Implementing UPGMA and NJ Method For Phylogenetic Tree Construction Using Hierarchical Clustering

[1]Sukhpreet Kaur, [2]Harwinder Singh Sohal, [3]Rajbir Singh Cheema
[1]Dept. of CSE, LLRIET Moga, Punjab, India
[2,3]Dept. of IT, LLRIET Moga, Punjab, India

Abstract
The research in bioinformatics has accumulated large amount of data. As the hardware technology advancing, the cost of storing is decreasing. 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. To construct a phyloenetic 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, distance using UPGMA and Neighbour Joining. 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 UPGMA. Distance data for human phylogenetic problem is considered for the present work. There are different approaches to construct tree. UPGMA and Neighbour Joining Methods are used to retrieve the results. The final trees give the anthromorphical information for the human being. The results are also shown in Hierarchical clustering form.

Keywords
Phylogenetics, Protein, Cladistic

I. Introduction

Phylogenetic is the study of evolutionary relationships. Phylogenetic analysis is the means of inferring or estimating these relationships. The evolutionary history inferred from phylogenetic analysis is usually depicted as branching, treelike diagrams that represent an estimated pedigree of the inherited relationships among molecules (‘‘gene trees’’), organisms, or both. Phylogenetic is sometimes called cladistics because the word ‘‘clade,’’ a set of descendants from a single ancestor, is derived from the Greek word for branch. However, cladistics is a particular method of hypothesizing about evolutionary relationships.

The basic tenet behind cladistics is that members of a group or clade share a common evolutionary history and are more related to each other than to members of another group. A given group is recognized by sharing unique features that were not present in distant ancestors. These shared, derived characteristics can be anything that can be observed and described from two organisms having developed a spine to two sequences having developed a mutation at a certain base pair of a gene. Usually, cladistic analysis is performed by comparing multiple characteristics or ‘‘characters’’ at once, either multiple phenotypic characters or multiple base pairs or amino acids in a sequence.

- There are three basic assumptions in cladistics: Any group of organisms is related by descent from a common ancestor (fundamental tenet of evolutionary theory).
- There is a bifurcating pattern of cladogenesis. This assumption is controversial.
- Change in characteristics occurs in lineages over time. This is a necessary condition for cladistics to work.

The resulting relationships from cladistic analysis are most commonly represented by a phylogenetic tree:

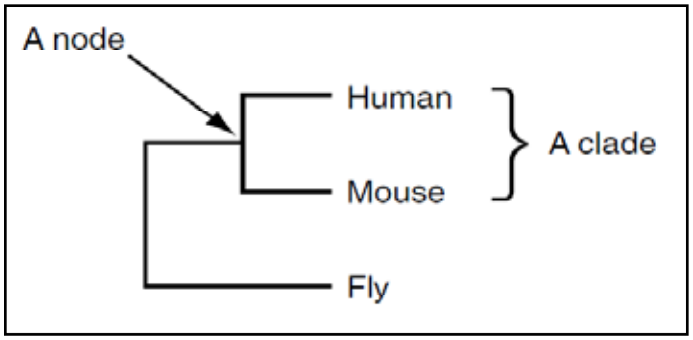

Fig. 1: Relationships of Clade and Node

Even with this simple tree, a number of terms that are used frequently in phylogenetic analysis can be introduced:

- A clade is a monophyletic taxon. Clades are groups of organisms or genes that include the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. Clade is derived from the Greek word ‘‘klados,’’ meaning branch or twig.
- A taxon is any named group of organisms but not necessarily a clade.
- In some analyses, branch lengths correspond to divergence (e.g., in the above example, mouse is slightly more related to fly than human is to fly).
- A node is a bifurcating branch point.

A. Neighbor Joining Method (NJ)

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:

- Based on the current distance matrix calculate the matrix Q (defined below).
- Find the pair of taxa for which Q(i,j) has its lowest value. Add a new node to the tree, joining these taxa to the rest of the tree. In the figure at right, f and g are joined to the tree by the new node u.
- Calculate the distance from each of the taxa in the pair to this new node.

- Calculate the distance from each of the taxa outside of this pair to the new node.
- Start the algorithm again, replacing the pair of joined neighbors with the new node and using the distances calculated in the previous step.

B. UPGMA Method

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative or hierarchical clustering method used in bioinformatics for the creation of phonetic trees (phonograms). UPGMA assumes a constant rate of evolution (molecular clock hypothesis), and is not a well-regarded method for inferring relationships unless this assumption has been tested and justified for the data set being used. UPGMA was initially designed for use in protein electrophoresis studies, but is currently most often used to produce guide trees for more sophisticated phylogenetic reconstruction algorithms. The algorithm examines the structure present in a pair wise distance matrix (or a similarity matrix) to then construct a rooted tree (dendrogram). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B is taken to be the average of all distances between pairs of objects "x" in A and "y" in B, that is, the mean distance between elements of each clusters.

