fig. 2. Alignments of the three conserved blocks of the kinase sequences for three successive database search runs over SwissProt with iteratively improved property pattern by adding found homologue sequences to the alignments. The results of the database searches for all three blocks and the three successive runs are given in the three rows of histograms showing the number of sequences at different deviation scores. Note the distance between significant scores for galactokinases, mevalonate and homoserine kinases and the onset of the insignificant scores (black).

vative blocks of the four known galactokinases from the SwissProt database (Bairoch and Boeckmann, 1991) to construct a property pattern using program MAKPAT, shown in Figure 1.

Search over SwissProt with that pattern using PROPATS led, at a threshold level of up to 30, to the property pattern alignments given in Figure 1(d). The statistics of this first run are discussed in Figure 2, which shows in the middle histogram of the first row (for the first run alignment) that all four galactokinases we had started with were redetected without deviation. An additional galactokinase GAL1$SALTY (and a protein shown to be related to galactokinases; Bajwa *et al.* 1988) have been found with deviation scores of 2 and 16 respectively.

Far more interesting are the medium deviation scores of a group of mevalonate kinases that are well under the deviation score of unrelated candidates ('noise') that appear at a level of 24. A similar behaviour is displayed by the two flanking conservative blocks. To discriminate this effect we add the next best four sequences, as found in our first run, to the first alignment. This creates, by iterative use of MAKPAT, a new, improved and more sensitive property pattern. The results of a second database pattern search with this improved pattern are given in the second row of histograms of Figure 2.

The improvement of the discriminating power of the pattern is striking. For all three conservative blocks one finds the galac-

Table I. Some recent applications of the property pattern approach

Applications	Examples	References
Cofactor binding sites	defining sequence and structural relations between several pyridoxal phosphate dependent enzymes	Bork and Rohde (1990)
Structural based sequence patterns	detection of an ATPase domain in prokaryotic cell cycle proteins and sugar kinases related to actin, hsc 70 and hexokinase	Bork *et al.* (1992)
Shuffled domains	recognition and collection of various domains in mosaic proteins	Bork (1992) and refs therein
Description of spatial motifs	characterization and prediction of α-helices terminated by glycine[a]	Bork and Preissner (1991)

[a]Only the database search algorithm was used; the description of these patterns is different.

Table II. Statistics of sequence database searches with galactokinases

	Gal_Ecoli[a]	Gal_Sacca	Gal_Strli	Gal_Klula	Gal_Salty	Gal3_Yeast	Kime_Yeast	Kime_Rat
Gal_Ecoli[a]	***	120(5,4)[b]	299(3, 3)	130(4, 38)	1660(2,2)	92(6,50)	**81(10,27)**	**60(79, 10)**
Gal_Sacca	120(5,4)	***	65(61, 7)	448(4, 3)	128(4,5)	1020(2, 2)	**86(7, 8)**	**101(6, 6)**
Gal_Strli	299(2,2)	68(68,4)	***	59(131,411)	278(3,3)	57(159,22)	**70(38,96)**	**50(222,370)**
Gal_Klula	130(4,4)	448(2,2)	86(7,68)	***	119(5,6)	379(3,3)	**70(23, 5)**	**66(36,205)**

[b]Values are given as following $A(B,C)$ where A is the FASTA optimized score, B the rank in a FASTA search sorted for optimized scores, and C the rank as usually given in the FASTA output (Person and Lipman, 1988).
[a]SwissProt codes are used, the respective items of mevalonate kinases are shown in bold. There are no significant hits for mevalonate kinases by searching with galactokinases. In addition, except for a search with galactokinase from *Sacharomyces cerevisae*, FASTA gives no indication for a putative homology.

tokinases and mevalonate kinases without any deviation, the 'noise' being distinctly separated. For the middle block an interesting new effect is the appearance of a group of homoserine kinases well under the 'noise' level (see Figure 2). In order to increase the sensitivity of the property pattern we included these sequences into the multiple alignment of the second (and only the second) block. In a third database run with this improved pattern the homoserine kinases with their local similarity to the second (middle) block of the galactokinases and the mevalonate kinases could be well discriminated from the 'noise', reflecting a common block that might be involved in ATP-binding.

In spite of progress in estimating the significance of blocks (Schuler *et al.*, 1991), a strict statistical theory assessing the significance of pattern matches for flexible patterns is still lacking. The discriminating power of a pattern (improved by iterative inclusion of similar patterns) may serve as a strong indicator of significant results. Groups of non-related sequences added to the multiple alignment should lead to patterns without distinct properties, resulting in weak discriminating power.

Another striking example of the sensitivity of our approach is the identification of additional members of the actin/hsp70/hexokinase family. The tertiary structures of the three functional diverse proteins can be well superimposed, but the corresponding sequences have no obvious similarity (Bowie *et al.* 1991; Doolittle, 1992). By selecting five functional and structural important motifs, our approach was not only able to identify in database searches all members of the family (actins, hsp70 family and hexokinases), but also other sugar kinases as well as some prokaryotic cell cycle proteins (Bork *et al.*, 1992).

Our property pattern approach has been used in a series of further different tasks, compiled in Table I, with comparable success.

Comparison with standard methods

The power of our approach of searching for short property patterns with flexible distances in between may be demonstrated by the results of accompanying FASTA runs. In Table II the

optimal score and the rank of FASTA database searches for galactokinases is shown. The best ranks of the mevalonate kinases are well under the level of acceptance, since the short, single conservative blocks of the galactokinases are below the threshold of longer sequences, while a combination of the conservative blocks is hampered by gaps of variable length between them.

