CABIOS

Vol.9 no.4 1993
Pages 397–402

Multiple alignment of sequences on parallel computers

Shashank Date, Rajendra Kulkarni, Bhavana Kulkarni, Urmila Kulkarni-Kale[1] and A.S.Kolaskar[1,2]

Abstract

A software package that allows one to carry out multiple alignment of protein and nucleic acid sequences of almost unlimited length and number of sequences is developed on C-DAC parallel computer—a transputer-based machine. The farming approach is used for data parallelization. The speed gains are almost linear when the number of transputers is increased from 4 to 64. The software is used to carry out multiple alignment of 100 sequences each of α-chain and β-chain of hemoglobin and 83 cytochrome c sequences. The signature sequence of cytochrome c was found to be PGTKMXF. The single parameter, multiple alignment score, S, has been used to categorize proteins in different subfamilies and groups.

Introduction

In the last few years the size of the protein sequence databank has increased considerably due to advancements in DNA and protein sequencing. These protein sequences are from various species, organs, tissues and have different functions. They provide a useful tool to get an insight in one of the major challenges in molecular biology, namely identification of amino acid sequence elements that are directly involved in the function, stability, antigenic property, etc., of the protein. Three-dimensional structures, obtained from either X-ray crystallography or high-resolution NMR studies, are ideal but this information is available for only a few proteins. Another approach, which is not only more popular but also seems to be providing very useful information regarding the function of the protein, is to determine signature sequences responsible for functions of the proteins. Site-directed mutagenesis help to confirm the role of these patterns and the extent of their involvement in various functions of the proteins. Earlier studies have pointed out the usefulness of obtaining signature patterns (Jongeneel *et al.*, 1989; Doolittle, 1990).

In order to obtain such patterns the first step is to align sequences optimally. The pairwise alignment of proteins using the algorithms of Needleman and Wünsch (1970), Sellers (1974) and Waterman *et al.* (1976) and methods based on the Wilbur and Lipman (1983) algorithm have helped to categorize proteins in superfamily, family, subfamily, group and entry. In NBRF-PIR this division is indicated by assigning a unique five-digit number to each entry (Barker *et al.*, 1990). In recent years, methods have been developed to carry out multiple alignment of sequences (Feng and Doolittle, 1987, 1990; Barton and Sternberg, 1987; Higgins and Sharp, 1988; Taylor, 1988. Lipman *et al.*, 1989; Vingron and Argos, 1989; Barton, 1990; Schuler *et al.*, 1991). One method attempts to generalize dynamic programming algorithm of Needleman and Wünsch (1970) to align more than two sequences (Jue *et al.*, 1980; Murta *et al.*, 1985; Johnson and Doolittle, 1986). Unfortunately, generalization in this way leads to an explosive increase in computer time and storage requirement with number of sequences considered for multiple alignment. Although these methods give very useful results, they are computer intensive.

The availability of transputer-based, general-purpose parallel computers has led to their extensive use to solve computer-intensive problems which are parallelizable (Lander *et al.*, 1989). Although there are many examples of data parallelization, there are very few known cases where algorithmic parallelization is carried out to achieve efficiency. This is particularly true for problems in biology. A computer program, 'Parallel Multiple Alignment of Sequences' (PRAS), in which data parallelization has been carried out to achieve multiple alignment of sequences on a parallel computer, indigenously developed at the Centre for Development of Advanced Computing (C-DAC) India, is discussed.

The method is applied to align 83, 100 and 100 sequences of cytochrome *c*, α-hemoglobin and β-hemoglobin respectively. An attempt is made to calculate the multiple alignment score (*S*), which indicates how closely the sequences under consideration are related. Our studies indicate that *S* can be used to pick up homologous sequences. The algorithm used and its implementation on a hardware platform are discussed, along with results obtained, in the following sections.