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y) \quad (1)$$

This method follows a clustering procedure:

- Assume that initially each species is a cluster on its own.
- Join closest 2 clusters and recalculates distance of the joint pair by taking the average.
- Repeat this process until all species are connected in a single cluster.

This algorithm is phenetic, which does not aim to reflect evolutionary descent. It assigns equal weight on the distance and assumes a randomized molecular clock. WPGMA is a similar algorithm but assigns different weight on the distances.

II. Related Work

Gabriel Robins and Tongtong Zhang [1] derived a phylogenetic tree reconstruction method that detects and reports multiple, topologically distant, low cost solutions. Our method is a generalization of the Neighbor-Joining (NJ) method of Nei and Saitou, and affords a more thorough sampling of the solution space by keeping track of multiple partial solutions during its execution. The scope of the solution space sampling is controlled by a pair of user-specified parameters the total number of alternate solutions and the number of alternate solutions that are randomly selected effecting a smooth tradeoff between run time and solution quality and diversity. This method can discover topologically distinct low cost solutions. In tests on biological and synthetic datasets using the least-squares distance or minimum-evolution criterion, the method consistently performed as well as, or better than, either the Neighbor-Joining heuristic or the PHYLIP implementation of the Fitch-Margoliash distance measure. In addition, the method identified alternative tree topologies with costs within 1 or 2% of the best, but with topological distances 9 or more partitions from the best solution (16 taxa); with 32 taxa, topologies were obtained 17 (least-squares) 22 (minimum-evolution) partitions from the best topology when 200 partial solutions were retained. Thus, the method can find lower cost tree topologies and near-best tree topologies that are significantly different from the best topology.

Erich Bohl, et al. [2] derived mathematical model for the analysis of phylogenetic trees is applied to comparative data for 48 species. The model represents a return to fundamentals and makes no hypothesis with respect to the reversibility of the process. The species have been analysed in all subsets of three, and a measure of reliability of the results is provided. The numerical results of the computations on 17,296 triple of species are made available on the internet. These results are discussed and the development of reliable tree structures for several species is illustrated. It is shown that, indeed, the Markov model is capable of considerably more interesting predictions than has been recognized to date.

Deni Khanafiah and Hokky Situngkir [3] The Innovation of Artifacts is somewhat can be seen as a process of evolution. The paper presents an endeavor to view the evolution of artifact by using evolutionary concept of memetics. We showed the ways to build a phlyomemetic tree based on memes constituting an artifact to infer or estimate the evolutionary history and relationship between artifacts. UPGMA algorithm and the Shortest Tree Method using Minimum Spanning Tree (MST) techniques are presented to construct the Phylomemetic tree of innovation. To show an implementation, we use innovation of cell phone as an example.

Chuang Peng, et al. [4] analysis the evolutionary histories of living organisms are represented by finite directed (weighted) graphs, in particular, directed (weighted) trees. There are basically two types of phylogenetic methods, distance based methods and character based methods. Distance based methods include two clustering based algorithms, UPGMA, NJ, and two optimality based algorithms, Fitch- Margoliash and minimum evolution. This paper focuses on distance based methods. The paper starts with some preliminary knowledge and definitions in the area, including finite directed graphs, directed trees and matrices. It discusses the verification of the metric property of distance matrices, including detections of errors if a distance matrix fails to satisfy the metric property, and then provides an algorithmic modifying the distance matrix to satisfy the metric property. The second part of the paper is a brief survey based on the excerpts from the references, on various frequently used distances based phylogenetic tree construction methods, both cluster-based and optimality base methods, including UPGMA, Neighbor Joining, and Fitch-Margoliash, and Minimum Evolution methods. Also, it discusses the assessment of the phylogenetic trees and some analysis of the algorithms

Martin Simonsen, et al. [5] derived the neighbor joining method reconstructs phylogenies by iteratively joining pairs of nodes until a single node remains. The criterion for which pair of nodes to merge is based on both the distance between the pair and the average distance to the rest of the nodes. In this paper, we present a new search strategy for the optimization criteria used for selecting the next pair to merge and we show empirically that the new search strategy is superior to other state-of-the-art neighbor joining implementations.

Naznin, F. et al. [6] described that In order to design life saving drugs, such as cancer drugs, the design of protein or DNA structures have to be accurate. These structures depend on multiple sequence alignment (MSA). MSA is a combinatorial optimization problem which is used to find the accurate structure of protein and DNA sequences from the existing sequences. In this paper, we have proposed a new iterative progressive alignment method, for multiple sequence alignment, which is a close variant of the MUSCEL algorithm. MUSCEL starts with the distance table. The

other steps of this algorithm include: generating a guide tree using UPGMA, multiple sequence alignments, l distance calculation from aligned sequences and new techniques to improve multiple sequence alignments. They introduced two new techniques in this research: the first technique is to generate guide trees with randomly selected sequences and the second is of shuffling the sequences inside that tree. The output of the tree is a multiple sequence alignment which has been evaluated by the sum of pairs method (SPM) considering the real value data from PAM250. To test the performance of our algorithm, we have compared with the existing well known methods: T-Coffee, MUSCEL, MAFFT and Probcon, using Bali Base benchmarks and NCBI based our own datasets. The experimental results show that the proposed method works well for some situations, where other methods face difficulties in obtaining better solutions.

