IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 31-36
www.iosrjournals.org

Tree Mining and Tree Validation Metrics: A Review

Swati Mittal[1], Ms. Geetika Munjal[2]
[1,2](CSE, SOET/ ITM University, India)

Abstract: *In this paper, various tree comparison metrices have been discussed, where tree is showing Structure information of species or other related data. Some algorithms enable us to find the distance between the two trees efficiently. This paper focuses on tree pattern mining and tree validation methods. While comparing species trees, we can even gain information about their evolution and also the relationship that exists between several organisms. A comprehensive comparison of various metrics is also shown taking common dataset of species.*

Keywords: *Maximum Agreement Subtree, Nodal Distance, Phylogenetic tree, Robinson Foulds-Distance*

I. INTRODUCTION

Tree mining is an important field of data mining where we can analyze data and extract useful information. Tree mining refers to finding frequent patterns in a forest of trees. In domains where we have to mine semi structured data like in web mining or bioinformatics, tree mining is very profitable. It is used to extract informative patterns from large sets of data but it is an expensive task. There are many algorithms which are helpful in finding frequent subtrees in a forest. In the domain of bioinformatics, tree mining can be used to analyze phylogenetic data sets and also in analyzing RNA structure [1]. We can discover common subtree patterns of several organisms when provided with several phylogenies(evolutionary trees) and come out with results based on their evolution as it is postulated that in the past all the biological species have some common ancestors by which they are linked. If we are able to find the relationships among different species, then this can be used to predict the functioning of genes. Comparing two trees is yet another important aspect where we can compare two trees either on the basis of their topology or calculating the difference between the various nodes in the tree etc. There are different methods used for the comparison that takes into account different techniques and features for their comparison. By comparing two trees we can come out with some useful information and data of our interest.

Phylogenetic trees can be compared and can be used to analyze the relationship between the different species. Analyzing relationships between different species might help us to understand the evolution of different organisms and can predict the function of genes. Comparing phylogenetic trees is a primitive task in the field of bioinformatics or specifically said "computational biology". The end result of comparing two phylogenetic trees can be the distance between them or the similarity or dissimilarity calculated in one way or the other. There are different tree comparison measures. Robinson-foulds distance, maximum agreement subtree, nodal distance algorithm, symmetric distance, finding frequent patterns in trees.

II. TREE COMPARISON METRICES

2.1 Nodal Distance Algorithm

It is the widely used method which is used to compare large set of phylogenetic trees and that too in small computational time. Here it is assumed that the trees which are to be compared must have the exact set of species or data. The Nodal Distance algorithm reflects the change that has arrived in the positions of several species present in the tree. Count of all the branches that occur in the path while going from one node to another makes up the Nodal Distance. After the calculation of all these values, Nodal Distance metric (ND) is calculated which is the sum of the differences of the nodal distances of the two trees. So, ND metric depicts the changes in the positions of the species [2].

The following steps are to be followed in order to compare the two trees using nodal distance algorithm: i) Find the nodal distance of the two trees and ii) Calculate the ND-metric.
Let us say we have two trees T1 and T2

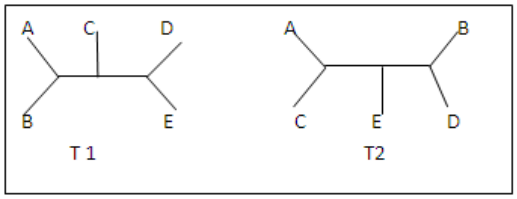

Fig.1. Two trees

Table 1.Nodal distance Table

Species	T1(Distance)	T2(Distance)	Difference
AB	2	4	2
AC	3	2	1
AD	4	4	0
AE	4	3	1
BC	3	4	1
BD	4	2	2
BE	4	3	1
CD	3	4	1
CE	3	3	0
DE	2	3	1

Table 1 calculates the distances of different species depicted in Fig 1 of T1 and T2. Now, we have to calculate the ND-metric which is the sum of the differences that is 10 in this case. Change in the positions of the species is represented by this value. A nodal distance algorithm is an efficient method for comparing two phylogenetic trees as it takes less computational time.

