Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Phylogenetic Tree Construction Revision

Manmeet Kaur
M-tech (CSE Dept)
ACET, Manawala, Asr, India

Rajbir Singh
Associate Professor CSE
ACET, Manavala, Asr, India

Abstract—The construction of phylogenetic tree is complex yet very important in the field of bioinformatics. A phylogenetic or evolutionary tree once constructed can help get insight into the evolution of different species. But there is computational complexity as the problem grows into large number of species that cannot be easily solved. So, new methods have been researched that are applied to phylogenetic tree construction and have provided some promising results. Various algorithms have been reviewed and are discovered by studying the patterns of nature.

Keywords— MSA, NJ, UPGMA, Phylogenetic tree.

I. INTRODUCTION

To start with, phylogenetics is the science that studies evolutionary relationship between species. To predict about relationships, phylogenetic trees are constructed that link species. The phylogenetic tree construction is regarded as an NP-complete which means it is in a very small class of the most difficult problems to solve.

Phylogeny is classified as relationship between two species. The resulting relationship is represented as binary tree. Two main types of trees are: 1) Rooted trees—where all nodes are derived from single node. And 2) Unrooted trees—where it is not clear that where the nodes originated from . The notation followed is standard graph theory notation where each species is represented by a node or leaf, the relationship is specified by an edge or branch, and the time estimate is represented by length of branches.

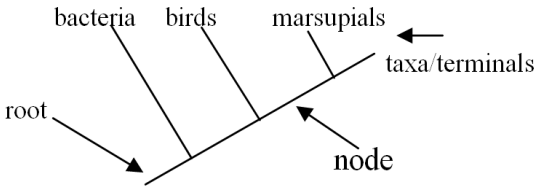

Figure 1.1: simple rooted tree with the root at the bottom and the tips at the top.

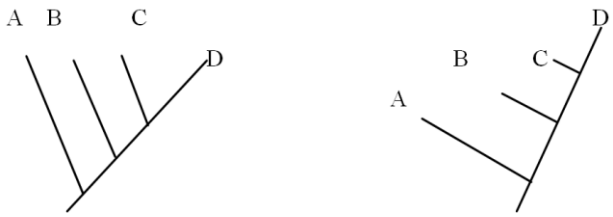

Figure 1.2: Two Trees with different patterns containing same information

The possible number of trees is large and thus it is important to distinguish the correct tree from which the species have evolved,hence an efficient algorithm is required to find the correct one.

II. METHODOLOGY

DISTANCE BASED APPROACHES

The classification for phylogenetic tree construction falls into two categories:distance based and character based. Commonly used distance based methods are the weighted pair group method using arithmetic averages(UPGMA);Neighbour joining and the fitch and margoliash algoritms that are based on creation of distance matrix initially. The character based methods such as Maximum Likelihood [ML] and Maximum Parsimony [MP] are alternatives to the distance based methods . The character based methods construct trees based on probabilistic approach.

A. UPGMA And Neighbour Joining

UPGMA and Neighbour joining use data mining procedure called clustering. These methods form a new node on a tree by clustering nodes at each stage. The tree grows upward starting at the bottom and adding new node at each step. At each step the length of branch is determined by the difference in the heights of the nodes at end of each branch. UPGMA assumes that all nodes are equally distant from the root and the tree is additive .UPGMA is not much applicable today;

since it poses biological issues by using "molecular clock" hypothesis. So NJ method is used as common approach than UPGMA.NJ creates a new distance matrix at each step and tree is constructed based on matrices. NJ directly calculates distance to internal nodes nad do not construct clusters. The first step of NJ algoritmis the creation of matrix using hamming distance between each node and taxa. The distance from the two nodes to the node that directly links them is calculated using minimal distance.Hence a new matrix is calculated and the new node is substituted by joining the original nodes. Since the distances are calculated directly .There is no assumption about the distance between the nodes.

III. NEIGHBOUR JOINING ALGORITHM

- T= Set of leaf nodes ,one for each given sequence and put L=T.
  ITERATION
- From L, pick a pair of i,j where i to j is the minimal distance.
- K=new node and set distance between k and m= half(distance between i and m+ distance between j and m + distance between i and j), for all M in L
- From L remove i and j and add K.
  TERMINATION
  When two leaves i and j contained in L and the remaining edges between i and j ,with length distance between i and j .

