Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 7, July 2019, pg.45 – 51

Multiple Sequence Alignment Based Method for Construction of Phylogenetic Trees

Priyanka Shaktawat; Parvati Bhurani
Govt. Women Engineering College, Ajmer
Email ID: priyankashaktawat14@gmail.com, parvatibhurani@gweca.ac.in

Abstract: Due to the importance of DNA (genetic material) and protein sequences, their comparison becomes the major part of biology. But the presence of large and complex datasets of biological information requires an efficient computational methodology to handle them. The sequence comparison facilitates identification of genes and conserved sequence patterns to infer the evolutionary relationship among different species. This paper uses Multiple Sequence Alignment (MSA) method that aligns multiple sequences at a time to depict phylogeny. The p53 protein sequences of ten different species are loaded from the NCBI (National Center for Biotechnology Information) databank in the FASTA format. Based on the evolutionary distances of these species, two phylogenetic trees are constructed for the two divided parts of this dataset. A single tree is generated by joining two trees using pruning method. To obtain an optimal alignment, each sequence in the pruned alignment is locally aligned with the consensus sequence. The minimum optimal alignment is obtained after performing left and right shift operations.

Keywords: Bioinformatics, Sequence Alignment, Phylogeny, Biotech data, Data Mining, NCBI data bank.

I. Introduction

Bioinformatics is a combined science that associates biological data with the computerized techniques of storage, distribution and analysis of data to support scientific research in multiple areas such as evolutionary studies, microbial genome applications, biomedicine and etc. It requires knowledge of many branches such as biology, mathematics, computer science, data mining and etc. to analyze biotech data. The mathematical, statistical, computing algorithms aim to solve the biological problems using DNA and amino acid sequences.

A. Sequence Alignment

A sequence alignment organizes DNA, RNA or protein sequences to identify similar regions that are a result of structural, functional or evolutionary relationships between different species. The sequence alignment is further classified on the basis of number of sequences as: pairwise sequence alignment and multiple sequence alignment. The pairwise sequence alignment aligns only two nucleotide sequences at a time while multiple sequence alignment methods align more than two nucleotide sequences at a time. Aligned sequences of nucleotides or amino acids are represented in the form of matrix. The gaps are introduced between the nucleotides so that similar characters can be aligned in successive columns. Mismatches can occur, if two or more sequences have a common ancestry and gaps as indels are introduced due to mutations. Due to the continuous increase in

© 2019, IJCSMC All Rights Reserved

number of biological data in the field of proteomics, multiple sequence alignment methods are required that deals with multiple sequences at a time. The biological data is available in different data types such as: genome database, sequence database, structure database, enzyme database and etc. The multiple sequence alignment methods take sequence databases for evolutionary study. There are different types of sequence databases such as: GenBank, EMBL: European Molecular Biology Laboratory (for nucleotides), Swiss-Prot, UniProt (for proteins) and etc.

B. Phylogeny

The phylogenetic tree is a form of graphical representation of relationship between organisms, species or genomic sequence. Phylogeny is the study of evolutionary history of different species. Each node in the tree represents a specific event of evolution. The structure of tree has terminal nodes (sequences) and internal nodes (Last Common Ancestor). There are mainly two types of phylogenetic trees:

1. Rooted tree: A tree in which all the objects (DNA, RNA or proteins) shares a common ancestor (the root). The evolution is shown by the path from root to leaf nodes.
2. Unrooted tree: A tree in which all the objects are related descendants and common ancestor is not specified in it.

Figure.1. Rooted and unrooted trees

C. Data Mining

The data mining is the process of extracting useful knowledge from large databases. It is of three types: classification, association, and sequences. The classification divides the data into classes which are further used to predict new unclassified classes. Associations are the patterns in the data that occur frequently sequences are the knowledge about the data where time is involved. The primary objectives of data mining are prediction and description. Prediction uses existing databases unknown variables of interest, and description extracts human readable pattern that describe data (Sita Rani, 2012) [1].

In bioinformatics, to analyze large amount of biological data available in the study of proteomics, genomics and etc. requires high computational methods. So data mining is used in analysis of biological data such as protein structure prediction, gene classification, analysis of mutations in cancer etc. The process of evaluating hidden patterns in large data set and to establish relation among them is called data mining.

