An Efficient Algorithm to Locate All Locally Optimal Alignments Between Two Sequences Allowing for Gaps

Geoffrey J. Barton

Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU
U.K.

Tel: 0865-275368
Fax: 0865-510454
e-mail: gjb@bioch.ox.ac.uk

Keywords: Smith-Waterman / Dynamic programming / locally optimal alignment / protein sequence

Running Head: Finding All Locally Optimal Alignments

This paper appeared in Computer Applications in the Biosciences, (1993), 9, 729-734

1 Abstract

An efficient algorithm is described to locate locally optimal alignments between two sequences allowing for insertions and deletions. The algorithm is based on that of Smith and Waterman (*J. Mol. Biol.*, **147**, 195–197, 1981) which returns the single best local alignment. However, the algorithm described here permits *all* non-intersecting locally optimal alignments to be determined in a single pass through the comparison matrix. The algorithm simplifies the location of repeats, multiple domains and shuffled motifs and is fast enough to be used on a conventional workstation to scan large sequence databanks.

2 Introduction

Dynamic programming algorithms that locate optimal alignments of two sequences are central techniques for the comparison of biological sequences [Needleman & Wunsch, 1970, Sellers, 1974, Smith & Waterman, 1981] or three-dimensional structures [Barton & Sternberg, 1988, Taylor & Orengo, 1989, Sali & Blundell, 1990, Russell & Barton, 1992]. The algorithms can be divided broadly into those that seek to find a *global* alignment between the sequences (e.g. [Needleman & Wunsch, 1970] and those that find *local* alignments (e.g [Erickson & Sellers, 1983, Smith & Waterman, 1981]). Global alignment methods optimize the score for alignment over the full length of both sequences, and are most appropriate when the sequences are known to be similar over their entire length. Local alignment methods allow the common sub-regions of the two sequences to be identified and are appropriate when it is not known in advance if the sequences being compared are similar. Local alignment methods are effective in locating common sub-domains between long sequences that otherwise share little similarity. This feature makes such algorithms suitable for scanning large sequence databanks for similarities to a newly determined sequence.

The Smith and Waterman [Smith & Waterman, 1981] algorithm is perhaps the most widely used local similarity algorithm for biological sequence comparison. The algorithm identifies the single highest scoring sub-sequence alignment and allows for gaps (insertions/deletions). However, it is often true that there may be more than one biologically important alignment between two sequences. For example, a protein domain may be repeated, or domains may be shuffled within multi-domain proteins. Waterman and Eggert [Waterman & Eggert, 1987] have shown how the Smith-Waterman algorithm may be extended to locate the second-best and subsequent local alignments with minimal recalculation, subject to the primary restriction that the different alignments should not intersect. Here, I describe an algorithm that allows all such locally optimal alignments to be determined *without* the need for re-calculation. The algorithm is similar in principle to that developed by Coulson *et al.* for the parallel processing Distributed Array Processor (DAP) [Coulson *et al.*, 1987]. However, the algorithm described here is general and has been implemented in C for widely available computers. The algorithm may be applied to any problem that is amenable to the Smith-Waterman dynamic programming algorithm.

3 Efficient Determination of Best Score

In order to simplify the explanation of the “all local alignment” algorithm I shall first recapitulate a well known method for efficient computer implementation of the Smith-Waterman [Smith & Waterman, 1981] algorithm. For a full introduction to dynamic programming algorithms in sequence comparison see [Kruskal, 1983].