Distance measure is the measurement of additivity.NJ even works in cases where the lengths are not additive but the tree is no longer guaranteed to be correct. To test additivity four point condition exists. Additivity is the sum of the lengths of two distances that must be greater than the third distance . For example;if distance dij is the distance from i to j and if there exist four nodes,then dij + dkl = dik + djl and is greater than dil+dik. Since the link between the two smaller clusters is common in two of the distance sums.it is efficient to execute and easy to understand NJ approach. Since tree construction takes the advantage of common clustering techniques. The tree produced by NJ is unrooted since it shows the relationship between sequences without assigning a root node from which all other sequences have been derived . An outgroup species is choosen that is distantly related to the remaining sequences in order to construct a tree. The place where new species connects to the recently constructed tree is a good indicator for the most likely location for the root of the tree. When is difficult to find outgroup other strategies to locate the root of the tree such as using mid point of the longest chain of consecutive edges and will indicate the root if the tree followed the molecular clock.

A. Maximum Parsimony

Maximum Parsimony(MP) till date is the most widely and accepted method for tree construction . It uses character based algorithm and hence is different from previous distance based methods. MP assigns a cost to each tree by searching through the possible tree structures. It assumes that the most likely tree is one that requires the fewest number of changes in order to explain about the data in the alignment.Since the nodes or taxa inherited the characteristics from a common ancestor, there is a premise that taxa or nodes are sharing common characteristics. The term homoplasy explain conflicts with this major assumption. Conflicts can be reversed in three ways:converg(completely independently unrelated taxa evolved the same characteristics), reverse(to reverse back to original state) and parallelism(development of characteristics in certain manner since different taxa may have similar mechanisms). The tree with the minimum lenghth or score; that is the number of changes summed along the branches, becomes the most parsimonious tree and that tree best represents the evolutionary pattern. The total overall length in terms of number of changes is taken in contrast to other methods that find branch lenghth and hence is different from other methods. Many times MP, finds two or more trees that tend to be equal and does not provide, proper answer to distinguish between the trees that represent the actual evolutionary tree and hence majority rule or strict consensus is used to solve this problem.

Recursion has been traditionally used by MP approach to search for the minimal number of changes within the trees. This is done by post order traversal that is starting at the leaf of the tree and working upwards the root. Another version called weighted parsimony links a cost factor to the algorithm and weighs certain scenarios accordingly, An artefact that m,ust be handled in parsimony is long branch attraction where branch lenghth is the number of substitutions between two nodes or taxa. Parsimony assumes that all taxa assumes the same amount of information and all taxa evolve at same rate. A phenomenon called long branch where rapidly evolving taxa are placed together as they have many mutations.

B. Maximum Likelihood

Felsenstein proposed this approach in 1981, ML is one of the most computationally intensive approach but is also the flexible one. Given a tree topology and a model of nucleotide evolution. ML optimizes the likelihood of observing the data under a specified model of evolution. ML finds the tree that explains the observed data with the greatest probability. ML is based on probability or likelihood and hence is different from other methods. The probability or likelihood for a tree are obtained using few equations. The likelihood , L = P (data | tree) which for a given tree is the probability of observing the data. The ability to make statistical comparisons between topologies and data sets is one of the big advantages of ML . Several equally likely trees can be returned by ML. Which may be a pro or con depending on the study. ML assumes that the model is accurate and assumes the method is inconsistent if the model does not accurately reflect the underlyning data set. ML has the advantage of being robust but breaching of assumptions can cause problems.

The disadvantage of ML is that there may be multiple maximum likelihood points for a given phylogenetic tree and it involved extensive computation.

C. Literature Survey

The research work involves Computational Cluster Model for Phylogenetic Tree Construction. The available literature has been reviewed in this context.

Geetika Munjal ,et. al. (2015): specified sequence analysis which includes the alignment based and alignment free methods of tree generation are reviewed and these find distance/similarity among the sequences of different species. Alignment free method based on tuple count and set theory is proposed and the results are compared with the guide tree obtained using alignment based method. The proposed method is tested on DNA sequence of length below 1000bp (dataset1) and Sequence of length above 16000bp (dataset2). It achieves the similar performance as that of the alignment based method but without the alignment phase.

Dega Ravi Kumar Yadav and Gunes Ercal (2015): performs multiple sequence alignment, an important way of which is the progressive alignment approach studied in this work.Progressive alignment involves three steps: find the distance between each pair of sequences; construct a guide tree based on the distance matrix; finally based on the guide tree align sequences using the concept of aligned profiles. Our contribution is in comparing two main methods of guide tree construction in terms of both efficiency and accuracy of the overall alignment: UPGMA and Neighbor Join methods. Our experimental results indicate that the Neighbor Join method is both more efficient in terms of performance and more accurate in terms of overall cost minimization.