Systems and methods

Hardware

Recently, the C-DAC India has developed reconfigurable, distributed-memory, MIMD-type parallel computers based on transputers called PARAM. These systems use T805 transputers (25 MHz) with a 4 Mbytes memory per transputer as the basic unit, called a node. It is designed to have 1–64 cards, each having four such nodes. PARAM is a backend supercomputing

Center for Development of Advanced Computing and [1]Bioinformatics, DIC, Department of Zoology, University of Poona, Pune 411 007, India
[2]To whom reprint requests should be sent

resource with a variety of industry standard hosts, e.g. IBM PC-AT compatibles, MicroVAX II, Sun workstations and other popular VME or Multibus II machines with Unix/Xenix environment. We have used an IBM PC-AT (80286) as a host end.

System software

The host operating system can be MS-DOS 3.2 or above. The programs were developed using version 1.30 of the parallel C toolkit (3LC Ltd, UK). This implementation of parallel C offers a special library of functions to exploit parallelism on the given architecture. It offers a very clean abstraction for implementation of the farming approach discussed below.

Algorithm

The algorithm used for multiple alignment is similar to that used earlier (Feng and Doolittle, 1987, 1990; Barton and Sternberg, 1987; Barton, 1990) and belongs to alignment methods by hierarchial clustering (Corpet, 1988; Higgins and Sharp, 1988; Taylor, 1988; Vingron and Argos, 1989). It can then be divided into three steps: (i) alignment of a pair of sequences using the dynamic programming approach; (ii) clustering; and (iii) profiling. Profiling is done by using the iterative procedure of Gribskov *et al.* (1988) and Vingron and Argos (1989).

Calculation of the multiple alignment score, S, for Z sequences is done using the following simple formula

$$S = \frac{P - Q}{Sd(Q)}$$

where P is the average real similarity score

$$P = \sum_{i=1}^{Z} \sum_{j=i+1}^{Z} P_{i,j}/K$$

$P_{i,j}$ is the real similarity score for sequence pair 'i' and 'j' and $K = B(B - 1)\ 0.5$.

Q is the average of all mean similarity score ($Q_{i,j}$) for a sequence pair i,j. Thus,

$$Q = \sum_{i=1}^{Z} \sum_{j=i+1}^{Z} Q_{i,j}/K$$

where

$$Q_{i,j} = \sum_{L=1}^{N} M_L/N$$

N is the number of random runs, and M_L is the similarity score for a randomized sequence pair at the Lth random run.

$Sd(Q)$ is the standard deviation associated with the mean scores (Q).

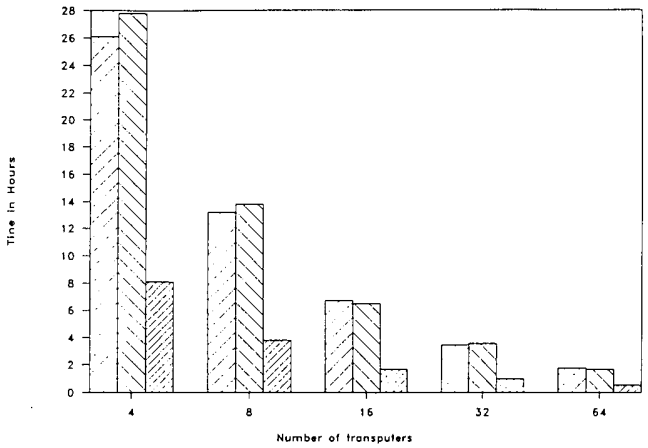


Fig. 1. A bar diagram showing benchmark timings of multiple alignment of 100, 100 and 83 sequences each of α-hemoglobin (▨), β-hemoglobin (▧) and cytochrome c (▨). See text for details.

Table I. Benchmark timings in hours:minutes

Sequences	Numbers of transputers				
	64	32	16	8	4
Cyto 83	00:35	01:01	01:50	03:50	08:10
% gain	86.40%	86.49%	81.25%	84.00%	86.72%
Hem α 100	01:49	03:25	06:42	13:08	26:07
% gain	90.90%	91.00%	88.00%	88.40%	87.60%
Hem β 100	01:58	03:39	06:33	13:54	27:55
% gain	92.48%	92.5%	83.6%	87.75%	88.12%

Gain is calculated as % gain = [(observed speedup)/(expected speedup)] 100.