Celine Vens, Eduardo Costa and Hendrik Blockeel [7] proposed a novel method for reconstructing phylogenetic trees. It is based on a conceptual clustering method that is an extension of the well-known decision tree learning approach. As such, the method differs from existing methods, both algorithmically and with respect to the information it uses and the assumptions it makes. It scales better in terms of the number of sequences, and as such can be used on large sets of sequences, beyond the point where other methods break down. By nature, it has a stronger tendency to construct trees in which sub- species are defined by particular polymorphisms. It show that its performance is comparable to that of state-of-the-art methods.

Robert Mc Lay, et al. [8] presents a phylogenetic analysis, is a critical area of modern life biology, and is among the most computationally intensive areas of the life sciences. Phylogenetics, the study of evolutionary relationships, is used to study a wider range of topics, including the evolution of critical traits, speciation, adaptation, and many other areas. In this paper, we present an implementation of a scalable approach to the construction of phylogenetic trees, derived from the neighbor joining algorithm for tree construction, and specifically upon the optimizations to this algorithm implemented in the NINJA software.

Martin Simonsen, et al. [9] discovered a neighbour-joining method which is a widely used method for phylogenetic reconstruction that scales to thousands of taxa. However, advances in sequencing technology have made data sets with more than 10,000 related taxa widely available. Inference of such large phylogenies takes hours or days using the Neighbour-Joining method on a normal desktop computer because of the O(n3) running time. Rapid NJ is a search heuristic which reduce the running time of the Neighbour-Joining method significantly but at the cost of an increased memory consumption making inference of large phylogenies infeasible. We present two extensions for Rapid NJ which reduce the memory requirements and allows phylogenies with more than 50,000 taxa to be inferred efficiently on a desktop computer. Furthermore, an improved version of the search heuristic is presented which reduces the running time of Rapid NJ on many data sets.

III. Problem Formulation

Phylogenetics Tree Construction is one of the most important goals pursued by bioinformatics and theoretical chemistry. The practical role of Tree structure construction is now more important than ever. 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.

A number of factors exist that make Phylogenetics Tree Construction a very difficult task. Phylogenetic tree construction uses distances between sequences and determining relations. To constructing a phylogenetic tree based on data from protein or nucleotide sequence comparisons first do a multiple alignments for the sequences and then calculate distance measure dij between all taxa. Phylogenetic trees among a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods such as neighbour-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model.

The number of possible alignments between two or more sequence to training data is extremely large and Phylogenetics tree is a structure in which species are arranged on branches that link them according to their relationship and/or evolutionary descent. The most popular and frequently used methods of tree building can be classified into two major categories phenetic methods based on distances and cladistic methods based on characters. The former measures the pair-wise distance/dissimilarity between two genes, the actual size of which depends on different definitions, and constructs the tree totally from the resultant distance matrix.

To generate a phylogenetic tree the main computational problems are threefold. Firstly is to determine and compute a distance metric between every genomic sequence. Secondly is to perform hierarchical clustering on the given data sets, utilizing the distance metric computations. Finally is to visualizing the resulting phylogenetic tree. Each problem comes with its own set of difficulties must be overcome, but all share the problem of computational efficiency. Different distance metrics have different time complexities, but none have running times faster than linear. Hierarchical clustering similarly has several different approaches each with their own time complexity, but again running times are linear. Visualizing does not have such strict running time requirements, but rather computational complexity requirements. These requirements stem from the fact that the main focus of the project is hierarchical clustering and ensuring that aspect is correct.

The proposed model uses data mining (Hierarchical clustering) as the retrieval strategy and Tree construction algorithm to identify the Tree of the given Phylogenetic Training data.

IV. Objectives

This work has been focused to achieve the following objectives

- To study the different techniques of Data Mining & selecting the appropriate one.
- To design the model for phylogenetic tree construction.
- To compute and implement the tree construction for the Phylogenetic Training Data.
- To compare the trees constructed through UPGMA and Neighbor Joining approaches by gaining information from multiple sequences.
- To design the model in Java and MATLAB R2011b for Phylogenetic Tree Construction.
- To evaluate the Distance dij between two or more sequence Training data in different species
- To Test the UPGMA and Neighbor Joining Method.

V. Methodology

A. Criteria for Phylogenetic Tree Construction

As more Nucleotide & Protein sequences become available,

multiple sequence and function can be better studied with more accuracy and efficiency. The goal of Hierarchical clustering of nucleotide & protein sequences is to get a biologically meaningful partitioning. Clustering a large set of nucleotide sequences offer several advantages: Nucleotide & Protein are usually grouped into families based on the sequence similarity clustering, which provides some clues about the general features of that family and evolutionary evidence of proteins; Clustering also helps to infer the biological function of a new sequence by its similarity to some function-known sequences. Moreover, protein clustering can be used to facilitate protein's three dimensional structure discoveries, which is very important for understanding protein's function.

