Agri Articles
(e-Magazine for Agricultural Articles)
Volume: 03, Issue: 01 (JAN-FEB, 2023)
Available online at http://www.agriarticles.com
©Agri Articles, ISSN: 2582-9882

Nucleotide and Protein Sequence Analysis

(*Farheen Ansari, Asit Jain, Mohan Singh Thakur, Vaishali Khare, Akansha Singh, S. K. Joshi and S. S. Tomar)
Department of Animal Genetics and Breeding, College of Veterinary Science & Animal Husbandry, Nanaji Deshmukh Veterinary Science University, Jabalpur (MP)
*Corresponding Author's email: farheen218@gmail.com

Abstract

The goal of sequence analysis is to investigate the patterns in DNA, RNA, and protein sequences. With the large variety of tools and methodologies available, analysis of the nucleotide and protein sequences is possible. Evolutionary analysis, gene/protein structure prediction, and reconstruction of DNA sequences can be carried out by sequence analysis. Similarities among multiple sequences is found using sequence alignment. Finding structurally or evolutionarily related positions in a group of amino acid sequences is the goal of protein sequence alignment. One of the most popular sequence analysis tools that is freely accessible is BLAST. Multiple sequence alignment is used to identify conserved sequence patterns or motifs throughout the whole sequence family.

Keywords: DNA or protein sequence, Local alignment, Global alignment, Pairwise alignment, Multiple sequence alignment

Introduction

The possibilities for study on proteins and their related domains have expanded thanks to the enormous library of protein sequences and the computational tools needed for protein sequence analysis. In order to measure the local and global sequences, similarity is detected between the query and the given sequences. Analyzed results of protein sequence have applications in classification of protein sequences, prediction and evolution of protein, analysis of expression, bioinformatics and genetic engineering. Analysis of the nucleotides and protein sequences can be done with the availability of wide range of tools and techniques.

Goals of Sequence Analysis

Once a genome is completely sequenced, the question arises that what sorts of analyses can be performed on it? Some of the goals of sequence analysis are mentioned below:

1. Identification of the genes.
2. Determination of the function of each gene.
3. Identification of the proteins involved in the regulation of gene expression.
4. Identification of sequence repeats.
5. Identification of other functional regions, for example origins of replication, pseudogenes, sequences responsible for the compact folding of DNA, and sequences responsible for nuclear anchoring of the DNA.

Sequence Alignment

The process of comparing and detecting similarities between biological sequences is called sequence alignment. Finding similarities in the sequences of proteins, DNA, and RNA via sequence alignment is crucial for analysing the evolutionary relationships between them. Annotation and assembly of the sequences, prediction of structure and function of genes/proteins, analysis of phylogeny and evolutionary relationship are some of the applications of alignments (Batzoglou, 2005). Alignment of DNA and protein sequences is the fundamental task of bioinformatics. Further whether the protein sequences are inherited from common ancestor can be known by analysis of evolutionary relationship. Alignments can be divided into two types: Pairwise Alignment and Multiple Sequence Alignment. Only two sequences are considered in the pairwise sequence alignment, whereas similarities and conserved sequences in more than two sequences are found using multiple sequence alignment.

Basis of Sequence Alignment

1. To find the relatedness of the proteins or gene, if they have a common ancestor or not.
2. To find mutation in the sequences, brings the changes or divergence in the sequences.
3. It can also reveal the part of the sequence which is crucial for the functioning of gene or protein.

Scoring model/matrix

It includes the mutation scores for DNA or protein and the gap penalty scores for insertions or deletions (indels or gaps). Algorithms which compare protein or DNA sequences rely on some mutation scoring. DNA scoring matrices are rather simple, usually counting a match as one and a mismatch as zero. Scoring for amino acids is complex as it has to reflect both the properties of amino acid residues, as well as the likelihood of certain residues being substituted among the homologous sequences. The empirical amino acid scoring matrices which include PAM (Dayhoff *et al.*, 1978) and BLOSUM (Henikoff and Henikoff, 1992) matrices, are derived from actual alignments of highly similar sequences. Frequencies of amino acid substitutions in these alignments can be analyzed and a scoring system can be developed by giving a higher score to more likely substitutions and correspondingly lower scores for rare ones.

Gap penalties

Alignment between sequences often involves adding gaps that represent insertions and deletions. Gap penalty values being set too low would result in numerous gaps or matching of non-related sequences with high similarity scores. Similarly, too high values of gap penalty make appearance of gaps difficult which would not allow creation of a reasonable alignment.

Pairwise Alignment

This method compares two biological sequences of either protein, DNA or RNA. The exact matches of similar sequences are revealed in this method.
It can be categorized into local or global alignment.

