Volume 5, Issue 6, June 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Improved Computational Methods for Phylogenetic Tree Construction using Cluster Analysis

Shaminder Kaur
Student of M.Tech
Department of CSE
SBBSIET, Jallandhar,
Punjab, India

Prof. Baldeep Singh
Assistant Professor
Department of CSE
SBBSIET, Jallandhar,
Punjab, India

Prof. Tajinder Kaur
Assistant Professor
Department of CSE
SBBSIET, Jallandhar,
Punjab, India

Abstract - Bioinformatics is a branch of biology science and information technology (Computational Technique) in the field of research and development. Phylogeny is the study of evolutionary relatedness amongst organisms based on genetic information codes. The genetic relationships between species can be represented using phylogenetic trees. To construct a phylogenetic tree is a very challenging problem. The main purpose of phylogenetic tree is to determine the structure of unknown sequence and to predict the genetic difference between different species. There are different methods for phylogenetic tree construction from character or distance data. There are different methods to compute distance which include the comparative distance from two sequences using computational methods. A method for construction of distance based phylogenetic tree using hierarchical clustering is proposed and implemented on different data sequences. The sequences are downloaded from NCBI databank. Evolutionary distances are calculated using computational methods. Multiple sequences are applied on different datasets. Trees are constructed for different datasets from available data using both the distance based methods and pruning technique. Computing distance from the available sequences is itself an intricate problem and each method has its own merits and demerits. In the present project work, distance is computed using comparative method (scoring using differences) and using distance based computational methods. Distance data of barley varieties are considered for phylogenetic problem. There are different approaches to construct tree. Computational Methods are used to retrieve the results. The final Phylogenetic trees give the anthromorphical information for the Barley. The results are also shown in Improved Clustered form.

Keywords: Computation Methods, Genetics, Hierarchical clustering, Phylogenetic tree, Sequences.

I. INTRODUCTION

A. Bioinformatics

Bioinformatics is the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied in molecular genetics and genomics [1]. The present role of bioinformatics is to aid biologists in gathering and processing genomic data to study protein function. Another important role is to aid researchers at pharmaceutical companies in making detailed studies of protein structures to facilitate drug design. DNA (Deoxyribonucleic acid) is a genetic material. It is the information contain in DNA that allows the organization of inanimate molecules into functioning, living cells and organisms that are able to regulate their internal chemical composition, growth and reproduction [2]. Genes themselves contain their information as a specific sequence of nucleotides that are found in DNA molecules. Guanine (G), Adenine (A), Thymine (T) and Cytosine (C) are four different bases used in DNA [3]. Each base is attached to phosphate group and a deoxyribose sugar to form a nucleotide. The specific order of its four nucleotides is transcribed by RNA polymerases into mRNA that are then translated by ribosomes into proteins. Twenty different amino acids are used to make proteins, and a specific order and composition of those building blocks play an important role in establishing and maintaining structure and function of enzymes. Some of the research areas in bioinformatics are given below:

- Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome.
- Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript.
- Determining effective protein-DNA, protein-RNA and protein-protein recognition codes.
- Infrastructure and education challenge: Education: development of appropriate bioinformatics curricula for secondary, undergraduate, and graduate education.
- Phylogenetic tree construction emphasizing the evolutionary relationships between different species.

B. Phylogenetic Tree Construction

Phylogenetic tree is typically a graphical representation of the evolutionary relationship among three or more genes or organisms [4]. It is also known as dendrogram, phylogenetic tree are made by arranging nodes and branches [5]. Phylogenetic trees are of two types: Rooted tree and unrooted tree [6].

1) Rooted Tree: In rooted trees a single node is designated as a common ansector, and a unique path leads from it through evolutionary time to any another node.

2) Unrooted Tree: Unrooted trees only specify the relationship between nodes and say nothing about the direction in which evolution occurred.

