Biological Forum – An International Journal **8**(2): 330-339(2016)

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Phylogenetic Tree Construction with Optimum Multiple Sequence Alignment

Pankaj Bhambri* and Om Parkash Gupta**
*Research Scholar, *Department of Computer Science and Engineering,*
I.K.G. Punjab Technical University, Kapurthala, (Punjab), INDIA
** *Department of Information Technology*
Punjab Agriculture University, Ludhiana, (Punjab), INDIA

(Corresponding author: Pankaj Bhambri, pkbhambri@gndec.ac.in)

(Received 25 August 2016, Accepted 13 October, 2016)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: **Bioinformatics concern is related to every living and non-living species on the planet. So, there are massive opportunities to work in this evolving discipline. Proteins are yielded from the DNA sequences after translation and transcription processes. Sequence alignment is a common approach to identification and categorization among the different residues. Through sequence alignment, prediction for the further changes in the existing structures could be identified and it enhances the further research oriented process like drug discovery. As the numbers of protein sequence are enhanced, the complexity for the multiple sequence alignment also increases. In this project, we put the efforts to optimize the multiple sequence alignment process and thereafter designed and developed the phylogenetic tree using distance based methods.**

Keywords: Phylogenetic Tree, Multiple Sequence Alignment, Distance based Methods, Proteins

INTRODUCTION

Bioinformatics involves the computer technology for the management of biological information. Computer systems are used to collect, store and analyze the information which can be then applied to different bioinformatics applications. It is basically an interdisciplinary research area between computer and biology. Java, XML, Perl, C, C++, PHP and MATLAB etc are some software tools and technologies used in the field of bioinformatics.

Challenges. The challenges in bio-informatics include the basic requirements to meet for the computation of results. With the enormous amounts of data, the challenge of bio-informatics is to store, analyze and interpret the sequence data. There should be an easy access to the information needed. Also, there should be a method for extracting only the information needed to answer a specific biological question. Thus, there is an urgent need for the new techniques and tools to be developed. The different techniques are being developed which are useful for handling these large sized databases (Xiong 2006).

Applications. The applications of Bio-informatics include: 1) Bioinformatics is used to organize biological data that help the researchers to access information, add new information about biomolecules and modify existing information in datasets. There are three types of data sets: genome sequence, macromolecular structure and data from functional genomics experiments. 2) Second level is used to develop tools and resources that help the researchers to analysis data. 3) The third level is to use these tools to analyze the data and interpret the result in a biological meaningful manner.

DNA. DNA or deoxyribonucleic acid is the basic building block of life. Basically it is the storage repository which constitutes of the information that is required for any cell to function. It is in form of a double-helix structure. DNA contains the instructions for making proteins. As shown in Fig. 1, it is composed of 4 types of nucleotides. It has two bases, purines (R) and pyrimidines (Y). Purines are of two types adenine (A) and guanine (G), and pyrimidines are cytosine (C) and thymine (T).

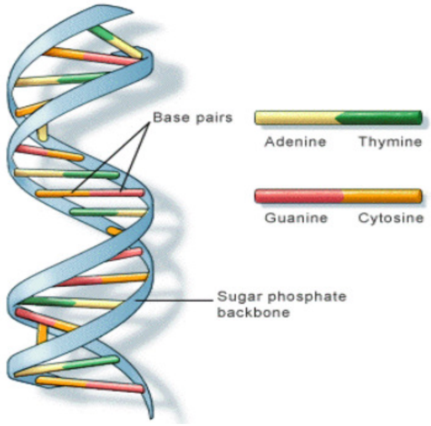


Fig. 1. Structure of DNA.

The significance of a DNA is that it takes the gene's sequence which is like a language that instructs cell to manufacture a particular protein.Further, an intermediate language, encoded in the sequence of Ribonucleic Acid (RNA), translates a gene's message into a protein's amino acid sequence (Xiong 2006).

RNA. The bases in RNA are adenine, guanine, uracial and cytosine. The various components are linked up as in DNA. RNA is very similar to DNA. RNA has properties of both DNA and proteins. Because of the dual functionality of RNA, life evolved from RNA alone, DNA and proteins having evolved later. Fig. 2 shows a RNA structure.

Amino Acid. Amino acids are the building blocks of proteins. There are 20 natural occurring amino acids that are found within proteins with a variety of chemical versatility. The amino acids are necessary for diet components. The best food source of these nutrients is protein. As amino acid band together in chains to form the stuff from which our life is born. It's a two-step process: Amino acids get together and form peptides and polypeptides. It is from these groupings that proteins are made. Commonly recognized amino acids include glutamine, glycine, phenylalanine, tryptophan, and valine.

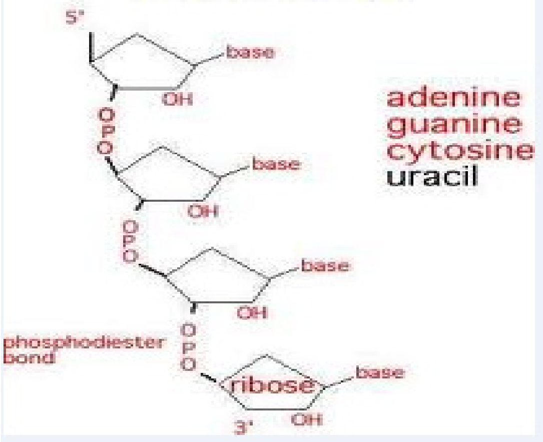


Fig. 2. Structure of RNA.

