2012 International Conference on Environment Science and Engieering
IPCBEE vol.32(2012) © (2012)IACSIT Press, Singapoore

In-Depth Assessment of Local Sequence Alignment

Atoosa Ghahremani and Mahmood A. Mahdavi[†]

Department of Chemical Engineering, Ferdowsi University of Mashhad, Azadi Square, Pardis Campus, 91779-48944, Mashhad, Iran

Abstract. The overall procedure of searching local similarities among amino acid sequences using local alignment tool has been described in detail. Different phases of the alignment and use of available scoring matrices are explained. The update of scoring matrices based on new sequence information regularly take place over time based on the overall procedure assembled here in a single document. This document assists young scientists who seek simple references to understand step-by-step local alignment.

Keywords: sequence alignment, local alignment, dynamic programming, modified local alignment, BLAST.

1. Introduction

Sequence alignment is a way for comparing two or more sequences in order to search a series of specific characters or character patterns with same order. It causes to identify positions and regions in protein sequences that provide insights into the function or structure of an uncharacterized sequence by predicting similarities to a protein that have been studied experimentally and ultimately infer homology between two real related sequences (share a common evolutionary ancestor). Computational approaches to sequence alignment generally ordered into two categories: global alignment (Needleman-Wunsch, 1970) and local alignment (Smith- Waterman, 1981). Two rapid alternative local alignment algorithms compared to the Smith-Waterman algorithm are FASTA (Pearson & Lipman, 1988) and BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990).

The biological reliable measure that discriminates homologs from unrelated sequences and indicates the degree of functional or structural similarity is score or expectation value. These values are computed from the alignment and give biologists a measure of the relatedness of two sequences. For protein sequence alignment, the 20 × 20 matrices are used to evaluate all possible combinations of amino acid pairs to calculate an overall score for the alignment of two sequences. Several types of scoring matrices have been proposed for protein sequences. Series of PAM (point accepted mutation) (Dayhoff et al., 1978) and BLOSUM (block substitution matrix) (Henikoff & Henikoff, 1992) matrices are the most popular and initial matrices for scoring all possible amino acid substitutions in pair-wise alignment during evolution.

Performance of sequence alignment tools is heavily upon the procedure scoring matrix is constructed. This procedure has been updated over time depending on the number of sequenced proteins and completed proteoms available in the literature. There is, however, an overall process of data gathering and calculations that need to be followed in any update of scoring matrices such as PAM. The overall procedure has not yet been fully assembled in a single document and different references have cited and explained partials of the procedure as required.

In this article, the basic concepts of local alignment algorithm and modified Smith & Waterman algorithm, which is applied for searching similarity in databases, are described and step by step calculation

[†] Corresponding author. Tel.: + 985118805008.
E-mail address: mahdavi@ferdowsi.um.ac.ir

operations are exemplified. Details of searching local similarities are explained and the procedure for statistical evaluation of a successful homology is outlined.

2. Methods

In local alignment, scoring each cell like i, j or H (i,j) is created by selecting maximum value from four possible values. For two sequences, $a=a_1\ a_2\ \dots\ a_n$ and $b=b_1\ b_2\ \dots\ b_m$, where $H\ (i,j)= H\ (a_1\ a_2\ \dots\ a_i,\ b_1\ b_2\ \dots b_j)$ then:

$$H_{i,j} = max\{H_{\mathrm{i-1,\ j-1}} + s(a,b), max(H_{i-x,j} - w_x), \max\ (H_{i,j-y} - w_y), 0\} \tag{1}$$

$$H_{x0} = H_{0y} = 0\ for\ 0 \le x\ \le n\ \ and\ \ 0 \le y \le m$$

$$1 \le i\ \le n\ \ and\ \ 1 \le\ j \le m$$

$$\mathrm{w_x} = 1 + \frac{1}{3\times\mathrm{x}}\quad ,\quad \mathrm{w_y} = 1 + \frac{1}{3\times\mathrm{y}} \tag{2}$$

In equation (1), $H_{i,j}$ defines the score at position" i" in sequence "a" and position "j" in sequence "b". S(a,b) shows the alignment score at position i,j which is 1 or -0.3. In equation (2), w_x calculates the penalty for inserted gap of length x in the sequence and w_y shows the penalty for a gap of length y in sequence b. This algorithm finds the optimal alignment between different possible alignments. For scoring the alignment different scoring matrices like BLOSUM62 or different PAM matrices can be used (Smith & Waterman, 1981).