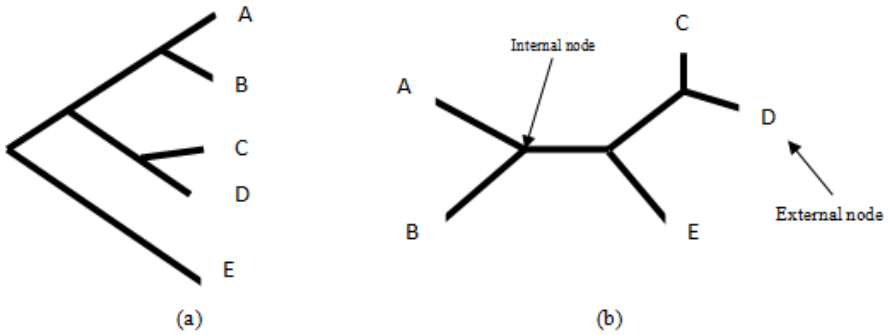

Fig.1 (a) Rooted Tree (b) Unrooted Tree

Traditionally, phylogenetic trees were built from morphological features (e.g., beak shapes, presence of feathers, number of legs, etc). Today, we use mostly molecular data like DNA sequences and protein sequences. Data can be classified into 2 categories:

1) Discrete characters: Each character has a finite number of states.

2) Comparative Numerical Data: These data encode the distances between objects and are usually derived from sequence data.

There are three major methods for constructing phylogenetic trees:

1) Distance methods: Evolutionary distances are computed for all OTUs and these are used to construct trees.

2) Maximum Parsimony: The tree is chosen to minimize the number of changes required to explain the data.

3) Maximum Likelihood: Under a model of sequence evolution, the tree which gives the highest likelihood of the given data is found.

C. Sequence Alignment

Sequence Alignment is a way of arranging the sequences of DNA, RNA or Protein. Alignment is done to identify the regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Homologous genes (genes with shared evolutionary origin) have similar sequences [7]. Multiple sequence alignment is extension of pair wise alignment to incorporate more than two sequences at a time. An MSA consists of at least three biological sequences.

Example:

S_1= GCGCATGGATTGAGCGA

S_2= TGCGCCATTGATGACC

A possible alignment:

S'_1= ATTGA-G

S'_2= ATTGATG

D. Cluster Analysis

Cluster analysis is the technique of data mining. Data mining refers to extracting or "mining" useful knowledge from large amounts of data. It is part of the larger process known as Knowledge Discovery. Clustering or Cluster Analysis is "the process of organizing objects into groups whose members are similar in some way". A Cluster is therefore a collection of objects which are "similar" between them and/or distance between them is small and "dissimilar" to the objects belonging to other clusters and/or distance between them is large [8][9]. Clustering is an unsupervised learning [9]. There are two main types of clustering techniques, those that create a hierarchy of clusters and those that do not.

1) Hierarchical Clustering: The hierarchical clustering techniques create a hierarchy of clusters from small to big. Hierarchical Clustering creates a hierarchy of clusters [8]. It produces a set of nested clusters and can be organized and visualized as a hierarchical tree. The advantage of hierarchical clustering methods is that they allow the end user to choose from either many clusters or only a few. Hierarchical clustering is of two types: Agglomerative & Divisive. In Agglomerative clustering techniques start with, as many clusters as there are records where each clusters containing just one record. It is also known as Bottom up Clustering Technique.

Divisive Clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into small pieces and then further into smaller pieces. It is also known as top down clustering technique.

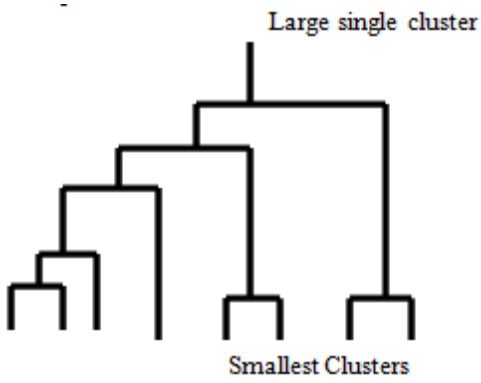

Fig.2 Hierarchical Clustering

*2) Non Hierarchical Clustering***:** Non hierarchical clustering does not create hierarchy of clusters or hierarchical Tree. Its techniques are very fast to compute on the database relative to hierarchical clustering.

