Implementing Phylogenetic Distance Based Methods for Tree Construction Using Hierarchical Clustering

Archi Kataria
University college of Engineering, Punjabi University
Patiala, Punjab, India
Email: archi.kataria@gmail.com

Dr.Amardeep Singh
University college of Engineering, Punjabi University
Patiala, Punjab, India
Email: amardeep_dhiman@yahoo.com

Abstract- Bioinformatics is a data intensive field of research and development. Key problem of knowledge discovery from large and complex databases is deal problem data mining. It is used to discover relationships and patterns in large databases to provide useful information. Clustering is the one of the main techniques for data mining. Phylogeny is the evolutionary history for a set of evolutionary related species. Diagrams that display the phylogeny of a set of taxa in a tree like manner are called phylogenetic trees. One approach on determining the evolutionary histories of a dataset are distance based methods. There are number of different distance based methods of which two are dealt with here: the UPGMA (Unweighted Pair Group Method using Arithmetic average) and Neighbor Joining. These two are clustering based methods. A method for construction of distance based phylogenetic tree using hierarchical clustering is proposed and implemented on different Oryza sativa 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. Extractions of closely related varieties are performed by applying threshold condition. Then, final tree is constructed using these closely related varieties.

Keywords: Bioinformatics, Multiple Sequence Alignment, Data mining, Clustering, Phylogenetic, UPGMA (Unweighted Pair Group Method using Arithmetic average) , Neighbor Joining,.

I. INTRODUCTION

Bioinformatics is the symbolic relationship between computational and biological sciences. The ability to sort and extricate genetic codes from a human genomic database of 3 billion base pairs of DNA in a meaningful way is perhaps the simplest form of Bioinformatics. Moving on to another level, Bioinformatics is useful in mapping different people's genomes and deriving differences in their genetic make-up. The genetic code actually codes for amino acids and thereby proteins and the specific role, played by each of these proteins controls the state of our health. The role or function of each of our genes in coding for a specific protein, which in turn regulates a particular metabolic pathway, is described as "functional genomics". The true benefit of Bioinformatics therefore lies in harnessing information pertaining to these genetic functions in order to understand how human beings and other living systems operate[1].Phylogenetics is the study of evolutionary relationships among groups of organisms. (e.g. species, populations), which are discovered through molecular sequencing data. It represents the evolutionary divergences by finite directed (weighted) graphs, or directed (weighted) trees, known as Phylogeny. A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics[2]. Data mining refers to extracting or "mining" knowledge from large amounts of data. Cluster analysis is one of the techniques of data mining. Clusters of objects are formed so that objects within a cluster have high similarity in comparisons to one another but are very dissimilar to in other clusters. There are two main types of clustering techniques, those that create a hierarchy of clusters called hierarchical clustering and those that do not are non-hierarchical clustering. The hierarchical clustering techniques create a hierarchy of clusters from small too big. The main reason for this is that clustering is an unsupervised learning technique, and as such, there is no absolutely correct answer. For this reason and depending on the particular application of the clustering, fewer or greater number of clusters may be desired. With a hierarchy of clusters defined, it is possible to choose the number of clusters that are desired. Exactly how many clusters should be formed is a matter of interpretation. The advantage of hierarchical clustering method is that they allow the end user to choose from either many clusters or only a few [7].

II. METHODS AND ALGORITHMS

The main distance-based tree-building methods are UPGMA (Unweighted Pair Group Method using Arithmetic average) and Neighbor Joining. Both rely on a different set of assumptions, and their success or failure in retrieving the correct phylogenetic tree depends on how well any particular data set meets such assumptions.

A. UPGMA

UPGMA stands for Unweighted Pair Group Method using Arithmetic average. Given a distance matrix, it starts with grouping two taxa with the smallest distance between them according to the distant matrix. A new node is added in the midpoint of the two, and the two original taxa are put on the tree. The distance from the new node to other nodes will be the arithmetic average. Then obtain a reduced distance matrix by replacing two taxa with one new node. Repeat this process until all taxa are placed on the tree. The last taxon added will be the root of the tree. More precisely, for any two clusters Ci and Cj, it define the distance between the clusters as:

a. Algorithm

Let *d* be the distance function between species, and define the distance *Di,j* between two clusters of species C_i and C_j the following:

$$D_{ij} = \frac{1}{|n_i||n_j|}\sum_{p\in Ci, q\in Cj} d(p,q), \qquad (1)$$

Where $n_i = |C_i|$ and $n_j = |C_j|$ in equation (1) denote the number of sequences in cluster i and j, respectively.