The best score for a local alignment between two sequences A and B of length m and n is determined by calculating the comparison matrix $H_{m,n}$ starting with $H_{1,1}$ and working forwards through the matrix column by column. The value of each cell $H_{i,j}$ is given by the equation:

$$H_{i,j} = \max \left\{ \begin{array}{c} H_{i-1,j-1} + w_{A_i,B_j} \\ H_{i,j-1} + w_{A_i,\Delta} \\ H_{i-1,j} + w_{\Delta,B_j} \\ 0 \end{array} \right\}$$

where w_{A_i,B_j} is the score for equivalencing A_i and B_j, and $w_{A_i,\Delta}, w_{\Delta,B_j}$ are the scores for aligning with a gap in B or A respectively. To calculate the value of $H_{i,j}$, it is only necessary to know the contents of the three predecessor cells ($H_{i-1,j-1}, H_{i,j-1}, H_{i-1,j}$). If just the best score is required, and no alignment, then only the previous column, C of the matrix, need be stored together with the score for the current cell, r and previous cell, p in the column. This is illustrated in Figure 1[1] for the comparison of A = C-C-A-A-T-C-T-A-C-T-A-C-T-G-C-T-T-G-C-A- G-T-A-C and B = A-G-T-C-C-G-A-G-G-G-C-T-A-C-T-C-T-A-C-T-G-A- A-C with $w_{A_i,B_j} = 10$ if $A_i = B_j$, $w_{A_i,B_j} = -9$ if $A_i \neq B_j$, and $w_{A_i,\Delta} = w_{\Delta,B_j} = -20$. This example is used here to allow a direct comparison of the “all local alignment” algorithm with the work of Waterman and Eggert [Waterman & Eggert, 1987].

Figure 1[2] illustrates the processing of $H_{7,5}$ and $H_{8,5}$. The cells of H that are stored are shown in large numerals. The cells shown in small numerals are discarded. The best overall score is updated if the value of the current cell is greater than the maximum so far. This is the simplest efficient implementation of the “best score only” local similarity algorithm, and may be coded to run very fast. It is also straightforward, by the addition of a further 1-dimensional array to

[1] Figure1.ps

[2] Figure1.ps

adapt the algorithm to cope with a gap-penalty function having the form $w_k = u_k + v$ where k is the gap-length [Gotoh, 1982].

4 All local alignment algorithm

Smith and Waterman [Smith & Waterman, 1981] identified the best local alignment by storing the entire H matrix, finding the maximum element, then tracing back through the matrix. Waterman and Eggert [Waterman & Eggert, 1987] described an algorithm to identify alternative locally optimal alignments by partial re-calculation of the H matrix subject to the condition that the alignment paths do not intersect, and that the first and last equivalenced pairs have a positive score. Here, I show that for a length-dependent gap-penalty the scores for all such locally optimal alignments can be obtained on a single pass through H. The essential observation is that since alignment paths are not allowed to intersect, it is only possible to have one path passing through each cell $H_{i,j}$. Therefore, when processing H it is simply necessary to maintain a record of the starting residues and best score for the alignment that passes throught the current cell $H_{i,j}$. The algorithm requires storage for the column scores C as for the single best score algorithm. In addition, the current maximum path scores M, and the start and end of the local alignment, S, E, where S and E hold the co-ordinates of the cells in H where the alignment starts and ends must also be stored.

Figure 2(a-i)[3] illustrates the processing of H to find the first, but not the highest scoring, locally optimal alignment between the sequences. For the sake of clarity, only the elements of C, M, S and E that are relevant to this alignment are illustrated in each figure. Figure 2a[4] shows the first cell in the alignment $H_{3,1} = 10$, this is also the maximum score for the alignment so far, and the alignment starts and ends with the same pair of residues $S_3 = E_3 = (3, 1)$. In Figure 2b[5], the alignment is continued, but since the current cell, $H_{4,2} = 1 (< [M_4 = 10])$, S_4, E_4 and M_4 are left unchanged. In Figure 2c[6], $H_{5,3} = 11$, so E_5 and M_5 are updated to 11 and (5,3) respectively. Similar

[3] Figure2.ps

[4] Figure2.ps

[5] Figure2.ps

[6] Figure2.ps

processing is shown in Figure 2(d-e)[7] where C, M, S and E are updated to the values 12, 21, (3,1), (6,4), while Figure 2f[8] illustrates termination of the path when $H_{8,5} = 0$. Now that this local alignment path has decayed to zero, the starting and ending coordinates are saved together with the alignment score. However, since we have not completed the processing of H, we do not yet know if this is the best possible score for an alignment starting in $H_{3,1}$. Figures 2f-i[9], show that the alignment cannot be extended further, so the maximum score of 21 for an alignment starting in $H_{3,1}$ stands. Had it been possible for the alignment to be extended, then the new best-score and end-point (M and E) would have replaced the 21, (6,4) currently saved on the results list. Once the entire matrix has been calculated (but not stored), the results list contains the score, start and end co-ordinates for all local alignments between the two sequences. Although not essential for the functioning of the algorithm, it is often desirable to set a minimum score threshold T_s such that only those local alignments where $M > T_s$ will be saved on the results list.

