# Phylogenetic Tree Generation using Different Scoring Methods

Rajbir Singh
Associate Prof. & Head
Department of IT
LLRIET, Moga

Sinapreet Kaur
Student of M.Tech
Department of CSE
LLRIET, Moga

Dheeraj Pal Kaur
Assistant Prof. (ECE)
Department of ECE
LLRIET, Moga

## ABSTRACT
Data Mining is a branch of knowledge discovery in the field of research and development. The biological data is available in different formats and is comparatively more complex. Knowledge discovery from these large and complex databases is the key problem of this era. Data mining and machine learning techniques are needed which can scale to the size of the problems and can be customized to the application of biology. Hierarchical Clustering is the one of the main techniques for data mining. Phylogeny is the evolutionary history for a set of evolutionary related species. One approach on determining the evolutionary histories of a dataset are scoring based methods. There are number of different distance based methods of which two are details with here: the UPGMA (Unweighted Pair Group Method using Arithmetic average) and Neighbor Joining. A method for construction of distance based phylogenetic tree using hierarchical clustering is proposed and implemented on different rice varieties. The sequences are downloaded from NCBI databank. Evolutionary distances are calculated using jukes cantor distance method. Multiple sequence alignment is applied on different datasets. Trees are constructed for different datasets from available data using both the distance based methods and pruning technique. SNAP calculates synonymous and non-synonymous substitution rates based on a set of codon aligned nucleotide sequences. The DNA Multiple sequences to calculate the GC content of eukaryotes, molecular weight, melting temperature and tree information. Extractions of closely related varieties are performed by applying threshold condition. Then, final tree is constructed using these closely related rice varieties.

## Keywords
Data Mining, DNA, Phylogenetics

## 1. INTRODUCTION
DNA or deoxyribonucleic acid is a molecule that encodes the genetic instructions which are used in the development of all living organisms . Along with RNA and proteins, DNA is one of the three major macromolecules essential for all known living species. Genetic information is encoded in the form of sequence of nucleotides as guanine (G), adenine (A), thymine (T), and cytosine(C). DNA is organized into long structures called chromosomes. Like DNA, RNA is also vital for living beings. The bases in RNA are adenine, guanine, uracil and cytosine. RNA is very similar to DNA All of the genetic information of any living species is stored in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), which are polymers of four simple nucleic acid units which are called nucleotides. Phylogenetics is the study that helps to find out the relationship among different group of species. The relationship is discovered by moleculer sequencing data. A phylogenetic tree is a branching diagram which shows the evolutionary relationship among different species. Gene phylogeny represents the evolutionary relationship which is derived from genes or protein sequences. Species phylogeny represents the evolutionary path of the species. Reconstruction of phylogenetic relationship using a DNA, RNA or amino acid sequences is a hierarchal process. . Early approaches were based on single processor computers. and phylogenetics is based on parallel and distributed computing. In general phylogenetic trees represents the relationships between taxa. There are two different methods for construction of phylogenetic tree that are distance based tree construction method and character based tree construction method. These two categories both offer a vast variety of options when constructing trees in two different directions. Trees with scaled edges are called phylograms while with non scale edges are called cladograms.

## 2. METHODOLOGY
The methodology for the work involves the use of distance based methods for phylogenetic tree construction. MATLAB and SNAP are the software tools used. Out of different Data Mining techniques, Hierarchical clustering is used. The data is taken from the databank of NCBI. SNAP (Synonymous Non-Synonymous Analysis Program) calculates synonymous and non-synonymous substitution rates based on a set of codon aligned nucleotide sequences and calculates pair wise synonymous and non-synonymous distances. If changes in the nucleotide sequences which does not change the sequences of amino acid of a proteins is synonymous substitution and if produce changes in amino acid sequences is termed as non synonymous substitution. Any biological information can be used to find out evolutionary relationship among taxas is called phylogenetic information maker. For this DNA sequences is used. Then multiple sequence alignment is done. For constructing DNA phylogenies, Jukes-cantor is used.. In this cluster-based methods are used. After selecting the appropriate methods and steps for tree construction, tree is constructed. Data used is taken from NCBI. sequences that are used, entered through the various formats. The various format available are: Plain Text format, FASTA format, GENBANK and Genetic Computer Group format (GCG). FASTA format is used for this problem and it is simple and easy to understand. Therefore DNA sequences are used and taken from NCBI in FASTA format for this work.

## 2.1 Distance Measuring

For constructing a phylogenetic tree based on data from protein or nucleotide sequence comparisons, first do a multiple alignment for the sequences and after that calculate distance measure of all taxa. Both methods use pair wise distances for determining the tree , here define distances $d_{ij}$ between each pair of sequences i, j in the given dataset.