- Initialization:

1. Initialize *n* clusters with the given species, one species per cluster.
2. Set the size of each cluster to 1: $n_i \leftarrow 1$
3. In the output tree *T*, assign a leaf for each species.

- Iteration:

1. Find the *i* and *j* that have the smallest distance D_{ij}.
2. Create a new cluster - (*ij*), which has $n_{(ij)} = n_i + n_j$ members.
3. Connect *i* and *j* on the tree to a new node, which corresponds to the new cluster (*ij*), and give the two branches connecting *i* and *j* to (*ij*) length $\frac{D_{i,j}}{2}$ each.
4. Compute the distance from the new cluster to all other clusters (except for *i* and *j*, which are no longer relevant) as a weighted average of the distances from its components:

$$D_{(ij),k} = \left(\frac{n_i}{n_i+n_j}\right) D_{i,k} + \left(\frac{n_j}{n_i+n_j}\right) D_{j,k} \qquad (2)$$

5. Delete the columns and rows in *D* that correspond to clusters *i* and *j*, and add a column and row for cluster (*ij*), with $D_{(ij),k}$ computed as in equation (2).
6. Return to 1 until there is only one cluster left.

B. Neighbor Joining

Neighbor Joining (NJ) works like UPGMA in that it creates a new distance matrix at each step, and creates the tree based on the matrices. The difference is that NJ does not construct clusters but directly calculates distances to internal nodes. The first step in the NJ algorithm is to create a matrix with the Hamming distance between each node or taxa. The minimal distance is then used to calculate the distance from the two nodes to the node that directly links them. From there, a new matrix is calculated and the new node is substituted for the original nodes that are now joined. The advantage here is that there is not an assumption about the distances between nodes since it is directly calculated. Additivity is a measurement that depends on the distance measure used.Neighbor Joining works even if the lengths are not additive but the tree is no longer guaranteed to be the correct tree.

a. Algorithm

- Initialization:

Same as in UPGMA

- Iteration:

1. For node *i*, its distance u_i from the rest of the tree is estimated using the formula:

$$u_i = \sum_{k\neq i} \frac{D_{i,k}}{(n-2)} \qquad (3)$$

2. Choose the *i* and *j* for which $D_{i,j} - u_i - u_j$ is smallest.
3. Join clusters *i* and *j* to a new cluster - (*ij*), with a corresponding node in *T*. Calculate the branch lengths from *i* and *j* to the new node as:

$$d_{i,(ij)} = \frac{1}{2}D_{i,j} + \frac{1}{2}(u_i - u_j),$$

$$d_{j,(ij)} = \frac{1}{2}D_{i,j} + \frac{1}{2}(u_j - u_i) \qquad (4)$$

4. Compute the distances between the new cluster and each other cluster:

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

5. Delete clusters *i* and *j* from the tables, and replace them by (*ij*).

6. If more than two nodes (clusters) remain, go back to 1. Otherwise, connect the two remaining nodes by a branch of length $D_{i,j}$ [3].

C. Jukes cantor Method

The Jukes and Cantor model is a model which computes probability of substitution from one state (originally the model was for nucleotides, but this can easily be substituted by codons or amino acids) to another. From this model, also derive a formula for computing the distance between 2 sequences[6]. The main idea behind this model 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:

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

Where N_d is the number of mutations (or different nucleotides) between the two sequences and N is the nucleotide length.

III. METHODOLOGY AND DEVELOPMENT OF ALGORITHM

The methodology for the proposed 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 NCBI databank.

A. Steps for Phylogenetic tree Construction

Step 1: Choosing an appropriate marker for the phylogenetic analysis:

In building molecular phylogenetic trees, either DNA, RNA, nucleotide or protein sequence data can be used, but the outcomes from the choice could be quite different. In the research work, DNA is selected as an information marker.

Step 2: Perform sequence alignment:

The second step in the phylogenetic construction involves the alignment of edited sequences. Aligning two sequences is known as pair-wise sequence alignment, while the alignment that includes more than two sequences is known as multiple sequence alignments (MSA). Multiple sequence alignment is done in proposed work.

