2011

Improving the Approximation Ratio of the Maximum Agreement Forest (MAF) on k trees and Estimating the Approximation Ratio of the Acyclic-MAF on k trees

Puspal Bhabak
University of Windsor

IMPROVING THE APPROXIMATION RATIO OF THE MAXIMUM AGREEMENT FOREST(MAF) ON K TREES AND ESTIMATING THE APPROXIMATION RATIO OF THE ACYCLIC-MAF ON K TREES

by

Puspal Bhabak

A Thesis
Submitted to the Faculty of Graduate Studies
through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2011

Improving the Approximation Ratio of the Maximum Agreement Forest(MAF) on k trees and Estimating the Approximation Ratio of the Acyclic-MAF on k trees

by

Puspal Bhabak

APPROVED BY:

Dr. Xiaobu Yuan
School of Computer Science

Dr. Myron Hlynka
Department of Mathematics and Statistics

Dr. Asish Mukhopadhyay
Advisor
School of Computer Science

Dr. Dan Wu
Chair of Defence
School of Computer Science

20 May, 2011

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Molecular phylogenetics has long been a well-established field of scientific research where the structure of the phylogenetic tree has been analysed to know about the evolutionary process of the organism. In biology, leaf-labelled trees are widely used to describe the evolutionary relationships. In this setting, the leaves of the tree correspond to extant species, and the internal vertices represent the ancestral species. However, for certain species, evolution is not completely tree-like. Reticulation events such as horizontal gene transfer (HGT), hybridization and recombination play a significant role in the evolution of the species. Suppose we have two phylogenetic trees each of which is for a gene of the same set of species. Due to reticulate evolution the two gene trees, though related, appear different. As a result, instead of the tree like structure, a phylogenetic network is widely viewed as a most suitable tool to represent reticulation. A phylogenetic network contains hybrid nodes for the species evolved from two parents. The distance between two phylogenetic trees can be computed with the help of a Maximum Agreement Forest (MAF) of those trees. The fewer components in MAF, the greater is the similarity between the two trees. This number of components in that agreement forest shows how many edges from each of the two trees need to be cut so that the resulting forest agree after all forced edge contractions. Recent research reveals that the MAF on k trees can be approximated within a ratio of 8. We have given a better approximation ratio for the MAF on k trees and also provide an approximation ratio for Maximum Acyclic Agreement Forest (MAAF) on k ($\geq$2) trees.

Dedication

to my

mother and father

and my loving wife

Acknowledgements

I express my sincere gratitude to my advisor, Dr. Asish Mukhopadhyay for giving me the opportunity to work under his supervision as well as for his guidance and support in my research. I would like to thank Dr. Xiaobu Yuan for his valuable suggestions and comments. I would also like to convey my sincere thanks to Dr. Myron Hlynka for his advice and help.

I am thankful to my friends and colleagues in the Univeristy of Windsor for their help and frequent assistance throughout my work and stay. Last but not the least, I would like to thank my family for their love and confidence in my abilities that helped me to progress.

Table of Contents

Page

Author's Declaration of Originality iii

Abstract iv

Dedication v

Acknowledgements vi

List of Figures ix

1 Introduction 1

2 Basic Concepts 5

2.1 Graph 5

2.2 Binary Tree 7

2.3 Phylogenetic network 7

2.3.1 Properties of a hybrid network 8

2.4 Agreement Forest 9

2.5 Subtree Prune and Regraft(SPR) 13

2.6 Tree Bisection and Reconnection(TBR) 14

2.7 Nearest Neighbor Interchange(NNI) 15

3 Literature review 16

3.1 Hein et al. 16

3.2 Allen and Steel 18

3.3 Baroni et al. 21

3.4 Bordewich et al. 22

3.4.1 Fixed-Parameter Tractable 23
3.5 Rodrigues et al. 23
3.6 Other work in this area 24
3.7 Literature review with $k(\geq 2)$ phylogeny trees 25
4 Our Contribution 32
4.1 Approximation Algorithms 32
4.1.1 Terms and Definitions 32
4.1.2 3-Approximation for MAF on k binary trees (rooted) 35
4.1.3 2-approximation for MAAF on k rooted binary trees 41
4.1.4 Heuristic 45
4.2 Implementation 46
5 Conclusion 50
References 52
Vita Auctoris 60

List of Figures

1.1 Evolutionary Tree 2

2.1 (a) Directed Graph (b) Undirected Graph 6
2.2 Hybrid Network 8
2.3 Three rooted phylogeny trees 10
2.4 Agreement Forest of trees in Fig 2.3 11
2.5 A rooted tree T with leaf-set [a,b,c,d,e] , subtree T(L) and restriction subtree $T|$L where L=[a,c,e] 12
2.6 rSPR Operation 13
2.7 TBR Operation 14
2.8 This operation exchanges between B and C. Another NNI operation is possible between B and D 15

3.1 Effect of a recombination. The genetic material (thick lines) that is now on one sequence was, just before the recombination, on two sequences, and the rest (thin lines) of the genetic material most likely does not have any descendent in the present sample. Recombination occurs at rp. [31] 17
3.2 Two rooted binary phylogenetic trees reduced under Rule 1 19
3.3 Two rooted binary phylogenetic trees reduced under Rule 2 20
3.4 T_1, T_2 are two phylogenetic trees and H is the hybrid network of T_1 and T_2 20

3.5 Topology of a tree with 3 leaves 25
3.6 Conflicted Set of Triples. Any one leaf can be deleted from each tree to remove the conflict 26
3.7 (i) A phylogeny tree T. The thick lines (r_U) show the edges being cut to get the maximal size edge-disjoint set of triples and is stored in M1. (ii) A set of maximal size edge-disjoint set of triples have been cut from T according to Lemma 3.12 [18] 29
3.8 If the maximal size edge-disjoint triple ((i,j),m) obtained from Fig 3.7 is a conflicted triple, s_U is cut to remove the conflict and is added to M1. The leaf-set (g,h,k,l) will form the combination of triples in M2 . 29
3.9 (i) U1 = ((g,h),k) and U2 = (h,(k,l)) and w in this case is an edge. (ii) U1 = ((g,h),x) and U2 = (x,(k,l)) and w is a vertex. In either case remove the thick edges in case of incompatible triples in U1 and U2 . 30
3.10 i) 2 phylogeny trees T1 and T2. ii) Components in a forest which are topologically compatible in both T1 and T2. iii) For Separability check the components are in conflict with T1. So the graph has been formed with conflicted edges. The vertex cover of this graph will give d. So remove d from ii to get MAF. 31

4.1 Minimum incompatible triple is $ab|c$ as $ab|c < xy|c$ 33
4.2 Overlap Components [[15], p:6] 34
4.3 Central Lemma 34
4.4 Layout of a minimal incompatible triple[[15],p:5] 36
4.5 MAF of T_1 and T_2 38
4.6 Directed Graph obtained from the MAF components in Figure 4.5. There is a cycle between t_1 and t_2. 40
4.7 MAAF obtained from the Directed Graph in Figure 5.6 41
4.8 An incompatible triple $ab|c$ 45

4.9 Trees with leaf-set [A,B,C] and the MAF 47
4.10 Trees with leaf-set [A,B,3,4,D,C] and the MAF 48
4.11 Trees with leaf-set [A,B,3,4,D,E,C] and the MAF 49

Chapter 1

Introduction

Phylogenetic trees, or evolutionary trees are used in evolutionary biology to represent the evolutionary history of biological entities such as present-day species or genes. In a rooted phylogenetic tree, the leaves are uniquely labelled by the extant species, while the internal nodes represent the ancestors. These are generally not labelled. The universal common ancestor of all the species is represented as the root of the tree. The out-degree of an internal node is the number of its children. The distance between two nodes in an evolutionary tree represents evolutionary distance such as time or number of mutations. In figure 1.1 the evolutionary history of the Protozoan Ancestors have been shown with the present-day species as the leaves of the tree.