To compare our method with a standard pattern search method we ran PROFILESEARCH by Gribskov *et al.* (1987) on two different VAX systems. Judging by the final z-score, defined by these authors, all galactokinases of the learning set had been found with $z > 35$, other galactokinases also rank with $z > 14$ well above the random background of unrelated sequences. Rat mevalonate kinase was found with $z = 7.41$, but the first false positive had already a z-score of 6.99. Yeast mevalonate kinase was found at rank 20 ($z = 5.14$) and the best homoserine kinase at position 28 with $z = 4.68$.

A better discrimination gave the length-dependent original scores where the homoserines also came out with a relatively high score (but a number of false positives too).

On a VAX 9000 (VAX 4000/200) PROFILESEARCH took (6219/38 566) s, whereas our approach took (245/1300) s—at least 25 times faster.

Conclusion

The property pattern approach as presented proves extremely useful for pattern search of short, flexible patterns separated by sequence regions of variable length. It combines a simple, but very sensitive amino acid description of the flexible pattern, allowing for gaps at each position of the pattern (in contrast to routines as FASTA, BLASTP/BLAST3), using an accelerated Needleman–Wunsch routine, which runs with (at least) comparable sensitivity markedly faster than PROFILESEARCH. The iterative use of pattern construction and database run improves the discriminating power of the method. The use of a histogram of the pattern deviations for the database run in order to discriminate significant matches from appearing insignificant 'noise' is the subject of further work.

Acknowledgements

The authors are grateful to J.G.Reich for helpful suggestions as well as critical reading of the manuscript, and thank G.Freudenberg as well as E.Wolf for technical assistance.

Note added in proof

Recently the sequence of phosphomevalonate kinase has been published (Tsay,Y.H. and Robinson,G.W. (1991) *Mol. Cell. Biol.*, **11**, 620–631) and an ATP binding site common to homoserine kinases and mevalonate kinases has been proposed in what we call pattern 2 (Figure 2).

References

Adams,C.W., Fornwald,J.A., Schmidt,F.J., Rosenberg,M. and Brawner,M.E. (1988) Gene organization and structures of the *Streptomyces lividans* gal operon. *J. Bacteriol.*, **170**, 203–212.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **214**, 1–8.

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

Bairoch,A. (1991) A dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19**, 2241–2245.

Bajwa,W., Torchia,T.E. and Hopper,J.E. (1988) Yeast regulatory genes GAL3: Carbon regulation. UAS_{GAL} elements in common with GAL1, GAL2, GAL7, GAL10, GAL80, and MEL1; encoded protein strikingly similar to yeast and *Escherichia coli* galactokinases. *Mol. Cell. Biol.*, **8**, 3439–3447.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. In Doolittle,R.F. (eds), *Methods in Enzymology* Academic Press, San Diego, Vol. 183, pp. 403–428.

Barton,G.J. and Sternberg,M.J.E. (1990) Flexible protein sequence patterns—a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, **212**, 389–402.

Bashford,D., Chothia,C. and Lesk,A.M. (1987) Determinants of a protein fold—unique features of the globin amino sequences. *J. Mol. Biol.*, **196**, 199–215.

Bork,P. (1992) Mobile modules and motifs. *Curr. Opin. Struct. Biol.*, **2**, 413–421.

Bork,P. and Grunwald,C. (1989) A method for property pattern searches in protein sequence data bases, demonstrated by detection of GTP-binding sites. *Stud. Biophys.*, **129**, 231–240.

Bork,P. and Grunwald,C. (1990) Recognition of different nucleotide-binding sites in primary structures using a property pattern approach. *Eur. J. Biochem.*, **191**, 347–358.

Bork,P. and Preissner,R. (1991) On alpha-helices terminated by glycine 2. Recognition by sequence pattern. *Biochem. Biophys. Res. Commun.*, **180**, 666–672.

Bork,P. and Rohde,K. (1990) Sequence similarities between tryptophan synthase beta subunit and other pyridoxal-phosphate-dependent enzymes. *Biochem. Biophys. Res. Commun.*, **171**, 1316–1325.

Bork,P. and Rohde,K. (1991) More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-sensitive calcium channel and inter-alpha-trypsin inhibitor. *Biochem. J.*, **279**, 907–911.

Bork,P., Sander,C. and Valencia,A. (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl Acad. Sci. USA*, in press.

Boswell,D.R. (1988) A program for template matching of protein sequences. *Comput. Applic. Biosci.*, **4**, 345–350.

Bowie,J.U., Luethy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 165–170.

Doolittle,R.F. (1992) Reconstructing history with amino acids. *Prot. Sci.*, **1**, 191–200.

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

MIPSX (1991) Protein sequence data base. In MIPSX Rel. 27, 15 Feburary 1991, MIPS at Max-Planck-Institut für Biochemistry, Martinsried.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search of 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.

Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. *J. Mol. Biol.*, **212**, 403–428.

Sankoff,D. (1972) Matching sequences under deletion insertion constraints. *Proc. Natl Acad. Sci. USA*, **69**, 4–6.

Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.

Sellers,H. (1974) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, **26**, 787–793.

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

Smith,H.O., Annau,T.M. and Chandrasegaran,S. (1990) Finding sequence motifs in groups of functionally related proteins. *Proc. Natl Acad. Sci. USA*, **87**, 826–830.

Smith,R.F. and Smith,T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences, *Proc. Natl Acad. Sci. USA*, **87**, 118–122.

Staden,R. (1988) Methods to define and locate patterns of motifs in sequences. *Comput. Applic. Biosci.*, **4**, 53–60.

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

Zvelebil,M.J., Barton,G.J., Taylor,W.R. and Sternberg,M.J.E. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957–961.

Received on April 14, 1992; accepted on July 31, 1992