ejpmr, 2019,6(2), 270-272

SJIF Impact Factor 4.897

EUROPEAN JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH

www.ejpmr.com

Research Article

ISSN 2394-3211

EJPMR

PAIRWISE SEQUENCE ALIGNMENT – *A GLOBAL ALIGNMENT APPROACH*

Roma Chandra*[12] and Sumit Kumar[1]

[1]Shri Venkateshwara University, Gajraula, Amroha, Uttar Pradesh, India.
[2]Department of Biotechnology, IILM College of Engineering & Technology, Greater Noida, Uttar Pradesh, India.

***Corresponding Author: Dr. Roma Chandra**

Shri Venkateshwara University, Gajraula, Amroha, Uttar Pradesh, India.

Article Received on 12/12/2018 **Article Revised on 01/01/2019** **Article Accepted on 22/01/2019**

ABSTRACT

Sequence alignment is a process to find similar regions when two or more than two sequences are aligned together. Motif or similar regions are supposed to be the consequence of structural, functional or evolutionary relationships between sequences. Sequences can be either nucleotide or protein sequences usually in fasta format. Alignment includes role of specific biological databases both for nucleotide and protein sequences like SWISSPROT, PDB, etc. for protein sequences and GenBank, EMBL,DDBJ ,etc. for nucleotide sequences. The two common types of alignment methods include pairwise and multiple sequence alignment which tends to find similar regions either of the two types - local or global type. Information obtained from sequence alignment is helpful and show applications including motif prediction, homology modeling, and phylogenetic tree construction. Dynamic programming method under global alignment approach produces optimal alignment results. The article includes study of global alignment process.

KEYWORDS: GenBank, EMBL,DDBJ, etc.

1. INTRODUCTION

Sequence alignment is a method of finding similar regions, also termed as motif regions when two or more sequences are aligned together. It is done basically to find similar region which are supposed to carry structural, functional and evolutionary information. To study evolutionary relation among organisms their nucleotide sequences are aligned together under multiple alignment method with phylogenetic tree construction. On the other hand protein sequences align giving information about the motif regions responsible to have characteristic information for a particular protein family, developing protein structures and drugs. Sequence alignment is of two types where under local type only motif regions are discovered whereas under global type both motif and gaps are discovered. There are basically two methods of sequence alignment – pairwise and multiple. Pairwise methods include alignment of two sequences while multiple methods include alignment of multiple sequences. Sequence alignment is an easiest method to compare two or more than two protein or nucleotide sequences finding out similarity among them. When aligned, the most similar regions are searched between the sequences. A hypothesis is made regarding the sequences to be aligned that at one generation they were exactly similar or identical and generation after generation mutational changes caused variation leading to changes both in structure and function. The matches represents the similar regions where as mismatches represents variation caused because of substitution of one amino acid by another or substitution of one base by another. On the other hand gaps are produced as a result of insertion or deletion. An alignment score is calculated for the aligned sequences based on the amount of similarity which is measured as the sum of matches, mismatches and gaps.

2. MATERIALS AND METHODS

Alignment is done finding either local regions or global regions and thus alignment is said to be of local or global types. Local alignment finds only the most common regions or the motif regions whereas global alignment tends to find maximal possible similar regions thus including mismatches and gaps too. Methods of alignment are based on aligning two sequences or more than two sequences i.e. pairwise or multiple types of alignments. Pairwise Sequence alignment includes the following methods:

- Dot Plot or Dot Matrix Method
- Dynamic Programming Method
- Similarity Search or K tuple Method.

Out of above three methods, dynamic programming method produces optimal alignments. The concept behind dynamic programming method is to divide a problem into sub problems and then finding solutions to those sub problems. The best solution is supposed to be the answer to the problem. Based on this concept the sequences are aligned finding out the maximum possible similar regions including indels & substitutions. The

dynamic programming method is based on finding either local regions or global regions of alignment. Smith Waterman algorithm is used to align sequences locally while Needleman Wunsch algorithm is used to align sequences globally. A two dimensional matrix is constructed based on this concept. The technique of dynamic programming can be applied to produce global alignments via the Needleman Wunsch algorithm, and local alignments via the Smith-Waterman algorithm. A two dimensional matrix is created for both local & global alignment trying to search the best possible pathway which is supposed to be the best solution of alignment for the given sequences. A two dimensional matrix is constructed for both Needleman Wunsch and Smith Waterman algorithms. The matrix is constructed following the three steps:

- Initialization
- Matrix filling
- Trace back

A scoring matrix is constructed for protein and nucleotide sequence alignment. In case of protein sequence alignments, 20 × 20 matrix is created for all the possible 20 amino acids. PAM & BLOSUM are such scoring matrices. The scoring values for matches & mismatches are taken from this scoring matrix. On the other hand the nucleotide sequence alignment i.e. for DNA and RNA, specific scoring values are considered for matches, mismatches & gaps. According to the algorithm, the first step initialization starts with the filling of first row and first column with the gap values in case of Needleman Wunsch algorithm and zero value with Smith Waterman algorithm. With the matrix filling step, we consider every block of the matrix as (i^{th}, j^{th}) block i.e. it belongs to i^{th} row and j^{th} column.