Thus, the calculated multiple alignment score, S, will have similar properties to those of an alignment score for a pair of sequences.

The formula used to calculate the variability index (VI) is the same as that given by Wu and Kabat (1970):

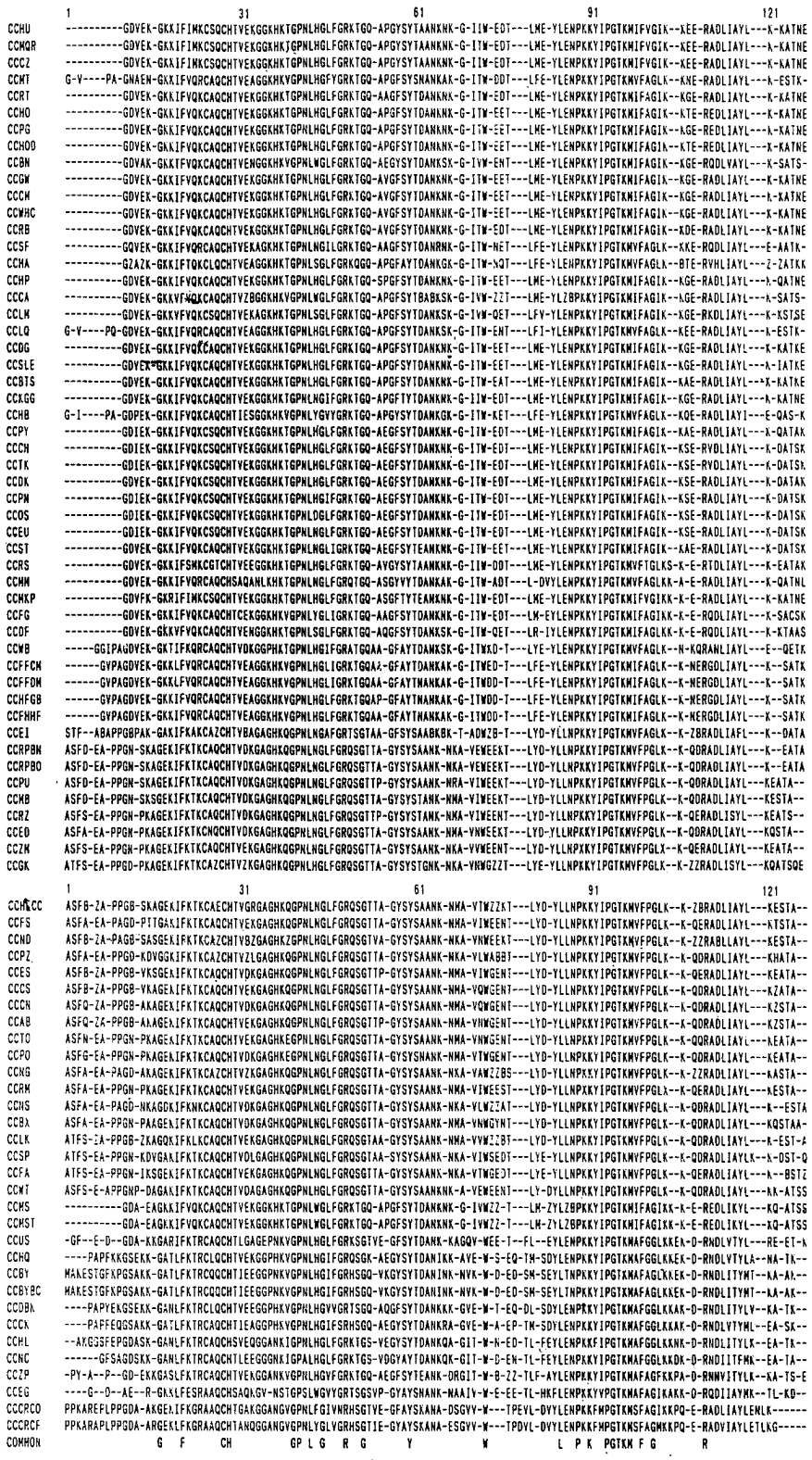
$$VI = \frac{\text{number of different amino acids at } i\text{th position}}{\text{frequency } (F) \text{ of the most common amino acid at } i\text{th position}}$$