## 2.2 Jukes Cantor method

The Jukes and Cantor model is a model which computes probability of substitution from one state to another. From this method, also derive a formula for computing the distance between 2 sequences. The main idea behind this method is the assumption that probability of changing from one state to a different state is always equal. As well, it is assumed that the different sites are independent. The evolutionary distance between two species is given by the following formula.

$$d = -\frac{3}{4}\ln\left(1-4\frac{4}{3}\frac{N_d}{N}\right)$$

$N_d$ is the number of mutations between the two sequences and N is the nucleotide length.

## 2.3 Molecular Clock

For assigning branch lengths to phylogenetic tree, must consider whether the evolutionary rate is constant. The determination of a phylogenetic tree for the evolution of species from amino acid or nucleotide sequence comparisons depends on measurements and assumptions concerning the rate of evolutionary change is molecular clock.

## 2.4 Mutation Rate

Mutation are exchange of nucleotide or any insertion or deletion of nucleotide.The mutation rate is a measure of the rate at which various types of mutations occur during some unit of time .It may be small or large scale insertions or deletions.

## 2.5 Functional constraint

Natural selection occurs when changes to gene that diminish the ability to survive of an organism and reproduce, are removed from gene pool. The portion of genes which are very important for survival of organism is under functional constraint, and changes occur very slowly over the evolution.

## 2.6 UPGMA Method

UPGMA stands for Unweighted Pair Group Method using Arithmetic average. It starts with grouping two taxa having smallest distance between them according to the distant matrix, then new node is added in the midpoint of the two, and the two original taxa put on the tree. The distance from the new node to other nodes will be the arithmetic average. Then by replacing two taxa with the new node is used to obtain a reduced distance matrix. Repeat the same process until all taxa are placed on the tree. The taxon added at the last is the root of the tree.

$$D_{(i,j),k} = \left(\frac{n_i}{n_i+n_j}\right)D_{i,k} + \left(\frac{n_j}{n_i+n_j}\right)D_{j,k}$$

## 2.7 Neighbor joining method

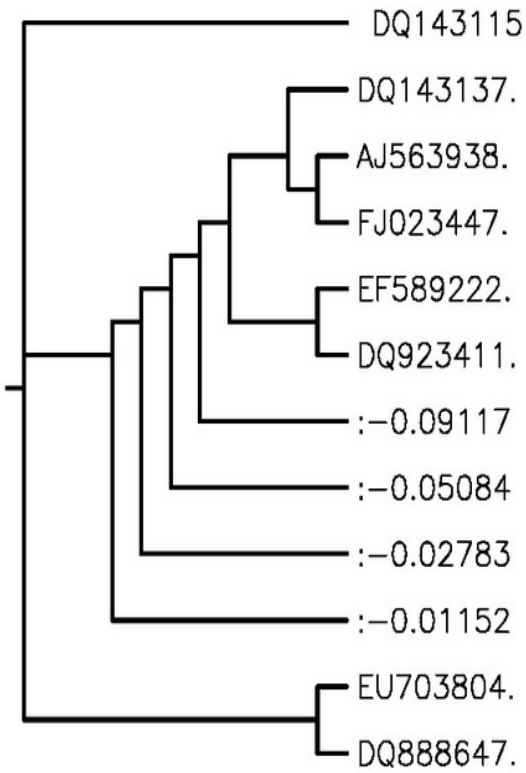
Neighbor Joining (NJ) works like UPGMA method. It creates a new distance matrix at each step, and creates the tree based on the matrices. NJ does not construct clusters and directly calculates distances to internal nodes,that is the difference between UPGMA and NJ methods. The first step of the NJ method is that to create a matrix with the Hamming distance between each node. The minimal distance that is calculated is then used to calculate the distance from the two nodes to the node that directly links them. Therefore a new matrix is calculated and the new node is substituted for the original nodes that are now joined. Neighbor Joining works even if the lengths are not additive but there is no guarantee that the tree is correct.