This kind of representation is appropriate for many groups of species which include the mammals. But it has been observed that not all groups follow the same distribution of evolutionary patterns. Sometimes reticulation events come into play. This type of evolution does not follow the tree like evolutionary process, rather the species under reticulation events form a composite of genes derived from different ancestors. The processes include hybridization, horizontal gene transfer and recombination.

This thesis concentrates primarily on hybridization. It has been found through re-

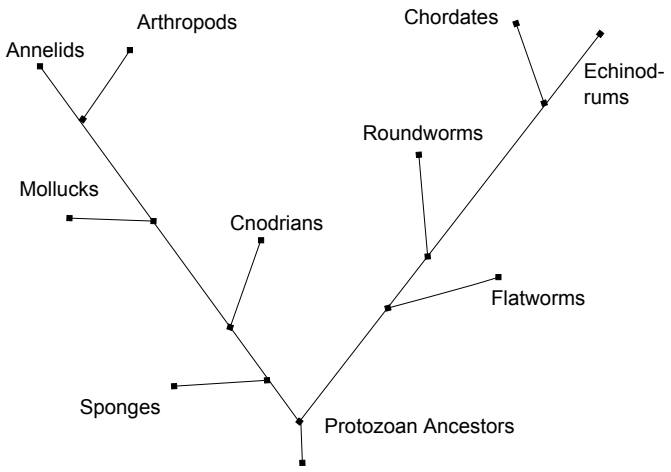

Figure 1.1: Evolutionary Tree

search over the years that the evolutionary history of Eukaryotes contains hybridization events that include certain groups of plants, birds and fish. Spontaneous hybridization events have also been reported in the evolutionary history of some mammals and even primates. Study on hybridization shows that at least 25% of plant species and 10% of animal species, mostly the youngest species are involved in hybridization events [43].

Several techniques have been devised to reconstruct phylogenetic trees from a given set of species. Biologists are interested in determining the 'distance' between two such trees. Distance metrics such as NNI (Nearest Neighbor Interchange), SPR (Subtree Prune and Regraft) and TBR (tree bisection and reconnection) have been proposed in [53] for measuring the distance between the two phylogenetic trees. In an pioneering paper Allen and Steel [2] proposed algorithms for estimating these distances. The hybridization number and the rooted SPR (rSPR) distance have proven to be a very useful tool in estimating the reticulation events that have occurred. Baroni et al. [5] showed that rSPR distance provides a lower bound on the number of reticulation events.

Computing hybridisation number, rSPR and TBR distances have been shown to be NP-hard problems [58]. Hence the interest in approximation algorithm and fixed parameter tractable algorithm. Hein et al. [32] came up with the idea of Maximum Agreement Forest(MAF) as a new tool to determine the distance between two phylogenies. They showed that a 3-approximation ratio algorithm exists for computing the MAF for 2 trees is 3. They proposed a NP-hardness proof for computing SPR distances. Allen and Steel [2] extended the NP-hardness idea of Hein et al. [32] to prove that maximum agreement forest problem is NP-hard. In fact they rectified certain errors in the paper by Hein et al. [32] in their paper and showed that the TBR distance between two trees is equal to the number of components in MAF. Rodrigues et al. [53] with the help of certain instances showed that approximation ratio for the size of MAF cannot be less than 4 which disproves the 3-approximation claim of Hein. The approximation ratio later has been improved to 3 by Bordewich et al [15] for the rSPR distance between two trees. Bordewich and Semple [16] showed that the SPR distance between two rooted trees is also equal to the number of components in MAF. Baroni et al. [5] introduced the concept of Maximum Acyclic Agreement Forest(MAAF) and showed that the hybridisation number of two trees is one less than the number of components in a MAAF. Chataigner [18] obtained an 8-approximation ratio for the maximum agreement forest on $k(\geq 2)$ trees.

In another approach to these problems, attempts have been made to find the Fixed Parameter Tractable (FPT) algorithm when the distance between the two trees is small. Allen and Steel [2] introduced certain tree reduction rules to obtain an FPT algorithm for the TBR distance and the running time is $O(k^{3k} + p(n))$ where k is the distance between two trees and $p(n)$ is a polynomial function of input size n. Bordewich et al. [15] proposed an FPT algorithm for computing the rSPR distance of the two trees whose complexity is in the order of $O(4^k k^4 + n^3)$ when k is small.

Our contributions in this thesis are the following:

A)A better approximation ratio in deriving the Maximum Agreement Forest on k rooted phylogenetic trees

B)Estimate the approximation ratio of Maximum Acyclic Agreement Forest on k trees.

C)An approximation algorithm for finding the rSPR distance between 2 trees, whose approximation ratio we conjecture to be 2.

D)We have implemented the algorithm given in [15] on the 3-approximation ratio for the SPR distance on two rooted binary phylogenetic trees in Java.

Chapter 2

Basic Concepts

This chapter introduces the basic concepts which have been used in this thesis.

2.1 Graph

This section deals with the prelimaries of graph theory. There are two kinds of graph: directed and undirected.

A ***directed graph*** (or digraph) G=(V,E) consists of a set of vertices V and a set of directed edges E such that for each edge $e \in E$ there is a pair of vertices u,$v \in V$ connected at the two end-points of e. Each e is an ordered pair (u,v) so that the roles of u and v are not interchangeable and we call u the tail of the edge and v the head.

In an ***undirected graph*** G=(V,E), the set of edges E consists of unordered pair of vertices instead of having the ordered pairs. If an edge $e \in E$ can be represented by the pair of vertices (u,v) where u,$v \in$ to the set of vertices V, then (v,u) also represents the same edge e.

Many definitions for the directed and the undirected graphs are the same though there are certain terms which have different meanings in these two contexts. If (u,v) is an

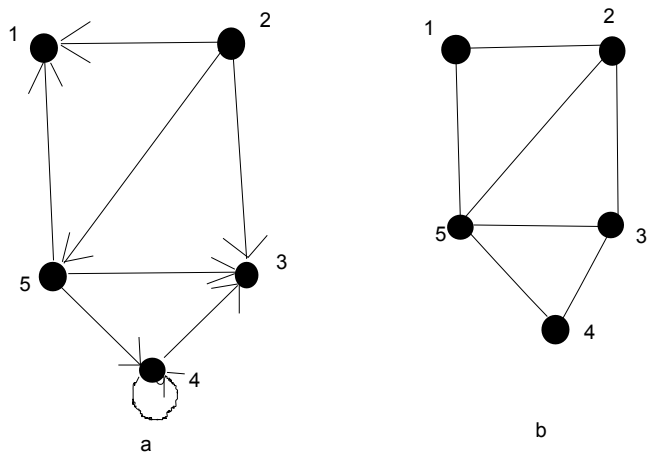

Figure 2.1: (a) Directed Graph (b) Undirected Graph

edge in a graph $G = (V,E)$ we say that vertex v is **adjacent** to vertex u. When the graph is undirected, the adjacency relation is symmetric. When the graph is directed this symmetry is not always true. In figure 2.1(a) and (b) vertex 1 is adjacent to vertex 2 since edge (2,1) belongs to both graphs. Vertex 2 is not adjacent to vertex 1 in figure 2.1(a) as the edge (1,2) does not belong to the graph.