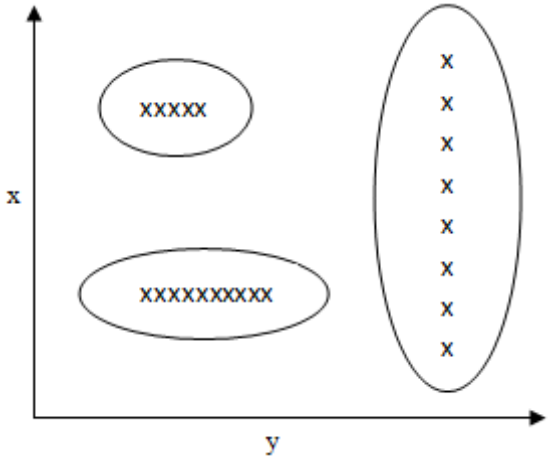

Fig.3 Non Hierarchical Clustering

II. METHODOLOGY

To construct a phylogenetic tree based on data from protein or nucleotide sequences comparisons, first do a multiple alignments for the sequences and then calculate distance measure d_{ij} between all taxa. Phylogenetic trees among a nontrivial number of input sequences are constructed using computational phylogenetic methods. Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model.

For this, we need firstly, to determine and compute a distance metric between every genomic sequence, secondly, to perform hierarchical clustering on the given data sets, utilizing the distance metric computations, finally, to visualizing the resulting phylogenetic tree. The method of Phylogenetic Tree Construction depends on assigning a set of Training Data values to a given sequence and then applying a UPGMA and neighbor-joining algorithm to those numbers.

A. Structure of Data Mining Model

Data mining model deals with the extraction of knowledge from the data. Knowledge means relationship and patterns between the data elements. The database is an essential component of this model. The growth and complexity of data in bioinformatics field emphasis the need of the system, which can help in analyzing this data. Main aim of this model is to manage, organize, access and analyze the biological data.

B. Data Mining Model

This model makes use of Hierarchical Clustering as the data mining method and uses conceptual clustering as the type of clustering. Clustering can be considered the most important unsupervised learning problem. The choice of clustering methods is just as critical as is the choice of an association measure. Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class. For Hierarchical clustering, we used the unweighted pair group method with arithmetic mean (UPGMA) algorithm. This algorithm takes a lower triangular n * n matrix whose entries are distance measurements between the entity defining the row and the entity defining the column. These entities are termed clusters, regardless of if they have one element or many elements (as arises during the algorithm). UPGMA first creates a working copy of the distance measurements matrix, and leaves the original for reference. Then while there is

still more than one cluster in the working copy, the lowest non-zero entry is located. The cluster that defines the row and the cluster that defines the column of that entry are grouped together to form a new cluster. This new cluster is added as a row and a column to the working distance matrix.

Cluster analysis is a class of statistical techniques that can be applied to Training data that exhibit "natural" groupings. Cluster analysis sorts through the raw data and groups them into clusters.

C. Computation Methods

All the biological sequences data were collected from the National Center for Biotechnology Information (NBCI), available from *www.ncbi.nlm.nih.gov.* Its genome databases are organized into six major organism groups: archaeam, bacteria, eukaryote, viruses, viroids and plasmids. It provides access to the complete genomes of over 3,200 organisms, and provides downloads of DNA sequences in FASTA format. This distance computes the number of insertions, deletions, or substitutions of single characters between two different sequences. The algorithm uses the dynamic programming approach to compute the minimal amount of operations needed.