where

$$F = \frac{\text{Number of occurrences of most common amino acid at } i\text{th position}}{\text{total number of sequences considered for multiple alignment}}$$

Implementation

The algorithm discussed above is implemented on a transputer-based system by carrying out data parallelization. The pairwise alignment and profiling processes are parallelized by using the farming approach. In this approach one processor generates work units by chopping the large data set into smaller subsets. The processor then passes this work unit to one of the other processors (the worker) in the farm. The work unit is collected only by a worker processor that is free. The farming approach thus consists of dynamic distribution of work. There is no direct



Note the total length increase is only about 20% even when cytochrome c sequences from 83 different species are aligned simultaneously.
NBRF-PIR IDCODES are given.

Fig. 2. Multiple alignment output of 83 cytochrome *c* sequences using PRAS. Note the conserved sequence region PGTKMXF at the C-terminal region, which is unique to cytochrome *c* proteins.

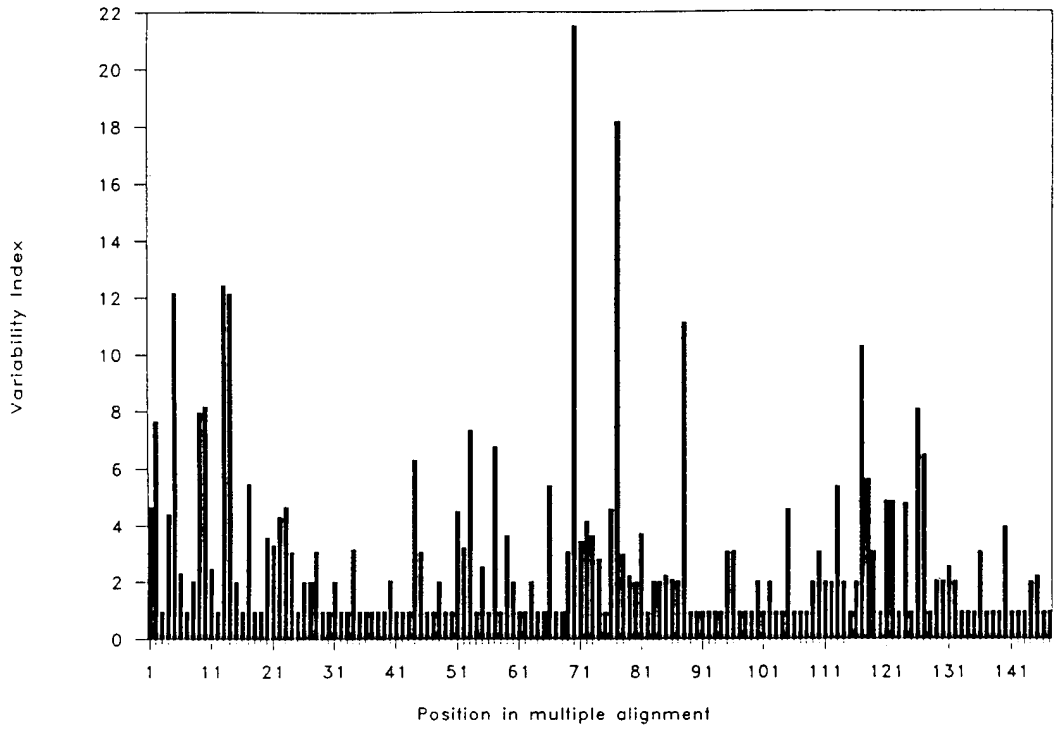


Fig. 3. Variability index plot obtained using multiple alignment output for β-hemoglobin sequences from 100 different species. Note several regions of $VI = 1$.