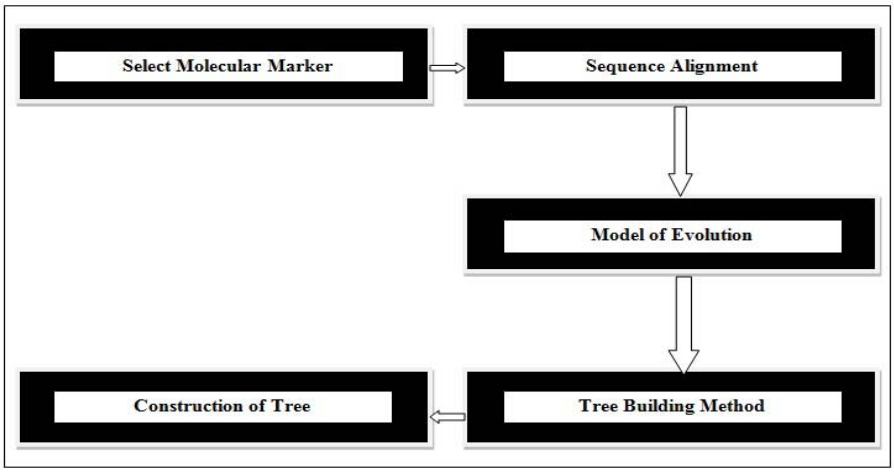

Fig 1: Steps for Phylogenetic Tree Construction

Step 3: Selection of an evolutionary model

For constructing DNA phylogenies, there have been Jukes-Cantor model and Kimural model. From these two, Jukes-cantor is used in present work.

Step 4: Determine a tree building method:

The algorithms of cluster-based include unweighted pair group method using arithmetic average (UPGMA) and neighbor joining (NJ) are taken in present work.

Step 5: Construct the Phylogenetic tree:

After selecting the appropriate methods and steps for tree construction, tree is constructed [4].

B. Development of Algorithm

The work includes the construction of the phylogenetic tree for oryza sativa rice varieties. The sequences for the varieties are loaded from the NCBI database. The phylogenetic distances are calculated based on jukes cantor method and trees are constructed based on UPGMA and Neighbor joining methods for different datasets. The closely related species are selected based on the threshold condition and the sequences which are not satisfying the threshold condition are needed to be pruned. The sequences are also aligned using multiple sequence alignment. The trees are compared and checked if they are similar or not. The steps are shown below:

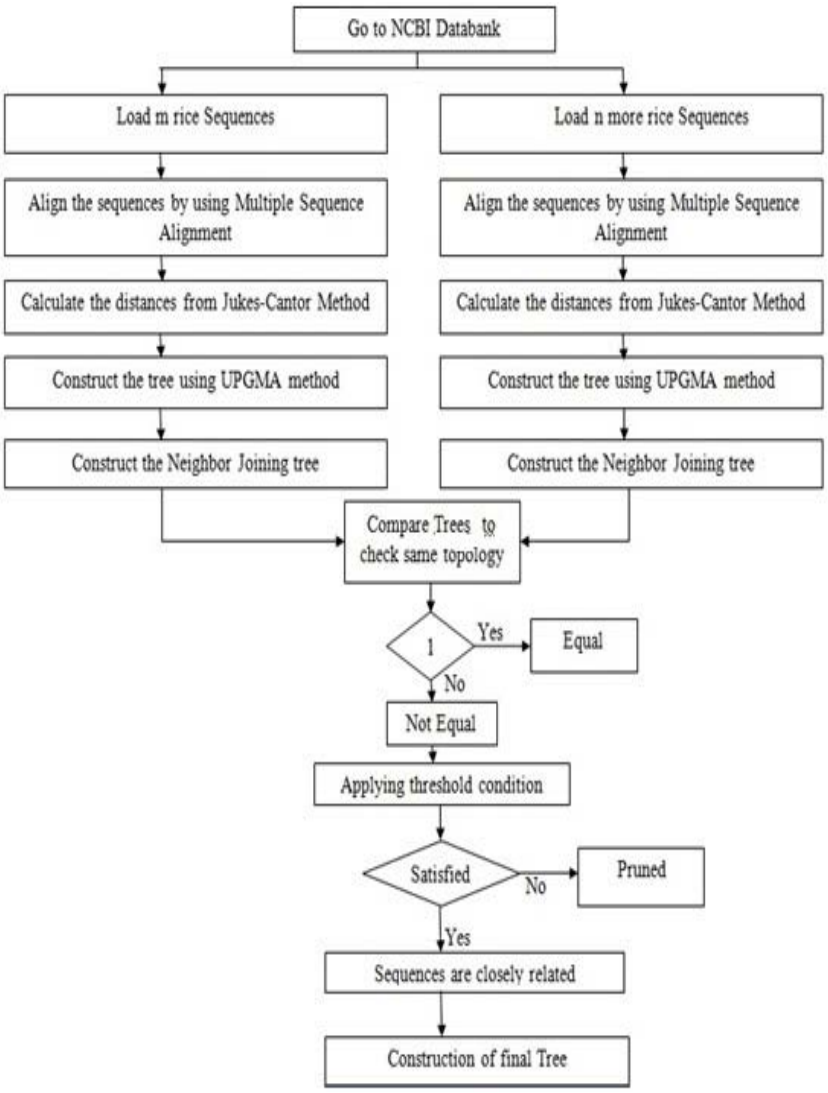

Fig 2: Flowchart for phylogenetic tree construction

C. Synonymous Non-Synonymous Analysis Program

A synonymous substitution (also called a silent substitution) is the evolutionary substitution of one base for another in an exon of a gene coding for a protein, such that the produced amino acid sequence is not modified. Synonymous substitutions and mutations affecting non-coding DNA are collectively known as silent mutations. A non-synonymous substitutions results in a change in amino acid that may be arbitrarily further classified as conservative (change to an amino acid with similar physiochemical properties), semi-conservative (e.g. negative to positively charged amino acid), or radical (vastly different amino acid).

If a gene has lower levels of non-synonymous than synonymous nucleotide substitution, then it can be inferred to be functional because $dN/dS < 1$ is a hallmark of sequences that are being constrained to code for proteins where dS and dN are the Jukes-Cantor correction for multiple hits of the proportion of observed synonymous and the Jukes-Cantor correction for multiple hits of proportion of observed non-synonymous respectively.

SNAP is one of the most important goals pursued by bioinformatics and theoretical chemistry. SNAP calculates synonymous and non-synonymous substitution rates based on a set of codon aligned nucleotide sequences [5].

IV. RESULTS AND DISCUSSIONS

Using SNAP the dS and dN values are calculated which are given as:

ds = 0.9086, dn = 0.1943, ds/dn = 4.3698

On the basis of these values dS and dN trees are constructed.

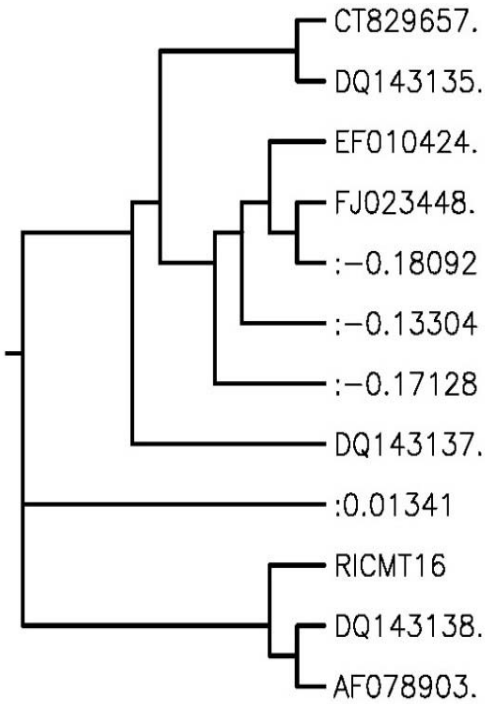

Fig 3: Construction of dS tree

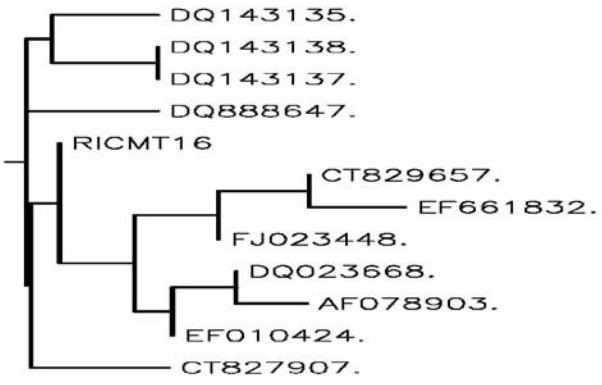

Fig 4: Construction of dN tree

The Oryza Sativa Rice varieties shown in table I are used as input for the present research work. The data is loaded from National Center for Biotechnology Information advances science and health (ncbi.nlm.nih.gov).

