A Review: Phylogeny Construction Methods

Priyanka Shaktawat, Parvati Bhurani

***Abstract:** In bioinformatics, the phylogenetics is the study of evolutionary pathway that gives the relationship between various organisms. Phylogenetics act as an important tool of bioinformatics as it handle huge biological data in the form of evolutionary trees. This relationship can be inferred on the basis of the heritable traits such as nucleotide or protein sequences. The phylogenetic tree can be constructed using different methods: WPGMA, UPGMA, neighbor joining, maximum parsimony or etc. Phylogeny has become the central part of bioinformatics and can be used in different fields to solve their respective problems such as forensics, drug discovery, and vaccine development etc.*

***Keywords:** Bioinformatics, WPGMA, UPGMA, Neighbor Joining, NCBI.*

I. INTRODUCTION

Phylogenetics is the study of evolutionary relationship between different organisms. The relations are inferred using phylogenetic inference methods, which depends on heritable traits such as genes, proteins or etc. The result of this inference is called phylogeny or phylogenetic tree. In other words, a phylogenetic tree represents the homologous characters of various organisms. According to Walter M. Fitch, homology is the relationships between two or more heritable characters from a common ancestor (Fitch, 2001) [1]. The homologous characters are similar in structure but perform different functions. The homology is classified as: morphological homology (structural similarity), ontogenetic homology (similar embryo development) and molecular homology (similarities in DNA, RNA or proteins).

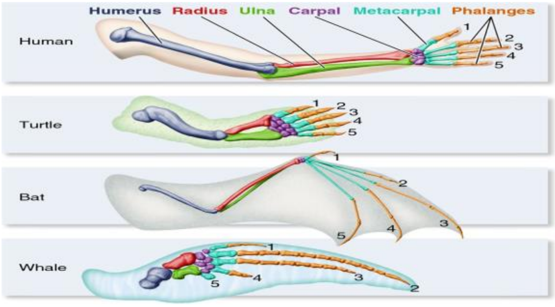

Figure 1.Forelimbs of organisms is adapted to perform different functions

The analogous traits have similar functions but different origin. For example, wings of insects, bats and birds are evolved independently. In the phylogenetic tree, each node represents a specific event of evolution. A tree has terminal nodes (represents molecular sequences and called OUT i.e. Operational Taxonomic Unit), internal nodes (represents inferred ancestral units i.e. Last Common Ancestor). There are mainly two types of phylogenetic trees:

1. Rooted tree (Cladogram): A tree in which all the objects (DNA, RNA or proteins) shares a common ancestor i.e. root node. The path from root to each leaf node represents evolution pathway.

2. Unrooted tree (Phenogram): A tree in which all the objects are related descendants, but the common ancestor is not specified.

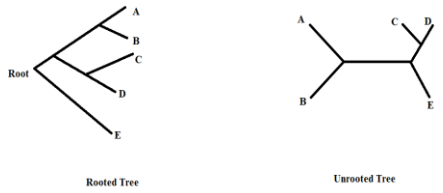

Figure2. Rooted and Unrooted tree

The branches of the phylogenetic tree can be grouped in different ways such as (SL, 2003 June) [2]:
i) Monophyletic group: It consists of an internal LCA node and all other OTUs arising from it. All the members are derived from a common ancestor and have inherited a set of similar characteristics.
ii) Paraphyletic group: It includes ancestor but not all descendants are present.
iii) Polyphyletic group: It is a collection of distantly related OTUs that are associated with similar traits but not directly descendant from common ancestor.

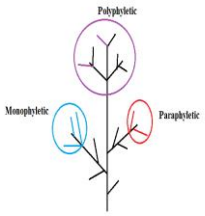

Figure3. Grouping of branches of tree

Revised Manuscript Received on July 25, 2019.

***Correspondence Author**

Priyanka Shaktawat* (M.Tech Scholar, Govt. Women Engineering College, Ajmer) Rajasthan, India

Ms. Parvati Bhurani (Assistant Professor, Govt. Women Engineering College, Ajmer) Rajasthan, India

The number of rooted and unrooted phylogenetic trees that can be constructed depends on the number of OTUs present. This can be calculated using the given formulas:

$$Nu = \frac{(2n-5)!}{2\textasciicircum(\mathrm{n}-3)(\mathrm{n}-3)!}$$

$$Nr = \frac{(2n-3)!}{2\textasciicircum(n-2)(n-2)!}$$

Here, n: number of OTUs.