1) Jukes Cantor Method: When substitutions are common, there were no guarantees that a particular site had not undergone multiple changes. Fig. 4 shows two possible scenarios that multiple substitutions at a single site would lead to underestimation of the number of substitutions that had occurred if a simple count were performed. To address the possibility, Jukes and Cantor assumed that each nucleotide was just as likely to change into any other nucleotide [J]. Using that assumption they created a mathematical model shown in Fig.5 in which Jukes and Cantor Assumed that all nucleotides changed to each of the three alternative nucleotides at the same rate α.

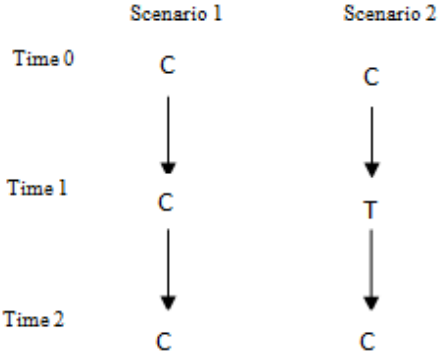

Fig.4 Two Possible Scenarios

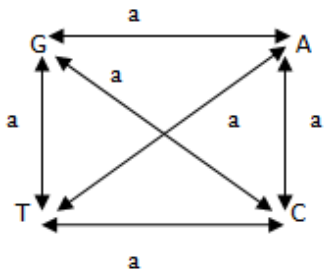

Fig.5 Three Alternative Nucleotide

From this model, the overall rate of substitution for any given nucleotide was 3α.If a site within a gene was occupied by a C at time 0, then the probability (*P*) of that site would still be the same nucleotide at time 1would be

$$PC\,(1) = 1\text{-}3\alpha \qquad (1)$$

Because a reversion to C could occur if the original C changed to another nucleotide in that first time span, at time 2 the probability would be:

$$PC\,(2) = (1\text{-}3\alpha)\,PC\,(1) + \alpha\,[1\text{-}\,PC\,(1)] \qquad (2)$$

At any given time (*t)* in the future, the probability of that site would contain a C was defined by the following equation:

$$p_{c(t)} = \frac{1}{4} + \left(\frac{3}{4}\right) e^{-4\alpha t} \qquad (3)$$

However, this model oversimplified the substitution pattern. We can derive the equation that yields an estimate of the true number of substitutions that have occurred between two sequences when only a pairwise counting of differences is available:

$$K = -\frac{3}{4}\ln\left[1 - \left(\frac{4}{3}\right)(p)\right] \qquad (4)$$

Where K is the actual number of substitutions per site, and p is the fraction of nucleotides that simple count reveals to be different two sequences.

2) UPGMA Method (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean) is a simple agglomerative or hierarchical clustering method used in bioinformatics for the creation of phonetic trees (phonograms). The algorithm examines the structure present in a pairwise distance matrix (or a similarity matrix) to then construct a rooted tree (dendrogram)[DENI]. At each step, the nearest two clusters are combined into a higher-level cluster. Let *d* be the distance function between species, we define the distance *Di, j* between two clusters of species C_i and C_j the following:

Where $n_i = |C_i|$ and $n_j = |C_j|$

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 $\mathbf{D_{ij}}$ **/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:

$$D_{(i,j),k} = \left(\frac{n_i}{n_i+n_j}\right) D_{i,k} + \left(\frac{n_j}{n_i+n_j}\right) D_{j,k} \qquad (5)$$

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 above.
6. Return to 1 until there is only one cluster left.

Complexity: The time and space complexity of UPGMA is $O(n^2)$, since there are *n*-1 iterations, with $O(n)$ work in each one.