Three of those: phenylalanine, tryptophan, and valine are essential amino acids for humans. There are two types of amino acids such as essential and non essential amino acids. The essential amino acids cannot be made by the body and it got from the food but in non-essential amino acid can be made by the body. Human can produce 10 of the 20 amino acids and rest all supplied from the food. If human fails to obtain enough amino acids that means degradation of the body proteins, muscles. Amino acid consisting of an amine group (-NH_2), an α-carbon and an acidic group or a carboxylic group (-COOH).

Side chain is attached to the α-carbon as shown in Fig. 3. This side chain varies with each amino acid. Structure of an amino acid consists of an amino group, a side chain and a carboxylic group as shown in Fig. 3.

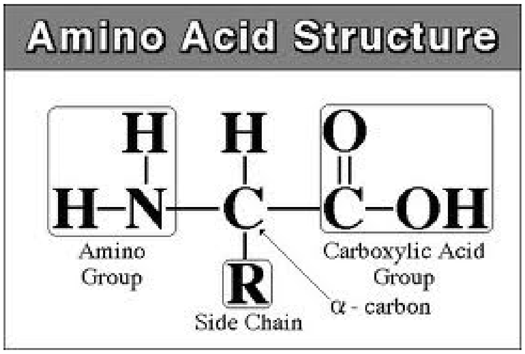


Fig. 3. Structure of an Amino Acid.

Proteins. Proteins are used for making the structure, function, and regulation of the biological functions. Each protein has unique functions. Proteins are enzymes that make the chemical reactions necessary for life possible; they provide the sensors that see, taste and smell; the effectors that make muscles move; they are the detectors that distinguish from others and create an immune response. Proteins are the building block of all life and essential for the growth of cells and repair of tissues. Proteins are made up of hundreds and thousands of units. To build a protein we need to build a long chain of amino acids. There are 20 different types of amino acids and each amino acid links together to perform the biological functions. Each amino acid shares a basic structure. Fig. 4 shows a basic structure of protein (Lesk 2002).A variety of databases include information on sequences sharing common properties that have been grouped together.

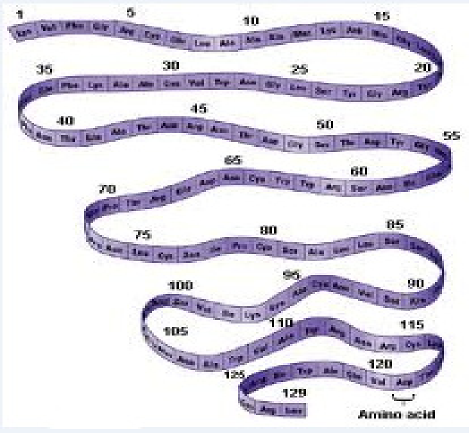


Fig. 4. Structure of Protein.

For example, the Protein Family (Pfam) database consists of several thousand families of homologous proteins. Structure databases contain information on the structure of proteins and other macromolecules. There are several protein databases from which information can be accessed. To understand full life of an organism, proteins are related within an organism and between organisms. It is done after completing the sequencing of the organisms. Further, tree of life can also be depicted from sequences of proteins and their families.

Genes. A gene is a segment of DNA representing nucleotides required for the production of a functional protein or a functional RNA molecule. Genes range in size from small to large. Genes contain not only the actual coding sequences but also adjacent nucleotide sequences required for the proper expression of genes. The vast majority of living organisms encode their genes in long strands of DNA. DNA (deoxyribonucleic acid) consists of a chain made from four types of nucleotide subunits, each composed of: a five-carbon sugar (2'-deoxyribose), a phosphate group, and one of the four bases adenine, cytosine, guanine, and thymine. The most common form of DNA in a cell is in a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine. The base pairing between guanine and cytosine forms three hydrogen bonds, whereas the base pairing between adenine and thymine forms two hydrogen bonds. The two strands in a double helix must therefore be complementary, that is, their bases must align such that the adenines of one strand are paired with the thymines of the other strand, and so on. Many prokaryotic genes are organized into operons, or groups of genes whose products have related functions and which are transcribed as a unit. By contrast, eukaryotic genes are transcribed only one at a time, but may include long stretches of DNA called introns which are transcribed but never translated into protein (they are spliced out before translation). In fig. 5, organization of a protein is being shown which constitutes of three steps, Transcription, Splicing and Translation.

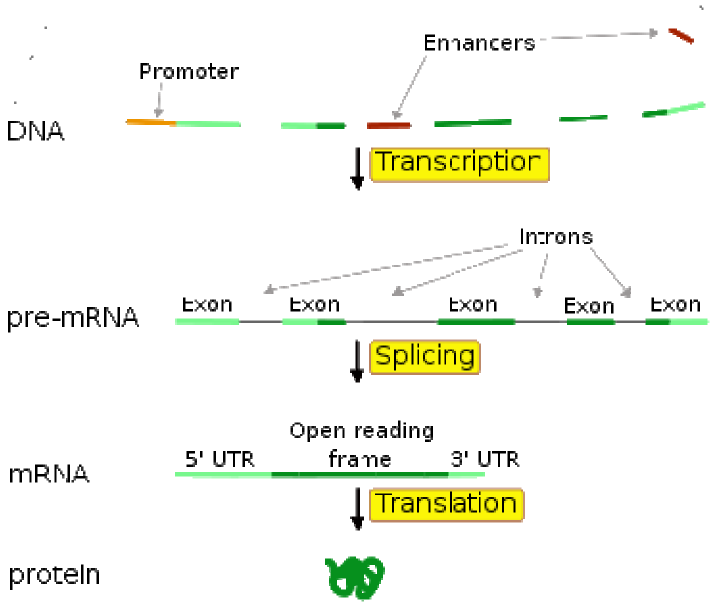


Fig. 5. Organization of a Gene.

