Reconstructing Phylogenetic Network with ReTF algorithm (Rearranging Transcriptional Factor)

Shamita Malik and Dolly Sharma

***Abstract*— The term Phylogentics is the study of evolutionary relationship between different species, organisms or genes. These relationships are depicted as branched, tree like diagrams that provide insight into the events that occurred during the evolution process. These trees may also have a root which is known as the common ancestor. Building the "Tree of Life" has been the prime objective of many researchers, until it was proved that the tree of life cannot be represented by a single? tree. Many evolutionary events cannot be represented with the help of a simple tree, hence phylogenetic networks came into picture. Phylogenetic networks can be classified into different categories. In this paper, an algorithm (ReTF) has been proposed which would improve the results of the current phylogenetic network reconstruction algorithms. The idea behind ReTF is rearranging the input sequences in a way that the new arrangement gives a better tree, since the order of input sequences affects the outcomes of phylogenetic network.**

***Keywords*— Phylogenetic Networks, Transcription Factors.**

I. INTRODUCTION

THE graphical representation of genetic data can be done in many different ways, such as Phylogenetic Trees, Split networks, Median networks, Reticulate networks, and `Other'. The `other' category describes trees which have at least one branch representing a connection between a DNA sequence incorporated directly from a non-ancestral DNA sequence. Reticulate networks are capable of displaying events where organisms gain additional genetic material as well as reticulation events within a population [2][6]. Split networks may be further classified into Median networks or Consensus networks. Although numerous computational techniques have been presented in this research area, they have not been thoroughly tested for their consistency and efficiency. Numerous algorithms have been designed in the past to reconstruct phylogeny from the available genetic data, but most of them do not consider the occurrences of recombination events within a population.

This work was supported by Amity Institute of Information Technology, Amity University, Uttar Pradesh, India.

S. Malik is Research Scholar in AIIT, Amity University (corresponding author; e-mail: smalik@amity.edu).

D. Sharma is an Assistant Professor in AIIT, Amity University (e-mail: dsharma10@amity.edu).

Huson *et.al.* [2] have defined "Phylogenetic network" as a graph used to represent evolutionary relationship (abstract or explicit) between a set of taxa, that label its node from a common ancestor to DNA sequences that are diverged. Each of these two sequences over the time start to accumulate nucleotide substitutions, commonly known as mutations, and the number of these mutations are used in analyzing evolution [3][13]. The degree of divergence between two sequences of length 'L' differing at 'n' sites is given by n/L or (n/L)*100%. The phylogenetic networks may be rooted or unrooted. An unrooted phylogenetic network 'N' on 'X' is any undirected graph whose leaves are bijectively labeled by the taxa in X. Split networks and quasi-median networks belong to this category [11]. Although unrooted phylogenetic networks can be used to study the relationship between different taxa, the evolutionary history is lost without the root. Rooted Phylogenetic networks are defined as follows: Let 'X' be the set of taxa, a rooted phylogenetic network 'N' on X is a rooted DAG (Directed Acyclic Graph) where the set of leaves is bijectively labeled by the taxa in X.

These can further be divided into clusters, hybridization networks and recombination networks. An alternate way to construct rooted phylogenetic networks is to first construct an unrooted network and then apply computational techniques to find the root [5].

II. BACKGROUND

When we are in the process of building phylogeny, we focus on two approaches, (i) Character based approach, and (ii) Distance based approach. In Distance based approach, the phylogenetic tree is created by analyzing the relationship between the distances among the taxa and the edge lengths of the corresponding tree. When we apply Character based approach, we take into consideration the changes in character states and attempt to find perfect phylogeny [9].