Pairwise alignment methods:

1. Dot matrix method: Listing of one sequence (A) is done at the top of the matrix and towards downwards of the matrix at the left side other sequence (B) is placed. From the first character in B the process of comparison starts moving across in the first row and a dot is placed where the character in A is similar. When the sequences are closely related, diagonal row of dots appear. Limitations include empty space or noise taking up most of the actual area of the plot.
2. Dynamic programming: Best or optimal alignment is generated by comparing every pair of characters in the two sequences. Both global and local alignments can be produced, the former is done by the means of Needleman-Wunsch algorithm, and the latter by Smith-Waterman algorithm. In typical usage, protein alignments use a substitution matrix to assign

scores to amino-acid matches or mismatches, and a gap penalty for matching an amino acid in one sequence to a gap in the other. DNA and RNA alignments may use a scoring matrix, but in practice often simply assign a positive match score, a negative mismatch score, and a negative gap penalty.

3. Word methods: They are also known as k-tuple methods. These are heuristic methods that are not guaranteed to find an optimal alignment solution, but are significantly more efficient than dynamic programming. Database search tools such as FASTA and the BLAST employ word methods.

Local Alignment

In local alignment, the highest priority is given to the stretches of sequence with the highest density of matches (which are usually the most biologically significant), which leads to one or more islands of matches found in the aligned sequences. It involves only parts of the sequences. This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences. The optimal local alignments of the sequences are found using Smith-Waterman algorithm (Mott, 2005).

FASTA: FAST-All which means fast comparison of proteins or nucleotides represents FASTA (Donkor *et al.,* 2014). It was developed in 1995 as an improvement over FASTP (Fast Protein). DNA searches and evaluation of statistical significance is performed by FASTA. FASTA programmes are available in different types. For DNA libraries TFASTA and TFASTAY programmes are used, while for proteins FASTAX and FASTAY are used. k-tuple sub-words are searched using the heuristic algorithm. In the first step FASTA identifies the regions which are identical and in the next step, rescoring of the best regions of the previous step is done by PAM-250 matrix. Diagonals having high scores are joined together with gaps. Smith-Waterman algorithm is used to generate a score of optimal alignment.

BLAST: BLAST stands for Basic Local Alignment Search Tool. Comparison of two different proteins or nucleotides is done. Different variations of BLAST exist. megaBLAST performs searches for nucleotide-nucleotide similarity sequences. Nucleotide-nucleotide distant sequences are found using BLASTN. Comparison between protein-protein sequences is done by BLASTP based on searches by BLASTX and TBLASTN. For a translated nucleotide query, searches on protein database are performed by BLASTX and vice versa by TBLASTX. Position-Specific-Scoring Matrix (PSSM) is determined using BLASTP which is then used to search protein sequences in a database in PSI-BLAST. PSSM is also used by RPSBLAST to search a protein query quickly across a database. Working of DELTA-BLAST is similar to RPSBLAST but it is faster. Setup, Preliminary Search and Traceback are the three phases through which local alignment by BLAST is done (Madden, 2013). A fixed length for a set of words as per the query given to search is produced in the setup phase. In the preliminary phase, matches which corresponds to the words are detected and score is generated. Computation of insertions and deletions is done by gapped extensions in traceback. As compared with the dynamic programming methods such as Waterman, Wunsch, BLAST works faster (Tatusova and Madden, 2006).

Global Alignment

Alignment of entire sequences, as many characters as possible is done here. This is done for the entire or whole sequence of a given DNA or protein sequence. Most of the bioinformatics studies use Needleman-Wunsch algorithm for sequence alignment (Harris, 2014). The optimum matches between any two similar sequences are determined using global alignment.

FOGSAA: FOGSAA (Fast Optimal Global Sequence Alignment Algorithm) results are similar to Needleman-Wunsch method but it takes less amount of time (Chakraborty and Bandyopadhyay, 2013). For highly similar nucleotide sequences the time gained is 70-90%

and for sequences of 30-80% similarity, it gains time of 54-70%. Higher number of matches are found in FOGSAA while the number of mismatches and gaps is lower. Optimal alignment is found using a tree. Given sequences are used to build a tree. If some other branch is found to be better than the current one at an intermediate point, it is further expanded. This procedure gets repeated until an optimal alignment is obtained.

Multiple Sequence Alignment (MSA)

More than two sequences are aligned using multiple sequence alignment. Conserved sequence regions among several sequences can be known through this alignment. Alignment of more than two sequences is obtained by inserting gaps ("-") into sequences such that the resulting sequences have all length L and can be arranged in a matrix of N rows and L columns where each column represents a homologous position. The principle is that multiple alignments are achieved by successive application of pairwise methods.