SEQUENCE ALIGNMENT

In bio-informatics, sequence alignment is a way of arranging the primary sequences of DNA, RNA and proteins to identify regions of similarity. These regions of similarity may be a consequence of functional, structural or evolutionary relationships between the sequences. To compare the proteins already existing in the database and the new proteins, sequence comparison is most important. The process used in comparison is called sequence alignment. It provides the study of evolution. If two sequences are similar, then they evolved from the same origin. The sequences are compared by searching for common character patterns among the related sequences. The sequence alignment reveals significant similarity among a group of sequences, are considered to be belonging to the same family i.e the protein family. There are two methods for assigning sequences. 1) Global Alignments – It attempts to align every character in every sequence when the sequences in the query set are similar.

2) Local Alignment – These are more useful for dissimilar sequences. Fig. 6 shows a typical example of a local and global sequence alignment. In this example, two sequences are taken and are categorised into local and global (Vijan *et al.* 2011).

There are two types of alignments used, Pair-wise Sequence Alignment and Multiple Sequence Alignment.

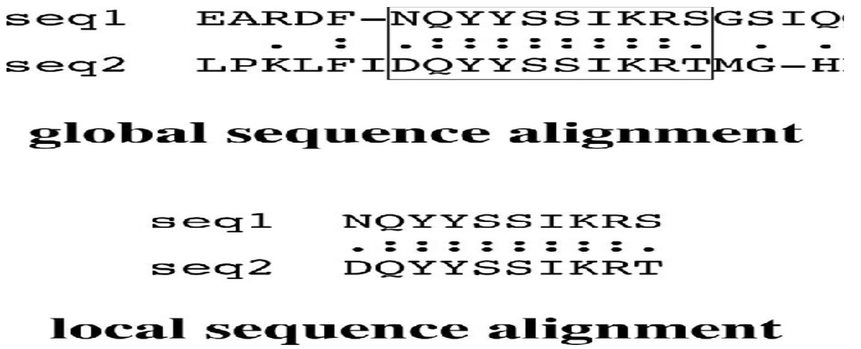


Fig. 6. Local and Global Alignment.

A. Pairwise Sequence Alignment

Pair-wise sequence alignment is the process of aligning two sequences and is the basis of database similarity searching. These can only be used between sequences at a time. These often use methods that do not require extreme precision. There are three primary methods of producing pair-wise alignments. These are Dot Matrix Methods, Dynamic Programming and Word Methods. Although each method has its individual strengths and weaknesses, all three methods have difficulty with highly repetitive sequences of low information content (low meaning of information) – especially the number of repetitions differ in the two sequences to be aligned. The best utility of given pairwise alignment has 'maximum unique match' (MUM) or longer subsequence occur in both query sequence. A longer MUM sequence reflects closer relatedness. Popular heuristic algorithms, such as *FASTA*or *BLAST*families are much faster than algorithms based on dynamic programming. Heuristic method may not be as accurate as dynamic programming, but it is fast and effective computational method. The problem exists when the no. of repetitions differ in the two sequences to be aligned (Rastogi 2007).

B. Multiple Sequence Alignment

Multiple sequence alignment is the extension of pair-wise sequence alignment. It is used to align multiple related sequences to figure out the optimal matching of the sequences. Multiple sequence alignments are very powerful as two sequences that may not align well to each other can be aligned via their relationship to a third sequence of any family. MSA allows the identification of conserved sequence patterns in the whole sequence family. It also gives more biological information than many pair-wise sequence alignments. The applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences are also a major part of MSA. The dynamic programming is not applied here. The heuristic approaches are most often used in MSA. One of the major applications of multiple sequence alignments in identifying the related sequences in databases is by construction of position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs) as well. It is basically an essential feature to do phylogenetic analysis of sequence families. The heuristic approach contains many algorithmic methods as well. The methods are: Progressive Alignment Method, Iterative Alignment Method, Block-based Alignment. These methods are mainly used to perform multiple sequence alignments. Visual examination of multiple sequence alignment tables is one of the most profitable activities that a molecular biologist can undertake away from the lab bench. A reasonable colour scheme (not the only one) is:

Table 1: Color Scheme.

Colour	Residue type	Amino acids
Yellow	Small Nonpolar	Gly, Ala, Ser, Thr
Green	Hydrophobic	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Magenta	Polar	Asn, Gln, His
Red	Negatively Charged	Asp, Glu
Blue	Positively charged	Lys, Arg

To be informative a multiple alignment should contain a distribution of closely- and distantly-related sequences. If all the sequences are very closely related, large amount of redundant information will be present, and few inferences can be drawn. If all the sequences are very distantly related, it will be difficult to construct an accurate alignment, and in such cases the quality of the results, and the inferences they might suggest, are questionable. Ideally, one has a complete range of similarities, including distantly-related examples linked through chains of close relationships. Fig. 7 shows the distinction between pairwise alignment and multiple sequence alignment.