Table I: Oryza Sativa Rice Varieties

Sr. No.	Variety Name	Accession Code
1.	Oryza Sativa Japonica	RICMT16
2.	Indica Cultivar	CT827907
3.	Oryza Sativa Indica	CT829657
4.	Cultivar Tetep	DQ023668
5.	Indica Cultivar 93-11	DQ143135
6.	Japonica Cultivar	DQ143138
7.	Cultivar Nipponbare	DQ143137
8.	Indica Dispersed	AF078903
9.	Indica Strain	EF010424
10.	Indica Dynein	DQ888647
11.	Cultivar Lemont	EF661832
12.	Cultivar Aromatic	FJ023448

The five rice varieties are chosen. The evolutionary distance is calculated with the help of Jukes Cantor method. The phylogenetic tree is created using UPGMA and Neighbor joining methods. The trees obtained are shown in figure 5 and 6 respectively.

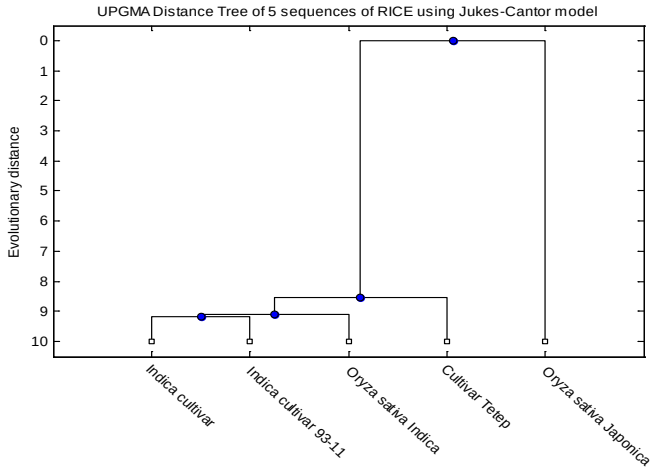

Fig 5: UPGMA tree for five rice sequence

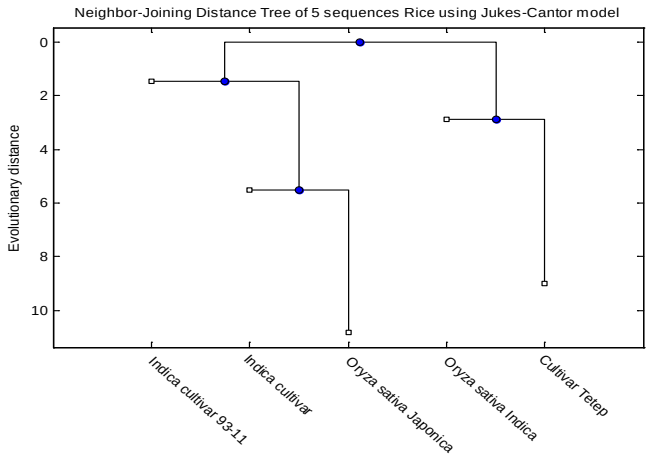

Fig 6: Neighbor joining tree for five rice sequence

Multiple Sequence Alignment is obtained for five varieties which is shown in figure 7.

Fig 7: Multiple Sequence alignment for five rice varieties

Then the different seven varieties are chosen. The UPGMA and Neighbor joining trees for seven varieties are constructed. These trees are shown in figure 8 and 9 respectively.

