CABIOS

Vol.4. no.1. 1988
Pages 35 – 40

A software tool for finding locally optimal alignments in protein and nucleic acid sequences

Jennifer D.Hall and Eugene W.Myers[1,2]

Abstract

We describe software for aligning protein or nucleic acid sequences based on the concept of match density. This method is especially useful for locating regions of short similarity between two longer sequences which may be largely dissimilar (e.g. locating active site regions in distantly related proteins). Our software is able to identify biologically interesting similarities between two sub-regions because it allows the user to control the matching parameters and the manner in which local alignments are selected for display. Furthermore, the collection and ranking of alignments for display uses a novel, highly efficient algorithm. We illustrate these features with several examples. In addition, we show that this tool can be used to find a new conserved sequence in several viral DNA polymerases, which, we suggest, occurs at a functionally important enzymatic site.

Introduction

Computer analysis is a powerful method for aligning protein and nucleic acid sequences. Alignment has been especially useful for discovering the functional and evolutionary relatedness of sequences. For example, because active site regions are frequently conserved in related proteins, amino acid similarities may indicate the locations of active sites and suggest which residues within these sites are functionally important.

Our particular concern has been to use computers to find regions of similarity in proteins which may show little overall similarity. Our interest stems from a study of viral DNA polymerases (Hall *et al.*, 1986; Hall, 1988) (including enzymes from herpes simplex type 1, Epstein Barr and adeno type 2 viruses and bacteriophage ϕ29 from *Bacillus subtilis*). These sequences have been shown by various methods to exhibit regions of strong sequence similarity. Furthermore, the conserved regions are known to encompass active site(s), since many mutations (in herpes simplex) which alter active site function map in these regions. Hence, alignments between these protein sequences provide valuable information for understanding the molecular basis of DNA replication.

The traditional method for sequence alignment, developed by Needleman and Wunsch (1970) and Sellers (1974), finds the minimum number of substitutions, insertions and deletions required to convert the entirety of one sequence into the entirety of another. While this method gives the best overall, or global, alignment, it may fail to discover short regions within the two sequences which match significantly above average or are separated by large discontinuities (see Goad, 1986, for review).

We encountered these problems in attempting to align the DNA polymerase sequences discussed above by this traditional method. First, while highly related sequences (e.g. from the herpes viruses, herpes simplex and Epstein Barr) showed extensive similarity, they frequently contained regions of strong similarity separated by large regions of no similarity. Second, sequences from very distantly related organisms (e.g. herpes simplex and adeno type 2 viruses and ϕ29 phage) showed strong similarity only in very short regions. Consequently, we wished to utilize a method which searches for locally optimal alignments.

Variations of the Needleman–Wunsch method have been developed (Smith and Waterman, 1981; Goad and Kanehisa, 1982; Sellers, 1984) which find subsequences within the matrix exhibiting locally optimal alignment by maintaining a minimum match density. It is the algorithm described by Sellers (1987) which we have utilized in the software presented here. Because this algorithm generates a large number of short, insignificant alignments, in addition to the significant ones, we augmented our software with a single-pass, post-analysis phase that eliminates undesirable alignments and displays the remainder in order of priority. Moreover, our software has many options which permit the user to adjust the nature of the scoring schemes, mismatch density and ranking features.

Algorithms

An alignment problem involves two sequences A and B and a scoring scheme d. Suppose $A = a_1a_2 \ldots a_m$ and $B = b_1b_2 \ldots b_n$ are sequences of length M and N over some alphabet S. Under the traditional model of Sellers (1974), $d(a,b)$ is the score of aligning a and b, $d(a,-)$ is the score for not aligning symbol a in A, where $-$ is a special symbol not in S, and $d(-,b)$ is the score for not aligning b in B. For example, suppose the score for aligning matching symbols is 0 and the score for mismatched, inserted and deleted symbols is 1. Then the alignment below has score 3:

Departments of Molecular and Cellular Biology and [1]Computer Science, University of Arizona, Tucson, AZ 85721, USA
[2]Present address: Department of Computer Science, The Pennsylvania State University, University Park, PA 16802, USA

$$\begin{array}{c} A - A G T \\ A G A - A \end{array}$$
$$d(A,A) + d(-,G) + d(A,A) + d(G,-) + d(T,A) = 0 + 1 + 0 + 1 + 1 = 3$$

An optimal alignment has the minimum score possible.