In character based methods, we assume that each character substitution is independent of its neighbors. Character based methods are further divided into maximum parsimony and maximum likelihood. In maximum parsimony, one tree would be built with the fewest number of changes needed to represent the variations determined within the information. This approach is best for similar sequences or with groups of sequences with small minimum discrepancy. For larger data sets, one has to opt for "branch and bound" methods. In maximum likelihood, methods work best on the model of

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

probability that accounts for the large amount of variations in the data sets. Analysis is done on each position of multiple alignments. In distance based methods, we assume molecular clock method, meaning that all the mutations are neutral and they can occur at random clock-like rates. But sometimes this assumption is not true because of the following reasons. Firstly, different environmental conditions affect mutation rates. Secondly, this assumption ignores selection issues which are different with different time periods. Distance matrix methods are also known as clustering methods because they try to place the correct positions of all the neighbors and aim to find correct branch lengths. Biological networks are representation of multiple interactions within a cell and are basis to understand underlying cellular behavior. Cells must keep on adapting to changing conditions by altering their gene expression pattern [4]. Genetic regulation is commonly understood and modeled through the use of transcriptional regulatory network (TRN) structures. The basic idea behind this concept is to understand the transcriptional regulatory interactions. These networks model the regulatory interactions between two key components: transcription factors (TFs) and the target genes (TGs) they regulate. Such networks are easier if they can be understood with the concept of modules. Each sub network is represented with a module, where nodes are connected physically or functionally. The nodes may represent the set of genes that share common regulatory TFs. We further studied such networks with the help of network motifs. Network motifs describe how single nodes connect with their neighbors [8].When two sequences found in two different organisms are mostly similar, we assume that they are derived from one common ascendant. The sequence alignment also reveals if the positions are preserved from the ascendant sequences. The progressive multiple alignment of a gaggle of sequences initially aligns the foremost similar try. Then it adds the distant pairs. However, all phylogenetic trees do not work like this. Most phylogenetic studies assume that every position in a sequence will modify several times from the opposite positions [10]. Gaps in alignments represent mutations in sequences like insertion, deletion, and genetic rearrangements. Gaps are generally treated differently, some studies ignore them, whereas some treat gaps by victimization sequences similarity scores because the base for phylogenetic analysis, instead of victimization alignment itself attempts to determine what happened at each of these position. The similarity scores supported by grading matrices (gaps scores) are utilized by distance ways [12] [19].

III. METHODOLOGY

We accessed database of genome sequences downloaded from National Centre for Biotechnology Information. In this paper we have taken DNA sequences of HIV-1 virus and fruitfly, *Drosophila melanogaster*. See Figure 1, for the current methodology. A set of multiple DNA sequences is represented as D = $\{s_1,s_2,s_3\ldots\ldots s_n\}$. Firstly we focus on sequence alignments. These alignments are required because it is the way of arranging the sequences of DNA, RNA or proteins and enable us to identify the regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences [7]. Alignments can be done either manually or by using built-in softwares. These built-in softwares can be categorized by algorithms. Sequence alignment can be performed by various common software tools including ClustalW2 and T-coffee. In this paper we have used ClustalW2 for alignment of HIV-1 virus and fuitfly sequences. After getting the set of aligned sequences, we can run the phylogenetic network reconstruction algorithms. These phylogenetic network algorithms are widely available on the internet. Network, Dendroscope Phylogenetic network reconstruction, SplitTree, MEGA, all can be used for drawing phylogenetic networks.

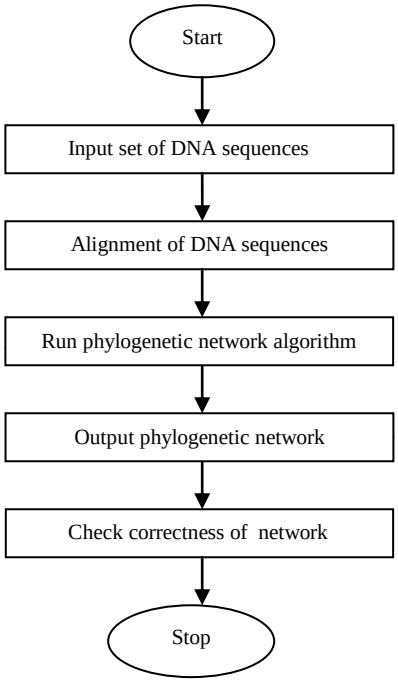

Figure 1: Flowchart for reconstructing phylogenetic network.