Multiple Sequence Alignment methods:

1. Progressive: Closely related sequences are aligned and then addition of remaining sequences which are less related is done. Multiple sequence alignment and construction of evolutionary tree is done by the iterative use of Needleman-Wunsch pairwise alignment.

CLUSTAL series: CLUSTAL family constitute very popular progressive alignment methods such as CLUSTAL-OMEGA (Sievers *et al.*,2011). CLUSTAL program consist of a combination of progressive alignment technique with memory-efficient dynamic programming (Chenna *et al.*, 2003). Using guide tree as a reference, a series of pair-wise alignment are used to construct progressive multiple sequence alignments. For the construction of a guide tree Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method is used. Choice of weight given to the matrix, position specific gap penalties and weighting of the sequence comprises the improved algorithm of CLUSTALW. In place of UPGMA method, neighbour joining method is used to construct dendogram. CLUSTALX provides a new window-based user interface to the CLUSTALW program. Vibrant multi-platform user interface development library is used by CLUSTALX which is developed by the National Center for Biotechnology Information (NCBI) as part of their NCBI software development toolkit.

MUSCLE: Multiple protein sequences are aligned using MUltiple Sequence Comparison by Log Expectation (MUSCLE) (Edgar, 2004). Initial alignment is built based on the similarities of paired alignments and then distance matrix is calculated and the rooted tree is generated. For aligned pair Kimura distance is used while for unaligned k-mer distance. Distance matrices are clustered using UPGMA that improve tree by recalculating similarities.

MAFFT: Quick identification of some of the more obvious regions of homology can be done by Multiple Alignment using Fast Fourier Transform (MAFFT) (Standley, 2013). Dynamic programming approaches are used to join these identified portions into a full arrangement. The initial version of MAFFT had the advantage of speed and was also one of the more accurate programs. It is available as a standalone or web interface. It returns many output formats, including interactive phylogenetic trees.

2. Iterative: This method is used to overcome the error which were obtained in progressive alignment. High quality alignment score is determined by realigning the previously aligned sequences. On convergence of score values the iteration stops.

T-Coffee: Computation of multiple alignments is done using a mixture of global and local pairwise alignments (Notredame *et al.*, 2000). Selection of pairwise alignments from the input library which fits best is done to construct multiple alignments. In this way the work is fast and errors are minimum.

DIALIGN: For those parts of sequences that are not related with each other local alignment is used. If the sequences are related over their entire length, then global alignment is done. Region of similarities is aligned in this method (Morgentern, 2004).

Conclusion

The study of the relationship between structure and function of proteins has made sequence alignment a crucial tool. Due to the rapid growth in the number of available sequences, alignments now offer a significant amount of information using a variety of computational techniques. Numerous recent research has shown significant advancements in the accuracy and scalability of alignment techniques.

References

1. Batzoglou, S. (2005). The Many Faces of Sequence Alignment. *Briefings in Bioinformatics*, **6**: 6-22.
2. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). A model for evolutionary change in proteins. In: Dayhoff, M.O. (eds.), Atlas of Protein Sequence and Structure, Washington DC: National Biochemical Research Foundation,5:345-352.
3. Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**: 10915-10919.
4. Mott, R. (2005). Smith-Waterman Algorithm, University of Oxford. DOI: 10.1038/npg.els.0005263, pp.1-5.
5. Donkor, E.S., Dayie, N.T.K.D. and Adiku, T.K. (2014). Bioinformatics with Basic Local Alignment Search Tool (BLAST) and Fast Alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, **6**: 1-6.
6. Madden, T. (2013). The BLAST sequence analysis tool. *National Center for Biotechnology Information*, 1-10.
7. Tatusova, T.A. and Madden, T.L. (2006). BLAST2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, **174**: 1-6.
8. Harris, S.R. (2014). New approaches to prokaryotic systematics. In: Goodfellow, M., Sutcliffe, I. and Chun, J. (eds.), Methods in Microbiology, Academic Press, 41: 348.
9. Chakraborty, A. and Bandyopadhyay, S. (2013). FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. *Scientific Reports*, **3**: 1746.
10. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D. and Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using CLUSTAL-OMEGA. *Molecular Systems Biology*, **7**: 539.
11. Chenna, R., Sugawara, H. and Koike, T. (2003). Multiple sequence alignment with CLUSTAL series of programs. *Nucleic Acid Research*, **31**: 3497-3500.
12. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**: 113.
13. Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4): 772-780.
14. Notredame, C., Higgins, D.G. and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**: 205-21.
15. Morgentern, B. (2004). DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acid Research*, **32**: 33-36.