The **degree** of a vertex in an undirected graph is determined by the number of edges incident to it. In figure 2.1(b) the degree of the vertex 5 is 4. In a directed graph, the **out-degree** of a vertex is the number of edges leaving the node and the **in-degree** of the vertex is the number of edges entering it. So, in figure 2.1(a) the out-degree of vertex 5 is 3 and the in-degree is 1.

A sequence of vertices $<v_0,v_1,v_2,...,v_k>$ such that $u=v_0$ and $v=v_k$ and $(v_{i-1},v_i) \in$ E, for $i = 1, 2, ..., k$ forms the **path** of length k from a vertex u to a vertex v in a graph $G= (V,E)$. The length of the path determines the number of edges in the path. If v is reachable from u by a path p we write $u \sim v$. In a directed graph a path $<v_0,v_1,v_2,...,v_k>$ forms a **cycle** if $v_0 = v_k$ and there is at least one edge in the

path. In an undirected graph a path $<v_0,v_1,v_2,...,v_k>$ forms a cycle if $v_0 = v_k$ and $v_0,v_1,v_2,...,v_k$ are distinct. A graph with no cycle is called an **acyclic graph**. We say an undirected graph is **connected** if every pair of vertices is connected by a path.

2.2 Binary Tree

A binary tree T is a data structure in which there are 3 disjoint set of nodes: a root node, the subtree immediately to the left of root called **left subtree** and the subtree immediately to the right of the root known as the **right subtree**. Every internal node in a binary tree other than the root is of degree 3. The nodes having degree 1 are called the leaves. The binary tree that contains no nodes is called a null tree. In a binary tree the edges are directed away from the root. This gives an idea about the parent-child relationship in a rooted binary tree [16]. Consider a node u in a rooted tree T with root X. Suppose v be any node on the unique path from X to u in T. Then v is known as the **ancestor** of u and u is the **descendant** of v. The length of the path from the root X to a node u is the **depth** of u in T. The largest depth of any node in T is the **height** of T.

2.3 Phylogenetic network

In a phylogenetic tree The ancient ancestor is at the root of the tree. The leaf set constitues the recent species and the internal nodes represent their ancestors. Due to reticulation events, instead of the tree like structure, phylogenetic network is mostly viewed as a tool to represent the reticulation which contains the hybrid nodes for the species evolved from two parents [5]. Figure 2.2 shows the hybrid nodes c and f (highlighted) originating from two different ancestors.

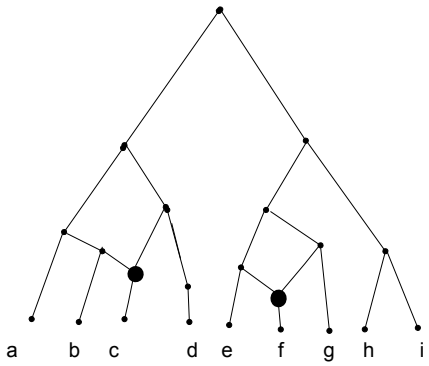

Figure 2.2: Hybrid Network

2.3.1 Properties of a hybrid network

For a digraph D and a vertex v of D, we denote the in-degree and out-degree of v by $d^-(v)$ and $d^+(v)$, respectively. A hybrid phylogeny on the set of present-day species X represented by the network H consisting of:

- a rooted acyclic digraph D in which the root has out-degree at least two and, for all vertices v with $d^+(v) = 1$, we have $d^-(v) \geq 2$
- the set of vertices of D with out-degree zero forms X.

It can be understood that, if $|X| = 1$, then the digraph consists of an isolated vertex X. The set X corresponding to the set of present-day species is called the label set of H and is denoted by $L(H)$. Vertices of in-degree at least two are called **hybridization vertices**. These vertices represent an exchange of genetic information between hypothetical ancestors. For a hybrid H on X with root ρ the hybridization number of H, denoted by $h(H)$ is:

$$h(H) = \sum_{v \neq \rho} (d^-(\text{v}) \text{ - } 1)$$

A rooted binary phylogenetic tree is a special type of hybrid phylogeny in which the root has degree two and all other interior vertices have at least degree three. For two rooted binary phylogenetic X-trees T and T' [5], we get

$$h(T,T') = min\ [h(H) : H \text{ is a hybrid on } X \text{ that displays } T \text{ and } T']$$

2.4 Agreement Forest

Let T and T' be two rooted binary phylogenetic trees. We denote the set of leaf labels of T as $L(T)$, the set of branches as $E(T)$. An agreement forest F for T and T' is a collection of rooted binary phylogenetic trees t_1, t_2,..., t_n such that:

- for any tree t_i, $L(t_i) \in L(T)$ and the union of $L(t_i)$ is equal to $L(T)$
- for each t_i, the minimal subtree connecting the nodes in $L(t_i)$, denoted as $S(t_i)$, is identical to t_i when nodes with degree two of $S(t_i)$ are contracted
- for any two trees t_i and t_j, $S(t_i)$ and $S(t_j)$ are node disjoint

The size of a forest is the number of trees in the forest. An agreement forest is obtained by cutting the same number of branches from both T and T' and after cleanup gives rise to the same set of trees [60]. Agreement forests are an invaluable tool for analyzing and understanding tree rearrangement operation. It can be observed that the deleted edges are those which do not agree in T and T' which suggest that they represent the different paths of genetic inheritance i.e. hybridization events. An agreement forest for T and T' is a Maximum Agreement Forest (MAF) if, amongst all agreement forests for T and T', it has the smallest number of components.

The paper in [5] made an observation on the hybridization number problem which they termed as Maximum Acyclic Agreement Forest (MAAF). This observation excludes agreement forests in which any vertex in the associated hybrid phylogeny inherits