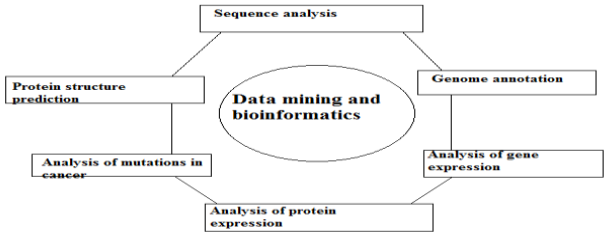

Figure.2. Applications of bioinformatics and data mining

© 2019, IJCSMC All Rights Reserved

II. Methodology

To implement this work progressive alignment technique is used that computes the alignment scores among various multiple sequences. On the basis of the alignment of multiple sequences the phylogenetic tree is constructed that depicts the relation among different species using neighbor joining method. The tumor protein p53 protein data of 10 different species has been taken from NCBI (National Center for Biotechnology Information) data bank.

A. Jukes-Cantor (JC) method

The jukes cantor method is a substitution model in which a sequence of symbols changes into different set of traits. The JC method computes the probability of such substitution from one sate to another (Geetika Munjal, 2015) [6]. For example, in cladistics (an approach of biological classification in which species is grouped on the basis common ancestors) each position in the biological sequence of an organism represents a specified property which can be present or absent in other organisms. This method is used to find the evolutionary distance between two sequences by using the formula:

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3} * \frac{Nd}{N})$$

Where,

N_d: number of mutations between two sequences

N: Nucleotide length (Krane, 2006)[2]

B. Algorithm

The proposed algorithm generates an optimal alignment of the most closely related species from the given set of different species. The main objective of this implementation is to construct a phylogenetic tree of the tumor protein p53 protein (its act as a tumor suppressor i.e. it regulates cell division by keeping cells from growing or proliferating too fast).The protein sequences has been taken from NCBI databank in the form of FASTA format. These data sequences can be accessed using an accession ID from the NCBI database. In FASTA format the sequence is described using a single line description that begins with the greater than sign (>) and the data sequence itself. The Multiple Sequence Alignment method is used to align the multiple sequences at a time which speed up the alignment process for the closely related species. Evolutionary distances are calculated on the basis of jukes cantor method and the trees are constructed on the basis of neighbor joining method (Joseph L. Staton, 2015) [5]. The final evolutionary tree is obtained by tree pruning method. The threshold condition is used to select the most closely related species. The optimal alignment is obtained by aligning the available sequence with the consensus sequence (fixed sequence). We can get the consensus sequence by aligning number of protein sequences those share a common structure or function. The consensus sequence is itself an optimal sequence which further improves the accuracy of the sequence alignment. The sequences are aligned on the basis of local alignment and five left and right shift operations are performed to obtain accurate multiple sequence alignment. The algorithm of constructing phylogenetic tree is implemented as:

- Load protein sequences of m different species from NCBI data bank (m can be any number, here m=10)
- Calculate the evolutionary distances using jukes cantor method of half of the total number of species
- On the basis of JC distance create a first matrix
- Select the smallest distance from the matrix. Half of that distance refers to the branch length from the first node to each of those two species
- Create a new reduced matrix with entries for the new node and unpicked species
- Select the smallest distance from the reduced matrix. If these two are unpicked taxa then they will form a new node with branch length half that of distance.
- Continue this process until all species are picked up. Now the first tree is generated.
- Build the second tree from remaining number of species
- First and second trees are joined using pruning method
- Select the species that are above threshold value 0.6
- Consider the consensus sequence for the p53 protein sequences

© 2019, IJCSMC All Rights Reserved

- Load the sequences that are above the threshold value
- Align the sequences with consensus sequence using local alignment
- Perform five left shift and right shift operations
- Consider the alignment with minimum score

III. Discussion of Result

The p53 protein sequences of ten different species from NCBI databank shown in table 1 are used as input for the present research implementation. These data sequences can be accessed from the data bank with their Accession ID having different sequence lengths and units written as amino acids.

