

Genome Sequence Analysis to Evaluate the Performance of Pair wise Statistical Significance of Solanum Lycopersicum

Uma kumari¹, Ashok Kumar Choudhary²

Department of Biotechnology, Jharkhand Rai University, Ranchi-835222, Jharkhand, India.

Department of Botany, Ranchi University, Ranchi-834008, Jharkhand, India.

Abstract: *Solanum lycopersicum* is an excellent plant for this approach considering the high quality and comparative analysis of genome sequence of tomato. Sequencing the genome of the crop *Solanum lycopersicum* will also help to identify beneficial genes in other plant relative of the tomato such as potato, pepper. All of these crops are members of the Solanaceae or nightshade family, one of the world's most important vegetable plant families in terms of both economic value and production volume. Developing better tomatoes will also contribute to the quest for global food security. As well as using this new genome information to develop a wide variety of beneficial traits, the (TGRD) the tomato genomic resources database is an online and interactive relational database developed using open source software. In sequence alignment is a way of arranging the sequence of DNA, RNA, or protein to identify the functional structural or evolutionary relationship between the sequence. If two sequences share a common ancestor, mismatches can be interpreted as a point mutation. FASTA format is a text-based format for representing either nucleotide sequence or peptide sequence, in which nucleotide or amino acid are represented using single letter code. Sequence homology is a general term that indicates evolutionary relatedness among sequences. NCBI provides a common data extraction platform for sequence analysis. Sequence similarity is a substitution with similar chemical properties. The ClustalW colored alignment also has the colour option in the output results. The colouring residue takes place according to the following physicochemical criteria (Red, blue, green, magenta, and grey colours). In addition to maintaining the gene bank nucleic acid, sequence database, NCBI provides a data retrieval system and computational resources for the analysis of gene bank data and variety of other biological data made available through NCBI.

Keywords: Bioinformatics, Genome database, TGRD (Tomato genomic resources database), Sequence analysis, Data compiled, Sequence alignment, *Solanum lycopersicum*.

I. Introduction

The aim of tomato genome sequencing is to reveal and explore the genetic variation availability in tomato. Tomato has been selected as a target crop because it is economically one of the most important crop species. The programme can run online from the EBI web server. The source code executables for Windows, Linux are available from EBI. The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequence and for preparing the phylogenetic trees. Taylor, Willie, Higgins, Des 2000, Bioinformatics. New features include NEXUS and FASTA format output, printing range numbers and faster tree calculations. ClustalW originally developed to run on local computers; numerous web servers have been set up, notably at the EBI (European Bioinformatics Institute). Tomato has been used extensively for genetic studies because of several reasons such as its diploid genome, short generation time, efficient transformation technology. The data can be submitted and accessed via the world wide web (Mount, David 2004). The tomato genome resources database is an interactive relational database developed using open source bioinformatics software. Sequence analysis created a huge impact in *Solanaceae* research. Using pairwise alignment to find the best matching in query sequences. FASTA format is a text-based format for representing either nucleotide sequence or protein sequence (Higgins, D. G.; Sharp, P. M. (1989)). The format originates from the FASTA software package. For DNA and Protein it is represented in one letter IUPAC nucleotide code and amino acid code. It finds the local similarity between the sequence and calculates the statistical significance of matches. Mismatches would be connected with a space. Using bioinformatics tools ClustalW is a widely used multiple sequence alignment in computer programs (Higgins, D. G.; Bleasby, A. J.; Fuchs, R. (1992)). An alignment will display by default the following symbols denoting the degree of conservation observed in each column. FASTA produces local alignment scores for the comparison of the query sequence to every sequence in the database. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. (1997). Sequence alignment or sequence comparisons lies at the heart of bioinformatics, which describe the way of arrangement DNA and RNA to identify the regions of similarity among them.

II. Materials And Methods

The national center of biotechnology information (NCBI) is a multidisciplinary research group that serves as a resource for molecular biology information, developing new methods to deal with the volume and complexity of data searching and methods that can analyze the structure and function of macromolecules, creating computerized systems for storing and analyzing data. The primary database retrieval system at NCBI, which links together several databases including GeneBank. FASTA is available as a part of a package of programs that construct local and global sequence alignment. For a more complete description of FASTA and related programs for identifying related DNA/RNA sequences, for evaluating the statistical significance of sequence similarities.

2.1 Database and Corresponding web services

Database name	Web services type: URL
NCBI	E—Utility web services (http://www.ncbi.nlm.nih.gov)
FASTA	www.ebi.ac.uk/tools
Clustal omega	http://www.ebi.ac.uk/Tools/msa/clustalw2/
EMBL/EBI	EMBL-EBI web services (http://www.ebi.ac.uk/tools/)
Uniprot KB	Programmatic access services (http://www.uniprot.org)
EBI/ftp site:	ftp://ftp.ebi.ac.uk/pub/software/clustalw2/

III. Results And Discussion

The FASTA file format, now largely used by other sequence database search tools, which takes input as nucleotide or protein sequence. The program (ClustalW) is a widely used multiple sequence alignment tool that manipulates existing alignments, profile analysis, and creates phylogenetic trees. Alignment can be done by two methods: slow/accurate, fast/appropriate. Clustal Omega is a new multiple sequence alignment program that uses a high-profile technique to generate alignments between two or more sequences. Local sequence alignment programs report alignment scores for the alignment constructed, and related (homologous) sequences will have higher alignment scores. The statistical significance of an alignment score is more widely accepted as a metric to comment on the relatedness of the two sequences being aligned. The ClustalW and ClustalX multiple sequence alignment programs have been completely rewritten in C++ (Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003)). This facilitates the further development of the alignment algorithms in the future and has proper portions of the program to the latest version of Linux/Windows operating systems. (Availability—the program can be run online from the EBI web server. <http://www.ebi.ac.uk/tools/clustalw2>). The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequences and preparing phylogenetic trees. Clustal was originally developed to run on a local computer, but numerous web servers have been set up, notably at the EBI (European Bioinformatics Institute). ClustalW improves the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and ClustalW as a data exploration tool rather than as a definitive analysis method.