Number of OTUs	Possible Number of	
	Rooted trees	**Unrooted trees**
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025

Table 1: Number of rooted and unrooted trees (Pevsner, 2003) **[3]**

II. TECHNIQUES TO CONSTRUCT PHYLOGENETIC TREES

There are many methods for tree construction, either phenetic (distance based) or cladistics (character based). Phenetics is the study of relationship between the various organisms on the basis of measure of similarity such as morphological or functional. It is also called distance based methods and can be implemented using clustering algorithms such as neighbor joining; UPGMA (unweighted pair group method with arithmetic mean) and WPGMA (weighted pair group method with arithmetic mean). The advantage of distance based methods is that they use model of evolution, so are simple and efficient.

1. WPGMA and UPGMA: Both the clustering methods are simple agglomerative approaches (bottom-up). A guide or phylogenetic tree is made in stepwise manner, by combining sequences or group of sequences i.e. OTU. If the genetic or evolutionary distance is small then the two respective OTUs are similar to each other so that, they can be grouped together to form new node. When the two OTUs are combined then they are considered as single unit or node. Now again select the pair whose distance is small and combine them. This step is repeated until two OTUs are left. In WPGMA, the averaging of distances doesn't depend on total number of OTUs in the cluster. For example, if node A, B and C are grouped to form new node 'u', now the distance from u to any node k is determined by:

$$duk = \frac{d(A,B),k + dCk}{2}$$

In UPGMA approach, the averaging of the distances is based on the number of OTUs present in respective clusters. So the evolutionary distance between any two nodes u and k can be calculated as:

$$duk = \frac{[N(AB) * d(A,B)k + Nc * dck]}{N(AB) + \ Nc}$$

Here, N_{AB}: Number of OTUs in cluster AB and N_C: Number of OTUs in cluster C

In WPGMA, the number of taxa in the clusters doesn't affect the distance matrix. While in UPGMA the averaging is weighted by the number of taxa in each cluster in each step. Both these methods assumes the constant rate of evolution (molecular clock hypothesis), but it not good for inferring phylogeny (Durbin R, 1999) [4].

2. Neighbor Joining Method: This method was proposed by Saitou and Nei (1987) and later modified by Studier and Kepler (1988). It doesn't assume molecular hypothesis. It creates updated distance matrix in each step and on the basis of this matrix phylogenetic tree is generated. The NJ algorithm starts with calculating net divergence for each terminal node. The rate corrected matrix calculated as:

$$Mij = dij - \frac{Ri + Rj}{N-2}$$

Here M_{ij} is the rate corrected distance matrix. R_i and R_j are the net divergence values for respective ith and jth taxa (R_i is the sum of all distances from node i to all the other nodes). If the M_{ij} is minimal, then group the respective OTUs to form a single node u. The branch length from new node to i and j are calculated as:

$$Siu = \frac{dij}{2} + \frac{Ri - Rj}{2(N-2)}$$

$$Sju = dij - Siu$$

Finally, the distance between new node and terminal node k is calculated by the given equation. These steps are repeated N-1 times.

$$dku = \frac{dik + djk - dij}{2}$$

The cladistics methods characterize the sequences on the basis of common traits. The organisms are grouped together on the basis of their evolutionary history. In other words, it is a study of pathway of evolution. It can be implemented using maximum parsimony and maximum likelihood.

1. Maximum Parsimony: It is a character based method that constructs a phylogenetic tree by minimizing the number of steps required to explain the given set of sequences.

In other words, the shortest evolutionary tree that explains the given set of sequences is considered as best. The problem of finding the optimal trees under parsimony method is divided into two sub problems:

- Determine the amount of mutations (character change) or length of the tree.
- Searching over all the trees that minimize its length.

2. Maximum Likelihood: It is a statistical method that determines the value of the unknown parameters of a probability model. In phylogenetics, the parameters such as mutations rates, tree itself and etc. A substitution model is required to determine the mutation probability. In other words, it finds a tree that maximizes the probability of the genetic data given in the tree.

III. DATA PREPARATION AND PHYLOGENETIC ANALYSIS

In this review paper the phylogenetic trees are constructed using UPGMA, WPGMA and neighbor joining methods for five genetically different varieties of Helianthus annuus (common sunflower) (National Centre for Biotechnology Information databank website) [6]. The molecular (DNA) sequences for different varieties of sunflower have been loaded from NCBI (National Centre for Biotechnology Information) databank. These DNA sequences can be accessed using their accession IDs which are given in table 1.