The UPGMA algorithm, to construct a tree, uses distances between sequences when determining relations. 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 dij 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 neighbour-joining algorithm to those numbers. The table of numbers is as follows:

Table 1: Label to Sequence Information

Labels	Name used in report	Common name	Identifier
A0	human	human	gb:AF542069.1
A1	human	human	gb:M12523.1
A2	cow	European taurine cattle	gb:AF542068.1
A3	cow	European taurine cattle	emb:Y17769.1
A4	cow	European taurine cattle	emb:X58989.1
A5	frog*	African clawed frog*	ref:NM_001004887.1
A6	frog	African clawed frog	gb:M21442.1
A7	frog	African clawed frog	gb:M18350.1
A8	frog	African clawed frog	ref:NM_001087775.1
A9	boar	wild boar	gb:AY663543.1
A10	human	human	gb:M12523.1
A11	salmon	Atlantic salmon	ref:NM_001123692.1
A12	wolf	gray wolf	dbj:AB090854.1
A13	mouse	house mouse	emb:AJ457860.1
A14	mouse	house mouse	emb:AJ011413.1

Labels	Name used in report	Common name	Identifier
A0	human haplogroup H2a1 E. Arfrica	human haplogroup H2a1 E. Africa	gb:FJ800808.1
A1	human haplogroup H3c sub Sarharan Africa	human haplogroup H3c sub Sarharan Africa	gb:FJ794693.1
A2	human haplogroup H1c1 W. & C. sub-Saharan Africa	human haplogroup H1c1 W. & C. sub-Saharan Africa	gb:FJ798928.1
A3	human haplogroup H5 Africa	human haplogroup H5 Africa	gb:FJ794473.1
A4	human haplogroup J2b Neolithic Greece	human haplogroup J2b Neolithic Greece	gb:FJ445408.2
A5	Mouse	House mouse	dbj:AP003428.1
A6	Hippopotamus	Hippopotamus	dbj:AP003425.1
A7	Giraffe	Giraffe	dbj:AP003424.1
A8	Bactrian Camel	Bactrian Camel	dbj:AP003423.1
A9	Blackbuck	Blackbuck	dbj—AP003422.1
A10	Fugu Puffer fish	Fugu Puffer fish	dbj:AP009536.1
A11	Oblong blow fish	Oblong blow fish	dbj:AP009535.1
A12	Eyespot Puffer	Eyespot Puffer	dbj:AP009534.1
A13	Norwegian pollock	Norwegian pollock	emb:AM489719.1
A14	Norwegian pollock	Norwegian pollock	emb:AM489718.1
A15	Haddock	Haddock	emb:AM489717.1
A16	Atlantic cod	Atlantic cod	emb:AM489716.1

Algorithm 1 Distance algorithm

```
Requires: String A of length n
Requires: String B of length m
 for i =1 to m do
    prev[i]← i
 end for
 for i = 1 to n do
    Curr[0] = i
 for j = 1 to m do
 if A[i] = = B[ j] then
   Score←0
 else
    Score←1
 end if
Curr[j] ← min (Curr [j -1] + 1, Prev[j] + 1, Prev [j -1] + Score)
 end for
 prev ← Curr
 end for
 return Prev[m]
```

The entries in the new row and column are computed as follows. Let the entry to be filled in location x, y, then let cluster A be the cluster that defines row x and cluster B be the cluster that defines column y. Then the value of the entry is

$$\frac{1}{|A| * |B|} * \sum_{x \varepsilon A} \sum_{y \varepsilon B} \text{distorig}(x, y)$$

where distorig (x, y) is the value in the original distance measurement matrix at location x, y. The value of x and y are all of the leaves in cluster A and B. The UPGMA algorithm also specifies when and how the tree is drawn. The tree is initialized with every cluster in the distance measurements matrix being a leaf. Then when the lowest entry is selected as described above, the cluster defining the row and the cluster defining the column are joined under a new root node. The total length from a leaf in a one cluster to a leaf in another cluster through the new root is the found minimal value. These lengths can be seen on our figures in this report. The preudocode for UPGMA is presented in algorithm 2. In the pseudo code the '*' operator indicates concatenating two clusters together, while the '×' operator indicates standard multiplication. The running time for this algorithm is order n2, due to the need to search the entire matrix for the smallest value.The final computational step that was required that was visualization. While the pseudo code for UPGMA does include steps on how to create the tree visualization, we decided to remove that complexity from our UPGMA implementation and instead do a post processing step on our output for the visualization.