It is well known that an alignment problem may be viewed as one of finding a least costly (lowest scoring) path between two vertices in a weighted and acyclic edit graph. This directed graph consists of a $(M + 1)$ by $(N + 1)$ matrix with vertices (i,j), where i has values 0 to M and j has values 0 to N. The designated source vertex is (0,0) and the sink vertex is (M,N). For all $i \geq 1$, there is a vertical edge from $(i - 1,j)$ to (i,j) of weight $d(a_i,-)$; for all $j \geq 1$, there is a horizontal edge from $(i, j - 1)$ to (i,j) of weight $d(-,b_j)$; and for all $i,j \geq 1$ there is a diagonal edge from $(i - 1,j - 1)$ to (i,j) of weight $d(a_i,b_j)$. Figure 1 gives the edit graph for the two sequences in the example above.

Every path from the source to sink describes an alignment where a_i is aligned with b_j for every diagonal edge to (i,j) in the path. Vertical or horizontal edges to (i,j) indicate that the symbol a_i or b_j respectively, is not aligned. For example, the highlighted path, in Figure 1:

$$(0,0) \rightarrow (1,1) \rightarrow (1,2) \rightarrow (2,3) \rightarrow (3,3) \rightarrow (4,4)$$

corresponds to the alignment in the example above. The cost of a path is the sum of the weights of the edges in the path and is also the score of the path's corresponding alignment. Finding an optimal global alignment is equivalent to finding a minimum cost path from source to sink in the edit graph.

We are interested in finding subpaths within the graph which are unusually low scoring and, hence, represent significant local alignments. A subpath from vertex (i,j) to $(i + a,j + b)$ gives a local alignment between the substrings $a_{i+1}a_{i+2} \cdots a_{i+a}$ and $b_{j+1}b_{j+2} \cdots b_{j+b}$. A number of methods have been proposed for identifying such subpaths, all based on the concept of mismatch density (Smith and Waterman, 1981; Goad and Kanehisa, 1982; Sellers, 1984). We have chosen to use Seller's more complicated but more refined method (Sellers, 1984).

To describe Sellers' algorithm, define the length of a subpath P from (i,j) to $(i + a,j + b)$ to be $a + b$. The mismatch density of P is the ratio of its score to length. Given a mismatch density threshold m, the algorithm prunes the edit graph by removing edges whenever the mismatch density of P exceeds m. This process continues in a series of forward and backward sweeps through the graph. As m decreases, more edges are eliminated and fewer paths remain. All surviving paths have the following properties: (i) the mismatch density of every prefix and suffix of a given path is less than or equal to m; and (ii) any two paths that intersect have the same mismatch density along the prefixes and suffixes leading to their intersection vertex.

We found that for a setting of m which preserves desired local homologies, many short and insignificant homologies also survive Sellers' pruning process. Therefore, we included a post-analysis phase that selectively displays those subpaths which

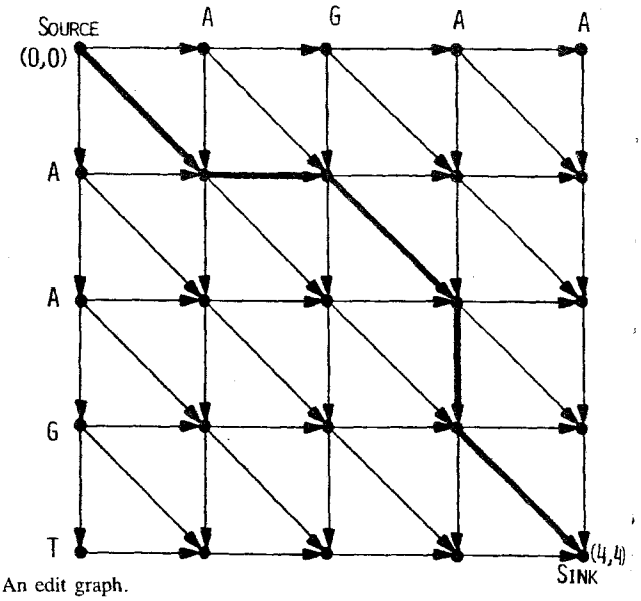

Fig. 1. An edit graph.

the user deems to be of interest. This procedure first selects a subpopulation of paths for analysis and then ranks them for display. The methods for selecting and ranking paths are independent of one another and user specifiable. These features provide greater flexibility when searching for biologically interesting local similarities.