(i^{th} -1, j^{th} -1)	(i^{th} -1, j^{th})
(i^{th}, j^{th} -1)	(i^{th}, j^{th})

In case of Needleman Wunsch algorithm, the maximum of the following three values is taken in this block:

- Value of (i^{th} -1, j^{th} -1) block + Match /Mismatch Value
- Value of (i^{th}, j^{th} -1) block + Gap Penalty Value
- Value of (i^{th} -1, j^{th}) block + Gap Penalty Value

On the other hand, in case of Smith Waterman algorithm, the maximum of the four values is considered. The three values are obtained by following above three conditions while the fourth value is zero. During the matrix filling step simultaneously arrows are plotted in the blocks pointing towards the block from which the information had been taken to fill each block. Under the trace back step these arrows are followed from the right bottom corner towards the left top corner, following all possible pathways finding out the maximal similar region of alignment.

3. RESULTS AND DISCUSSION

For example: For the following sequences S1: ATGC and S2: ATCC the Global alignment is done as presented in the matrix. Considering match = +1, mismatch = -1, gap penalty = -2.

	-	A	T	G	C
-	0	-2 ←	-4 ←	-6 ←	-8 ←
A	-2 ↑	+1 ↖	-1 ←	-3 ←	-4 ←
T	-4 ↑	-1 ↑	+2 ↖	0 ←	-2 ←
C	-6 ↑	-3 ↑	0 ↑	+1 ↖	+1 ↖
C	-8 ↑	-5 ↑	-2 ↖	-1 ↖	+2 ↖

On back tracing starting with the last block in the above matrix the following path is obtained that represents the best possible /optimal alignment with alignment score equal to +2.

A T G C
A T C C
+1 +1 -1 +1 = +2

Tools available for pairwise alignment

- dotmatcher
- dotpath
- dottup
- est2genome
- matcher
- needle
- needleall
- polydot
- stretcher
- seqmatchall
- supermatcher
- water
- wordfinder
- wordmatch

REFERENCES

1. Azzedine Boukerche, Jan M. Correa, Alba Cristina M.A de Melo, Ricardo P. Jacobi, "A Hardware Accelerator for the Fast Retrieval of DIALIGN Biological Sequence Alignments in Linear Space", IEEE Transactions on Computers, 2010; 59(6): 808-821.
2. Altschul S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. "Basic local alignment search tool". J. Mol. Biol., 215: 403-410.
3. Changjin Hong, Ahmed H. Tewfic, "Heuristic Reusable Dynamic Programming: Efficient Updates of Local Sequence Alignment", IEEE/ACM

Transactions on Computational Biology and Bioinformatics, 2009; 6(4): 570-562.
4. Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., Dubchak, I. 2003. "Strategies and tools for whole-genome alignments". Genome Res. (this issue).
5. Hassan Mathkour, Muneer Ahmad, "A Comprehensive Survey on Genome Sequence Analysis", IEEE International Conference on Bioinformatics and Biomedical Technology, 2010; 14-18.
6. Khaled Benkrid, Ying Liu, AbdSamad Benkrid, "A Highly Parameterized and Efficient FPGA-Based Skeleton for Pairwise Biological Sequence Alignment", IEEE Transactions on Very Large Scale Integration Systems, 2009; 17(4): 561-570.
7. Mayor C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. "VISTA: Visualizing global DNA sequence alignments of arbitrary length". Bioinformatics, 16: 1046-1047.
8. Miller W. 2001. "Comparison of genomic DNA sequences: Solved and unsolved problems". Bioinformatics, 17: 391-397.
9. Morgenstern B., Frech, K., Dress, A., and Werner, T. 1998. "DIALIGN: Finding local similarities by multiple sequence alignment." Bioinformatics, 14: 290-294.
10. Morgenstern B., Rinner, O., Abdeddaïm, S., Haase, D., Mayer, K., Dress, A., and Mewes, H-W. 2002. "Exon discovery by genomic sequence alignment." Bioinformatics, 18: 777-787.
11. 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.
12. Sanghamitra Bandyopadhayay, Ramakrishna Mitra, "A Parallel Pairwise Local Sequence Alignment Algorithm", IEEE Transactions on Nano Bioscience, 2009; 8(2): 139-146.
13. Schwartz S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. "PipMaker—A web server for aligning two genomic DNA sequences". Genome Res., 10: 577-586.
14. Zhang, Z., Schaffer, A., Miller, W., et al. 1998. "Protein sequence similarity searches using patterns as seeds." Nucleic Acids Res., 26: 3986–3990.