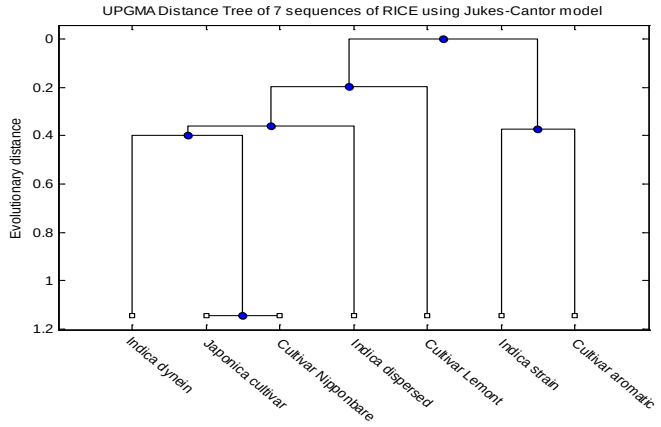

Fig 8: UPGMA tree for seven rice varieties

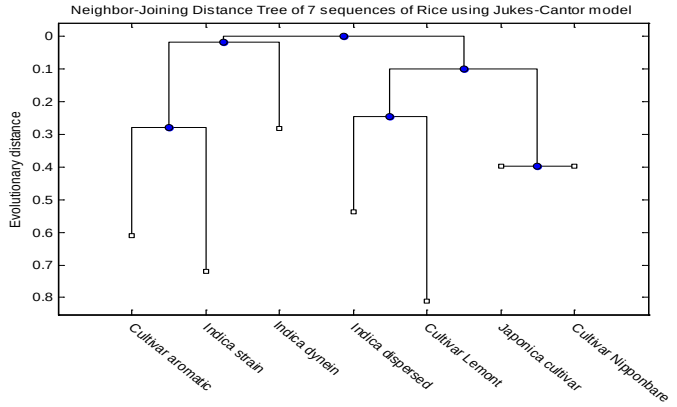

Fig 9: Neighbor joining tree for seven rice varieties

Multiple Sequence Alignment is obtained for seven rice varieties which is shown in figure 10.

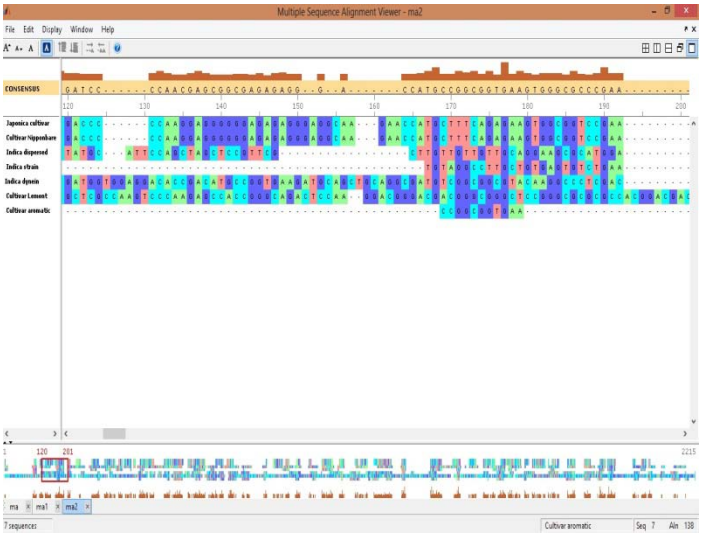

Fig 10: Multiple Sequence alignment for seven rice varieties

The final trees for twelve rice varieties are shown in figure 11, 12 and figure 13 shows the multiple sequence alignment for these varieties.