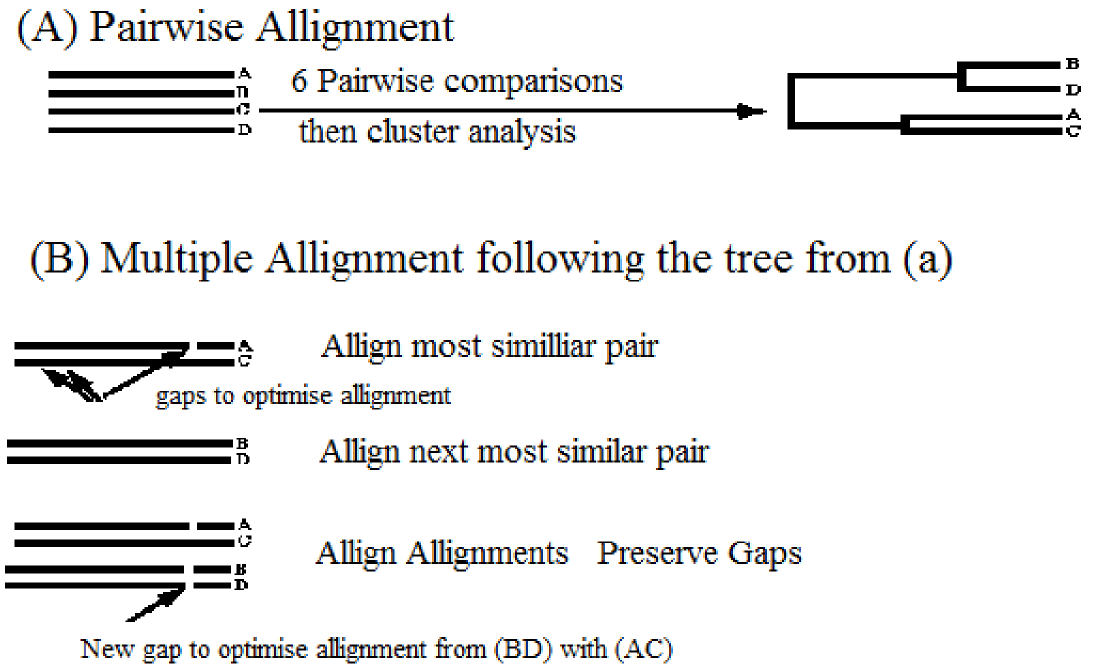

(A) Pairwise Allignment
A
B
C
D
6 Pairwise comparisons
then cluster analysis
B
D
A
C
(B) Multiple Allignment following the tree from (a)
A
C
Allign most similliar pair
gaps to optimise allignment
B
D
Allign next most similar pair
A
C
B
D
Allign Allignments Preserve Gaps
New gap to optimise allignment from (BD) with (AC)

Fig. 7. Process of Multiple Sequence Alignment.

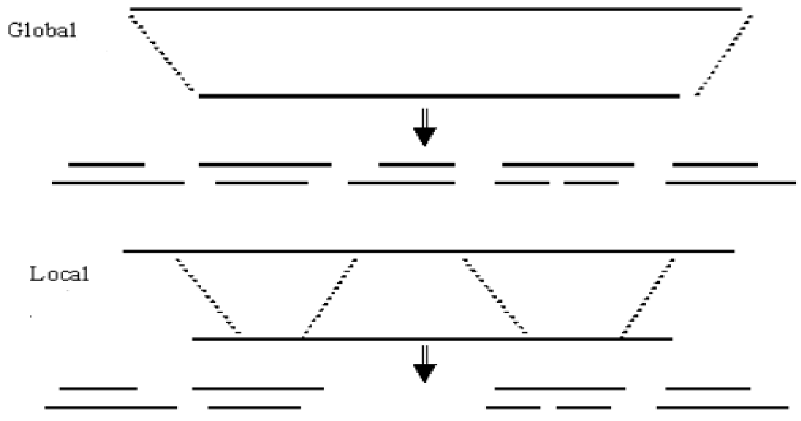


Fig. 8. Types of Alignment.

PHYLOGENETIC TREE

The Development of a biological form from other pre-existing forms or its origin to the current existing forms through some modifications is known as evolution. The study of evolutionary history of some organisms using tree-like diagrams is known as phylogenetic tree construction or phylogenetic analysis. These constructions are being used to visualize similarities and divergence among related biological sequences which is done through sequence alignment. Molecular phylogenetics is the fundamental aspect of bioinformatics.

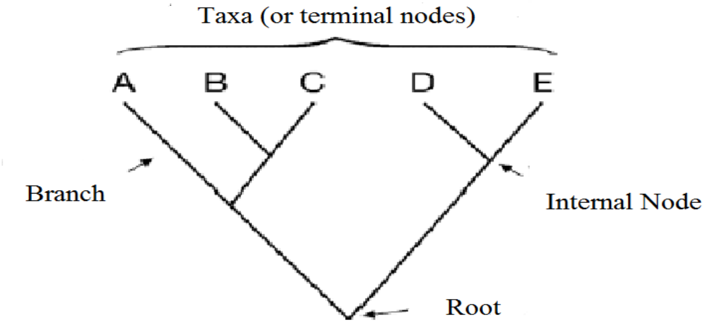


Fig. 9.A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches.

The different distance-based methods used for tree building are:

A. UPGMA

The Un-weighted Pair Group method with Arithmetic Mean (UPGMA) follows a clustering procedure where initially each species is a cluster on its own. This process is being repeated until all species are connected in a single cluster. This method is simple, fast and has been extensively used in literature.

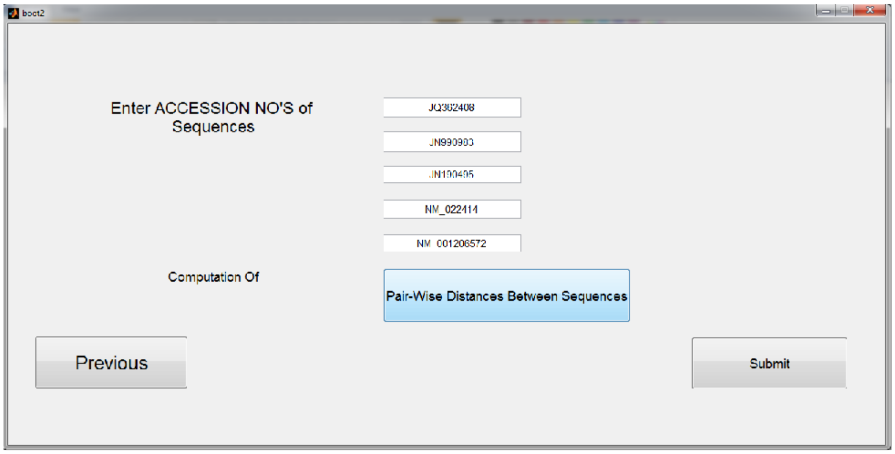
B. NJ

Neighbor Joining (NJ) It begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. It generally gives better results than UPGMA method. Under some condition, this method yields a biased tree.