communication among worker processors in the farm and the communication between the processors will be always routed through the farmer processor. Such an approach is most useful under the following conditions: (i) the time required to process a single work unit is significantly larger than the time required to generate and distribute the work in the farm; (ii) the order in which the results are obtained should have no effect on the final solution. These prerequisites are satisfied in the present problem of multiple sequence alignment by carrying out data parallelization at both the pairwise sequence alignment and the profiling stages. During the multiple alignment of Z sequences, one will have to carry out $Z(Z - 1)/2$ pairwise alignments. Further, during every optimal alignment of the pair of the sequences, to calculate the maximum average similarity score (M) one generates a large number of shuffled sequences and repeats the alignment process. This process of aligning the real and shuffled sequences of a pair is considered as one single work unit as it requires significantly more computer time than that required to generate and distribute the work. The results obtained are order independent. The length of the sequences under consideration may not be identical and thus the pairs formed can take different times; thus dynamic load balancing, which is an essential component of the farming approach, has been utilized during implementation to achieve maximum gain factor. In the profiling stage, the profile matrix requires two segments of the given sequence. These segments form the work unit. Calculation of each element in the profile matrix is treated as an independent process. In other words clustering, collapsing of sequences and calculation of the VI are carried out by the farmer processor. The scores $P_{i,j}$, $Q_{i,j}$ calculated during pairwise alignment are stored and used in calculation of S.

Discussion

During each pairwise alignment, to obtain the average similarity score (Q_{ij}) per pair, 100 runs using randomized sequences were carried out for each pair. The sequences are randomized using a pseudorandom number generator which works with a fixed seed on every worker, keeping the length and composition of the sequence the same. This makes the multiple alignment of the same set of data repeatable across many runs and processors. The time given for multiple alignments in Figure 1 thus includes the time required for calculations of M and Q. The parameters such as bias (B) and penalty (P) can be varied and in the present study we have used $B = 6$, $P = 6$. Similarly, one can use various types of matrices such as the unitary matrix, the mutation data matrix (Dayhoff *et al.*, 1979), the genetic code matrix (Sellers, 1974; Smith *et al.*, 1981), the alternate amino acid matrix (McLachlan, 1971), the conformational similarity weight matrix (Kolaskar and Kulkarni-Kale, 1992), etc. We have chosen the mutation data matrix in the present study. It can be seen from Figure 1 that as the number of transputers increases from 4, 8, 16, 32 to 64, the time required for multiple alignment decreases. The gain factor and the time required for the multiple alignment of the sequences (Table I) indicate that the data parallelization used in the present study

is near to the ideal and the algorithmic parallelization will not improve the gain factor considerably. Further, the farming approach, which provides dynamic load balancing, keeps every processor at work during almost the entire alignment process if the number of sequences is larger than the number of nodes.

The multiple alignment output for 83 cytochrome *c* sequences is given in Figure 2. It can be seen from Figure 2 that residues at positions 17, 21, 28, 29, 40, 41, 43, 45, 49, 52, 60, 73, 86, 89, 91, 94–98, 100, 102, 111 are invariant. It was reported earlier that CXXCH is the consensus sequence for the cytochrome *c* sequences (Dickerson, 1971; Doolittle, 1990). However, it can be seen from Figure 2 that in species *Euglena gracilis* (CCEG), *Crithidia oncopelti* (CCCRCO) and *Crithidia fasciculata* (CCCRCF) the pattern is AXXCH. Further, the CXXCH pattern occurs in the denitrification system component nirT (O4PSZ), adrenodoxin (AXBO), putidaredoxin (TXPSEP), nitrate reductase β-chain (RDECNB), thymidine kinase (KIVZSW), kinase-related transforming protein-erbB (TVCHLV), epidermal growth factor receptor (GQFFE), phosphotransferase (QQBEJ5), DNA-directed RNA polymerase (RNBY3L), etc., in addition to the cytochrome *c* sequences. On the other hand, the pattern ^{94}P G T K M X F^{100} seems to be the signature for cytochrome *c* sequences. This pattern occurs only in cytochrome *c* sequences and in no other protein listed in either SwissProt, release 21 (February 1992) or NBRF-PIR, release 32 (March 1992).