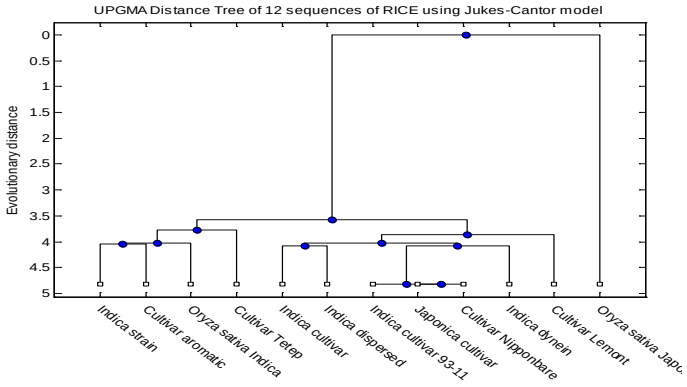

Fig 11: UPGMA tree for twelve rice varieties

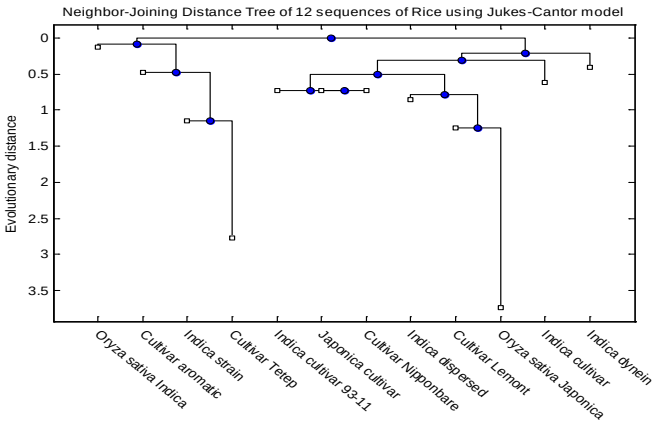

Fig 12: Neighbor joining tree for twelve rice varieties

Fig 13: Multiple Sequence alignment for twelve rice varieties

The values calculated by jukes cantor method are shown column wise as well as in graphical form. Graph plot for Jukes Cantor values is shown in figure 43. The distance values for twelve varieties shown column wise are shown below:

```
Columns 1 through 9

  1.6504    2.4673   34.2415    1.6680    1.6668    1.6668    1.2992    2.9411    1.7341

Columns 10 through 18

  1.0849    2.6219    0.8955    1.6445    0.8276    0.8319    0.8319    0.7506    0.9962

Columns 19 through 27

  0.7540    1.0144    0.9165    1.1381    0.9170    0.9256    0.9256    1.1130    0.7956

Columns 28 through 36

  0.8188    1.4111    0.8068    1.6323    1.6335    1.6335    2.0747    0.9450    1.5626

Columns 37 through 45

  2.6454    1.0528    0.0048    0.0048    0.7717    1.0387    0.7461    0.9588    0.9491

Columns 46 through 54

       0    0.7847    1.0426    0.7451    0.9588    0.9580    0.7847    1.0426    0.7451

Columns 55 through 63

  0.9588    0.9580    1.2693    0.7794    0.8568    1.1218    0.9502    1.6960    0.7695

Columns 64 through 66

  1.0049    0.8741    1.5299
```

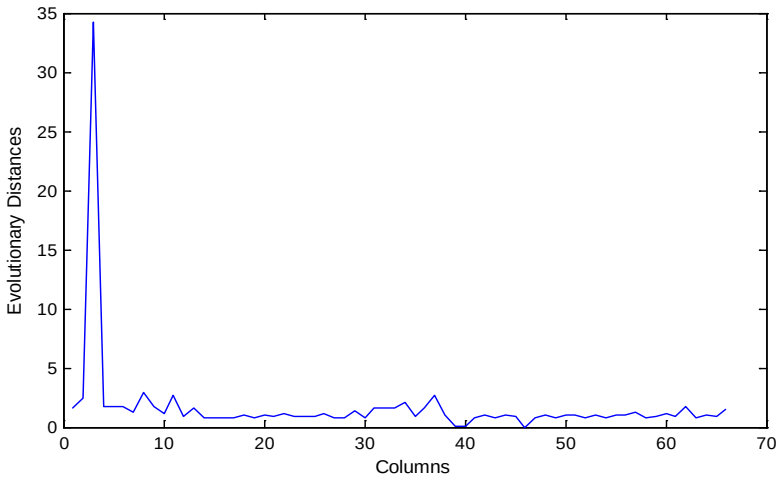

Fig 14: Graph plot for Jukes Cantor values

The threshold condition is applied to obtain the most closely related varieties from the set of twelve varieties. Tree is pruned and constructed using only closely related varieties. The resultant tree of most commonly related varieties are shown in figure 15.