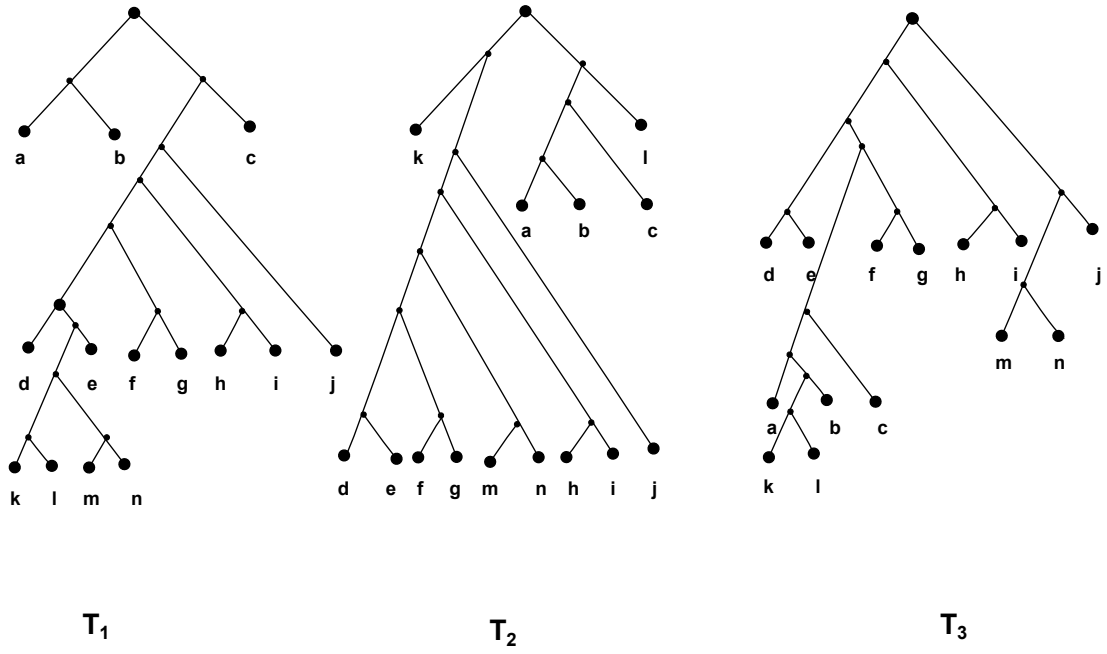

Figure 2.3: Three rooted phylogeny trees

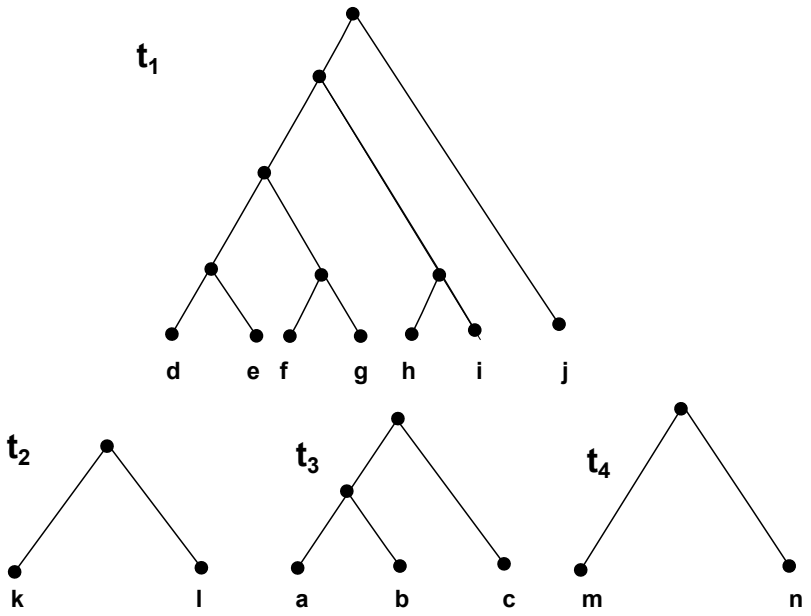

Figure 2.4: Agreement Forest of trees in Fig 2.3

genetic information from its own descendants.

Let $F_A = [T_1, T_2, . . . , T_k]$ be an agreement forest for T and T'. Let G_F be the directed graph whose vertex set is F_A and there is an edge from T_i to T_j if $i \neq j$ and either

- the root of $T(L(T_i))$ is an ancestor of the root of $T(L(T_j))$, or
- the root of $T'(L(T_i))$ is an ancestor of the root of $T'(L(T_j))$.

Since F_A is an agreement forest, the roots of $T(L(T_i))$ and $T(L(T_j))$, and the roots of $T'(L(T_i))$ and $T'(L(T_j))$ are not the same. We say that F_A is an acyclic-agreement forest if G_F is acyclic. If F_A contains the smallest number of components over all acyclic-agreement forests for T and T', we say that F_A is a Maximum Acyclic Agreement Forest (MAAF) for T and T'. So, intuitively we can say a forest is a MAAF if the forest is a MAF and acyclic.

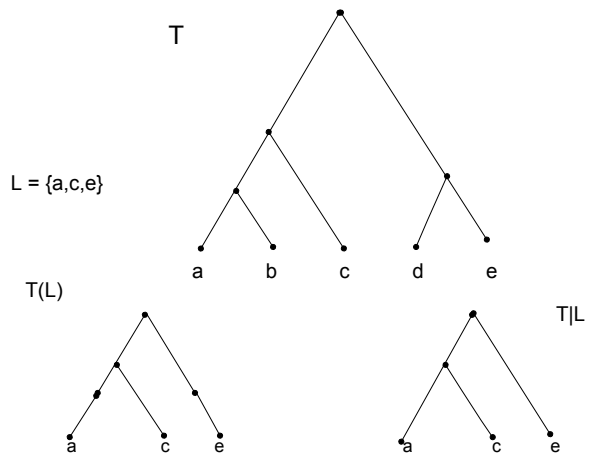

Figure 2.5: A rooted tree T with leaf-set [a,b,c,d,e] , subtree T(L) and restriction subtree $T|$L where L=[a,c,e]

In [5], Baroni et al. established and proved a fundamental relation between hybridization number and MAAF by the following theorem.

The hybridization number of T and T' is equal to the size of the MAAF for T and T' minus one.

Hence, it is essential to estimate the MAAF of two phylogenetic trees in order to know their hybridization number.

Definition 1. For a tree T and a subset L of the leaf-set of T, the *subtree* induced by L, noted as $T(L)$ is the tree defined as the smallest connected subgraph of T containing L. The *restriction* of T to L denoted as $T|_L$ is obtained from $T(L)$ by suppressing any degree-two vertices other than the root. (See Figure 2.5)

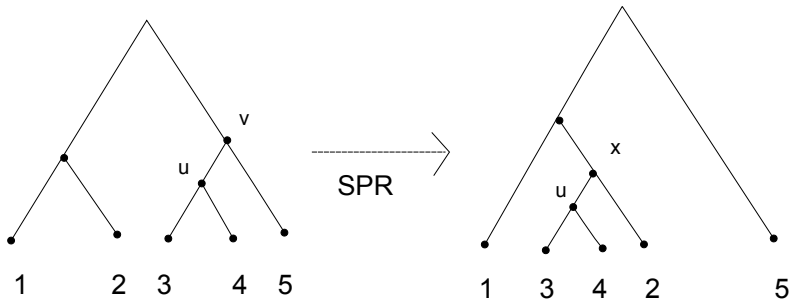

Figure 2.6: rSPR Operation

2.5 Subtree Prune and Regraft(SPR)

Another very important tool used to understand reticulate evolution is the subtree prune and regraft (SPR) [15] approach which measures the distance between phylogenies. If there are two phylogenies having the similar set of species but reticulation has occurred, then this inconsistency in the parent-child relationship between the two trees can be explained by the subtree prune and regraft operation. Given a subtree T, an SPR operation on a particular edge $e = [u,v]$ in T divides the tree into two subtrees T_u and T_v having the vertices u and v respectively. In order to reattach one subtree say T_u to a different edge in T_v it bisects another edge $f = [u',v']$ of T_v at x and adds an edge between u and x. Finally the degree-two vertex v is contracted. This kind of operation can take place in both rooted and unrooted trees. SPR(T,T') measures the minimum SPR operations required to transform T to T'. Figure 2.6 illustrates the SPR operation in a rooted tree.