C. Heuristic

Heuristic techniques help in the construction of phylogenetic trees. These heuristic searches begin by constructing an initial tree and find shorter trees using the initial tree as the starting point. A decision tree is a set of simple rules. Decision trees do not require any assumptions. These techniques have an advantage of producing accurate results.

RESULTS


boot2
Enter ACCESSION NO'S of Sequences
Computation Of
Pair-Wise Distances Between Sequences
Previous
Submit

JQ362408

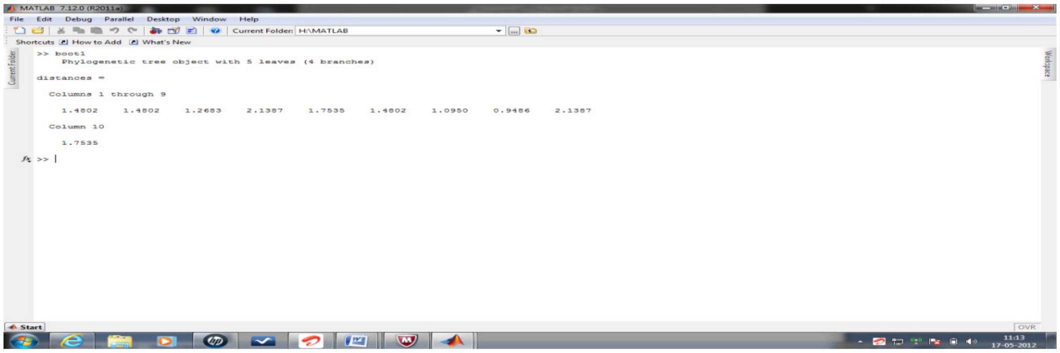
JN990903

JN190495

NM_022414

NM_001206572

Fig. 10. Entering accession no. of the sequences and calculating the paiwise distance among them.


MATLAB 7.12.0 (R2011a)
File Edit Debug Parallel Desktop Window Help
Current Folder: H:\MATLAB
Shortcuts How to Add What's New
Current Folder
Workspace
Start

```
>> boot1
    Phylogenetic tree object with 5 leaves (4 branches)

distances =

  Columns 1 through 9

    1.4802    1.4802    1.2683    2.1387    1.7535    1.4802    1.0950    0.9486    2.1387

  Column 10

    1.7535

>> |
```

Fig. 11. Showing the distances.

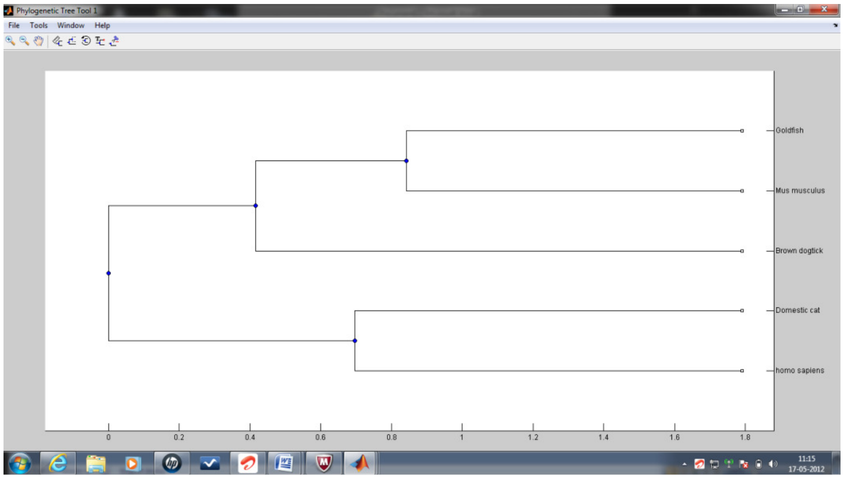


Fig. 12. Phylogenetic Tree being created.