The removal of edges disconnects the edit graph into a number of connected and acyclic subgraphs that we call zones. The post-analysis phase first identifies these zones and selects from each zone one of the potentially many paths from a source to a sink vertex as the zone representative. For each representative, three attributes—length, score and significance—are recorded. The significance of a path is the number of standard deviations that its score is above the mean in a distribution of scores. This distribution is the one that arises assuming every choice of the two strings aligned by this path is equally likely. The zone representative may be selected as either the longest or shortest source to sink path. Selection is on the basis of length, because (i) all paths in a zone have the same score, and (ii) selection of the most significant path is computationally intractable, because it requires examining every one of the potentially exponential number of representatives. Moreover in practice, the shortest path is usually the most significant.

Once the zone representatives have been determined, the post-analyzer selects a subpopulation by discarding any representatives which do not meet a cut-off criterion. This user-specified criterion may be any programmable predicate based on the three available path attributes: significance, score and/or length. The paths in the selected subpopulation are then ranked according to a user-specified ranking function, also based on the path attributes. Finally, the K best are displayed where K is a user-specified integer.

Table I. The most significant alignment at different mismatch densities

Maximum mismatch density	
0.35	F D W P F
	F D W P Y
0.40	V T G Y N I I N F D W P F L L
	V T G Y N V A N F D W P Y I L
0.45	V L E F D S E F E M L L A F M T L V K Q Y G P E F V T G Y N I I N F D W P F L L
	V Y E F P S E L D M L Y A F F Q L I R D L S V E I V T G Y N V A N F D W P Y I L
	+ +

We show alignments between DNA polymerase protein sequences from herpes simplex virus type 1 (Gibbs *et al.*, 1985) (residues 436–476) (upper sequence) and Epstein Barr virus (Baer *et al.*, 1984) (residues 349–389) (lower sequence). The scoring option for substitutions was -sD (Dayhoff). Other program default values were used, except we varied the -m (mismatch threshold) option, and we used -zL. + below the sequence indicates residues which match exactly.

A novel algorithmic aspect of this software tool is incorporated into the post-analysis phase. Namely, it was discovered that it is possible to detect zones and their representatives in a single $O(MN)$ sweep of the pruned edit graph. Rather than using the obvious search-and-mark connectivity-based approach, the 'scan-line' paradigm frequently used in computational geometry led to a superior algorithm for planar and grid-like edit graphs. In essence, the algorithm sweeps left to right, column by column, through the pruned edit graph, incrementally retaining those zone subparts and potential representatives that have been seen in the 'swept' region. The $O(MN)$ efficiency of the resulting post-analysis phase is such that the dominant cost of running the program is due to Sellers' pruning module. Moreover, this idea may ultimately lead to a complexity improvement in Sellers' algorithm itself.

Implementation

This software tool is called LOCAL. It is written in C, and should be portable to systems with C compilers. It presently operates on a VAX 8600 with the Unix 4.3 BSD operating system. LOCAL may be invoked with a number of options as follows:

local [-s{D/C/U} -mreal -z{S/L} -cfile] file_A[range] file_B[range]
-greal -l{D/P} -dint -rfile

Sequence format

File A and file B are text files containing the sequences to be compared. Within these files, new lines, tabs, spaces and entire comment lines beginning with > in column 1 are ignored. All other characters are considered to be part of the sequence. The program accepts sequences in single letter amino acid code or using the symbols A, T, G or C for nucleic acids. By convention, the sequences should read amino- to carboxy-terminus (proteins) or 5′ to 3′ (nucleic acids).

A substring of either sequence (from residue x to y) may be used by following its file name with an optional range specification. A range has the syntax, "file[x,y]".

Table II. The most significant matches using different scoring options

Scoring matrix	
Dayhoff	R I I Y
	R I I Y
Exact match	R I I Y G D T D S I F V L C R G L T
	R I I Y C D T D S I _H L _T G _T

We show alignments between DNA polymerase protein sequences from herpes simplex virus type 1 (residues 880–900) (upper sequence) and ϕ29 bacteriophage (Yoshikawa and Ito, 1982) (residues 450–470) (lower sequence). The default values for the program were employed, except we varied the -s (scoring) option and used -zL. The significance of the overlined residues is discussed in the text.