In this paper, we have used MEGA (Molecular Evolutionary Genetics Analysis) software for network construction [18]. After phylogenetic network construction, we would check the reliability of network by bootstrap test, which is evaluated using Efron's bootstrap resampling technique [15].

Suppose there are *a* sequences, each with *b* nucleotides (or codons or amino acids), a phylogenetic tree can be reconstructed using some tree building method. From each sequence, *b* nucleotides are randomly chosen with replacements, giving rise to *a* rows of *b* columns each. These

now constitute a new set of sequences. A new tree is constructed like this and it is compared with the original sequences. Then interior branches that are different from bootstrap tree, the sequence it partitions is given a score of 0; all other interior branches are given the value 1. This procedure is repeated several hundred times and the percentage of times each interior branch is given a value of 1 is noted. This is called as the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 65% or higher, then the topology at that branch is considered "correct" [15]. This ReTF algorithm, see Figure 2 works basically on the principle of number of mutations counted with predefined thresholds. Using the selected motif we assign score to each sequence, based on how conserved the motifs are in those sequences. On the basis of scores we further rearrange the sequences, assuming that ancestral sequence will be the one with the lowest number of mutation counts and the following nodes will have increased mutation counts.

After rearranging the sequences on the basis of ReTF algorithm, we run phylogenetic reconstruction program MEGA followed by bootstrap process for reconstruction of phylogenetic network after implementing sequence rearrangement process done by ReTF algorithm.

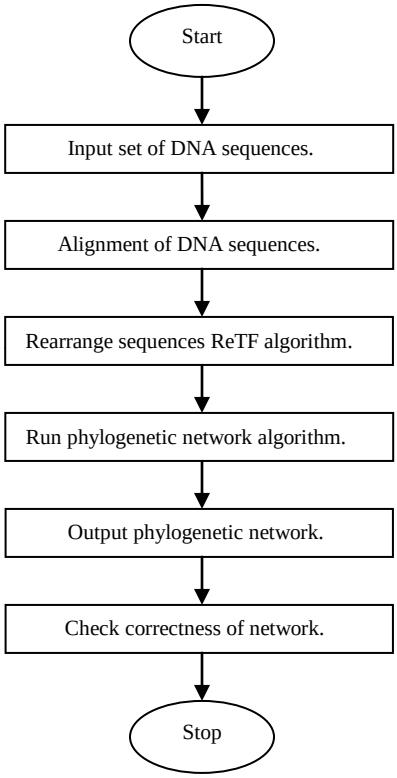

Figure 2: Flowchart for reconstructing phylogenetic network after introducing ReTF algorithm.

ReTF Algorithm

Input: A set of n DNA sequences, D = {s_1, s_2 …. S_n}.

Output: A set of rearranged sequences, D' = {s_1', s_2'… S_n'}, such that each sequence is assigned a new sequence number.

Construct the list L of all motifs in the database with length from 4 to 20.

For each x in L

Calculate P-value of particular character and based on that we come up with desired motif.

Sort L based on the P-value.

Output a set of motifs with highest P-Value.

IV. RESULT AND ANALYSIS

The ReTF algorithm was implemented in Java on an Intel Centrino processor with 3 GB RAM and 2.6GHz speed. The following four algorithms were used for the testing: Minimum Evolution algorithm, UPGMA algorithm, Maximum Parsimony algorithm, Neighbor Joining algorithm. It was observed that in most of the cases, we get an improvement in the Bootstrap score for the DNA sequences. In other cases, ReTF algorithm makes no difference in the results. The algorithm was tested on HIV 1 virus and fruitfly.

We used MEGA software for reconstructing the Phylogenetic network. To search for the motifs, we used PMS3 algorithm [16][17].