S.No	Name of species	Accession ID	Sequence Length
1.	African green monkey	P13481	393 aa
2.	Human	AAD28535	393 aa
3.	Crab eating Macaque	P56423	393 aa
4.	Northern Tree Shrew	Q9TTA1	393 aa
5.	Woodchuck	O36006	391 aa
6.	Pig	NP_998989 NP_999310	386 aa
7.	Cat	P41685	386 aa
8.	Dog	NP_001003210	381 aa
9.	Sheep	P51664	382 aa
10.	Rabbit	Q95330	391 aa

Table 1: Dataset of ten different species from NCBI databank

The p53 protein sequences of first five species are loaded from the NCBI databank. MSA for these sequences is obtained by aligning them with the consensus sequence which is a fixed and obtained from the most frequently occurring residues. The dataset of first five sequences is aligned as shown in figure.3.

Figure3: Multiple Sequence Alignment of first five species

© 2019, IJCSMC All Rights Reserved

The evolutionary distance between the species is found using Jukes - Cantor method and on the basis of these distances phylogenetic tree is constructed by the neighbor joining method as shown in figure.4.

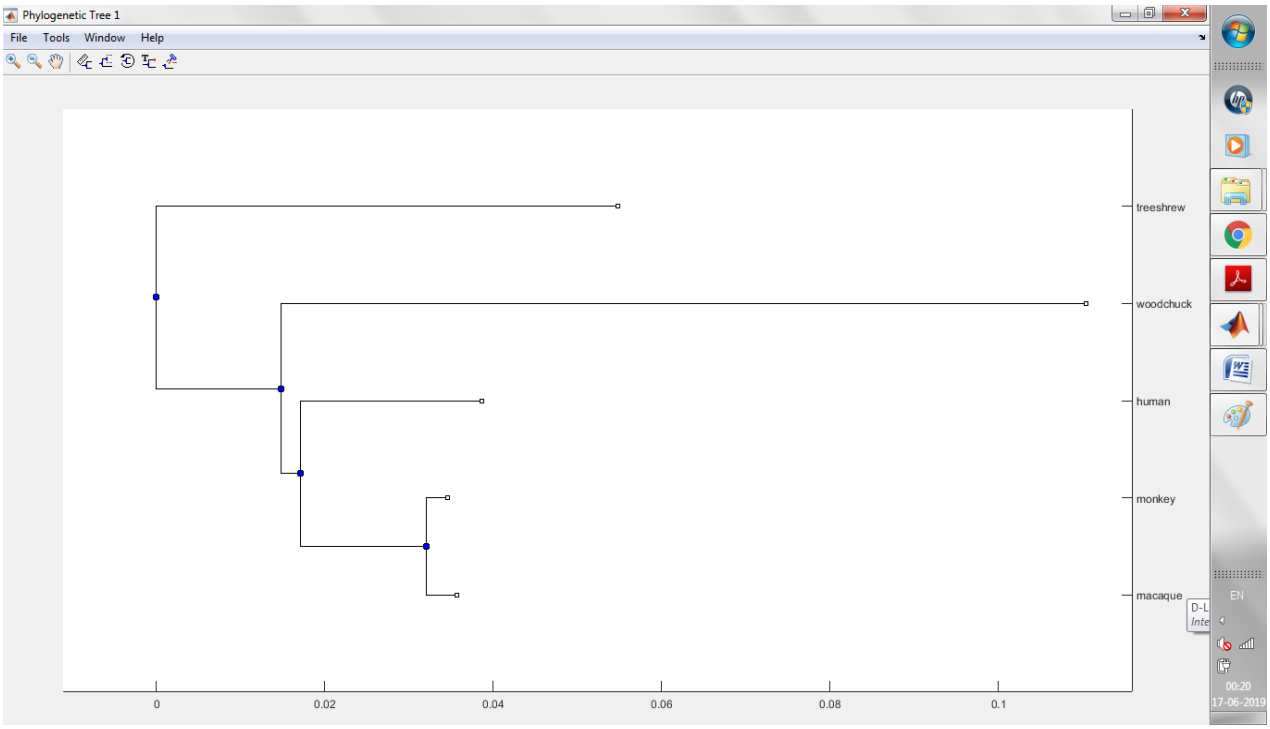

Figure 4: Phylogenetic tree of first five species

The same procedure is repeated for next five protein sequences. The MSA and phylogenetic tree for next five species are shown in figure.5 and figure.6 respectively. Now we get two phylogenetic trees which are further joined using pruning method. The pruning method reduces the size of the tree by combining most relevant sequences. The pruned tree contains all the species whose evolutionary distance is above the threshold value 0.6.