Sequence length restrictions

The maximum sequence length is 2048 and the maximum number of displayable homologies is 512.

Scoring options

The -s option specifies a substitution scoring matrix. There are three choices, each denoted by a letter.

D: A scaled version of the Dayhoff matrix (Dayhoff *et al.*, 1978) is used. The scores have been scaled between 0 and 1.45 with the linear transform $y = (173 - x)/251$. Substitutions weighted close to 0 are favored, those around 1 unbiased, over 1 are biased negatively.

C: Substitutions between amino acids are scored between 0 and 1 according to the distance between their DNA triplets or codons.

U: Substitutions are scored 1 except if the symbols match exactly in which case their score is 0.

The D and C matrices can only be used on proteins, while U can also be used on nucleic acids.

The -g option sets the score for insertions and deletions which by default is 1.0. Note that this value is an appropriate penalty for all three substitution options.

Table III. The most significant alignment using different insert/delete penalties

Insert/delete penalty	
0.75	A P G P Y S M R I I Y G D T D S I F V L C R G L T A A G L T A M G D K M A S H I S R A L F L P P I K L E C ____ E K T F T K L L L I A K K K Y
	A Q A C Y _D R I I Y C D T D S I H ______ L T __G _T E I P D V I K D _I V D ______ P K K L G Y W A H E S T F _K R ____ A __K Y
	+ +
1.25	R I I Y G D T D S I
	R I I Y C D T D S I

We show alignments between DNA polymerase protein sequences from herpes simplex virus type 1 (residues 870–950) (upper sequence) and ϕ29 bacteriophage (residues 440–500) (lower sequence), using the exact match scoring method. We employed the other program default values, except we varied the -g (insert/delete) option and used -zL. + indicates residues which match exactly.

Table IV. The three most significant alignments obtained with different ranking methods

Rank by significance	Signif. = 7.505 F H V Y D I L E N V F H V Y D I L E T V	Signif. = 6.980 W W	Signif. = 6.976 F Y N P Y L A P F Y N P F L R P
Rank by score	Score = 9.800 I T P T G T V I T L L G V V P C G I V I K L L G	Score = 6.940 F H V Y D I L E N V F H V Y D I L E T V	Score = 5.440 F Y N P Y L A P F Y N P F L R P
Rank by length	Length = 24 I T P T G T V I T L L G V V P C G I V I K L L G	Length = 22 P T Q R H T Y Y S E C P S D K Q G Y V V P C	Length = 20 F H V Y D I L E N V F H V Y D I L E T V

We show alignments between DNA polymerase protein sequences from herpes simplex virus type 1 (residues 1–200) (upper sequence) and Epstein Barr virus (residues 1–200) (lower sequence). The scoring option for substitutions was -sD (Dayhoff), and the mismatch density threshold was 0.45. Other program default values were used, except we varied the -r (ranking) option and used -zL.

Table V. Partial alignment of several DNA polymerase protein sequences

hsv	L A A R G L P T P V V L E F D S E F E M L L A F M T L V K Q Y G P E F V T G Y N I I N F D W P F L L A K L T D I Y K V P L D G Y
vzv	M K D A G L P E P T V L E F D S E F E L L I A F M T L V K Q Y A P E F A T G Y N I V N F D W A F I M E K L N S I Y S L K L D G Y
ebv	G V _E _V _Y E P F S E L D M L Y A F F Q L I R D L S V E I V T G Y N V A N F D W P Y I L D R A R H I Y S I N P A S L
vac	D Y V V T F N G H N F D L R Y I T N R L
ad2	L S E H G L S S P E E L T Y E _E L K K L P S I K G _T P R F L E L Y I V G H N I N G F D
	+ +

We aligned the four lower sequences [varicella zoster (Davison and Scott, 1986), vzv; Epstein Barr, ebv; vaccinia, vac; adeno type 2, ad2] to that from herpes simplex type 1 (hsv). In each case the first 500 residues were used and matched using the -sD (Dayhoff) scoring option. Other program default values were employed, except we used -zL. Very short alignments were discarded using the -c option and the five best scores collected. The alignments shown appeared within this group. + indicates the appearance of the same amino acid in three or more sequences.