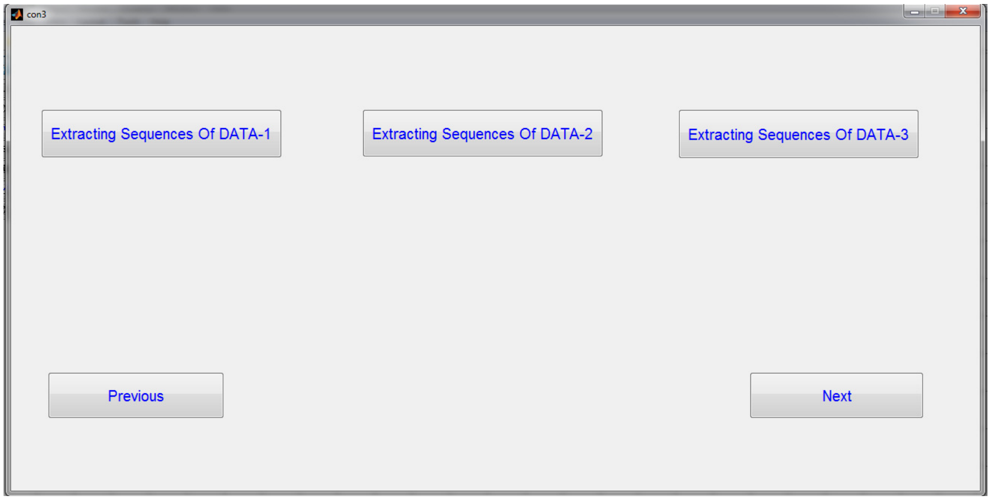
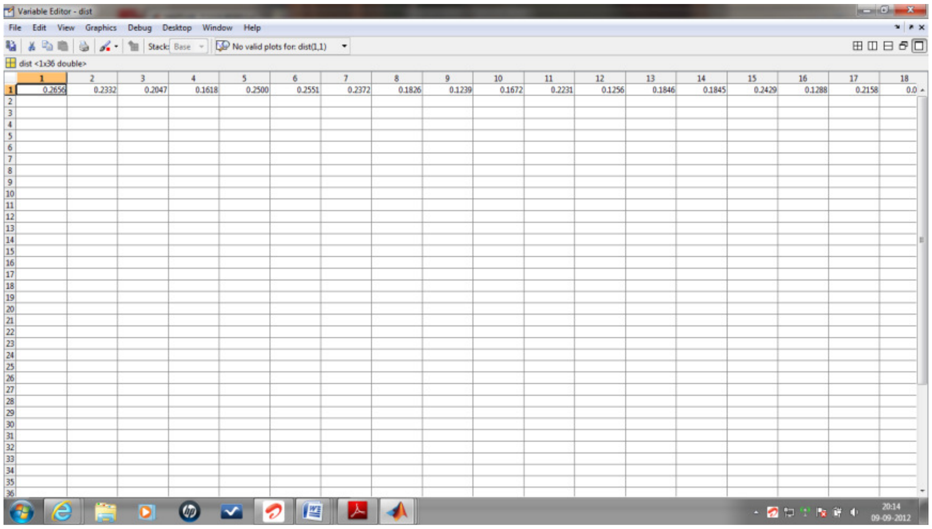

con3
Extracting Sequences Of DATA-1
Extracting Sequences Of DATA-2
Extracting Sequences Of DATA-3
Previous
Next

Fig. 13. Sequences are extracted of all the data from GENBANK.


Variable Editor - dist
File Edit View Graphics Debug Desktop Window Help
Stack: Base
No valid plots for: dist(1,1)
dist <1x36 double>
20:14
09-09-2012

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.2656	0.2332	0.2047	0.1618	0.2500	0.2551	0.2372	0.1826	0.1239	0.1672	0.2231	0.1256	0.1846	0.1845	0.2429	0.1288	0.2158	0.0

Fig. 14. The distances are being calculated of all the sequences.

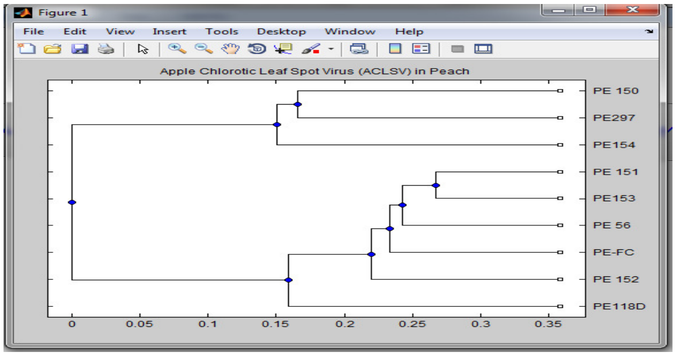


Fig. 15. Phylogenetic Tree for DATA-1.

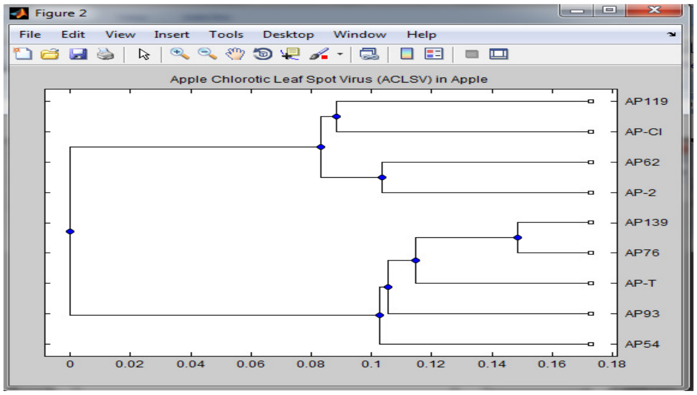


Fig. 16. Phylogenetic Tree for DATA-2.

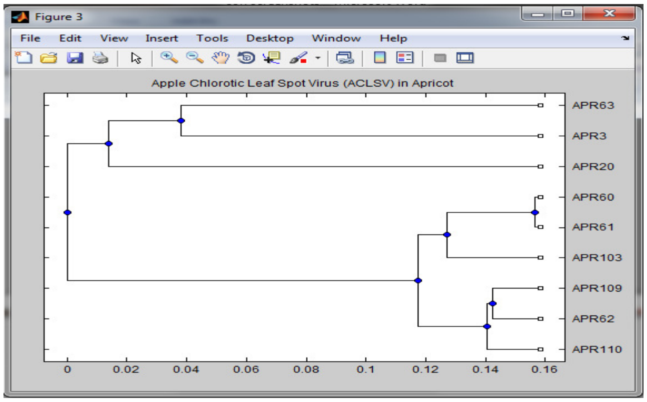


Fig. 17. Phylogenetic Tree for DATA-3.