3.1 Pairwise statistical significant estimation

Consider the pairwise statistical significance described in obtainable by the following function: where sequence 1 and sequence 2, and sc is the scoring scheme (substitution matrix, gap penalties), and N is the number of shuffles.

```
>gi|1050193310|ref|NM_001329952.1| Solanum lycopersicum plastid-specific 50S ribosomal protein 5-like (LOC101252938), mRNA
TCCAGAAAACCAAACCTCAAACCTGGAGAGATGGCTCTCCTTATCTTCACTGCAACAACCTCCCTCT
GTTC
TCCTCTCATCTCAATCTTCTCTACTTCACAAGCTTCTGCATTCCTGCTTCGTTATCCTCCAGGTTCTG
TG
CAACAATCACTTTACCCTGACACCTAAGTCTTATGCCAATGGTTATATTCAAGCACCTTTTATCTTCCAA
CAA
AGGAGAGGTGCATTGATTGCTACAGCGGCTGCAGACATTGATAGTGTCGGTTCAGATAATCCTGAGCCTT
GCCTT
CACCAGAAAAAAGGAGGAAAGTGTGCCTGTTGAGAATCTCCCTCTGGAGTCTAAGCTTCAAGAGAAAGCT
AAGCT
TGAACAGAAGATGAAGATGAAATTGGCAAAAAAGCTTAGACTACGGAGGAAGAGACTCGTTAGGAAGCGC
AAGCGC
CACCTAAGGAAGAAAGGACGATGGCCACCTTCAAAGATGAAGAAGAACAAGAATGTCTAACTTAACTTAA
AACCTG
```

AAATGCCTTGCAAGTGTCTCGTTTTTCTCGTAGTCTTTATAATATCGAAATACTGTAATCTCTGAG
ATC
ATTTTCTTCAACCTGTACCTGATACCTTATGAAATTGATTAGATTTTTTCCCGAAAAAAAAAAAA


IV. Conclusion

An application of pairwise statistical significance to empirically determine the effective gap opening penalties for protein local sequence alignment. Analysis of the sequence has been developed with the objective of providing a single platform for customizable data from the some of the major biological database .Larger list with more low scoring hits can be reported based on quality of alignment (the score) and size of the database by applying the sequence alignment method and bioinformatics tools.

Acknowledgement

We extended our sincere thanks to Dr.Savita sengar “Vice Chancellor “of Jharkhand Rai University, Ranchi, India for kindly providing me the platform to carry out the research.

References

- [1]. Altschuen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David (1990). "Basic local alignment search tool". *Jol, Steprnal of Molecular Biology* 215 (3): 403–410
- [2]. Andreas D.Baxevanis,B.F.Fracis Quellet,"A practical guide to the analysis of Gene and protein .3RD Edition october 2004.Published by Wiley,John and Sons.
- [3]. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: Update. *Nucleic Acids Research*, 2004, vol 32, Database Issue: D23-D26. 
- [4]. "ClustalW / ClustalX: Multiple Sequence Alignment". Retrieved 1 October 2013.
- [5]. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003). "Multiple sequence alignment with the Clustal series of programs". *Nucleic Acids Res* 31 (13): 3497–3500. doi:10.1093/nar/gkg500. PMC 168907. PMID 12824352.
- [6]. Giovannoni, J. 2001, Molecular biology of fruit maturation and ripening, *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, 52, 725-749. LINK
- [7]. Higgins, D. G.; Sharp, P. M. (1989). "Fast and sensitive multiple sequence alignments on a microcomputer". *Computer Applications in the Biosciences (CABIOS)* 5 (2): 151–153. doi:10.1093/bioinformatics/5.2.151. PMID 2720464.
- [8]. Higgins, D. G.; Bleasby, A. J.; Fuchs, R. (1992). "CLUSTAL V: Improved software for multiple sequence alignment". *Computer Applications in the Biosciences (CABIOS)* 8 (2): 189–191. doi:10.1093/bioinformatics/8.2.189. PMID 1591615.
- [9]. Higgins DG, Thompson JD, Gibson TJ. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, 266, 383-402.
- [10]. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem Sci.*, 23, 403-405.
- [11]. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.
- [12]. Madden T. (2002).The NCBI handbook, 2nd edition, Chapter 16, The BLAST Sequence Analysis Too.
- [13]. Mount .David 2004,Bioinformatics:-sequence \$ Genome Analysis", published by Cold spring Harbour laboratory press.
- [14]. NCBI Resource Coordinators (2012). "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Research* 41 (Database issue): D8–D20.
- [15]. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. (1994). "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* 22(22): 4673–4680. doi:10.1093/nar/22.22.4673. PMC 308517. PMID 7984417.
- [16]. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. (1997)."The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools". *Nucleic Acids Research* 25 (24): 4876–4882. doi:10.1093/nar/25.24.4876. PMC 147148. PMID 9396791.
- [17]. Taylor Willie, Higgins Des 2000,Bioinformatics :Sequence structure and database practical approach “,1st Edition October 2000 ,Published by Oxford university press.
- [18]. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.