To describe in details the performance of each three phases in modified local alignment, we give an example. For a query sequence as follows:

C I N C I N N A T I

The length of the sequence is n=10 and the threshold score that the default scoring matrix (BLOSUM62) considers is T=11. Since the considered sequence is a protein sequence, algorithm divides the sequence into 3-letter words. The number of words with w=3 are calculated as follow:

N= n-w+1

Thus, in the first stage, N for the given query sequence is calculated (here N=8) and three-letter words of the sequence are:

CIN(1) INC(2) NCI(3) CIN(4) INN(5) NNA(6) NAT(7) ATI(8)

Then, pairwise alignment of each query word with 8000 key words which are available in a hash table is performed. For example, when CIN which is located at position 1 and 4 is aligned with the 8000 words at the key table, 54 words with equal or greater score than 11 are obtained. Similarly, in pairwise aligning of ATI at position 8 with key words, only three pairs obtained score 11 and higher. Finally, preprocessing stage detects 204 entries of 8000 key words which obtained the threshold score in pairwise alignment with each query word. The next phase is scanning the target string or reference sequence successively to find exact matches to one of the words in query index. Given the following sequence as a target:

P R E C I N C T S (N=7)

The three-letter words are including:

PRE(1) REC(2) ECI (3) CIN(4) INC(5) NCT(6) CTS(7)

The word NCT in position 6 of the reference string generates two hits with two words in position 3 and 7 of query sequence. These two hits are defined as (3, 6) and (7, 6). The third phase begins when the locations of exact matches are found. Each hit, in turn, is extended to the right and left until the alignment score increases. Here, the alignment which is developed for the found hits at query position 3 and target position 6 is:

```
- - - c i N C I N n a t i
P r e c i N C T S - - - -
```

And the final alignment for hit (3, 6) will be:

```
C I N C I N
C I N C T S
```

The score of locally obtained alignment is calculated by BLOSUM62 scoring matrix as follows:

$$S_{CC}+ S_{II} + S_{NN}+ S_{CC}+ S_{IT}+ S_{NS}= 9+ 4+ 6+ 9+ (-1) + 1= 28$$

The alignment score of hit (7, 6) is no longer increased by extending to either left or right. Thus, the optimal alignment in this example will be the only obtained alignment. (Dwyer, 2003)

The most important measurement for assessing the statistical significance of the HSPs is E-value. This parameter is obtained as follows:

$$E= Kmne^{-\lambda s}$$

where, K is a constant value, m is the effective length of the query sequence, n is the effective length of the random sequence, λ is the scaling factor, and s reflects the similarity score of the pairwise alignment. Karlin and Altschul (1990) calculated the K and λ parameters by aligning 1000 random amino acid sequences whose lengths were variable. They aligned random sequences using Smith-Waterman algorithm and a combination of the scoring matrix and some suitable set of gap penalties for the matrix. For database similarity searching the cutoff number of E-value must be defined. Based on the defined threshold, if the E-values of the obtained alignments are lower than threshold they are considered as "significant" and the sequences with significant similarity are called "hits". Consequently, the database is categorized into two subsets, "hits" and "non-hits". These two groups of subsets are compared to the categories of true and false positives and negatives (which are created by experimental determination of protein structure and function) to recognize true positive subsets from true negative (non-hit with no biological background to the query sequence) and false negative (non-hit with a biological relationship in reality) subsets from false positives. There are two criteria for distinguishing hits from non-hits called sensitivity and specificity. Sensitivity (S_n) is the proportion of the real biological relationship in the database that is detected as hits and is shown as follow:

$$S_n = \frac{n_{tp}}{(n_{tp}+n_{fn})} \tag{3}$$

where n_{tp} is the number of true positives and n_{fn} is the number of false negatives. The specificity (S_P) is the proportion of the hits corresponding to the real biological relationships and is obtained as follows:

$$S_p = \frac{n_{tp}}{(n_{tp}+n_{fp})} \tag{4}$$

In order to obtain more accurate response in database searching, both S_n and S_p must be as close as to 1(Pearson, 1998).

3. Results and discussion

For measuring the similarity between sequences two approaches generally are used; global alignment and local alignment. In global similarity algorithms, the entire lengths of sequences are subject for representing the best alignment. This method will produce an alignment with large stretches of low similarity. Local similarity algorithms search and align the subsequences with the highest degree of identity or similarity. This method is a suitable way for aligning sequences which are similar along some regions of their lengths but are different in other regions. Furthermore, local similarity algorithms are useful for aligning sequences with different lengths, and sequences that share conserved regions. For database searches, the local similarity algorithms are most popular, because biological sequences often are not similar over entire lengths, but are similar just in some particular regions (Pearson & Miller, 1992).