Baye Wodajo, et. al. (2015) : analysed Safflower, Carthamus tinctorius, L. an oilseed crop that belongs to the family Asteracea. The genus Carthamus is comprised of 25 species including the only cultivated species of Carthamus tinctorius. So far, the characterization of safflower using molecular markers has been limited. The objective of this study was to examine the cluster analysis of safflower accessions collected from different regions of Ethiopia using ISSR molecular markers. For this purpose, seeds of seventy land race accessions collected from four administrative regions of Ethiopia (Amhara, Oromia, Tigray and SNNPR) were obtained from the EIB and grown in green house at Addis Ababa University, Faculty of Life Science. Four primers were selected. The four selected ISSR primers produced a total of 43 bands across the 70 safflower accessions. The number of amplified fragments with ISSR primers ranged from 6 to 15 per primer with varied in size of 100 to 1000 base pairs. The cluster analysis based on ISSR data Safflower individuals assembled from different localities and regions observed to be spread all over the trees without forming strict grouping based on their geographic origin.

J.Jayapriya, et. al. (2015): This paper illustrates an enhanced algorithm based on one of the swarm intelligence techniques for constructing the Phylogenetic tree (PT), which is used to study the relationship between species. The main scheme is to formulate a PT, an NP- complete problem through an evolutionary algorithm called Artificial Bee Colony (ABC). The tradeoff between the accuracy and the computational time taken for constructing the tree makes way for new variants of algorithms. A new variant of ABC algorithm is proposed to promote the convergence rate of general ABC algorithm through recommending a new formula for searching solution. In addition, a searching step has been included so that it constructs the tree faster with a nearly optimal solution. Experimental results are compared with the ABC algorithm, Genetic Algorithm and the state-of-the-art techniques like unweighted pair group method using arithmetic mean, Neighbour-joining and Relaxed Neighbor Joining.

Prof. Baldeep Singh, et. al. (2015): discussed that 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. In the present project work, distance is computed using comparative method (scoring using differences) and using distance based computational methods.

Ms. Bharti Goel, et. al. (2015): suggested that Data mining is an essential tool to discover the hidden data and extract patterns from a large data set. The biological data is available in different formats and is comparatively more complex.. Data mining methods have been used to study molecular phylogeny to discover the degree of relationship within a group of organisms. Phylogenetic trees are constructed from the molecular sequences of the different living organisms. These are actually needed to evaluate the relation between the different species and also the different time gaps from the actual origin. Sequence alignment is one of the applications of bioinformatics.

IV. CONCLUSIONS

The job is to find an optimal alignment and reducing the complexity of sequences involved and constructing tree for given sequence with improved accuracy. The biggest advantage of distance based methods is that it makes use of large number of models to correct distances.

REFERENCES

[1] Geetika Munjal, Madasu Hanmandlu and Deepti Gaur (2015) " *A New Alignment Free Method for Phylogenetic Tree Construction*",International Journal of Database Theory and Application Vol.8, *pp*.111-124.

[2] Dega Ravi Kumar Yadav and Gunes Ercal (2015) *"a comparative analysis of progressive multiple sequence alignment approaches using upgma and neighbor join based guidetrees*", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 5, No.3/4.

[3] Baye Wodajo, Feysal Bushira Mustefa and Kassahun Tesfaye (2015) "*Clustering Analysis of Ethiopian Safflower (Carthamustinctorius) Using ISSR Markers*",International Journal of Scientific and Research Publications, Vol 5.

[4] J. Jayapriya and Michael Arock (2015**) "**Enhanced Bio-Inspired Algorithm For Constructing Phylogenetic Tree",ICTACT Journal On Soft Computing, Vol 06.

[5] Shaminder Kaur ,Prof. Baldeep Singh and Prof. Tajinder Kaur (2015)**"** Improved Computational Methods for Phylogenetic Tree Construction using Cluster Analysis", International Journal of Advanced Research in Computer Science and Software Engineering,Vol 5, *pp. 378-385* .

[6] Rajbir Singh,Ms. Bharti Goel and Dheeraj Pal Kaur (2015) " *Improved Distance based Phylogenetic Tree Construction using Bootstrapping Method*",International Conference on Information Technology and Computer Science .