Algorithm 2: UPGMA algorithm

```
Requires: lower triangular n × n matrix, A, of pairwise Levenshtein distances
      cpy ← A
      for all c in cpy do
create node in tree of cluster c
end for
while |cpy| > 1 do
x ← lowest value in cpy
c_clus ← cpy[0][x]
r_clus  ← cpy[x][0]
create new cluster (c_clus * r_clus)
connect c_clus and r_clus to new node in tree with branch lengths
```

of x/2
delete row containing r_ clus in cpy
delete column containing c_clus in cpy
for all c in cpy do
if c ≠ c_clus && c ≠ r_clus then
cpy[(c_clus * r_clus)][c] ← _____ 1_____
|c.clus * r_clus|×|c| × ∑x€c ∑y€
(c_clus*r_clus) A[x][y]
end if
end for
end while

The UPGMA and Neighbor Joining Algorithm Contains the Following Steps:

1. UPGMA Algorithm

Let d be the distance function between species, we define the distance $D_{i,j}$ between two clusters of species C_i and C_j the following:

$$D_{i,j} = \frac{1}{n_i + n_j} \sum_{p \in C_i} \sum_{q \in C_j} d(p,q)$$

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

2. Initialization

- Initialize n clusters with the given species, one species per cluster.
- Set the size of each cluster to 1: $n_{i \leftarrow 1}$
- In the output tree T, assign a leaf for each species.

3. Iteration

- Find the i and j that have the smallest distance D_{ij}.
- Create a new cluster - (ij), which has $n_{(ij)} = n_i + n_j$ members.
- 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 D_{ij} /2 each.
- 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:

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

- 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.
- Return to 1 until there is only one cluster left.

4. 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.

Neighbour-Joining Algorithm

Initialization: same as in UPGMA.

1. Iteration

- For each species, compute $u_i = \sum_{k \neq i} \frac{D_{i,k}}{(n-2)}$.
- Choose the i and j for which $D_{i,j} - u_i - u_j$ is smallest.
- 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:

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

- Compute the distances between the new cluster and each other cluster:

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

- Delete clusters i and j from the tables, and replace them by (ij).
- If more than two nodes (clusters) remain, go back to 1. Otherwise, connect the two remaining nodes by a branch of length $D_{i,j}$.

VI. Results

The implementation of the following classes and algorithms for phylogenetic trees: The UPGMA and Neighbour Joining algorithm for constructing rooted phylogenetic trees. The algorithm takes as input the lower triangle of a symmetric distance matrix and contructs a rooted tree in the form of an UPCluster object and NJCluster object. The algorithm should construct an unrooted tree, but this version arbitrarily adds a root node with appropriate edges to the last two active leaves in the construction. The algorithm checks that the distance matrix satisfies the triangle inequality, it is an ultrametric or additive. This applet permits interactive experimentation with distance matrices and the resulting UPGMA and neighbour joining trees.
UPGMA and Neighbour Joining algorithm for constructing rooted phylogenetic trees to take input size 8, then sequence is counted and chose Random data values acc to this table 6.1 and to construct a UPGMA, Neighbour Joining Tree with UPCluster or NJCluster object shown.

Table 2: UPGMA and Neighbour Joining Random Values Data

	1	2	3	4	5	6	7	8
1	0.0	74.3	65.1	50.4	16.0	32.9	63.1	58.2
2	74.3	0.0	45.8	24.5	82.4	41.7	63.5	16.2
3	65.1	45.8	0.0	35.5	63.6	46.8	92.2	41.6
4	50.4	24.5	35.5	0.0	57.9	19.0	57.1	9.7
5	16.0	82.4	63.6	57.9	0.0	43.8	78.9	66.9
6	32.9	41.7	46.8	19.0	43.3	0.0	48.3	25.6
7	63.1	63.5	92.2	57.1	78.9	48.3	0.0	54.4
8	58.2	16.2	41.6	9.7	66.9	25.6	54.4	0.0