3) Neighbour Joining Method: In bioinformatics, Neighbor Joining is a bottom-up clustering method for the creation of phenetic trees (phenograms), created by Naruya Saitou and Masatoshi Nei. Usually used for trees based on DNA or protein sequence data, the algorithm requires knowledge of the distance between each pair of taxa (e.g., species or sequences) to form the tree. Neighbor joining takes as input a distance matrix specifying the distance between each pair of taxa. The algorithm starts with a completely unresolved tree, whose topology corresponds to that of a star network, and iterates over the following steps until the tree is completely resolved and all branch lengths are known.

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. For each species, compute.

$$u_i = \sum_{k \neq i} D_{\frac{i,j}{n-2}} \qquad (6)$$

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:

$$d_{i,(ij)} = \frac{1}{2} D_{i,j} + \frac{1}{2}(u_i - u_j) \qquad (7)$$

$$d_{j,(i,j)} = \frac{1}{2} D_{i,j} + \frac{1}{2}(u_j - u_i) \qquad (8)$$

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

$$D_{(ij),k} = \frac{D_{i,k} + D_{j,k} - D_{i,j}}{2} \qquad (9)$$

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 $D_{i,j}$.

III. DATA SETS

A. The forms of data sets

In order to successfully submit a job, it is important to understand what the various sequence formats are used for describing biological sequences and what their basic structure is. There are various formats of Nucleotide & Amino acid sequences, and each has its own set of characters and utility. To get a deeper understanding and better results it is essential to choose a valid input format. The various formats are: Plain text format, FASTA format, Genetic Computer Group Format (GCG), PHYLIP, NEXUS and NBRF & PIR. Plain text was chosen for the present problem. The plain text format is comparatively simple than the other formats available. The plain text sequence format is typically generated by word processors (saved as text file with line breaks). Twelve Barley varieties are taken as data sets for construction of phylogenetic tree. The data is downloaded from National Center for Biotechnology Information advances science and health (ncbi.nlm.nih.gov) with their accession code. Barley varieties are shown in the Table I.

Table I Barley varieties

Sr. No.	Variety Name	Accession Code
1.	Hordeum vulgare subsp(vulgare Morex barley)	AF509748
2.	Hordeum vulgare subsp (vulgare cultivar Oregon Wolfe Barley)	EU007829
3.	Hordeum vulgare subsp (vulgare cultivar Oregon Wolfe Barley)	AY750996
4.	Hordeum vulgare subsp (vulgare Peatland barley)	AF509756
5.	Hordeum vulgare subsp (vulgare Kindred barley)	AF509755
6.	Hordeum vulgare subsp (vulgare Leger barley)	AF509754
7.	Hordeum vulgare subsp (vulgare Chevron barley)	AF509753
8.	Hordeum vulgare subsp (vulgare Bowman barley)	AF509752
9.	Hordeum vulgare subsp (vulgare 80-TT barley)	AF509751
10.	Hordeum vulgare subsp (vulgare Q21861 barley)	AF509750
11.	Hordeum vulgare subsp (vulgare cultivar Vogels Gold note six-rowed barley)	EU886294
12.	Hordeum vulgare subsp (vulgare cultivar Ragusa note six-rowed barley)	EU886293

IV. RESULTS AND DISCUSSION

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 data sets, utilizing the distance metric computations, Finally, to visualizing the resulting phylogenetic tree. The idea behind the UPGMA algorithm is based on the number of variables that are similar between samples. By changing the similarity criterion in a stepwise, a hierarchical group structure develops, and can be displayed by a dendrogram.

The five barley varieties are chosen (shown in table I). The evolutionary distance is calculated with the help of Jukes Cantor Method.

The Jukes Cantor values for these varieties are shown below.