Figure.5. Multiple sequence alignment for the next five species

© 2019, IJCSMC All Rights Reserved

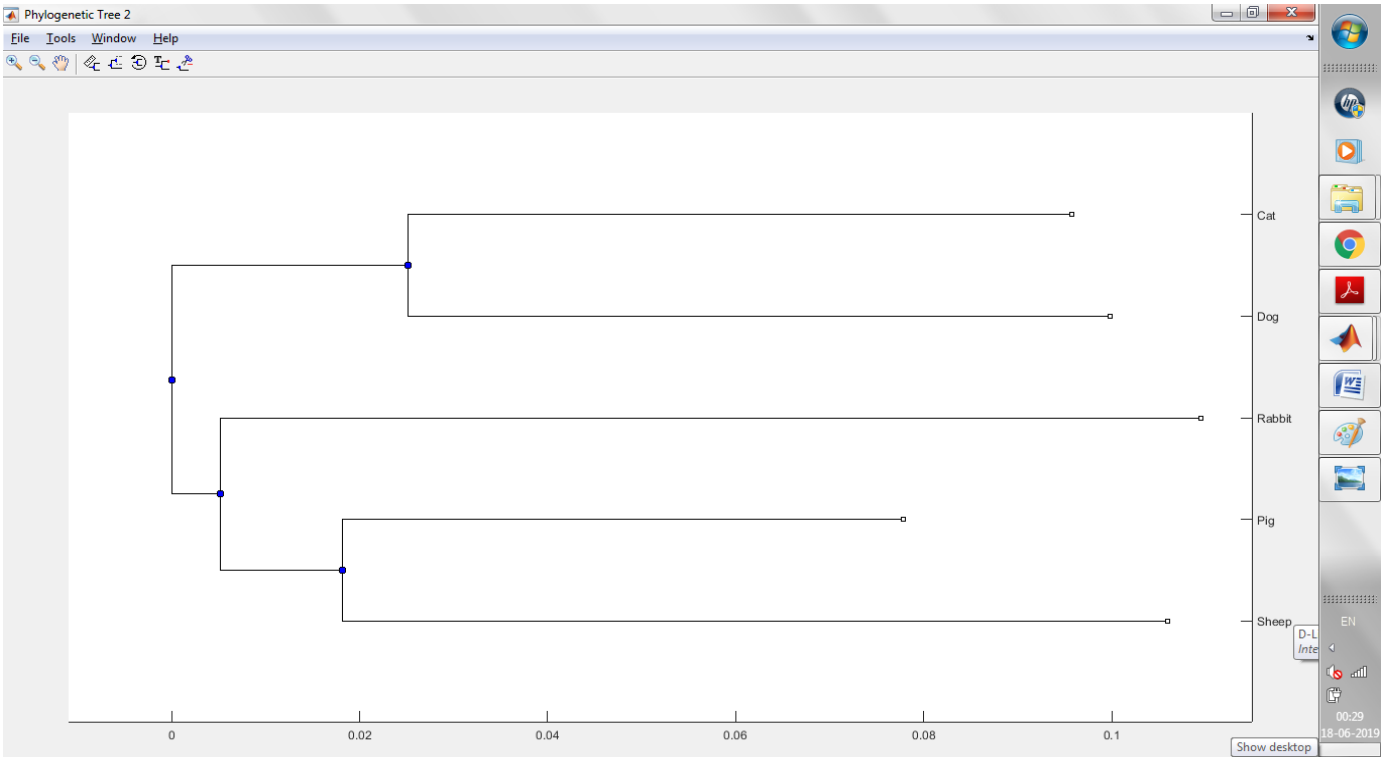

Figure.6. Phylogenetic tree of next 5 species

The multiple sequence alignment of ten sequences of p53 proteins is shown in figure.7 and the final phylogenetic tree of all ten species is shown in figure.8. It shows the complete alignment of all the sequences.