Density options

The -m option sets the mismatch density threshold, i.e. the ceiling on the score/length ratio. The -l flag provides two options for the length of an alignment.

P: The length is the sum of the number of columns in the display.

D: the length is the sum of the number of symbols in the aligned subsequences.

Display options

Within each zone of intersecting paths remaining in the graph either the shortest, -zS, or the longest, -zL, alignment is chosen as the zone representative.

The -c option eliminates representatives from consideration for display. The remaining representatives are ranked according to the -r function. The -d option determines the number of alignments displayed, starting, in order, with the highest ranking representative. The displayed alignments are given in lexicographical order along the length of the first sequence file in the alignment.

Every representative is attributed with its length, LEN, its score, SCR, and its significance, SIG. To eliminate representatives with the -c (cut-off) option, give a C-expression in terms of variables LEN, SCR or SIG in a specified file. This expression is applied to every representative. If non-zero, the representative is eliminated, if zero it is kept. To rank representatives with the -r (rank) option, give a C-expression in terms of variables LEN1, LEN2, SCR1, etc. Representative 1 will be considered higher ranking than representative 2, if and only if the expression is non-zero.

If the -c or -r option is not given, LOCAL first looks for

a file named CUTOFF or RANK in the current directory and, if present, uses the expressions within them. If these files do not exist, LOCAL uses the default expression given below.

Default settings

local -sD -g1.0 -m0.4 -lD -zS -cCutoff -rBetter -d50 file__A file__B
Cutoff = '0' 'All candidates are considered'
Better = 'SIG1 > SIG2' 'Candidates are ranked by significance'

Discussion

This program is especially useful for finding alignments between sequences where the overall similarity is weak. In this case, LOCAL is able to identify meaningful alignments from among the majority of small, insignificant ones, because it allows the user to control its ranking and display features. In addition, the options for mismatch density threshold and for the penalty values for substitutions, insertions and deletions also allow detection of alignments which might otherwise not be considered significant. We illustrate some of these flexible features with several examples. In addition, we use this software tool to make the novel finding of a conserved sequence in several distantly related DNA polymerase sequences (see discussion of Table V).

Mismatch density option

The mismatch density option sets the mismatch density threshold in a given aligned region. When few mismatches are tolerated, the program finds short, highly matched regions, while if a high mismatch density is allowed, longer, less stringently matched regions are detected. Table I illustrates this feature. In this example, only at relatively high mismatch density does the alignment on the right (centered around sequence FDWPF) extend across a region of low similarity to include the alignment at the left. Consequently, the choice of mismatch density can be used to identify the most highly conserved regions in two sequences as well as flanking sequences showing less similarity.

Scoring matrix option

Three methods for weighting substitutions in protein alignments are provided: (i) the exact match matrix which scores only identities; (ii) the codon matrix which uses the genetic code as a basis for weighting; and (iii) the Dayhoff matrix which was derived from observed amino acid replacements in related proteins (Dayhoff *et al.*, 1978). We plan to add options which score according to chemically similar amino acids or by the method of Argos (1987) which uses both chemical and structural characteristics of amino acids.

An example of the difference between the exact and Dayhoff matrices is shown in Table II. The sequence YGDTDS (overlined) is highly conserved in several animal virus and bacteriophage DNA polymerase sequences (Hall, 1988). Because it is not present in other known protein sequences (Hall, 1988), it may be functionally significant for this class of polymerases. Under the conditions in Table II, this sequence is aligned with a high score when comparing sequences from herpes simplex and ϕ29 by the exact match matrix. However, it is not found by the Dayhoff scoring matrix, presumably due to the highly unfavorable alignment of cysteine and glycine required. In contrast, others have found that the Dayhoff matrix is the most effective for global alignments of distantly related proteins (Feng *et al.*, 1985). Consequently, our choice of scoring options allows identification of alignments which may only appear significant with certain scoring schemes.

Insertion/deletion penalties

Since the scoring method for substitutions can be specified, it is important that insertion/deletion penalty also be flexible. Use of this option finds short, highly matched regions when the penalty is high and weaker alignments, containing gaps, when the penalty is low, as illustrated in Table III.