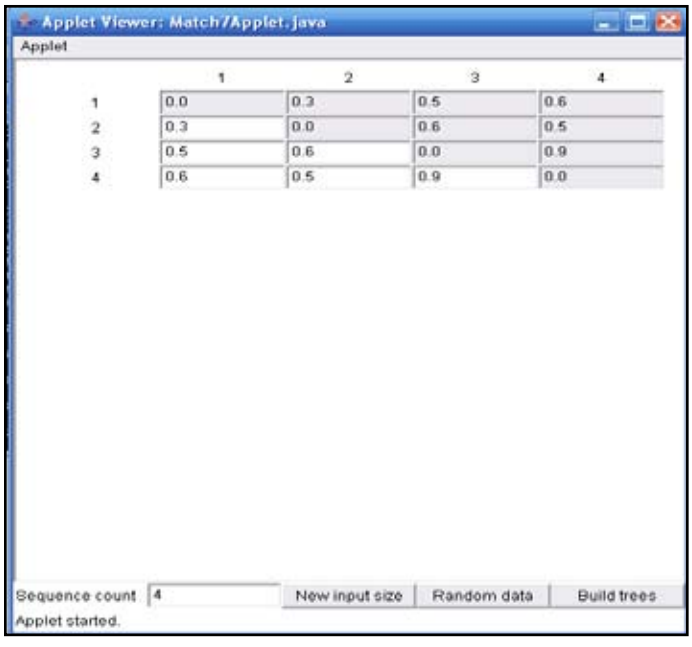

Fig. 1: First Display Screen of distance matrix of four sequences

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

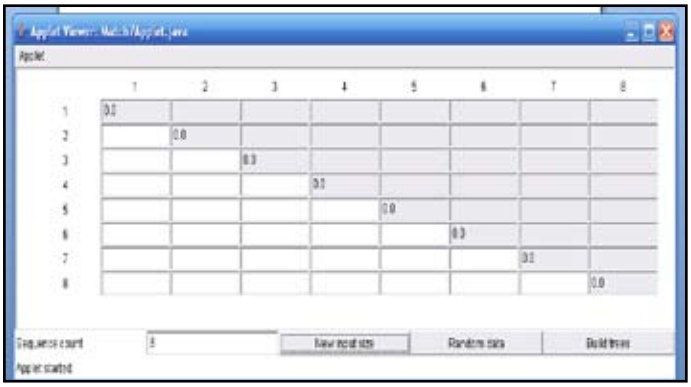

Fig. 2: Display the Input Size 8

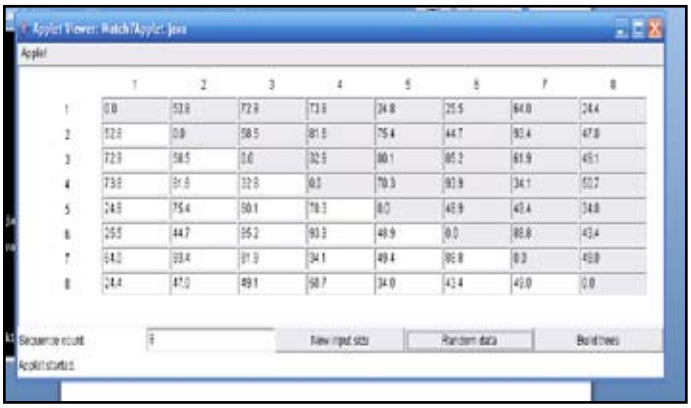

Fig. 3: Display the Random Values of Input Data for Distance Matrix

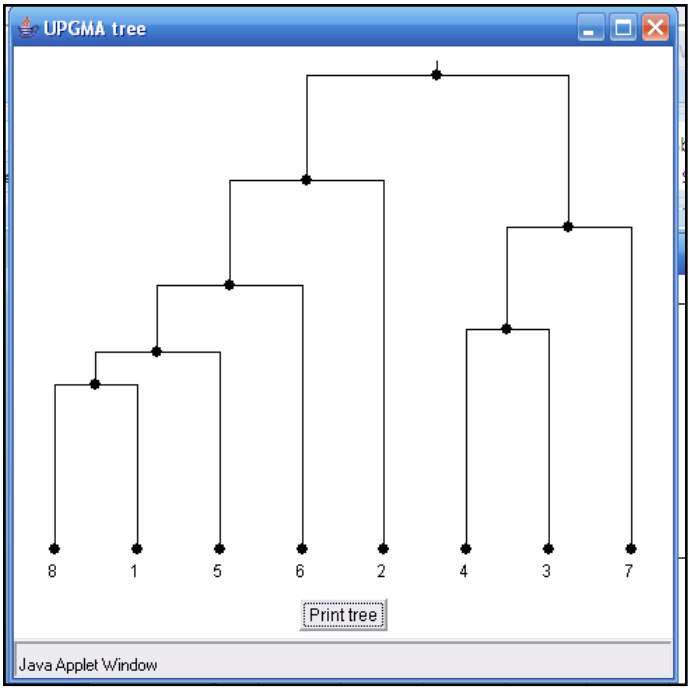

Fig. 4: Display the UPGMA Phylogenetic Tree and UP Cluster.