Sequences	Accession ID	Number of base pairs
Sequence 1	BU015365.1	457
Sequence 2	BU036497.1	240
Sequence 3	BU036493.1	164
Sequence 4	BU036490.1	152
Sequence 5	BU036487.1	498

Table2: DNA sequences with their accession ID [6]

Now the evolutionary distances between each pair of taxa are calculated using Jukes – Cantor method. It is a substitution model that computes the probability of mutations from one state to another. It represents the evolutionary distances in the form of matrix. The distance between two molecular sequences is calculated as (Philippe Lemey, 2009) [5]:

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

Here, N_d: number mutations between two sequences

N: sequence length

The multiple sequence alignment technique is used to align all the sequences progressively. This method generates an alignment stepwise, starting with the most similar sequences and progressively adding the divergent sequences. It first constructs a guide tree that gives the information of the order in which the sequences must be added progressively. The alignment of five sequences is given in figure 4. This is build using the progressive alignment method. The gaps are inserted between the sequences to align homologous characters of all the sequences, so that they represent an evolutionary pathway.

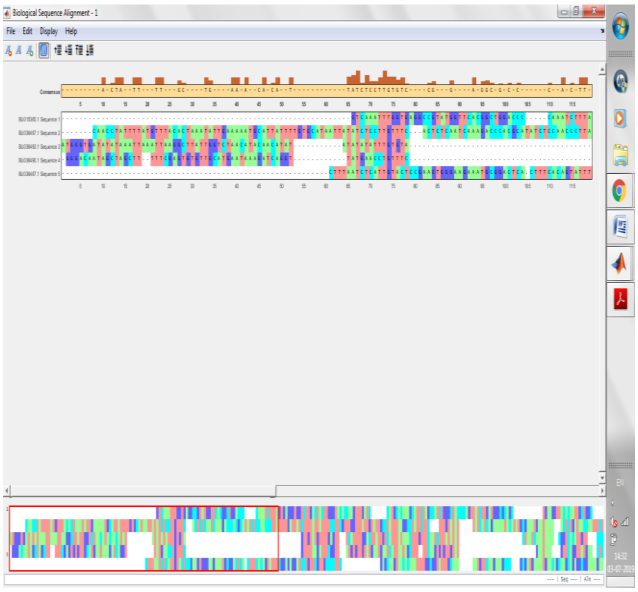

Figure4: Multiple Sequence Alignment

The phylogenetic tree in figure 5 is generated using UPGMA method. In this method averages are weighted by the number of taxa in each cluster at each step. The final output is affected by each evolutionary distance. This makes calculation slightly complicated.

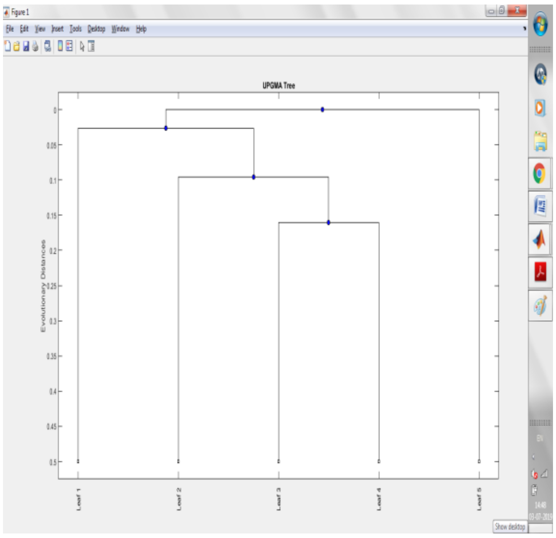

Figure5: UPGMA tree

Retrieval Number: E2283076519/19©BEIESP
DOI: 10.35940/ijese.E2283.078219
Journal Website: www.ijese.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In figure 6, phylogenetic tree is constructed using WPGMA method. This method calculates the distance between clusters as simple averaging method. It is computationally easier and if the number of taxa is unequal in the clusters, the distance of matrix do not contribute to intermediate calculations.