$$D_{(i,j),k} = \frac{D_{i,k} + D_{j,k} - D_{i,j}}{2}$$

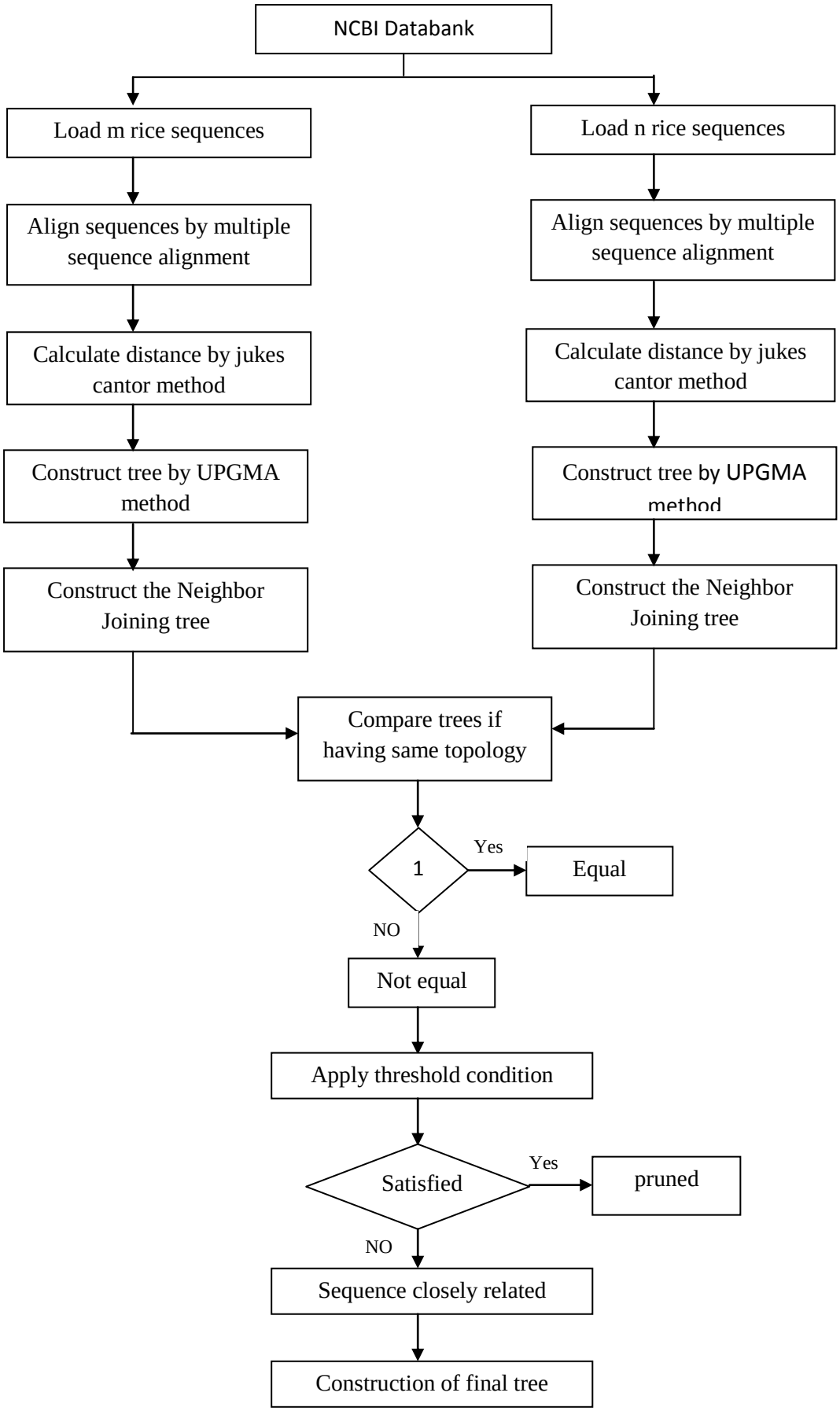
# 3. RESULTS AND DISCUSSION

Phylogenetic Tree Construction is one of the most important goals pursued by bioinformatics. To generate a phylogenetic tree the main computational problems are threefold. Firstly, to determine, and compute, a distance metric between every genomic sequence. Secondly, to perform hierarchical clustering on the given datasets, utilizing the distance metric computations. Two distance matrices that are compatible input for Jukes Cantor model are created, based on either the ds or dn values.