Multiple alignment of 100 α-hemoglobins consisting of α-I, α-II and α-A was carried out. In such a set there is no long pattern of conserved amino acids. However, the *VI* at several positions is small. On the other hand, the alignment output of only β-chain of hemoglobin from 100 different species gives $VI = 1$ for several residues, showing that the amino acids in these positions are conserved (Figure 3). The high variability is seen at positions 70 and 76. The multiple alignment output of β-chain of hemoglobin shows that ^{89}L S E L H C^{94} and ^{133}K V V X G V A X A L A^{143} are two long, conserved regions. It was noticed that the patterns LSELHC and KVVXGVAXALA occur in the β, β-I, β-II, β-A, β-C, δ, ε, θ chains of hemoglobins. No other protein has these patterns. These examples show that multiple alignment of a large number of sequences provides very useful information.

Multiple alignment score

In pairwise sequence alignment studies, the alignment score (*A*), the maximum real similarity score (*P*), the average similarity score (*Q*) and the standard deviation (*Sd*) associated with *Q* are used by Barker *et al.* (1990) and Dayhoff *et al.* (1979) to assign superfamily, family, subfamily, group, etc. Recently, Wu *et al.* (1992) used a neural net to categorize the given sequence. We have attempted here to show that the single-parameter, multiple-alignment score (*S*) can be used to categorize sequences with high accuracy and resolution if the size of the family is large. The example of α-hemoglobin is discussed.

Table II. Effect on multiple alignment score (*S*) with addition of proteins from different groups

Data set	Multile alignment score (*S*)
10 α	298.4
10 α + 1 β	253.8
10 α + 2 β	143.2
10 α + 3 β	92.9
10 α + 1 globin *v*	92.7
10 α + 2 globin *v*	57.5
10 α + 3 globin *v*	39.6
10 α + 1 myoglobin	56.3
10 α + 2 myoglobin	32.4
10 α + 3 myoglobin	24.5
10 α + 3 β + 3 myog + 3 glob *v*	18.5

It can be seen from Table II that $S(\alpha) = 298$ (for 10 α-hemoglobins), and decreases significantly and continuously as one, two and three β-hemoglobins are aligned with ten α-chains. Further, the decrease in *S* is very large when a single myoglobin is aligned with the set of 10 α-hemoglobins. A decrease in the value of *S* is observed when two and three myoglobins are aligned. This indicates that sequences of α-chains are closer to β-chains of hemoglobin than to myoglobins, in agreement with the experimental studies. Alignment of globin *v* sequences along with 10 α-chain sequences shows that $S(\alpha + \text{globin } v) < S(\alpha)$, but correspondingly that $S(\alpha + \text{globin } v) > S(\alpha + \text{myoglobin})$. One can thus draw the conclusion that the set of α-chain sequences is similar to β-chain, globin *v* and myoglobin sequences, in that order. This example shows that if a new sequence belongs to a different group, *S* decreases considerably and thus the position of the new sequence can be asigned by comparing the change in *S* among various groups. In short, the parallelization approach discussed above brings into focus the use of multiple alignment studies for categorizing protein sequence data.

Acknowledgements

We acknowledge financial support from the Department of Biotechnology, the Government of India and the Center for Development of Advanced Computing, India.