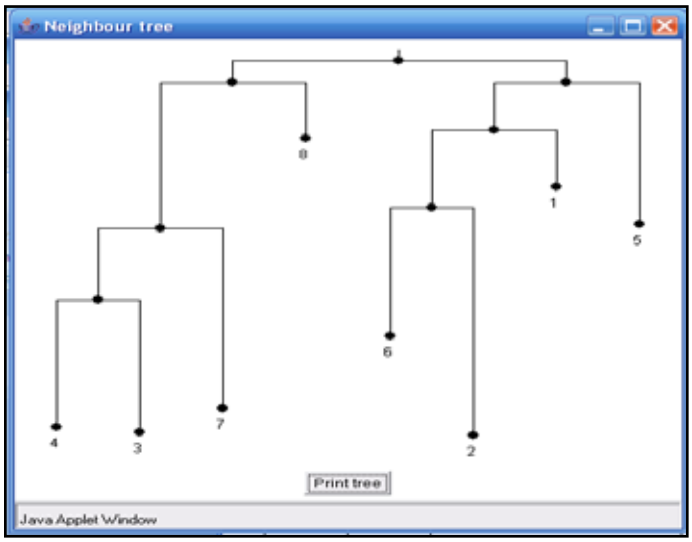

Fig. 5: Display the Neighbour Joining Phylogenetic Tree and NJCluster

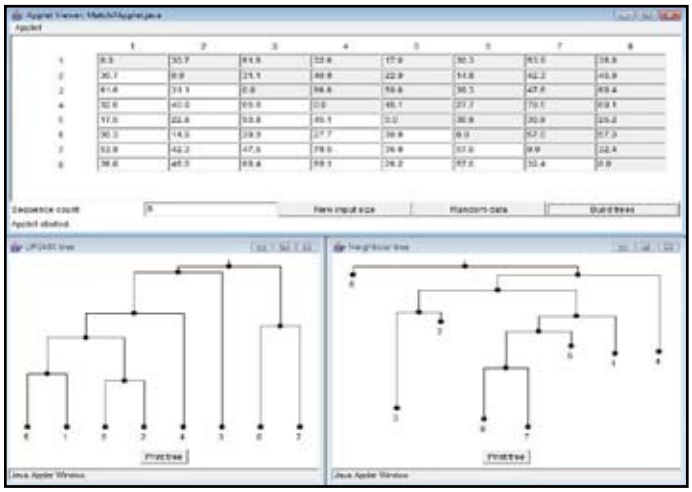

Fig. 6: Display the Overall Result in Java

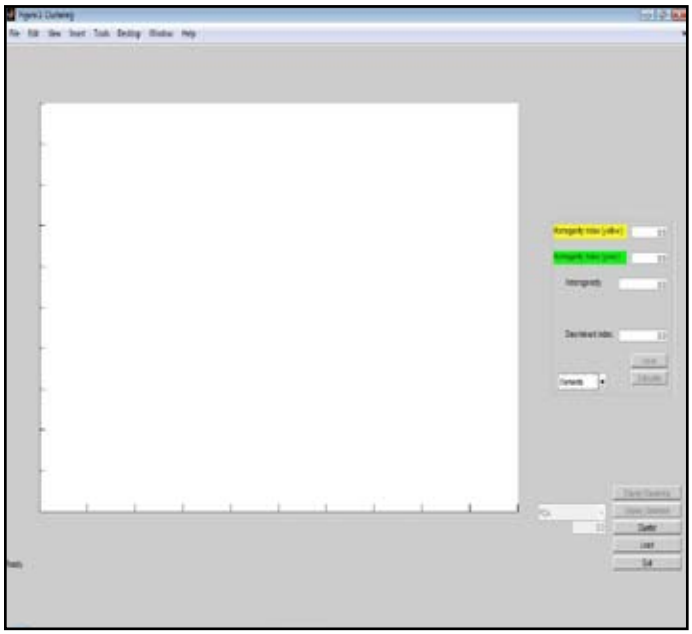

Fig. 7: Display the Clustering Screen and Load Database Files in MATLAB R2011b

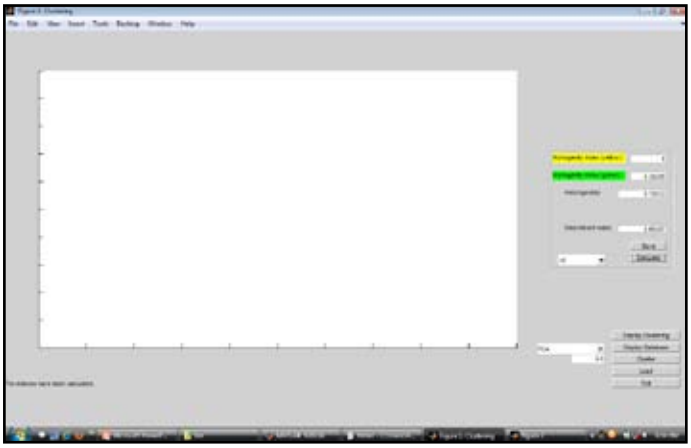

Fig. 8: Display the Homogeneity, Heterogeneity and Discriminate Indexes

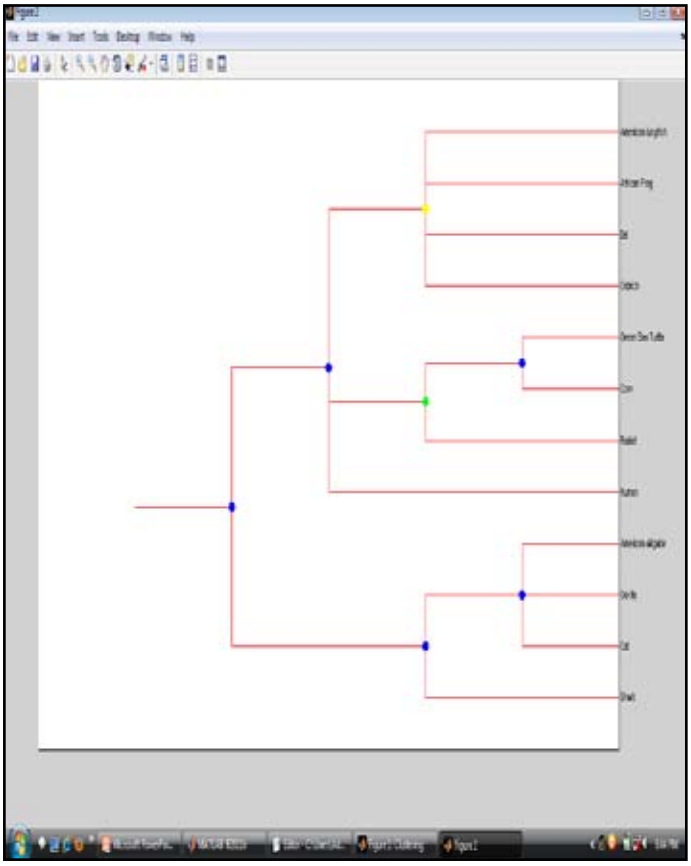

Fig. 9: Display the Final Result in Hierarchical Clustering

VII. 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 contructs 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.

VII. Future Work

Scope for further work on the data mining model:

Phylogenetic tree construction of butyrylcholinesterase has been contributed to further research studies for homology comparison of functional proteins in species.

To examine the load balancing issues with constructing phylogenetic trees using neighbor-joining algorithm.

To optimize the extend number of taxa that can be processed using the UPGMA custom computing machine.

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.

Algorithmic steps can be reduced to improve function and accuracy.

Further research on this technique can be implemented in the context of the larger protein or codon data sets.

References

[1] Gabriel Robins, Tongtong Zhang,"Generalized Neighbor-Joining: More Reliable Phylogenetic Tree Reconstruction", University of Virginia, Charlottesville, VA 22908, pp. 1 – 46, 1996.

[2] Erich Bohl, Peter Lancaster,"Implementation of a markov model for phylogenetic trees", Journal of Molecular evolution, pp. 1 - 16, 2005.

[3] Deni Khanafiah, Hokky Situngkir,"Visualizing the Phylomemetic Tree", Dept. Computational Sociology, Journal of Social Complexity, Vol.2, No.2, 2006, © 2006 Bandung Fe Institute, pp. 20-30, 2006.

[4] Chuang Peng, et al.,"DISTANCE BASED METHODS IN PHYLOGENTIC TREE CONSTRUCTION", 2007