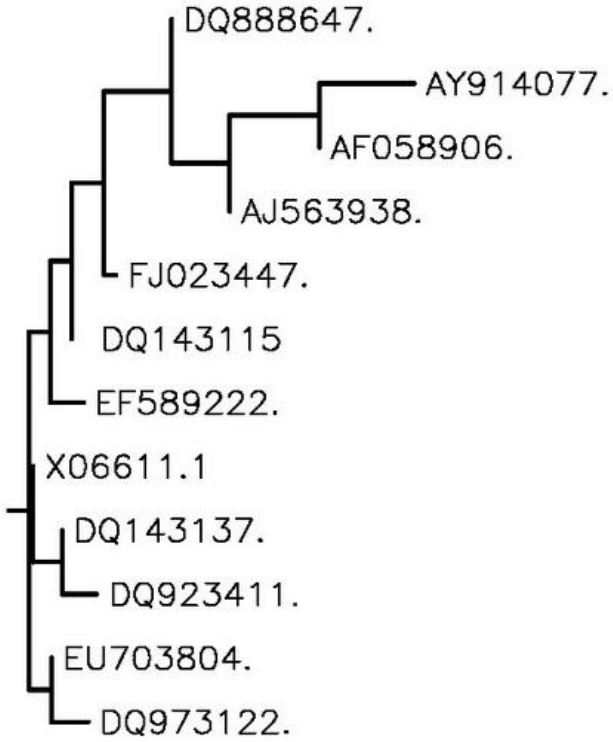
## 3.1 Construction of dS tree



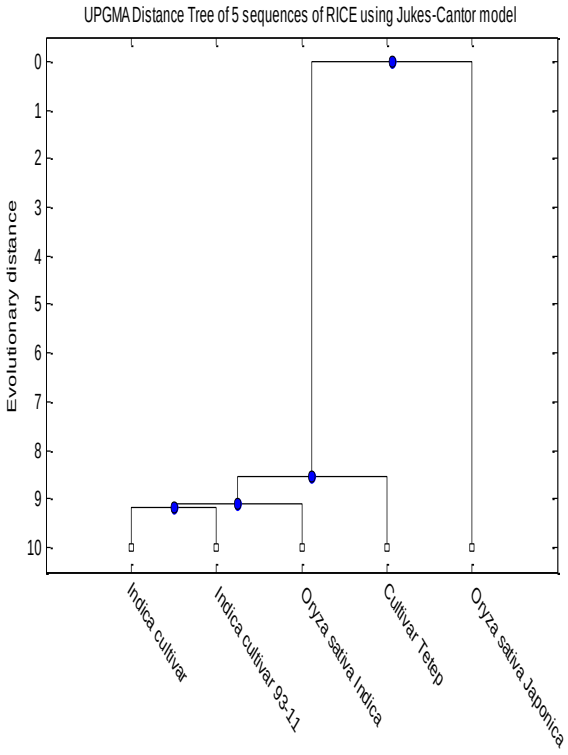
### 3.2 Flow Chart of present work



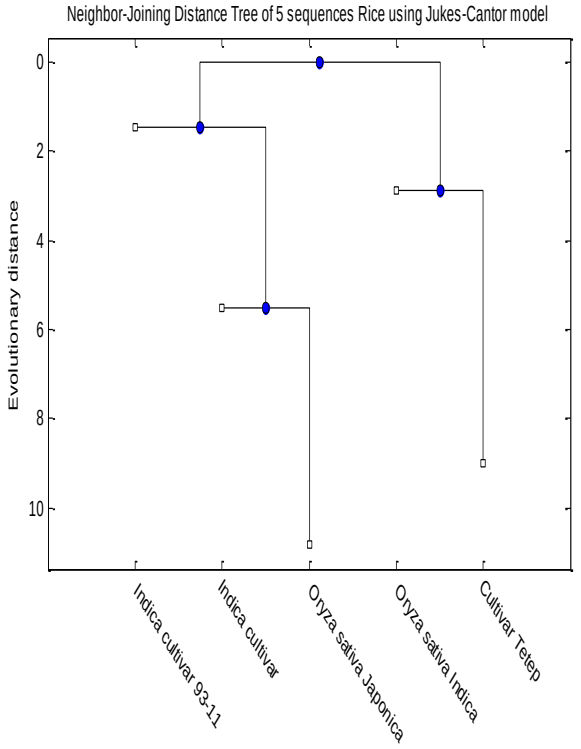
### 3.3 Construction of dN tree



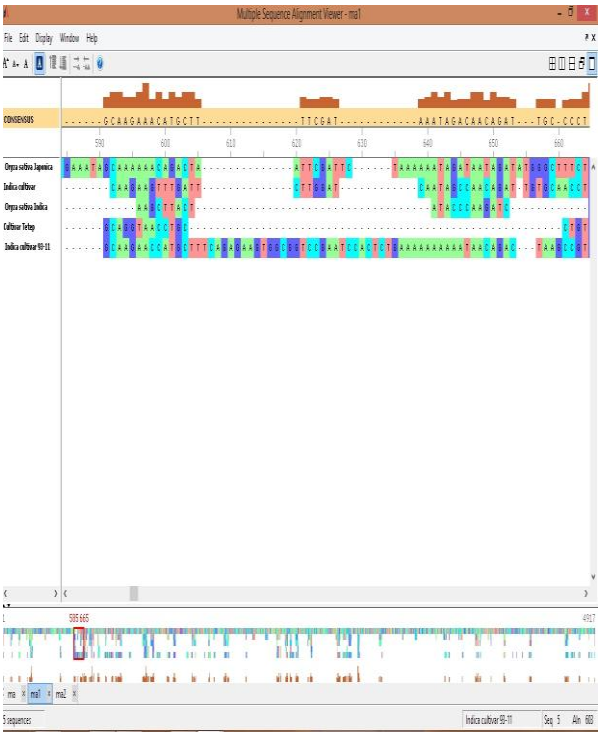

### 3.4 Phylogenetic tree of five rice varieties