Local alignment or Smith-Waterman algorithm seeks the local matches between two or more than two DNA or protein sequences with the highest score. This algorithm like global alignment algorithm constructs a matrix, but unlike global matrix, it has an extra row along the top and an extra column on the left side which are filled in with zeros. Therefore, for two sequences of length "m" and "n", the matrix dimension will be "m+1" and "n+1". In the first stage the match and mismatch letters are scored. Match letters are scored (+1) while mismatch letters are scored (-0.3). Then the score of each cell is given as the maximum of four possible values. In this algorithm if a negative value is created in each cell, it is changed to zero. After creating scoring matrix, the local maximum alignment can be detected. It begins and ends everywhere in the scoring matrix until the linear order of the two considered sequences are not violated. The trace-back procedure begins the alignment from the highest number in the matrix and proceeds diagonally up to the left until it reaches to the cell with a value of zero which is defined the start of the alignment.

The modified local alignment algorithm is an algorithm which is developed for database similarity searching. This algorithm involves three steps. These steps are including, compiling a list of high scoring words in a table called the "query index", scanning the database for finding hits, and extending the hits. In the first stage, the query string is divided into words of length w=3 for protein sequences and w=12 for DNA sequences (Altschul et al., 1990). The created words are scored in a pairwise alignment with 20^3 possible three-letter words and the query words which gain a score of at least T (threshold score) are selected for the list. For amino acid sequence comparison generally substitution matrices which are used for scoring word pairs are BLOSUM62 and different PAM matrices. Threshold score is selected for reducing the number of possible matches. For example, the default T is usually defined 11 by BLOSUM62 scoring matrix. For DNA sequence comparisons usually matched DNA word pairs are scored +5 and mismatch DNA word pairs are scored -4. In the second phase, database sequences are also divided into w-mers, and scanned in order to detecting an exact match to one of the words in the query index. In the third step, the found hits are extended, in turn, from right and left direction to find segment pairs whose score is better than the cutoff score. The process of extending is terminated in each direction once segment pair score decrease from the cutoff score. In the last phase, High scoring Segment Pairs (HSP which has a greater score than the original word) are found (Dawid W.M., 2001; Pevsner, 2003).

The modified Smith-Waterman algorithms such as FASTA and BLAST are heuristic algorithms which require less time for performing an alignment compare to dynamic programming algorithms, because these heuristic algorithms restrict the search by scanning the database for finding likely matches before performing the actual alignment. Unlike dynamic programming algorithms, these algorithms do not guarantee to find optimal alignments.

4. Conclusion

Different steps of local alignment algorithm and modified local alignment algorithms are described in detail. It has assigned that the two types of dynamic programming algorithms (local and global) guaranteed to find the optimal alignment, but modified version of Smith-Waterman algorithms, like BLAST, do not guarantee to find optimal alignment. These modified algorithms are suitable for aligning a query sequence against a database and recognizing conserved regions. These algorithms require less time to perform an alignment. After developing the alignments in each kind of algorithms, the statistical significance of the resulted alignments must be evaluated. Statistical tests help the biologists to detect real relationships from random.

5. References

[1] Altscul, S.F.; Gish, W.; Miller, W.; Myers, E.W. & Lipman D.J. (1990). Basic Local Alignment Search Tool. *J. Mol. Bio*l, Vol. 215, (15 May 1990), pp. (403-410).

[2] Dawid, W.M. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, USA.

[3] Dayhoff, M.O. (1978). *Atlas of Protein Sequence and Structure*. Vol. 5, suppl. 3, Nat. Biomed. Res. Found., Washington, DC.

[4] Dwyer, R.A. (2003). *Local Alignment and the BLAST Heuristic, In: Genomic Perl from Bioinformatics Basics to Working Code*, pp. (93-108), Cambridge University Press, 521-80177, UK.

[5] Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, Vol. 89, (November 1992), pp. (10915-10919).

[6] Karlin, S. & Altschul S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.*, Vol. 87, No. 6, (March 1990), pp. (2264-2268).

[7] 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*., Vol. 48, pp. (443-453).

[8] Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci., USA*, Vol. 85, No. 8, (April 1988), pp. (2444-2448).

[9] Pesrson, W.R. & Miller, W. (1992). Dynamic programming algorithm for biological sequence comparison. *Method Enzymol*., Vol. 210, PP. (575-601).

[10] Pearson, W.R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol*., Vol. 276, No. 1, (February 1998), pp. (71-84).

[11] Pevsner, J. (2003). *Pairwise Sequence Alignment, In: Bioinformatics and Functional Genomics*, pp. (41-84), John Wiley & Sons, ISBN: 0-471-21004-8, USA.

[12] Smith, T.F. & Waterman, M.S. (1981). Identification of Common Molecular Subsequences. *J. Mol. Biol*., Vol. 147, pp. (195-197).