2.2 Robinson-Foulds Distance

Phylogenetic trees are unrooted trees. Sometimes there is a need to convert one tree into another [3]. Minimum number of operations that are required to transform one tree into another is known as edit distance. Computing edit distance is NP-hard. So, we focus on a distance measure, Robinson-Foulds Distance that takes into account the characteristics of two trees rather focusing on the transformations. It highlights the differences between the two trees based on outcomes rather on transformations. This metric is also known as Partition Metric**.** RF distance counts the number of edges present in one tree and compare it with that of the other tree and that too in linear time. Hence, we can say that RF distance focuses on the dissimilarity between the two phylogenetic trees. One of the major advantages of Robinson-Foulds distance measure is that, it does not rely on any tree editing operations like NNI, SPR or TBR it just depends on the present characteristics of the two trees. Let T be a tree with some leaves, internal edge is represented by e then E (T) is the set of all internal edges in T. A non-trivial bipartition is defined by $\prod$, then the set of bipartitions is represented by $\{\prod e \mid e \in E(T)\}$ and this way a tree is uniquely represented [2]. The Robinson-Foulds distance between two trees T1 and T2 is given by

$$Drf(T1,T2) = 1/2((|r(T1) - r(T2)|) + (|r(T2) - r(T1)|)) \ldots\ldots (1)$$

It gives the count of the bipartitions present in one tree and not the other. If there are n leaves in a tree then the bipartitions induced will be n-3 which is the largest possible RF distance between the two trees. RFdistance is very sensitive, because even small changes made in a tree will maximize the distance [4].

In the following Fig 2, we have two unrooted trees where e1 , e2 and e3 are the three edges that divides both the trees into three non-trivial bipartitions represented by{AC|FEBD,ACF|BD,ACEF|BD} in tree T3 and {AB|CDEF,ABC|DEF,ABCD|EF} in tree T4.

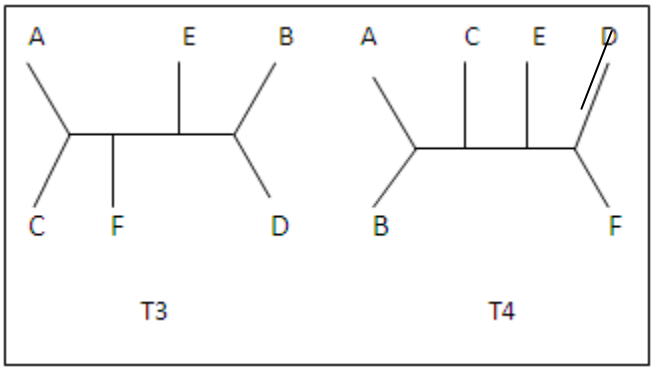

Fig.2. Two Unrooted Trees

By using equation 1, number of unique nodes in T3 is 6 and number of unique nodes in T4 is 6. So, RF-distance between the two trees can be calculated as 6+6/2 i.e. 6.

2.3 Maximum Agreement Sub Tree (MAST)

Maximum agreement subtree approach extracts maximum species about which we have confidence [5]. To extract the maximum agreement subtree of the two trees, the two trees must be rooted and their leaves must be drawn from the same set of species (or items).The MAST problem can be used for two evolutionary trees of species to find the consistency between them. In a rooted tree, the leaves represent the taxa and ancestor information is represented by the internal nodes [6].