using UPGMA method



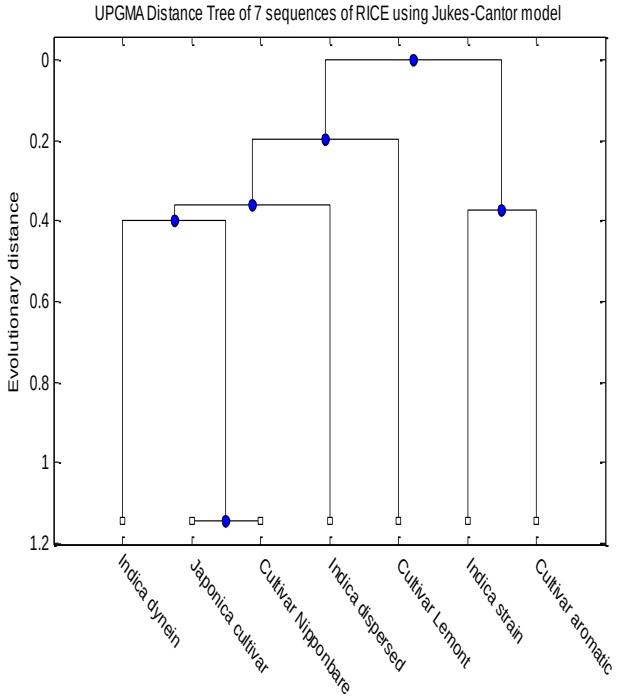

### 3.5 Phylogenetic tree of five rice varieties using NJ method



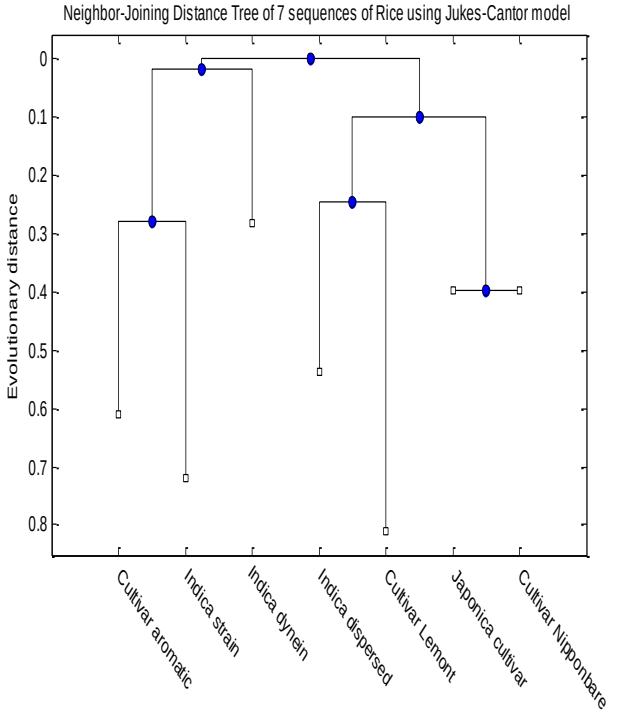

### 3.6 MSA of five rice varieties



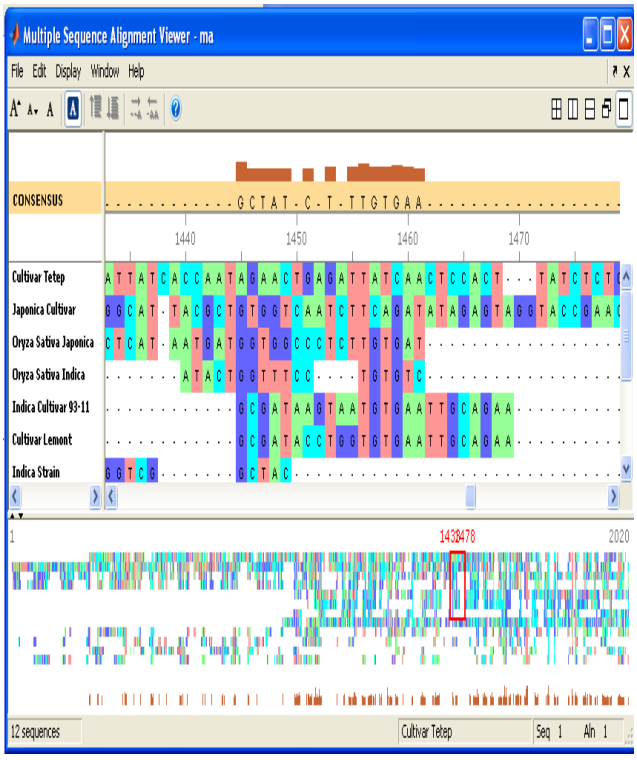
## 3.7 Phylogenetic tree of seven rice varieties Using UPGMA method