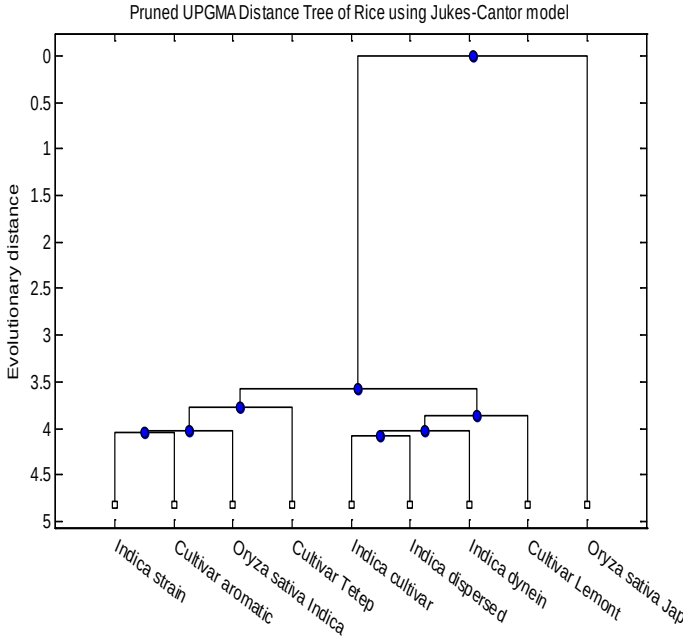

Fig 15: Pruned UPGMA tree for nine closely related varieties

Table II. shows the name of nine closely related rice varieties out of twelve selected.

Table II: Closely Related Rice Varieties

Sr. No.	Variety Name	Accession Code
1.	Oryza Sativa Japonica	RICMT16
2.	Indica Cultivar	CT827907
3.	Oryza Sativa Indica	CT829657
4.	Cultivar Tetep	DQ023668
5.	Indica Dispersed	AF078903
6.	Indica Strain	EF010424
7.	Indica Dynein	DQ888647
8.	Cultivar Lemont	EF661832
9.	Cultivar Aromatic	FJ023448

Phylogenetic tree of all the varieties showing varieties pruned are constructed by phylogenetic tree tool is shown in figure 16.

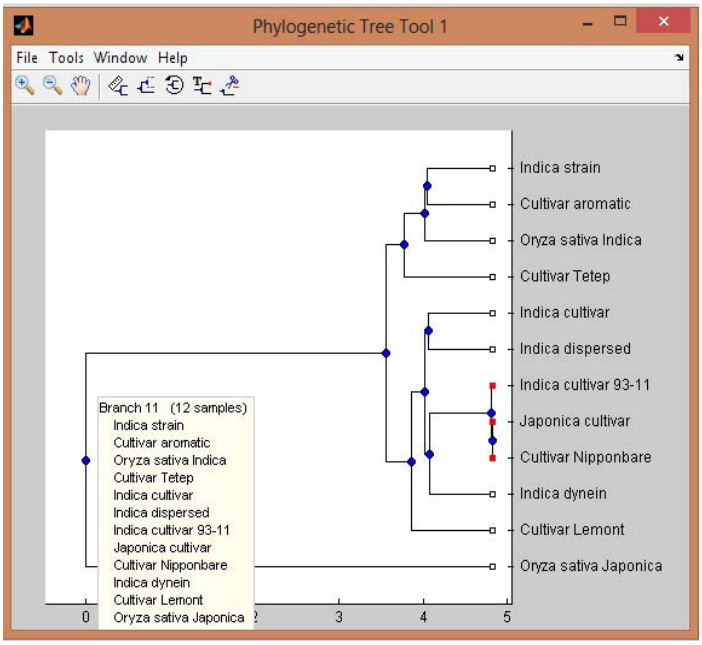

Fig 16: Final Phylogenetic tree

V. CONCLUSION AND FUTURE SCOPE

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.

Future Scope

- ✓ The model can be extended for protein sequence alignment and micro array gene expression analysis.
- ✓ Various other Data Mining techniques can be used to determine an optimum result.
- ✓ The future Work can contain two or more sequences with different length and then comparing their hierarchical results to obtain final conclusion.

REFERENCES

[1] Watson,The Structures of DNA and RNA, http://biology.kenyon.edu/courses/biol63/watson_06.pdf.
[2] Muhammad Sardaraz, Muhammad Tahir1, Ataul Aziz Ikram1, Hassan Bajwa, "Applications and Algorithms for Inference of Huge Phylogenetic Trees: a Review", American Journal of Bioinformatics Research, Vol. 2 No. 1, pp. 21-26, 2012.
[3] http://www.cs.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec08/node21.html
[4] Niranjan Reddy B P. Basics for the Construction of Phylogenetic Trees, WebmedCentral BIOLOGY, pp.1-11, 2011.
[5] http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html
[6] Sita Rani and Simarjeet Kaur, "Cluster Analysis Method for Multiple Sequence Alignment", International Journal of Computer Applications, Volume 43– No.14, pp.19-25, April 2012.
[7] http://www.cbcb.umd.edu/confcour/CMSC858W-materials/Sequence_clustering.pdf.