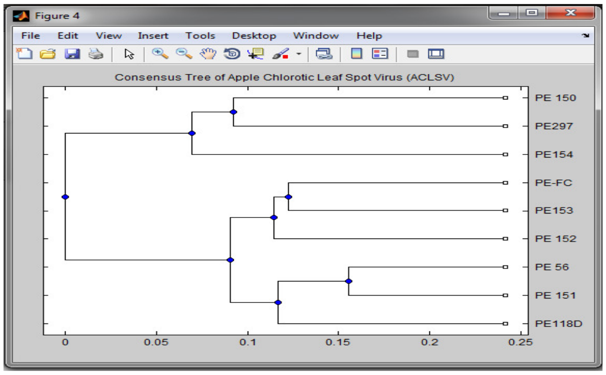
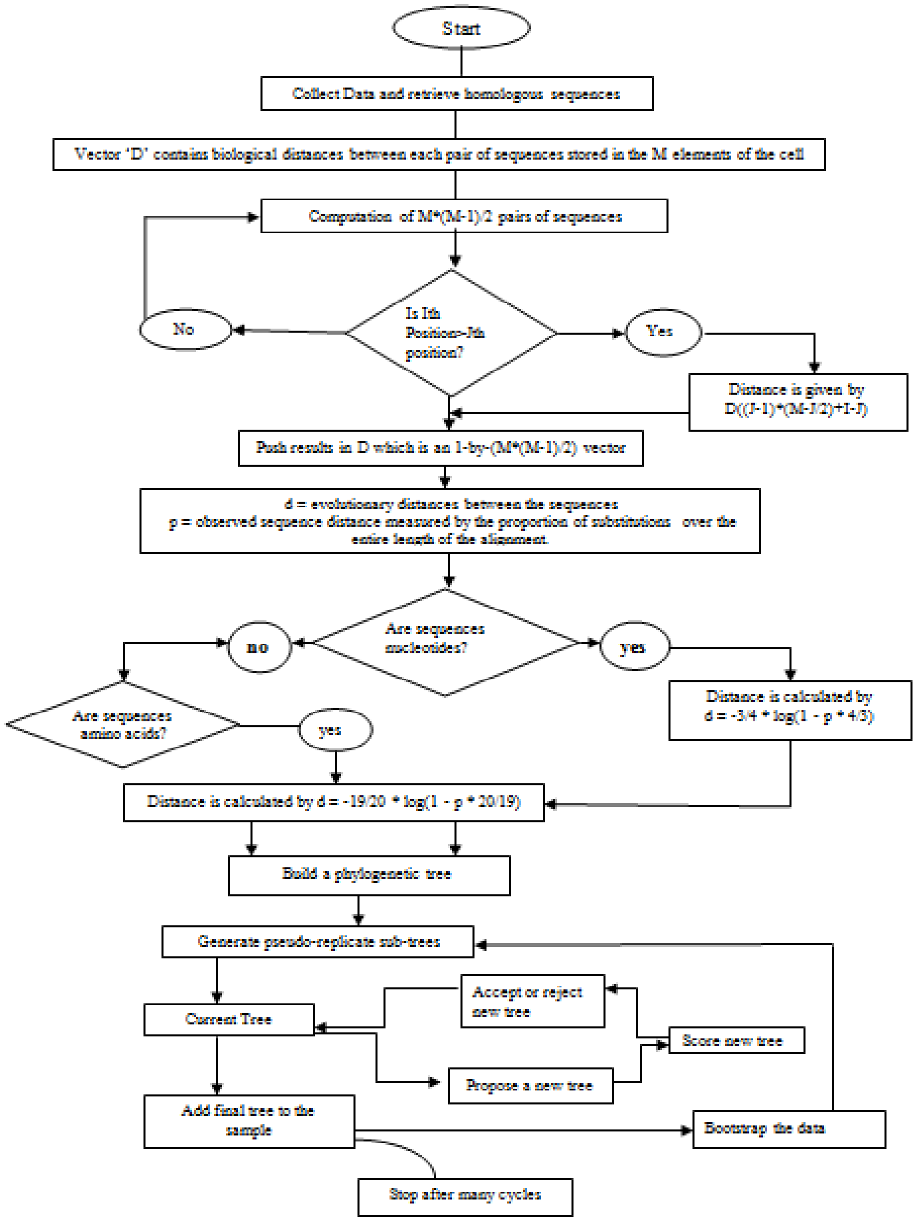


Fig. 18. A consensus phylogenetic tree.


Start
Collect Data and retrieve homologous sequences
Vector 'D' contains biological distances between each pair of sequences stored in the M elements of the cell
Computation of M*(M-1)/2 pairs of sequences
Is Ith Position=Jth position?
No
Yes
Distance is given by D((J-1)*(M-J/2)+I-J)
Push results in D which is an 1-by-(M*(M-1)/2) vector
d = evolutionary distances between the sequences
p = observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.
Are sequences nucleotides?
no
yes
Distance is calculated by d = -3/4 * log(1 - p * 4/3)
Are sequences amino acids?
yes
Distance is calculated by d = -19/20 * log(1 - p * 20/19)
Build a phylogenetic tree
Generate pseudo-replicate sub-trees
Current Tree
Accept or reject new tree
Score new tree
Propose a new tree
Add final tree to the sample
Bootstrap the data
Stop after many cycles

Annexure I: Methodology with Workflow Diagram

REFERENCES

Kakiuchi, I. and Kimura, M. (2011). "Characterization of a new neighborhood determined by three parameters", Technical Report of the NAS, pp. 1-11.

Vijan, S. and Mehra, R. (2011). "Biological Sequence Alignment for Bioinformatics Applications Using MATLAB", *IJCSET,* vol. **2**, 5, pp. 310-315.