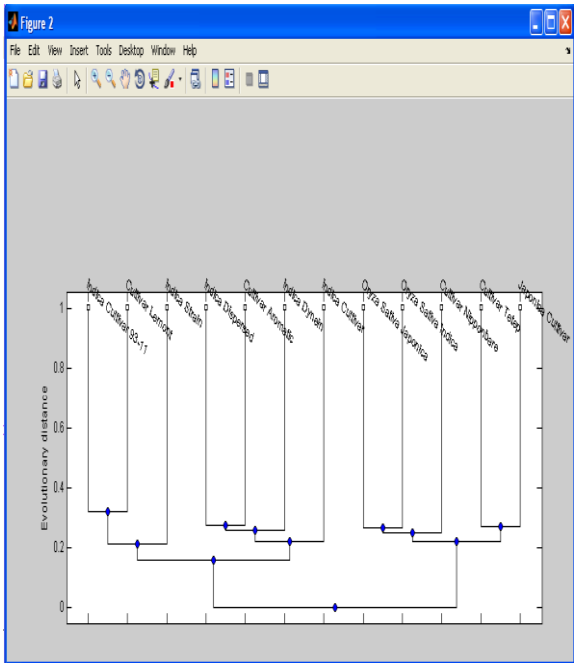

## 3.8 Phylogenetic tree of seven rice varieties Using NJ method



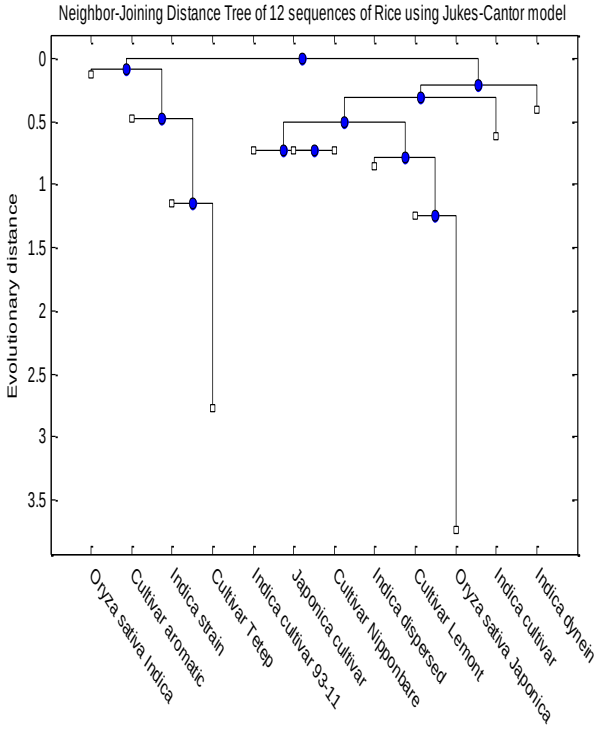

## 3.9 MSA of seven rice varieties



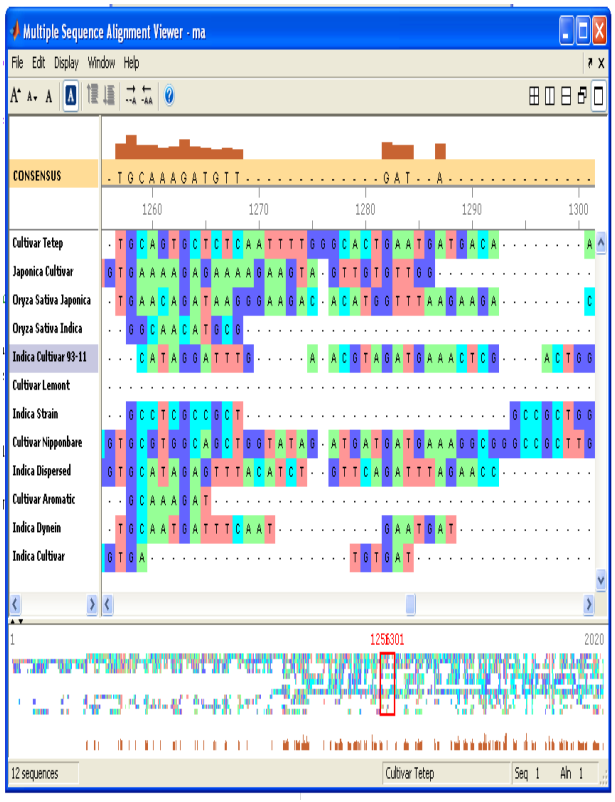

## 3.10 The final UPGMA tree for twelve rice varieties



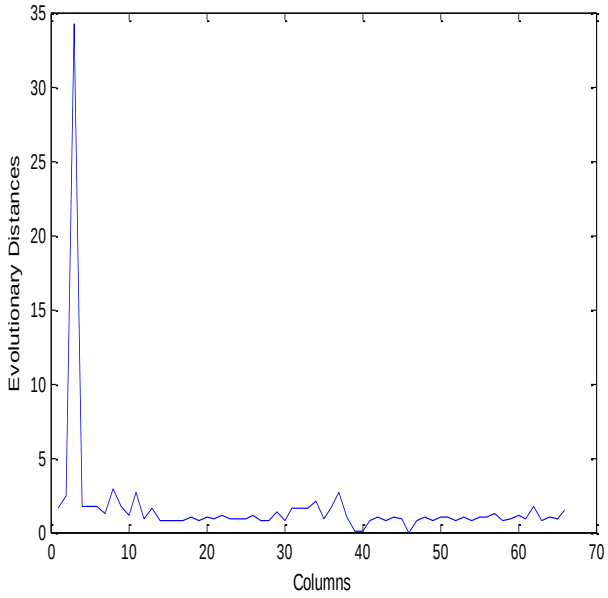
### 3.11 The final Neighbor joining tree for twelve rice varieties.