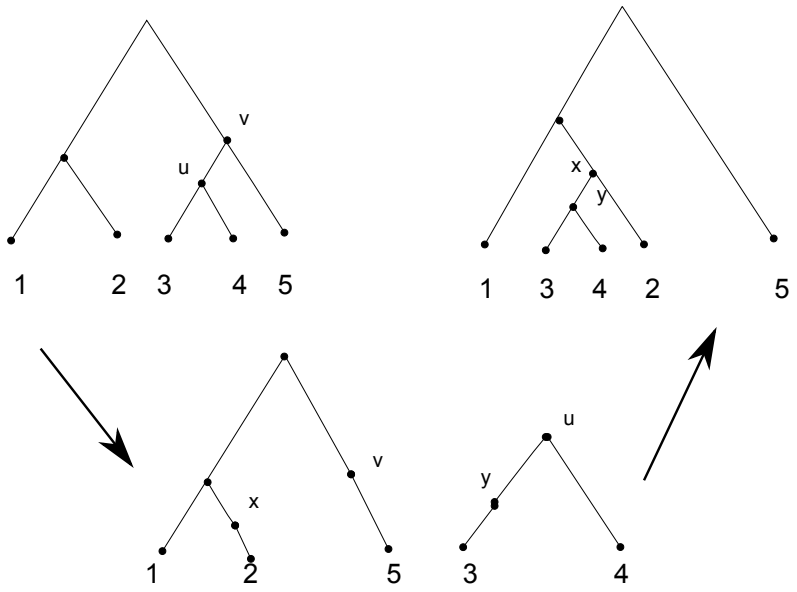

Figure 2.7: TBR Operation

2.6 Tree Bisection and Reconnection(TBR)

This method used to measure the distance between phylogenies works similar to SPR with a slight modification. Given a subtree T, a TBR operation on a particular edge $e = [u,v]$ in T divides the tree into two subtrees T_u and T_v having the vertices u and v respectively. In order to reattach one subtree say T_u to a different edge in T_v it bisects an edge in each of T_u and T_v at x and y respectively and adds an edge between x and y. Finally the degree-two vertices u and v are contracted [2]. This kind of operation can take place in both rooted and unrooted trees. TBR(T,T') measures the minimum TBR operations required to transform T to T'. Figure 2.7 illustrates the TBR operation in a rooted tree.

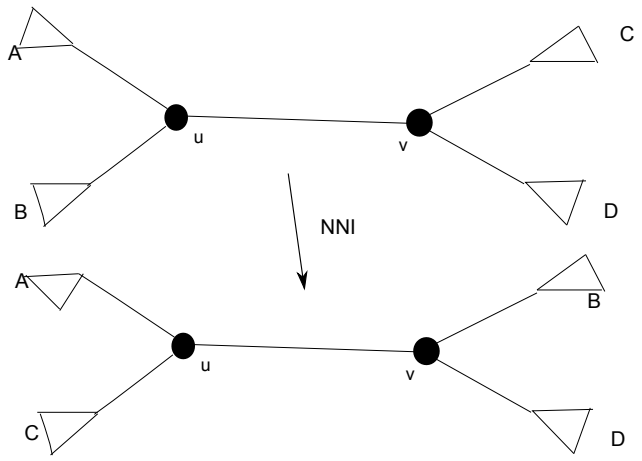

Figure 2.8: This operation exchanges between B and C. Another NNI operation is possible between B and D

2.7 Nearest Neighbor Interchange(NNI)

This metric for measuring the distance between phylogenies has been introduced in [44] and [52]. In a NNI operation two subtrees which are separated by an internal edge can be swaped. By an internal edge (u, v) in a tree we mean neither u nor v is a leaf of the tree. The operation only operates on the internal edge. NNI(T,T') measures the minimum NNI operations required to transform T to T' [20]. Figure 2.8 explains an NNI operation.

Each of the metrics, TBR distance, rSPR distance and hybridization number, have been proved to be one less than the number of components in a Maximum Agreement Forest (MAF) [58].

Chapter 3

Literature review

In this chapter we deal with the work done about the computational aspect on phylogeny trees. To the best knowledge of our survey in this area, Hein (1993) [31] started this area with his heuristic method to reconstruct the history of sequences subject to recombination. Sections 3.1 to 3.6 contain the survey of the work done with 2 phylogeny trees. In Section 3.7 we have discussed about the work done on k trees.

3.1 Hein et al.

Hein (1993) [31] in his paper presented a heuristic method to reconstruct the history of sequences due to recombination. In his paper he has shown the pictorial representation of the recombination. It has been proposed in his paper that the evolution of a sequence with k recombinations could be described by k recombination points and $k+1$ trees describing the evolution of the $k+1$ intervals, where two neighboring trees were either identical or differed by the transfer of one subtree within the whole tree.

The heuristic algorithm in [31] generates trees that are one recombination away from a given tree. The algorithm recursively visits all possible subtrees by visiting all internal edges. For every edge visited there will be a left and a right subtree. The left subtree can be moved to all possible edges in the right subtree producing all trees

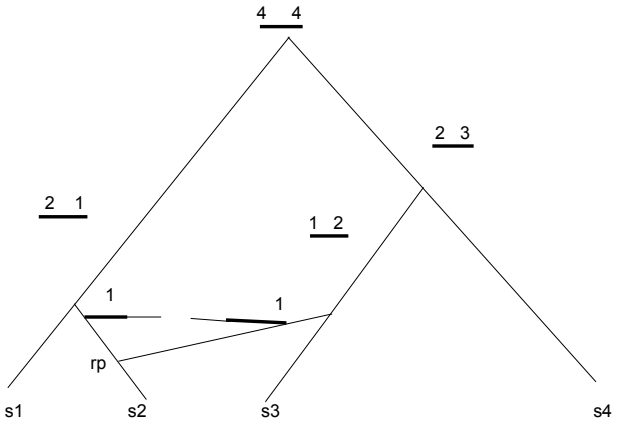

Figure 3.1: Effect of a recombination. The genetic material (thick lines) that is now on one sequence was, just before the recombination, on two sequences, and the rest (thin lines) of the genetic material most likely does not have any descendent in the present sample. Recombination occurs at rp. [31]

one recombination away from a given tree.

Hein et al.(1996) extended their work [32] to show that computing the subtree-transfer distance between two evolutionary trees is NP-hard and gave an approximation algorithm with performance ratio 3. The idea of Maximum Agreement Forest came from their work but it was not clearly defined. The basic idea behind this algorithm is to select a pair of sibling leaves (a, b) in the first tree T_1 at a time. If the pair a and b are siblings in the second tree T_2, this pair is replaced with a new leaf labeled (a, b) in both the trees. Otherwise, T_2 is being cut until a and b become siblings or separated. This has been handled by considering 5 different cases. They have proved a very useful relation which is stated below:

Lemma 3.1 [32]: *The size of a MAF of T_1 and T_2 is one more than their subtree-transfer distance*

3.2 Allen and Steel

Allen and Steel (2000) in their paper discusses the problem of determining how far apart two reconstructed trees are from each other or from the true historical tree. The metrics they investigated are Nearest Neighbour Interchange (NNI), Subtree Prune and Regraft (SPR), and Tree Bisection and Reconnection (TBR).