Figure.7.The pruned sequences are aligned with the consensus sequence

© 2019, IJCSMC All Rights Reserved

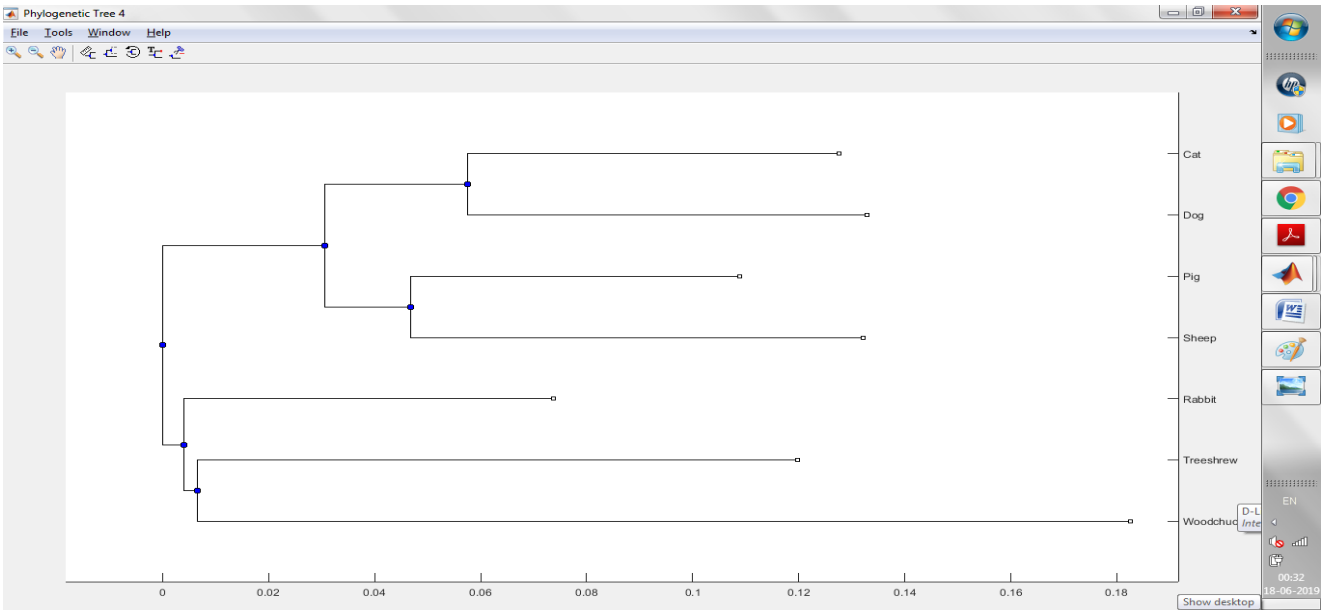

Figure.8. Phylogenetic tree obtained by tree pruning method

IV. Conclusion and Future Scope

The multiple sequence alignment is obtained by aligning the p53 protein sequences available in the FASTA format. The model is implemented in the MATLAB software with the threshold value 0.6. On the basis of this threshold condition, the closely related species are selected. This model is capable of providing an optimal multiple sequence alignment using consensus sequence. In future, this model can be implemented for different dataset with different threshold value. The evolutionary distance calculation method can be kimura-2 parameter while scoring matrix like Blossom, PAM and etc can be used in place of GONNET matrix.

References

1. Sita Rani and Simarjeet Kaur, **"Cluster Analysis Method for Multiple Sequence Alignment",** International journal of Computer Applications, Volume: 43, Issue: 14 April, 2012.
2. Krane, D. and Raymer, **"Fundamental Concepts of Bioinformatics",** Pearson Education Publishers, 2006, pp. 1-98.
3. Manmeet Kaur and Navneet Kaur, **"Multiple Sequence Alignment of Phylogenetic tree Construction",** International Journal of Application or Innovation in Engineering and Management, Volume: 5, pp. 126-130, 2016.
4. Manmeet Kaur and Navneet Kaur Bawa, **"Alignment Based Method Using Clustering and Distance Evaluation for Construction of Phylogenetic Tree"**, International Conference on Information Technology – the Next Generation IT Summit, 2016.
5. Manmeet Kaur and Rajbir Singh, **"Phylogenetic Tree Construction Revision"**, International Journal of Advanced Research in Computer Science and Software Engineering, Volume: 6, pp. 517 – 520, 2016.
6. Geetika Munjal, Madasu Hanmandlu and Deepti Gaur, **"A New Alignment Free Method for Phylogenetic Tree Construction"**, International Journal of Database Theory and Application, Volume: 8, pp.111-124, 2015.
7. Joseph L. Staton, **"Understanding phylogenies: Constructing and Interpreting Phylogenetic trees"**, Journal of the South Carolina Academy of Science, Vol. 13, 2015.

© 2019, IJCSMC All Rights Reserved