Ranking option

A useful display feature is the choice of method for ranking the alignments. Three options are available: ranking by significance, score or length. Where the alignment is strong, these methods tend to give the same result. However, when the alignment is weaker or when a scoring method other than the exact match matrix is used (e.g. the Dayhoff matrix), these methods can give different results, as illustrated in Table IV. This difference occurs because the score of the path increases with length (see Table IV), while the significance of the alignment depends on both the mismatch density and length. The ranking options are, therefore, useful for identifying important alignments which would receive a low ranking if only the default (by significance) method were available.

Use of LOCAL to find new functional domains in DNA polymerases

A major use for this program is to identify conserved domains within proteins which may not show extensive overall homology. We have found this approach fruitful in detecting conserved regions in DNA polymerases.

A particular group of viral DNA polymerases has recently been found to exhibit substantial amino acid homologies (Gibbs *et al.*, 1985; Quinn and McGeoch, 1985; Argos *et al.*, 1986; Davison and Scott, 1986; Hall *et al.*, 1986; Hall, 1988; Larder *et al.*, 1987). These viruses include several herpes viruses, vaccinia virus, adeno type 2 virus and ϕ29 bacteriophage. Sequence comparisons reveal extensive homology within the herpes virus group, strong similarity between the herpes and the vaccinia virus sequences and shorter regions of strong homology with the adeno and phage sequences. Several mutations in herpes

simplex which affect substrate (deoxynucleoside triphosphate) recognition map at or near some of these highly conserved sites (Hall, 1988). Consequently, these conserved regions appear to be located at the substrate binding site and may be important for enzyme function. This result further predicts that structural and mechanistic similarities occur within the enzymes in this group. Hence, these conserved sequences may hold the key for understanding the function of this class of DNA polymerases.

At least three of these conserved regions have been described previously (Argos *et al.*, 1986; Earl *et al.*, 1986; Hull *et al.*, 1986; Larder *et al.*, 1987). Based on our studies with LOCAL, we suggest that a fourth region is also important. This region had been found previously to be conserved among the highly related sequences from herpes simplex virus type 1, Epstein Barr virus and vaccinia virus (Earl *et al.*, 1986; Hall *et al.*, 1986). We now show in Table V that this homology extends to adeno virus type 2 and a third herpes virus, varicella zoster. The fact that this region is conserved in such a distantly related virus as adeno type 2 suggests that this region is indeed functionally important.

Acknowledgements

The authors thank D.Mount for many helpful discussions. This work was supported by National Institutes of Health grant GM32646 and National Science Foundation grant DCR-8511455.

References

Argos,P. (1987) *J. Mol. Biol.*, **193**, 385–396.
Argos,P., Tucker,A.D. and Philipson,L. (1986) *Virology*, **149**, 208–216.
Baer,R., Bankier,A.T., Biggen,M.D., Deininger,P.L., Farrell,P.J., Gibson,T.J., Hatfull,G., Hudson,G.S., Satchwell,S.C., Seguin,C., Tuffnell,P.S. and Barrell,B.G. (1984) *Nature*, **310**, 207–211.
Davison,A.J. and Scott,J.E. (1986) *J. Gen. Virol.*, **67**, 1759–1816.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedial Research Foundation, Washington, DC, Vol. 5, suppl. 3, pp. 345–358.
Earl,P.L., Jones,E.V. and Moss,B. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3659–3663.
Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) *J. Mol. Evol.*, **21**, 112–125.
Gibbs,J.S., Chiou,H.C., Hall,J.D., Mount,D.W., Retondo,M.J., Weller,S.K. and Coen,D.M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 7969–7973.
Goad,W.B. (1986) *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 79–95.
Goad,W.B. and Kanehisa,M.I. (1982) *Nucleic Acids Res.*, **10**, 247–263.
Hall,J.D. (1988) *Trends Genet.*, in press.
Hall,J.D., Gibbs,J.S., Coen,D.M. and Mount,D.W. (1986) *DNA*, **5**, 281–288.
Larder,B.A., Kemp,S.D. and Darby,G. (1987) *EMBO J.*, **6**, 169–175.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
Quinn,J.P. and McGeoch,D.J. (1985) *Nucleic Acids Res.*, **13**, 8143–8163.
Sellers,P.H. (1974) *J. Combinator. Theory*, **16**, 253–258.
Sellers,P.H. (1984) *Bull. Math. Biol.*, **46**, 501–514.
Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
Yoshikawa,H. and Ito,J. (1982) *Gene*, **17**, 323–335.

Received on August 17, 1987; accepted on November 17, 1987