Liu, C. and Wang, F. (2012). "Pair-wise sequence alignment algorithm in bioinformatics", *EEESym.2012*, pp. 36-38.

Shehab, S.A., Keshk, A. and Mahgoub, H. (2012). "Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics", *IJCA,* vol. **37**, 7, pp. 54-61.

Mukunthan, B., Nagaveni, N. and Pushpalatha, A. (2011). "Identification of unique repeated patterns, location of mutation in DNA finger printing using AI technique", *Journal of Bioinformatics and Sequence Analysis*, vol. **3**, 6, pp. 100-115.

Naznin, F., Sarker, R. and Essam, D. (2010). "DGA: Decomposition with genetic algorithm for multiple sequence alignment", CIBCB 2010, pp. 1-8.

Lesk, A.M. (2002). "Introduction to Bioinformatics", Oxford University Press, pp 180-200.

Lin, X., Meng, Z., He, X., Liu, Q., Liu, Y., Li, J. and Zhou, Y. (2010), "A solution to integrate the phylogenetic tree's generation based on web", ICBBE2010, pp. 1-3.

Huson, D.H. (2009). "Drawing Rooted Phylogenetic Networks", *IEEE Transactions on computational biology and bioinformatics*, vol. **6**, 1, pp. 103-109.

Miyazawa, S. (2011). "Advantages of a Mechanistic Codon Substitution Model for Evolutionary Analysis of Protein-Coding Sequences", *PLoS ONE,* vol. **6**, 12, pp. 54-69.

Mount, D.W. (2004). "Bioinformatics: Sequence and Genome Analysis", 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.. ISBN 0-87969-608-7.

Bogdanowicz, D. and Giaro, K. (2012). "Matching Split Distance for Unrooted Binary Phylogenetic Trees", *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. **9**, 1, pp. 150-160.

Nakhleh, L. (2010). "*A Metric on the Space of Reduced Phylogenetic Networks", IEEE/ACM Transactions on computational biology and bioinformatics,* vol. **7**, 2, pp 1-5.

Sundaravardhan, N., Patnaik, D., Marwah, M., Shah, A. and Ramakrishnan, N. (2012). "System and method for tree discovery*",* United States Patent Application Publication, Pub. No.: US 2012/0185421 A1

Krane, D. and Raymer, M. (2006). "Fundamental concepts of bioinformatics", Pearson Education Publishers.

Pevsner, J. (2009), "Bioinformatics and Functional Genomics", A john wiley & sons, Inc. Publication, pp 215-221.

Rastogi, S.C., Mendiratta, N., Rastogi, P. (2007). "Allignment of Multiple Sequences and Phylogenetic Analysi-Bioinformatics Methods and Applications", 3rd edition, PHI publication, pp. 5-120.

Heard, E. J., Riechmann, L. J., Creelman, A. R., Ratkliffe, O., canals, D. R., Repetti, P. P., Kumimoto, W. R. and Libby, M. J. (2010). *"Plant transcriptional regulators"*, United States Patent Application Publication, Pub. No.: US 2010/0175145 A1

Torres, M., Dias, G., Gonçalves, G. and Vieira, C. (2011). "Tool that Integrates Distance Based Programs for Reconstructing Phylogenetic Trees", *IEEE latin america transactions*, vol. **9**, 5, pp. 895-901.

Gronau, I., Moran, S. and Yavneh, I. (2010). "Adaptive Distance Measures for Resolving K2P Quartets: Metric Separation versus Stochastic Noise", *Journal of Computational Biology*, vol. **17**, 11, pp. 1509-1518.

Guo, P., Chen, G. and Wang, Y. (2011). "Constructing phylogenetic tree based on three-parameter model", Key Engineering Materials, vol. **474-476**, pp. 2193-2197.

Sul, J., S. and Williams, L., T. (2009). "An experimental Analysis of consensus tree algorithms for large-scale Tree Collections*",* ISBRA, LNBI 5542, pp 100-111.

Rwahnih, A. M., Turturo, C., Minafra, A., saldarelli, P., Myrta, A., Pallas, V. and savino, V., "Molecular variability of apple chlorotic leaf spot virus in different hosts and geographical regions*",journal of plant pathology*, pp 117-122.

http://www.chemguide.co.uk/organicprops/aminoacids/dna1.html

http://www.odec.ca/projects/2004/mcgo4s0/public_html/t3/RNA.html

http://www.thefoodadvicecentre.co.uk/reference/protein

http://www.sci.sdsu.edu/DNA/protein.html

http://en.wikipedia.org/wiki/Gene

http://mcclintock.generationcp.org/pairwisesequenceallignment

Zhou, J., Sander, J., Cai, Z., Wang, L. and Lin, G. (2010). "Finding the Nearest Neighbors in Biological Databases Using Less Distance Computations", *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. **7**, 4, pp. 669-680.

Zimek, A., Buckwald, F., Frank, E. and Kramer, S. (2010). "A Study of Hierarchical and Flat Classification of Proteins", *IEEE/ACM Transactions on computational biology and bioinformatics,* vol. **7**, 3, pp. 563-571.

Wang, Y., E. and Sheth, D., N. (2011). "Methods and collective reasoning framework for complex decision making*"*, United States Patent Application Publication, Pub. No.: US 2011/0320374 A1

Wang, L.S., Leebens-Mack, J., Wall, P.K., Beckmann, K., DePamphilis, C.W. and Warnow T. (2011). "The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation", *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. **8**, 4, pp. 1108-1119.

Xiong J., (2006). "Essential Bioinformatics", United States of America, Cambridge University Press, New York.