If it is necessary to generate the alignments rather than just report the best scores and end-points, then a direction matrix $e_{m,n}$ [Smith *et al.*, 1981, Gotoh, 1982] is saved during the calculation of H such that each element $e_{i,j}$ indicates which of $H_{i-1,j-1}$, $H_{i,j-1}$ or $H_{i-1,j}$ contributed to the value of $H_{i,j}$. Accordingly, $e_{m,n}$ may be a compact data type of as few as 3-bits per element, so the memory overhead for finding all locally optimal alignments between long sequences is modest compared to algorithms that require storage of H. Extension of the algorithm to allow gap-penalties of the form $u_k + v$ is straightforward if only best scores and end-points are required, but generation of the corresponding alignments will require two passes through H [Gotoh, 1982].

Figure 3[10] shows the 28 locally optimal alignments that are found between the sequences A and B. Alignments of length 1 are normally uninteresting and so are excluded. The two alignments illustrated by Waterman and Eggert [Waterman & Eggert, 1987] are ranked 1 and 2, with a further six alignments scoring ≥ 30. A total of 15 alignments score ≥ 20 with the remaining 13 optimal alignments scoring ≤ 12. Figure 4[11] illustrates the full H matrix with the paths highlighted that correspond to the 15 alignments scoring ≥ 20.

[7] Figure2.ps

[8] Figure2.ps

[9] Figure2.ps

[10] Figure3.ps

[11] Figure4.ps

5 Implementation and Efficiency

The “All local alignment” algorithm has been implemented in C. This language allows the C, M, S and E arrays to be grouped into a single data structure array of length m. The resulting code is faster than when C, M, S and E are coded separately, since sequentially accessed values are adjacent in memory. The results list is also of length m where each element points to a dynamically allocated “ragged” list of structures containing values for M, S and E.

As a check of the relative efficiency of the “all local alignment” algorithm, the protein sequence of human α– haemoglobin (141-residues, PIR code HAHU) was compared to a small sequence databank (PIR 14.0: 6,858 sequences, 2,080,148 amino acids) using the Dayhoff MDM250 matrix [Dayhoff *et al.*, 1978] and gap-penalty of 8 ($w_{A_i,\Delta} = w_{\Delta,B_j} = -8$). When run on a Sun SPARCstation 2, the efficient “best score only” algorithm determined one optimal local alignment for each databank sequence, and required 270 seconds to complete the scan. Another implementation of the Smith-Waterman algorithm SSEARCH [**?**] which also only returns the top scoring alignment for each sequence pair, required 1,100 seconds for the same scan. In contrast, the “all local alignment” algorithm described here with $T_s = 35$ examined 30,690 local alignments in only 1,000 seconds.

The value of T_s has little effect on the execution time, except when set very small such that large numbers of local alignments must be stored and sorted for each sequence comparison. For example, comparison of HAHU (141 residues) to one of the longest protein sequences known, Twitchin from *Caenorhabditis elegans* (PIR code S07571 - 6048 residues), finds 32,299 alternative local alignments when $T_s = 0$ and requires 16 sec CPU time. Setting $T_s = 35$ reduces the number of alignments to 58 and takes only 2 sec. The algorithms described here have also been implemented to work with *pscan* [Barton, 1991] to allow distributed processing on a network of workstations. *pscan* permits an approximately linear decrease in elapsed scan time with increasing numbers of processors.