Let T={T1,T2,T3,....Tn} is a set of trees .T5 and T6 be the two tress as shown in Fig 3 and L ={P,Q,R,S,T,U} be the set of labels then,

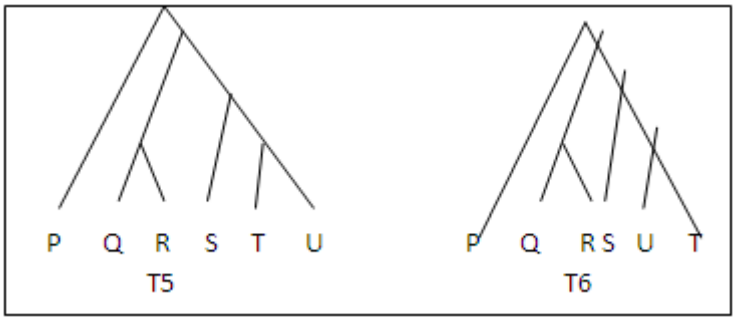

Fig.3. Two Rooted Trees

We can obtain maximum agreement tree from these two trees as follows in Fig 4:

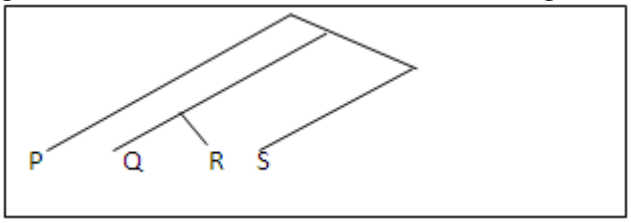

Fig.4. Aggreement Tree

So, we can define maximum agreement subtree as among all the agreement subtrees that can be formed with T, the subtree that has the maximum length is chosen as the maximum agreement subtree[7].

2.4 Frequent Patterns in Trees

The frequent tree mining aims at discovering all frequent subtrees from a database of trees represented by D. This large database of trees can also be referred as forest. Mining frequent patterns is a data mining technique where the goal is to find the complex interactions between the entities. Mining tree like patterns is the main focus of this paper. In the field of bioinformatics, discovering frequent patterns from different phylogenies

can help us to know about the evolutionary history of various organisms [8]. *Support* of a subtree S is defined as the total number of trees in the database D in which there must be atleast one appearance of subtree S. Also, the weighted support of subtree S is total number of occurrences of subtree S among all trees in the database D[9]. A minimum threshold is defined by the user and a subtree whose support is greater than the user defined threshold than that tree is said to be frequent. There are various frequent pattern mining algorithms as shown in Fig 6. All the algorithms follow the strategy used in well known Apriori Algorithm that is based on iterative pattern mining where we break each iteration into two phases:

i) Candidate Generation: Frequent patterns discovered in one iteration are used to generate potentially frequently candidates. We can merge two patterns whose size is k and consist of k-1 elements to generate candidates whose size is k+1.
ii) Support Counting: In this phase find the support of the frequent candidates, ignore the less frequent candidates and keep the actually frequent candidates.

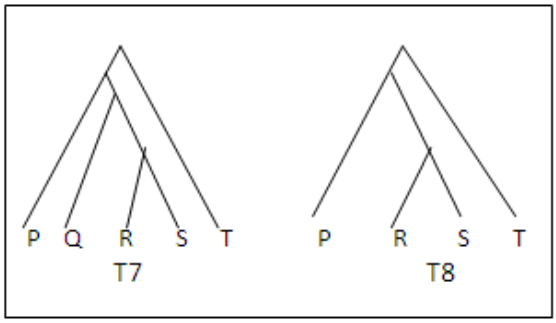

Fig.5. Rooted Trees

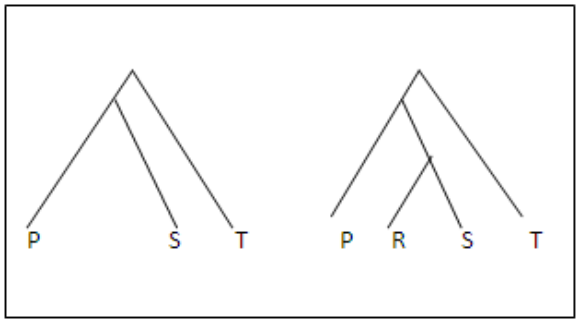

Fig 6: Frequent patterns of T7 and T8

III. DISCUSSION AND RESULTS

We have applied all the four tree comparison metrices on the two phylogenetic trees constructed from real data set of some organisms [10]. The two trees are as follows in Fig 7:

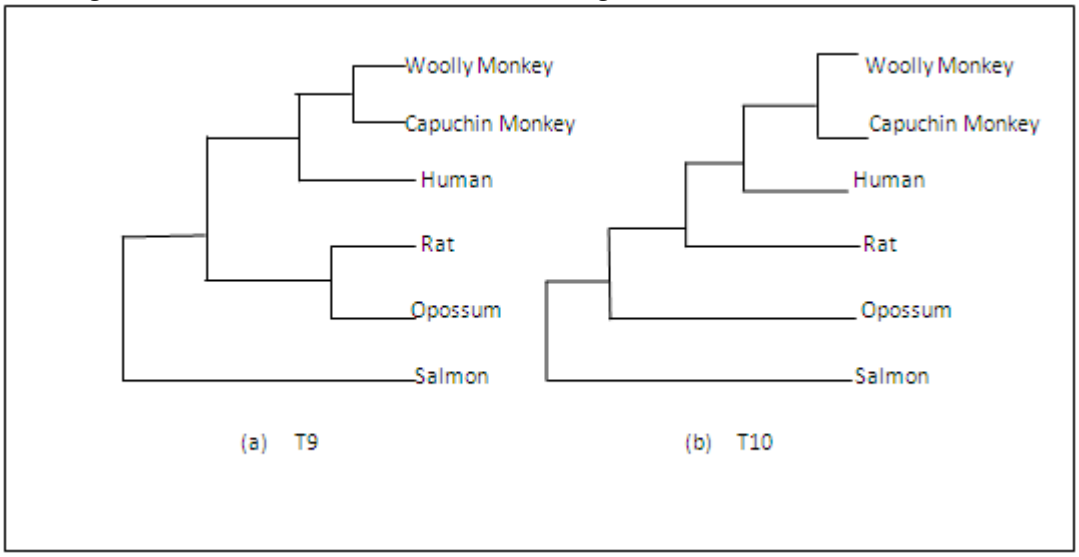

Figure 7

Fig.7. (a) Tree constructed using UPGMA method Of MEGA5 package
(b) Tree constructed using UPGMA method in MatLab

Results obtained are as follows:
Nodal Distance-metric for T9 and T10- 16,
Robinson Foulds for T9 and T10-Distance- 10
Maximum Agreement Sub Tree for T9 and T10 is shown in fig 8

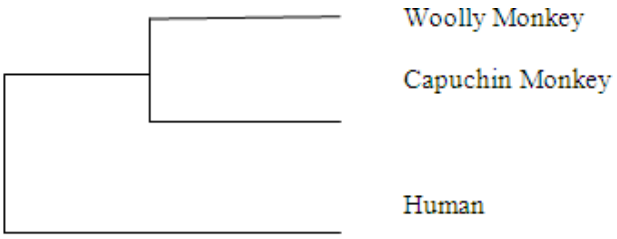

Fig.8. Maximum Agreement Sub Tree

Frequent Patterns for T9 and T10 is shown in Fig9

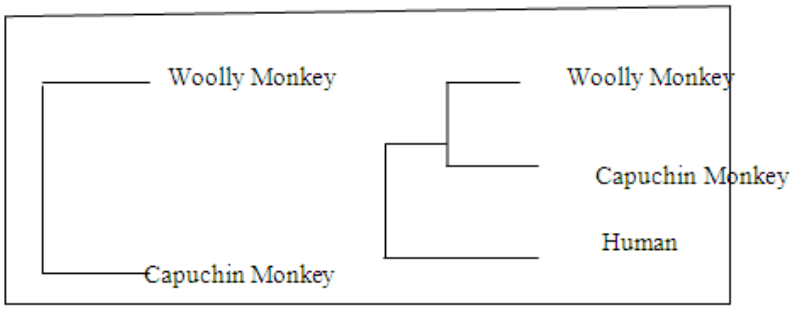

Figure 9: Frequent Patterns

A comparative study of all the tree comparison metrices studied above have been listed in Table 2.