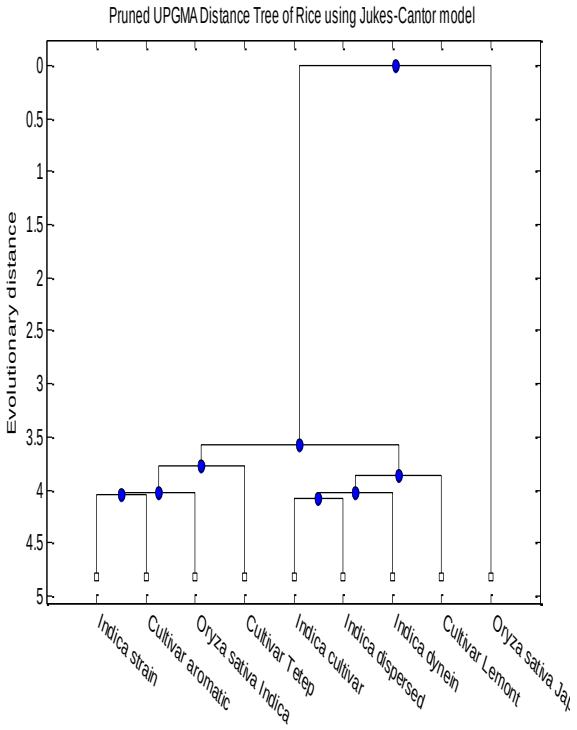

### 3.12 Multiple Sequence alignment for twelve rice varieties



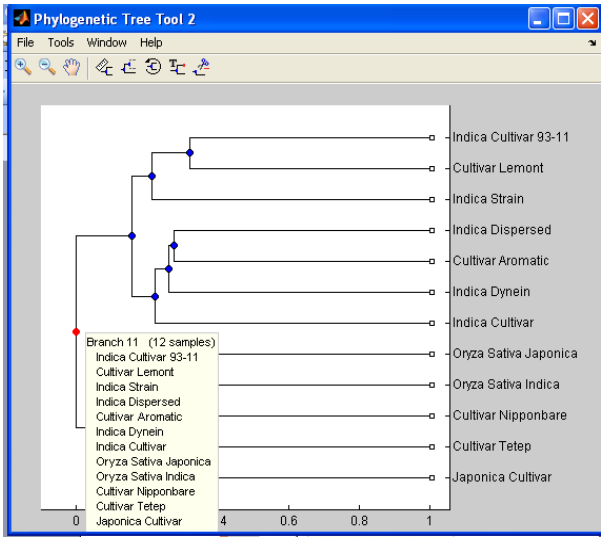

### 3.13 Graph plot for jukes cantor calculated values




### 3.14 Pruned UPGMA distance tree



### 3.15 Final phylogenetic tree for all varieties




## 4. CONCLUSIONS

Phylogenetic tree construction is a complex yet important problem in the field of bioinformatics. Once constructed, a phylogenetic or evolutionary tree can lend insight into the evolution of different species. While in general the topology in phylogenetic trees represents the relationships between the taxa, assigning scales to edges in the trees could provide extra information on the amount of evolution divergence as well as the time of the divergence. The phylogenetic trees with the scaled edges are called phylograms, while the non-scaled phylogenetic trees are called cladograms. The issue is that for a large number of species the problem grows to a computational complexity that is not easily solved. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a clustering method for building phylogenetic trees. Clustering algorithms attempt to repeatedly cluster the data by grouping the closest elements. The result is a rooted tree with original sequences at the leaves whereas the neighbor joining algorithm is essentially a so-called agglomerative hierarchical clustering algorithm: starting from one cluster per sequence, it iteratively merges clusters of sequences (merging the most similar ones first) until a single cluster is obtained.

The most frequently used distance methods are cluster based. The major advantages is that they are computationally fast and are therefore capable of handling databases that are deemed to be too large for any other phylogenetic methods. Model is constructed for aligning DNA sequences of different Rice varieties. There are different sequence formats available from which FASTA format is utilized. Jukes- Cantor method is used for finding the Evolutionary distances. Two phylogenetic trees are constructed using UPGMA for different datasets. Then by using the advanced pruning techniques, trees are combined to obtain the final tree for complete dataset. The closely related sequences are extracted based on threshold condition. Two phylogenetic trees are constructed using Neighbor Joining for different datasets. Trees constructed using UPGMA and neighbor joining are compared. Cluster analysis (Hierarchical clustering) is used as data mining model to retrieve the result. The result of this research work is the tree construction of a given sequence with improved accuracy. The overall advantage of all distance based methods has the ability to make use of large number of substitution models to correct distances and these algorithms are computed on the sequences of different Rice varieties.