6 Availability

An ANSI-C subroutine library that includes the “all local alignment” code and utility routines is available from the author. The files are quite small and can be e-mailed. Send requests to gjb@bioch.ox.ac.uk. Alternatively, please send a DOS formatted 3.5 inch disk with suitable packaging and a return address label.

7 Acknowledgements

I thank the Royal Society and Prof. L. N. Johnson for support, and a referee for pointing out the similarities between this work and that of Coulson *et al.* , [Coulson *et al.*, 1987].

References

[Barton, 1991] Barton, G. J. (1991). Scanning the protein sequence databank using a distributed processing workstation network. *Comput. Appl. Biosci.* **7**, 85–88.

[Barton & Sternberg, 1988] Barton, G. J. & Sternberg, M. J. E. (1988). Lopal and scamp: techniques for the comparison and display of protein structures. *J. Mol. Graph.* **6**, 190–196.

[Coulson *et al.*, 1987] Coulson, A. F. W., Collins, J. F. & Lyall, A. (1987). Protein and nucleic acid sequence database searching: a suitable case for parallel processing. *The Computer Journal,* **30**, 420–424.

[Dayhoff *et al.*, 1978] Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. matrices for detecting distant relationships. In *Atlas of protein sequence and structure,* (Dayhoff, M. O., ed.), vol. 5, pp. 345–358. National biomedical research foundation Washington DC.

[Erickson & Sellers, 1983] Erickson, B. W. & Sellers, P. H. (1983). Recognition of patterns in genetic sequences. In *Time warps, string edits and macromolecules: the theory and practice of sequence comparison*, (Sankoff, D. & Kruskal, J. B., eds), pp. 55–91. Addison Wesley.

[Gotoh, 1982] Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708.

[Kruskal, 1983] Kruskal, J. B. (1983). An overview of squence comparison. In *Time warps, string edits and macromolecules: the theory and practice of sequence comparison*, (Sankoff, D. & Kruskal, J. B., eds), pp. 1–44. Addison Wesley.

[Needleman & Wunsch, 1970] Needleman, S. B. & 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.

[Russell & Barton, 1992] Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct., Funct., Genet.* **14**, 309–323.

[Sali & Blundell, 1990] Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403–428.

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

[Smith & Waterman, 1981] Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

[Smith *et al.*, 1981] Smith, T. F., Waterman, M. S. & Fitch, W. M. (1981). Comparative biosequence metrics. *J. Mol. Evol.* **18**, 38–46.

[Taylor & Orengo, 1989] Taylor, W. & Orengo, C. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–21.

[Waterman & Eggert, 1987] Waterman, M. S. & Eggert, M. (1987). A new algorithm for best subsequence alignments with application to trna-rrna comparisons. *J. Mol. Biol.* **197**, 723–728.

8 Figure legends

8.1 Figure 1 - Efficient Determination of Best Score

Figure 1[12]

Two steps in the processing of the H matrix for the comparison of sequences A and B (see text). Only the values shown in large numerals are stored using a single vector C and two scalar values p and r.

8.2 Figure 2a-i - Finding the first locally optimal alignment without recalculation of H

Figure 2a-i[13]

Calculation of the H matrix. The vector C shown in Figure 1 is joined by a further five vectors to store the maximum score M on the current path, and the start S and current end-point E for the optimal alignment. To simplify the figure, only one element of each vector is illustrated. The sub-figures show the building up of the score, start and end point for the first locally optimal alignment to be found when processing the H matrix. In each sub-figure, the heavy-boxed cells of H have been assigned to the optimal alignment. The lightly boxed cells lie on the alignment path, but may follow the current maximum cell M. In Figure 2f, the score, start and end points of the locally optimal alignment has been stored in the results list, indexed by the row in which the alignment starts (3).

8.3 Figure 3

Figure 3[14]

The 28 locally optimal alignments found between A and B. The boxed alignments all score ≥ 20 and their paths are shown in Figure 4.

8.4 Figure 4

Figure 4[15]

[12] Figure1.ps

[13] Figure2.ps

[14] Figure3.ps

[15] Figure4.ps

The completed H matrix with the paths for the top 15 alignments highlighted.