TABLE I
COMPARISON OF ALGORITHMS FOR HIV-1 AND FRUITFLY

Algorithm	HIV	Fly
Neighbour Joining	Improved score on 2 sequences.	Improved score on 5 sequences.
Minimum Evolution	Improved score on 2 sequences.	Improved score on 7 sequences.
Maximum Parsimony	Improved score on 3 sequences.	Improved score on 12 sequences.
UPGMA	Improved score on 1 sequence.	Improved score on 11 sequences.

V. CONCLUSION

It was observed that ReTF algorithm improves the bootstrap score for the tested databases in many of the cases. In other situations, there were no improvements as many of the scores were already 100. In other situations, where

improvements were not observed, the result stayed the same. In the complete testing procedure, there were only few occurrences in the case of Maximum Parsimony where one of the sequences had a lesser Bootstrap score.

VI. ACKNOWLEDGEMENT

The authors thank Dr. Ashok K. Chauhan, Founder President, Amity University, for his support and encouragement along with providing us with the necessary infrastructure for research.

REFERENCES

[1] D,Huson and H.D. Bryant, "Application of Phylogenetic Networks in Evolutionary Studies". Mol. Biol. Evol. , 2006, pp. 254-267.
[2] R.R,Hudson, “Gene genealogies and Coalescent Process”, Oxford Surveys in Evolutionary Biology, Vol. 7, 1991, pp.1-44.
[3] J. Hein, “A heuristic method to reconstruct the history of sequences subject to recombination”, J. Mol. Evol., 1993, pp. 396-405.
[4] J.M. Smith, “Analyzing the mosaic structure of genes”, J. Mol. Evol., 1992.34, pp. 126-129.
[5] D, Huson, and C ,Scornavacca; “A survey of combinatorial methods for phylogenetic networks”,Genome Biology and Evolution.3(2011), pp.23-26.
[6] W.F,Doolittle,“Phylogenetic classification and the universal tree”. Science, 1999, pp. 2124–2128.
[7] D. M.Mount,’Sequence and Genome Analysis “ Bioinformatics 2(2004), Cold Spring Harbor Laboratory Press ,pp. 1-564.
[8] A,Blais and B,Dynlacht David, “Constructing transcriptional regulatory networks”In Genes Dev. 2005.
[9] D. Huson and M.H., Steel, “Phylogenetic trees based on gene content”. Bioinformatics 20(13): 2044-2049 (2004)
[10] A.Phillips, D, Janies and W, Wheeler , “Multiple sequence alignment in phylogenetic analysis” Molecular Phylogenetics and Evolution [2000, 16(3):317-330]
[11] D, Gusfield. “Efficient algorithms for inferring algorithms for evolutionary trees”. Networks,1991,vol. 21,pp. 19–28
[12] B. M. E. Moret, U. Roshan, and T.Warnow. “Sequence-length requirements for phylogenetic methods”.In R. Guig´o and D. Gusfield, editors, Proceedings of the Second International Workshop on Algorithms in Bioinformatics, WABI 02, volume 2452 of LNCS, pages 343–356, Berlin, 2002. Springer Verlag.
[13] D. Sankoff. Minimal mutation trees of sequences. SIAM J. Appl. Math., 28:35–42, 1975.
[14] W. Fitch and E. Margoliash”. Construction of phylogenetic trees. Science”, 155:279–284, 1967.
[15] B. Efron and R. J. Tibshirani. “An Introduction to the Bootstrap.” CRC Press, Boca Raton, FL, 1998.
[16] D. Sharma and S. Rajasekaran, “A simple algorithm for (l, d) Motif search”, CIBCB,2009.
[17] S. Rajasekaran, S. Balla, and C.-H. Huang, "Exact Algorithms for Planted Motif Problems", Journal of Computational Biology, Oct 2005, Vol. 12, No. 8: 1117-1128.
[18] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar, “MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance and Maximum Parsiomny methods”, Molecular Biology and Evolution, 2011.