[5] Martin Simonsen, Thomas Mailund,"Rapid Neighbour-Joining", Bioinformatics Research Center (BIRC), University of Aarhus, Denmark, K.A. Crandall and J. Lagergren (Eds.): WABI 2008, LNBI 5251, © Springer-Verlag Berlin Heidelberg 2008, pp. 113–122, 2008.

[6] Naznin, F., Sarker, R., Essam, D.,"Iterative Progressive Alignment Method (IPAM) for multiple sequence alignment Computers & Industrial Engineering, 2009. CIE 2009. International Conference on Digital Object Identifier: 10.1109/ICCIE. 2009. pp. 536 – 541

[7] Celine Vens, Eduardo Costa, Hendrik Blockeel,"Top-down phylogenetic tree reconstruction", 2009.

[8] Robert McLay, Dan Stanzioney, Sheldon McKayz, Travis Wheeler,"A Scalable parallel Implementation of the Neighbor Joining Algorithm for Phylogenetic Trees", University of Texas at Austin.

[9] Martin Simonsen, Thomas Mailund, Christian N. S. Pedersen,"BUILDING VERY LARGE NEIGHBOUR-JOINING TREES", Bioinformatics Research Center (BIRC), Aarhus University, C. F. Møllers All e 8, DK-8000 Arhus C, Denmark.

Er. Sukhpreet Kaur received her M-Tech in Degree from Punjab Technical University Jalandhar (Punjab). Her Field of Interests is Bioinformatics, Cloud Computing, and Gene Expressions Sequence Alignment. She has attended many conferences.

Er. Harwinder Singh Sohal received his B-Tech & M-Tech Degree from Punjab Technical University Jalandhar. He is working as an Assistant Professor in LLRIET Moga, Punjab. His research interests are in the fields of Congestion Control, Cloud Computing, Routing Algorithms, Routing Protocols, Load Balancing and Network Security. He has published many national and international papers.

Er. Rajbir Singh Cheema is working as a Head of Department in LLRIET Moga, Punjab. His research interests are in the fields of Data Mining, Open Reading Frame in Bioinformatics, Gene Expression Omnibus, Cloud Computing, and Routing Algorithms. He has published many national and international papers.