The main contributions in their paper are:

Lemma 3.2 [2]: *1. NNI* $\subseteq$ *SPR* $\subseteq$ *TBR*
2.If $d_\theta(T_1,\ T_2)$ *denotes the minimum number of* θ *operations required to transform the unrooted binary trees* T_1 *to* T_2 *where* $\theta \in$ *[NNI, SPR, TBR] then*
a. $d_{TBR}(T_1,\ T_2) \leq d_{SPR}(T_1,\ T_2) \leq d_{NNI}(T_1,\ T_2)$
b. $d_{SPR}(T_1,\ T_2) \leq 2 * d_{TBR}(T_1,\ T_2)$

Allen and Steel [2] rectified certain errors in Lemma 3.1 [32] and showed that this Lemma does not hold true for unrooted trees and also for SPR transformations. But it is true if TBR operations are taken into consideration. This fact is established in the following lemma:

Lemma 3.3 [2]: *Suppose we have two unrooted binary trees* T,T' *with* $L(T) = L(T')$. *Then,*
1. $d_{TBR}(T,T') = m(T,T')$
2. $d_{SPR}(T,T') \geq m(T,T')$ *where* $m(T,T')$ *is the size of MAF*(T,T') *- 1.*

They introduced the concept of Fixed Parameter Tractable (FPT) in measuring the distance between two trees and showed the running time for the TBR distance is $O(k^{3k} + p(n))$ where k is the distance between two trees and $p(n)$ is a polynomial

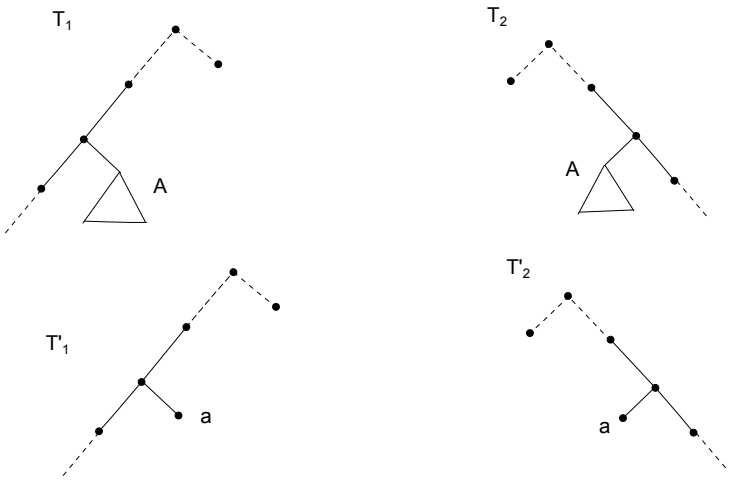

Figure 3.2: Two rooted binary phylogenetic trees reduced under Rule 1

function of input size n. In order to establish the FPT, they have kernalised the problem, that is the size of the problem has been reduced in such a way so that the answer to the reduced problem is same as that of the original problem. In order to kernalize the size of the problem in measuring the SPR or TBR distance they proposed to apply the following 2 rules repeatedly:

Rule 1: Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label.

Rule 2: Replace any chain of pendant subtrees that occurs identically in both trees by three new leaves with new labels correctly orientated to preserve the direction of the chain.

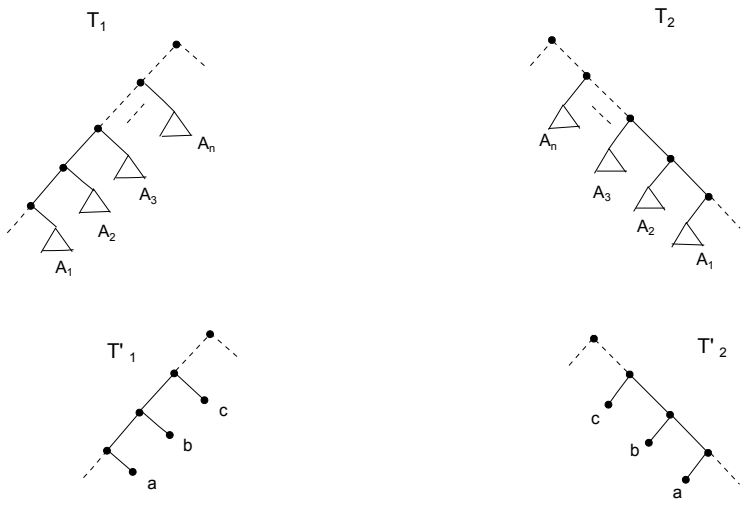

Figure 3.3: Two rooted binary phylogenetic trees reduced under Rule 2

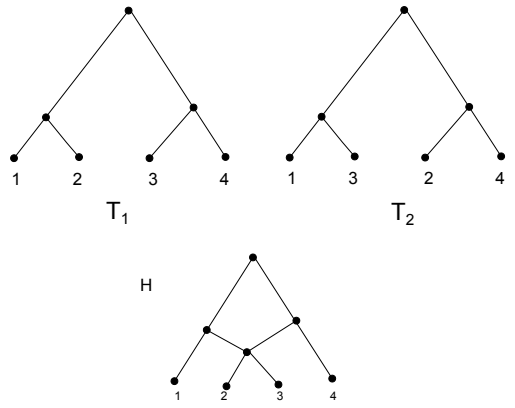

Figure 3.4: T_1, T_2 are two phylogenetic trees and H is the hybrid network of T_1 and T_2

3.3 Baroni et al.

Baroni et al.(2004) in their paper [6] analyzed Acyclic directed graphs (ADGs) which have been viewed as more appropriate for representing certain evolutionary relationships, and have developed a framework for the analysis of these graphs which are termed as hybrid phylogenies. Their work determines a hybrid phylogeny from a given set of phylogenetic trees which shows the smallest number of hybridisation events. They derived a very important equation:

$h(H) \geq |V|$ - $2|\chi| + 1$
where $h(H)$ is the hybridisation number, V is the vertex set and χ is the leaf-set.

Baroni et al.(2005) in their paper [5] gave a very clear definition about the concept of Maximum Agreement Forest and the conditions which need to be satisfied in order to get a MAF between two rooted binary phylogenetic trees. The paper also introduced for the first time the idea of Maximum Acyclic Agreement Forest(MAAF) and the way to determine the MAAF from MAF by removing the cycles. The most significant contribution of this paper has been the mathematical derivation of the following Theorem:

Theorem 3.4 [5]: *Let T and T' be two rooted binary phylogenetic trees. Then $h(T,T') = m_g(T,T')$,*
where $h(T,T')$ is the hybridisation number and $m_g(T,T')$ is the number of components in $MAAF(T,T')$ minus one

The paper also established the upper and lower bounds for $h(T,T')$.

Theorem 3.5 [5]: *Let $|\chi| = n$ and let T and T' be two rooted binary phylogenetic*

χ-trees. Then for all $n \geq 2$,
$d_{rSPR}(T,T') \leq h(T,T') \leq n\text{-}2$

3.4 Bordewich et al.

Bordewich et al.(2004) [16] has showed that computing the rooted subtree prune and regraft distance between two rooted binary phylogenetic trees on the same label set is NP-hard. In this paper they have established the relation between rSPR distance and the MAF of two rooted binary phylogenetic trees.