Columns 1 through 15

2.2538 2.5324 25.2015 2.4231 2.3426 2.2317 3.5423 4.3214 25.4321 1.2342 2.3421 0.7865 1.0567 0.0169 0

Columns 16 through 30

0.0302 0.8173 0.8855 1.6676 1.0540 0.0164 1.0039 0.7750 0.7881 0.7570 0.5893 0.8210 1.3579 1.3532 0.7690

Columns 31 through 45

1.0652 1.0897 1.0690 0.8495 0.8649 0.8760 2.1564 1.0322 0.0163 0.0422 0.8209 0.8863 1.6587 1.0786 0.0031

Columns 46 through 60

0.0432 0.9484 0.9966 1.5686 1.0680 0.0151 0.9082 0.9737 1.7875 1.0985 0.0500 0.8219 1.1671 1.6803 0.9227

Columns 61 through 66

0.8672 1.8765 0.9973 3.2340 1.7899 1.0834

The threshold condition is applied to obtain the most closely related varieties from the set of twelve varieties. The most commonly related varieties are:

- Hordeum vulgare subsp(vulgare Morex barley)
- Hordeum vulgare subsp (vulgare cultivar Oregon Wolfe Barley)
- Hordeum vulgare subsp (vulgare cultivar Oregon Wolfe Barley)
- Hordeum vulgare subsp (vulgare Peatland barley)
- Hordeum vulgare subsp (vulgare Kindred barley)

The phylogenetic tree of these five related varieties is created using neighbor joining technique. The tree obtained is shown in Fig 6 and multiple sequence alignment of these varieties are shown in Fig 7.

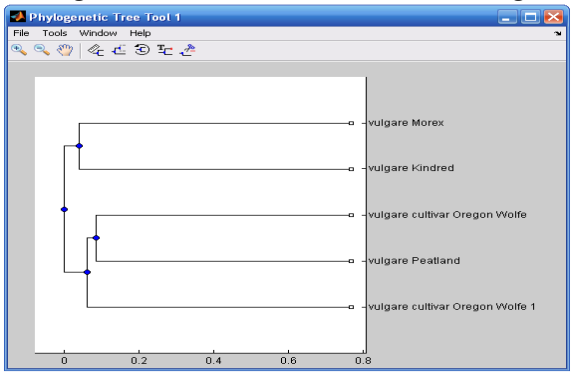

Fig.6 Phylogenetic tree of five closely related varieties

Fig. 7 Multiple sequence alignment of five closely related varieties

Then the different seven varieties are chosen. The tree is constructed by the same method and is combined with the first tree by using tree pruning techniques. The final tree for twelve barley varieties is shown in Fig. 8.