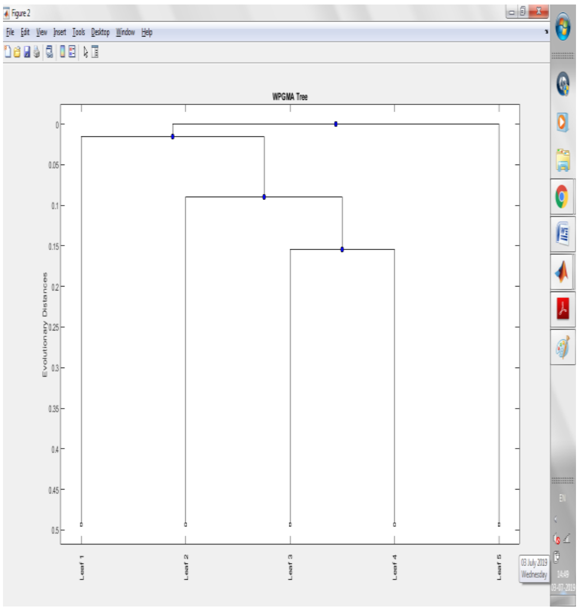

Figure6: WPGMA tree

The phylogenetic tree shown in figure 7 is constructed using neighbor joining method. It produces unrooted trees and does not assume constant rate of evolution.

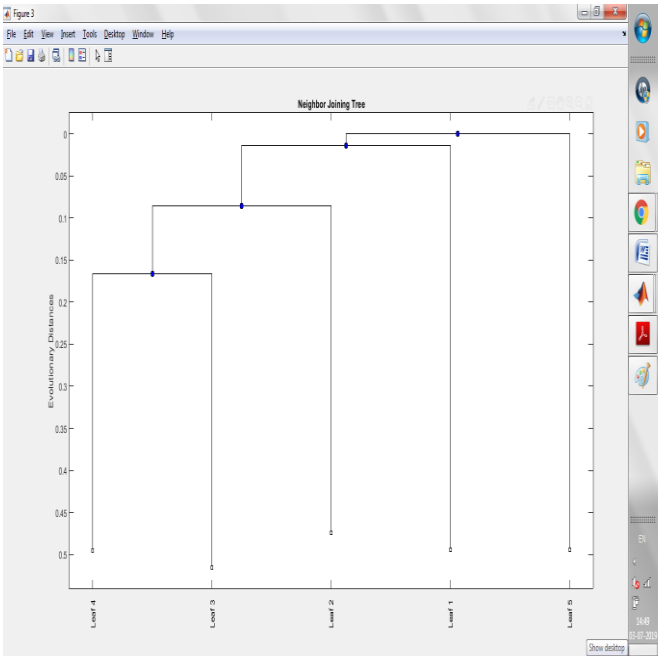

Figure7: Neighbor Joining Tree

IV. CONCLUSION

The phylogenetics trees are used to infer evolutionary relationship between various organisms. The construction of tree is a miscellaneous problem to solve on increasing the number of molecular sequences. The phylogeny methods are reviewed for this paper. The comparative study concludes that in most cases neighbor joining method is used over WPGMA and UPGMA. The NJ method allows the unequal rates of evolution, so that the branch lengths proportional to degree of mutations between the sequences.

FUTURE SCOPE

Nowadays, with the increase in number of molecular data the complexity of the construction methods also increase so as to get an optimal output. This prompted the researchers to evolve efficient algorithms to construct phylogenetic trees which give complete information of ancestor pathway. As Artificial Intelligence is emerging in the field of optimization techniques so new algorithms can be discovered using ant colony or particle swarm optimization methods.

REFERENCES

1. Walter M. Fitch, “Distinguishing Homologous from Analogous Proteins”, Systematic Zoology, Volume: 19, 2001.
2. Baldauf SL, “Phylogeny for the faint of heart: a tutorial”, Trends in Genetics, 2003 June.
3. J. Pevsner, “Molecular Phylogeny and Evolution”, 2003.
4. Durbin R., S. Eddy, A. Krogh and G. Mitchison, “Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acid”, Cambridge University Press, 1999.
5. Philippe Lemey, Marco Salemi and Anne- Mieke Vandamme, “The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing”, ISBN 978-0-51171963-9, Cambridge University Press, 2009.
6. National Centre for Biotechnology Information databank website (https://www.ncbi.nlm.nih.gov/).

AUTHORS PROFILE

Priyanka Shaktawat (M.Tech Scholar, Govt. Women Engineering College, Ajmer)
Email ID: priyankashaktawat14@gmail.com

Ms. Parvati Bhurani (Assistant Professor, Govt. Women Engineering College, Ajmer)
Email: parvatibhurani@gweca.ac.in