Theorem 3.6 [16]: *Let T and T' be two rooted binary phylogenetic trees. Then $d_{rSPR}(T,T') = m(T,T')$, where $m(T,T')$ is the size of MAF(T,T') - 1*

Using the two reduction rules as mentioned in [2] they have shown that computing the rSPR distance between two rooted binary phylogenetic trees is fixed parameter tractable considering the rSPR distance itself to be the parameter.

Proposition 3.7 [16]: *Let T_1 and T_2 be two rooted binary phylogenetic trees. Let T'_1 and T'_2 be two rooted binary phylogenetic trees obtained from T_1 and T_2 respectively by applying either Rule 1 or Rule 2. Then $d_{SPR}(T_1,\ T_2) = d_{SPR}(T'_1,\ T'_2)$*

Lemma 3.8 [16]: *Let T_1 and T_2 be two rooted binary phylogenetic χ-trees. Let T'_1 and T'_2 be two rooted binary phylogenetic χ'-trees obtained from T_1 and T_2 respectively by applying either Rule 1 or Rule 2 repeatedly until no further reduction is possible. Then $|\chi'| \leq 28 d_{SPR}(T_1,\ T_2)$*

Bordewich and Semple(2007) in their work [17] has shown that computing the hybridization number of two phylogenetic trees is Fixed-Parameter Tractable using the

two reduction rules.

Lemma 3.9 [17]: *Let T and T' be two rooted binary phylogenetic χ-trees and let P be an empty collection of 2-element subsets of χ. Let S and S' be two weighted phylogenetic χ'-trees obtained from T and T', respectively, by repeatedly applying Rules 1 and 2 until no further reduction is possible. Then $|\chi'| < 14h(T, T')$*

Bordewich et al.(2007) derived a 3-approximation algorithm SPR-APPROX(T,T') for the subtree distance between phylogenies [15]

Theorem 3.10 [15]: *Let T and T' be two rooted binary phylogenetic X-trees and let $|X| = n$. Let (F,k) be the output of SPR-APPROX(T,T'). Then F is an agreement forest for T and T' and k is a 3-approximation for d_{rSPR}(T,T').*

Theorem 3.10 is the main contribution of this paper.

3.4.1 Fixed-Parameter Tractable

The running time of the SPR-EXACT algorithm [15] to compute the fixed-parameter on the rSPR distance has been improved to $O(4^k k^4 + n^3)$ by using the kernelisation of Bordewich and Semple [17]. As a result upon the completion of the kernelisation the resulting two rooted binary phylogenetic trees T and T' have leaf sets of size at most $28d_{rSPR}(T, T')$.

3.5 Rodrigues et al.

Rodrigues et al. [53] with the help of certain instances showed that approximation ratio for the size of MAF cannot be less than 4 which disapproves the 3-approximation claim of Hein et al. [32]. They have claimed to present two new 3-approximation

algorithms for this problem.

3.6 Other work in this area

Hallett and McCartin(2006) in their paper [29] have given an efficient fixed-parameter tractable (FPT) algorithm for the MAF problem for 2 unrooted trees, and have claimed to make a significant improve on an FPT algorithm given in [2]. The running time has been improved from $O(k^{3k} + p(|L|)$ in [2] to $O(4^k.k^5)+ p(|L|)$ where k bounds the size of the agreement forest and L is the leaf label set.

Hickey et al.(2008) in their paper [34] have shown that the unrooted SPR distance computation is NP-Hard and has verified which techniques from related work can and cannot be applied. They have also presented an efficient heuristic algorithm for this problem and experimented with it on a variety of synthetic datasets. They have claimed to provide an algorithm that computes the exact SPR distance between unrooted tree. With the help of the reduction rules to give a FPT approach to this problem the running time of their algorithm is $O(n^{k+2})$.

Bonet and St.John (2010) in [13] have shown that subtree prune and regraft (uSPR) distance on unrooted trees is fixed parameter tractable with respect to the distance. They have claimed to make progress on a conjecture of Steel [2] on the preservation of uSPR distance under chain reduction and have improved on lower bounds of Hickey et al. [34].

Wu and Wang (2010) in [60] have presented a new practical method to compute the exact hybridization number. Their approach is based on an integer linear programming formulation.

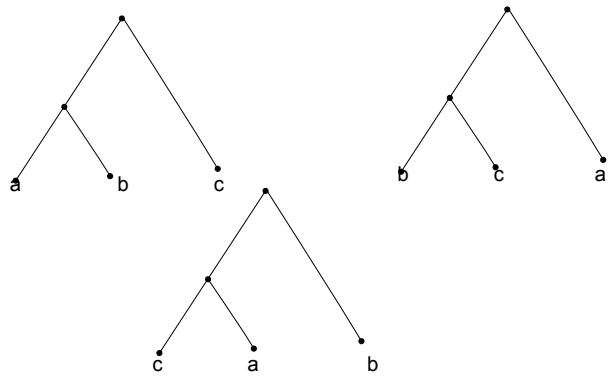

Figure 3.5: Topology of a tree with 3 leaves

3.7 Literature review with $k(\geq 2)$ phylogeny trees

In this chapter we describe Chataigner's approximation ratio for finding the MAF on k rooted binary trees. As per our knowledge this is the only work done in the field regarding the computational aspect on k trees [18].

Definition 1. Given 3 leaves a,b,c of the tree T, we call the subtree connecting the three leaves in T the triple in T. The topology of a tree T is the unique binary tree of which T is a subdivision. For example, if T is a tree with three leaves [a, b, c], there are 3 possible topologies for T, which can be represented as ((a, b), c), ((b, c), a) and ((a, c), b) respectively. See figure 3.5.

Definition 2. For a tree T and a subset L of $L(T)$, the subtree induced by L, noted $T|_L$ is the tree defined as the smallest connected subgraph of T containing L. Given two trees A and B, a triple of leaves $U = [a, b, c]$ is a conflict if the induced subtrees $A|_U$ and $B|_U$ have different topologies. See Figure 3.6.

Lemma 3.11 [18] *Two trees with the same set of leaves have the same topology*

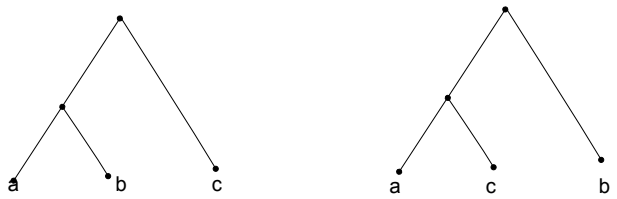

Figure 3.6: Conflicted Set of Triples. Any one leaf can be deleted from each tree to remove the conflict

iff they do not have any conflict.

A forest F can be obtained from a given set of trees T_i having similar species by the partition of the leaves (species being represented by leaves) L such that:

$\mathbf{P}_{topo}$: for each $j \in [1..m]$, the subtrees $T_i|_{Lj}$ induced by L_j have the same topology, and the partition induces a subforest on T_1
$\mathbf{P}_{sep}$: for each $i \in [1..k]$, the subtrees $T_i|_{Lj}$ are vertex-disjoint.