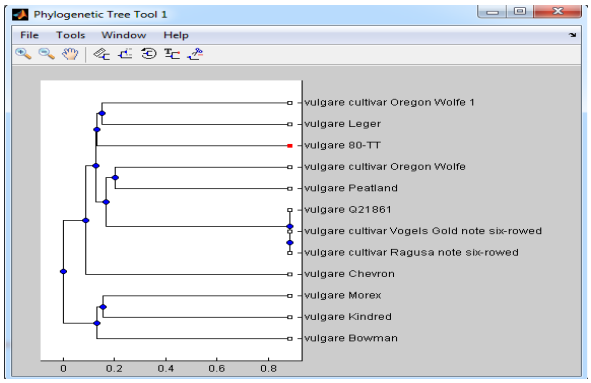

Fig.8 Phylogenetic tree of twelve barley varieties

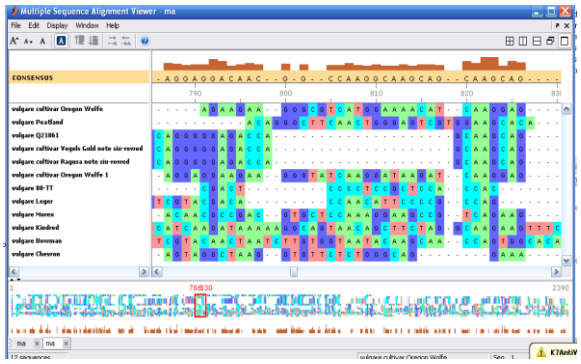

Fig. 9 Multiple Sequence alignment of twelve barley varieties

IV. CONCLUSION

Phylogenetic Tree Construction can be useful in analyzing the distances between more than two sequences and Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple species sequence alignments, are simplest to implement. The algorithm takes as input the lower triangle of a symmetric distance matrix and constructs a rooted tree in the form of an UPCluster object. The algorithm should construct an unrooted tree, but arbitrarily adds a root node with appropriate edges to the last two active leaves in the construction. The present work applies hierarchical clustering on the given data sets, utilizing the distance metric computations and to visualizing the resulting phylogenetic tree. The Data Mining Model for Phylogenetic Tree Construction is useful for determining and comparing the Branch level of structural similarity between different species sequences. The present model computes the distance of the given species sequences. By applying data mining to the input data an optimized result is produced i.e. the tree structure predicted is of high conformation and has less variance. Cluster analysis 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 a large number of substitution models to correct distances and these algorithms are computed within the polynomial time.

ACKNOWLEDGEMENT

I wish to express his sincere gratitude and indebtedness to my Supervisor, Prof. Baldeep Singh (Assistant Prof. Department of Computer Science) for his valuable guidance, attention-grabbing views and obliging nature which led to the successful completion of this study. I lack words to express my cordial thanks to the members of Research and Development Committee (R&DC) 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 the author to express her deep sense of gratitude towards my family members for their moral and financial support and encouragement without which the author would not have been able to bring out this.

REFERENCES

[1] Mohammed J. Zaki, Jason T. L. Wang, Hannu T.T. Toivonen, *BIOKDD01: Workshop on Data Mining in Bioinformatics* 7th ACIM SIGKDD International Conference on Knowledge Discovery & Data Mining volume 3 Issue 2- page 71, 26 August,2001.

[2] Nabeel Ahmad, AkhilesT. P. Speed, K. J. Kechris, and S. K. McWeeney, *Molecular Evolution, Substitution Models, and Phylogenies*, No. 1(1997) pp. 1-12**,** posted to todo_mendeley by casallas45 on 2010-10-23.

[3] Bind and Sanjiv Kumar Maheshwari, *Bioinformatics New Era: Introduction and Overview*, Journal of Computational Intelligence in Bioinformatics ISSN: 0973-385X Volume 4 Number 1 pp. 7–17, 2011.

[4] Robert Jones, *Introduction to Bioinformatics*, Published on MacDevCenter 6 Nov. 2011

[5] C.Peng, *Distance Based Methods in Phylogenetic Tree Construction*, Department of Mathematics, Morehouse College, Atlanta, GA 30314, 2007.

[6] Deni Khanafiah, Hokky Situngkir, *Innovation as Evolutionary Process,* Dept. of Computational sociology, Journal of Social Complexity Vol.2, No.2, pp.20- 30, 2006.
[7] Mark Craven, *Distance-Based Approaches to Inferring Phylogenetic Trees*, BMI/CS 576 Fall 2011.
[8] Andreas D. Baxevanis, B.F. Francis Ouellette, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition*, Copyright _ 2001 John Wiley & Sons, Inc. ISBNs: 0-471-38390-2 (Hardback); 0-471-38391-0 (Paper); 0-471-22392-1 (Electronic).
[9] Jianfu LI, Jianshuang LI and Huaiqing HE , *A Simple and Accurate Approach to Hierarchical Clustering*, Journal of Computational Information Systems 7: 7 2577-2584, Binary Information Press, pp. 2577 – 2584, July, 2011
[10] Lior Rokach, Oded Maimon, *Clustering Methods*, Department of Industrial Engineering, Tel-Aviv University.
[11] J.Felsenstei, *Taking variation of evolutionary rates between sites into account in inferring phylogenies, Department* of Genetics, University of Washington, Box 357360, Seattle, Washington.
[12] Kirsi Louhisuo, Constructing *phylogenetic trees with UPGMA and Fitch-Margoliash* on May 4, 2004, pp. 2 – 11.