Table 2.Comparative Study

Tree comparison metric	**Features**	**Advantages**	**Disadvantages**
Robinson-Foulds Distance	Finds the topological distance	Independent from any tree editing model	Sensitive to small changes
MAST	Finds the largest subtree	More descriptive	Complexity is NP hard
Nodal Distance Algorithm	Calculates nodal distance metric	Takes less computational time and is applicable to large datasets	Trees must have the same set of species
Frequent Pattern Miming	Finds all the frequent patterns occurring in the dataset	Finds complex interactions between the entities	Extracted patterns may be irrelevant

IV. CONCLUSION

Based on the Review of Various Tree metrics it can be concluded that RF is less discriminative than MAST that means MAST is more clearer in comparing the two trees. It distinguishes trees in a better way. On the other hand RF distance is sensitive as it responds to small changes made in a tree. We can compare the methods according to the features they incorporate. RF metrics finds the topological distance between the two trees without making any changes in the trees. By comparing the distances, we can say that which of the two trees are closer than the other depending on their structures. Since, RF- distance is sensitive to very small changes, as making small changes in a tree can maximise the distance. RF- distance uses binary weighting scheme, for example an edge (x, y) has weight 1 if the bipartions of both the trees are different otherwise it has weight 0. In order to extract more information through RF- distance measure, a different weighting scheme can be used.

REFERENCES

[1] Mohammed J. Zaki, Member, Efficiently Mining Frequent Trees in a Forest: Algorithms and Application**s,** IEEE, August 2005, Volume 17

[2] John Bluis and Dong-Guk Shin, Nodal Distance Algorithm: Calculating a Phylogenetic Tree Comparison Metric, Computer Science and Engineering University of Connecticut Storrs, CT 06269-3155, USA, Bioinformatics and Bioengineering, IEEE 2003, pp. 87-94.

[3] D. F. ROBINSON, L. R. FOULDS, Comparison of Phylogenetic Trees, MATHEMATICAL BIOSCIENCES 53,1981, pp.131-141
[4] Yu Lin,Vaibhav Rajan,and Bernard M.E.Moret, A metric For Phylogenetic Trees Based On Matching, IEEE, July 2012, Volume 9, pp.1014-1022
[5] Hong Huang and Yongji Li, MASTtreedist: Visualization of Tree Space based on Maximum Agreement Subtree, Journal of Computational Biology,Issue:Jan 7,2013,pp.42-49
[6] Vincent Berry & Franc ois Nicolas, Maximum agreement and compatible super trees, Journal of Discrete Algorithms, September2007,Volume 5,pp. 564-591
[7] Daniel M. Martin a, Bhalchandra D. Thatte,, The maximum agreement subtree problem, Discrete Applied Mathematics,2013,pp.1805-1817
[8] Tatsuya Asai1, Hiroki Arimura1, Takeaki Uno2, and Shin-ichi Nakano3,Discovering Frequent Substructures in Large Unordered Trees, Conference on discovery science, 2003, pp.47-61
[9] Aída Jiménez Fernando Berzal Juan-Carlos Cubero, Mining Different Kinds of Trees: A Tree Mining Overview, Department of Computer Science and Artificial Intelligence ETSIIT, pp.343-352
[10] Manoj Kumar Gupta, Rajdeep Niyogi, and Manoj Misra, A framework for Alignment-free methods to perform similarity analysis of biological sequence, IEEE, 2013, pp.337-342