## 5. ACKNOWLEDGMENTS

I wish to express my sincere gratitude and indebtedness to my Supervisor, Prof. Rajbir Singh (Assoc. Prof. & Head, Department of Information Technology) for his valuable guidance, obliging nature and attention-grabbing views which led to the successful completion of this study. I lack words to express my cordial thanks to the members of Departmental Research Committee (DRC) for their useful comments and constructive suggestions during all the phases of the present study as well as critically going through the manuscript.

Words fail to express the deep sense of gratitude towards my family members for their moral and financial support and encouragement without which would not have been able to bring out this thesis.

## 6. REFERENCES

[1] Amanda J. Garris,(2005) "*Genetic Structure and Diversity in Oryza sativa L.*", Oxford Journals, pp.1631-1638.

[2] Archak S. and Nagaraju J., (2007) "*Computational Prediction of Rice (Oryza sativa) miRNA Targets*", Genomics Proteomics & Bioinformatics, Vol. 5 No. 3–4, pp. 196-206.

[3] Arthur M., (2002) "*Introduction to bioinformatics*", oxford university press, pp. 25-28

[4] Bergeron, B. (2003) "*Bioinformatics Computing*", Pearson Education, pp.110-160.

[5] David J. HAND, (1998) "*Data Mining: Statistics and More?* ", The American Statistician, Vol. 52, No. 2, pp.112-118.

[6] Gronau I. and Moran S., (2007) "*Optimal Implementations of UPGMA and Other Common Clustering Algorithms*", Information Processing Letters, Volume 104, Issue 6, pp.205-210.

[7] Jacques Cohen (2004) "*Bioinformatics An Introduction for Computer Scientists*", ACM Computing Surveys, Vol. 36, No. 2, pp. 122–158.

[8] Jose C. Clemente et al., (2006) "*Phylogenetic reconstruction from non-genomic data*" Oxford University Press, Vol. 23, pp. e110–e115.

[9] Khalid R. (2012) "*Application of Data Mining in Bioinformatics*", Indian Journal of Computer Science and Engineering, Vol. 1 No 2, pp.114-118.

[10] Mai S. Mabrouk et al. (2006) "*BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using Matlab*", proc. cairo international biomedical engineering conference, pp.1-9.

[11] Nair Achuthsankar S., "Computational *Biology & Bioinformatics: A Gentle Overview*", Communications of the Computer Society of India, January 2007.

[12] Rakshit S. et al., (2007) "Large-scale DNA polymorphism study of Oryza sativa and O.rufipogon reveals the origin an divergence of Asian rice", Springer, pp. 731-743.

[13] Rani S. and Kaur S. (2012) "*Cluster Analysis Method for Multiple Sequence Alignment*", International Journal of Computer Applications, Vol. 43– No.14, pp.19-25

[14] Singh Harmandeep (2013) "*Implementing Hierarchical Clustering method For Multiple Sequence Alignment and Phylogenetic Tree Construction*", International Journal of Computer Science, Engineering and Information Technology, Vol.3, No.1, pp.1-12..

[15] Usama Fayyad et al., (1996) "*From Data Mining to Knowledge Discovery in Databases*", American Association for Artificial Intelligence, Volume 17 Number 3, pp.37-54.

## 7. AUTHOR'S PROFILE

**Rajbir Singh** is an Associate Professor & Head, Department of Information Technology of Lala Lajpat Rai Institute of Engineering & Technology Moga India. He received his B.E (Honor) degree in Computer Science and Engineering from MD University, Rothak, Haryana and M-Tech degree in Computer Science and Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). He has authored 03 books on Computer Science. His main field of research interest is Bio-Informatics and Data mining. He works on the Gene Expression, Phylogenetic Trees and Prediction of Protein Sequence & Structure.

**Sinapreet Kaur** is student of M.Tech Department of Computer Science & Engg., Lala Lajpat Rai institute of engg. & Tech, Moga (sinapreet@gmail.com), Punjab, INDIA. She received her B.Tech degree in Information and Technology from Punjab Technical University, Jalandhar, Pb. (INDIA). Her research interest includes Bio-Informatics. She worked on the phylogenetic tree generation using different scoring methods.

**Dheerajpal Kaur** is a Faculty with the Department of Electronics & Communication Engineering of Lala Lajpat Rai Institute of Engineering & Technology Moga, India. She received her B.E in Electronics & Communication Engineering and M-Tech degree in Electronics & Communication Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). Her research interests include Neural Networks, Genetics Algorithm and Data Mining. She works on the Antenna Propagation using Neural Networks in MAT Lab.