References

Barker,W.C., George,D.G. and Hunt,L.T. (1990) Protein sequence database. *Methods Enzymol.*, **183**, 31–49.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.*, **183**, 403–427.

Barton,G.J. and Sternberg,M.J.E. (1987) A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.*, **198**, 327–337.

Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.

Dayhoff,M.O. (1976) Survey of new data and computer methods of analysis. *Atlas of Protein Sequence and Structure Data*, Vol. 3, pp. 1–6.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1979) A model of evoutionary change in proteins. Matrices for detecting distant relationships: computer

methods and results. *Atlas of Protein Sequence and Structure Data*, Vol. 5, Suppl. 3, pp. 345–358.
Dickerson,R.E. (1971) Sequence and structure homologies in Bacterial and mammalian type cytochromes. *J. Mol. Biol.*, **57**, 1–15.
Doolittle,R.F. (1990) Searching through sequence database. *Methods Enzymol.*, **183**, 99–110.
Feng,Da-Fei and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic tree. *Methods Enzymol.*, **25**, 351–360.
Feng,Da-Fei and Doolittle,R.F. (1990) Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.*, **183**, 375–387.
Gribskov,M., Homyak,M., Edenfield,J. and Eisenberg,D. (1988) Profile scanning for three-dimensional structural patterns in protein sequences. *Comput. Applic. Biosci.*, **1**, 61–66.
Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
Johnson,M.S. and Doolittle,R.F. (1986) A method for the simultaneous alignment of three or more amino acid sequences. *J. Mol. Evol.*, **23**, 267–278.
Jongeneel,C.V., Bouriver,J. and Bairoch,A. (1989) A unique signature identifies a family of zinc-dependent metallopeptidases. *FEBS Lett.*, **242**, 211–214.
Jue,R.A., Woodbury,N.W. and Doolittle,R.F. (1980) Sequence homologies among *E.coli* ribosomal proteins: evidence for evolutionary related groupings and internal duplications. *J. Mol. Evol.*, **15**, 129–148.
Kolaskar,A.S. and Kulkarni-Kale,U. (1992) Sequence alignment approach to pick up conformationally similar protein fragments. *J. Mol. Biol.*, **223**, 1053–1061.
Lander,E., Mesirov,J. and Washington,T,,IV (1989) Study of protein sequence comparison matrices on the connection machine CM-2. *J. Supercomput.*, **3**, 255–269.
Lipman,D.L., Altschul,S.F. and Kececioglu,J.D. (1989) A tool for multiple sequences alignment. *Proc. Natl. Acad. Sci. USA*, **86**, 4412–4415.
McLachlam,A.D. (1971) Tests for comparing related amino acid sequences. *J. Mol. Biol.*, **61**, 409–424.
Murata,M., Richardson,J.S. and Sussman,J.L. (1985) Simultaneous comparison of three protein sequences. *Proc. Natl. Acad. Sci. USA*, **82**, 3073–3077.
Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet.*, **9**, 180–189.
Sellers,P.H. (1974) Evolutionary distances. *SIAM J. Appl. Math.*, **26**, 787–793.
Smith,T.F., Waterman,M.S. and Fitch,W.M. (1981) Comparative biosequence matrices. *J. Mol. Evol.*, **18**, 38–46.
Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, **28**, 161–169.
Vingron,M. and Argos,P. (1989) A fast and sensitive multiple sequence alignment algorithm. *Comput. Applic. Biosci.*, **5**, 115–121.
Waterman,M.S., Smith,T.E. and Beyer,W.A. (1976) Some biological sequence matrices. *Adv. Math.*, **20**, 367–387.
Wilbur,W.J. and Lipman,D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*, **80**, 726–730.
Wu,T.T. and Kabat,E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Med.*, **132**, 211–250.
Wu,C., Whitson,G., McLarty,J., Ermongkonchai,A. and Chang Tzu-Chung (1992) Protein classifiation artificial neural system. *Protein Sci.*, **1**, 667–677.

Received on May 29, 1992; accepted on November 10, 1992