P_{topo} takes care of the first 2 rules for the formation of agreement forest described in section 2.4 and P_{sep} suffices the 3rd rule in the agreement forest formation.

$\mathbf{P}_{topo}$ Check

Lemma 3.12 [18] *Let T be a rooted binary tree, and let $\Re$ be a set of triples in $L(T)$. A collection with maximal size of edge-disjoint subtrees $T|_U$, where $U \in \Re$, can*

be found in polynomial time.

This collection of the maximal size of edge-disjoint subtrees can also be found through Integer Linear Programming(ILP) and this gives the lower bound in detrmining the approximation ratio of MAF on k trees in their paper.

We start with a set of triples $\Re$ in L(T) and another set M which contains the minimal set of conflict edges. A decision variable d_e has been defined to take the value based on the edges present in M. The main object is to have minimum number of edges in M so as to maximise the agreement forest.

min $\sum_{e \in E(T)}$ d_e, E(T) is the edge-set of T

a $\in$ U, $\sum_{e : e \in T|_a}$ $\mathrm{d}_e \geq 1$, U $\in \Re$

$\mathrm{d}_e \in \{0; 1\}, d_e = 1$ if e $\in$ M

Relaxing this ILP to a linear program and taking the dual the maximal flow in the tree has been determined.

max $\sum_{a \in U)}$ f_a

e $\in$ E(T), $\sum_{a : e \in T|_a}$ $\mathrm{f}_a \leq 1$

$\mathrm{f}_i \geq 0$, f_i is the flow

This linear program can be strengthened to an integer program where $\mathrm{f}_i \in [0; 1]$. So it can be written as:

Max integral flow ≤ max fractional flow = min fractional multi-cut ≤ min multi-cut

Max integral flow (Υ) becomes the lower bound of the solution where $\Upsilon \subset \Re$

Lemma 3.12 suggests it is possible to cut a collection of maximal size of edge-disjoint set of triples Υ. For each triple $U = [a, b, c]$, with topology ((a, b), c) for $T|_U$, let r_U denote the root of $T|_U$, defined as lca(a, c) [lca means lowest common ancestor] in T, and s_U denote the node lca(a, b). Let $M1$ be the set of edges entering the nodes r_U and s_U , for all triples $U \in \Upsilon$. Let us consider the set of triples $\Re_1 \in U$ such that $T|_U$ does not contain an edge in $M1$. Let the set be $M2$ which contains at least one edge from $\Re_1$. It has been illustrated in figures 3.7 and 3.8.

Lemma 3.13 [18] *a triple $U \in \Re_1$ sharing nodes with a triple $V \in \Upsilon$ satisfies $r(T|_U) \in T|_V$ where $r(T|_U)$ is the root of $T|_U$*

Lemma 3.14 [18] *a triple $U \in \Re_1$ shares nodes with at most one triple from Υ.*

For each triple V = [a, b, c]$\in \Upsilon$, let $\Re_V$ be the set of triples in $U \in \Re_1$ such that $T|_U$ shares nodes with $T|_V$ = ((a, b), c). According to Lemma 3.13 [18] and Lemma 3.14 [18], $\Re_1$ is a disjoint union of the U_V.

Lemma 3.15 [18] *if $U_1, U_2 \in \Re_V$ are such that $T|_{U1} \cap T|_{U2} \neq \phi$, then there exists $w \in T|_V$ such that $w \in T|_{U1} \cap T|_{U2}$.*

We now split $\Re_V$ in two parts. Let d be the parent of s_V and let W_1 be the set of triples with edges in common with $T|_{a,b}$ and W_2 the set of triples with edges in

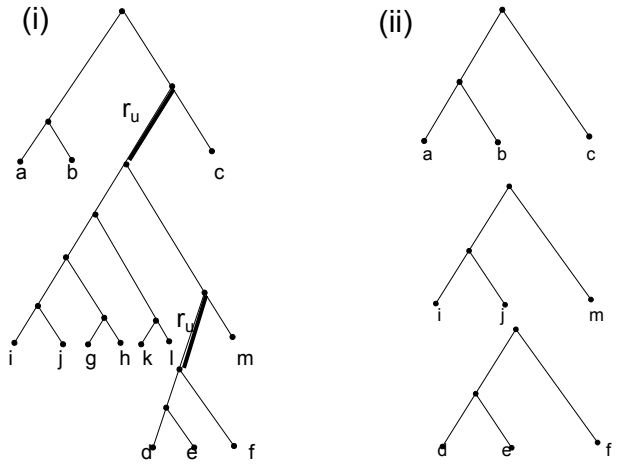

Figure 3.7: (i) A phylogeny tree T. The thick lines (r_U) show the edges being cut to get the maximal size edge-disjoint set of triples and is stored in M1. (ii) A set of maximal size edge-disjoint set of triples have been cut from T according to Lemma 3.12 [18]

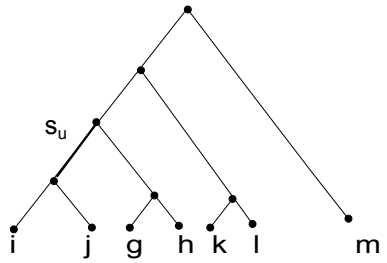

Figure 3.8: If the maximal size edge-disjoint triple ((i,j),m) obtained from Fig 3.7 is a conflicted triple, s_U is cut to remove the conflict and is added to M1. The leaf-set (g,h,k,l) will form the combination of triples in M2

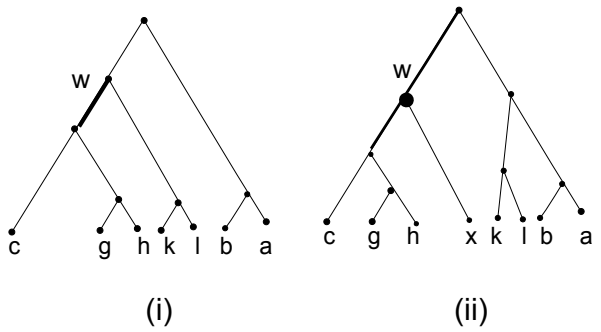

Figure 3.9: (i) U1 = ((g,h),k) and U2 = (h,(k,l)) and w in this case is an edge. (ii) U1 = ((g,h),x) and U2 = (x,(k,l)) and w is a vertex. In either case remove the thick edges in case of incompatible triples in U1 and U2

common with $T|_{d,c}$. By construction, $\Re_V = W_1 \cup W_2$

Lemma 3.16 [18] *Either* $W_1 = \phi$ *or* $W_2 = \phi$

From Lemma 3.15 [18] and Lemma 3.16 [18] we can say any 2 triples in either W_1 or W_2 have some elements w common with $T|_V$. w can be either an edge in which $M2$ contains that edge or it can be a node in which 2 edges need to be added to $M2$. This is illustrated in Fig 3.9. So altogether the size is bounded by $4|\Upsilon|$. Hence if the size of the MAF is $m*$, this step produces at most $4m*$ connected components. The cut is being made on a single tree say T_1.

$\mathbf{P}_{sep}$ Check

For every edge a of the forest F thus obtained from the above step, we consider the path $P_i(a)$ in the original tree T_i corresponding to the ends of a in F for every

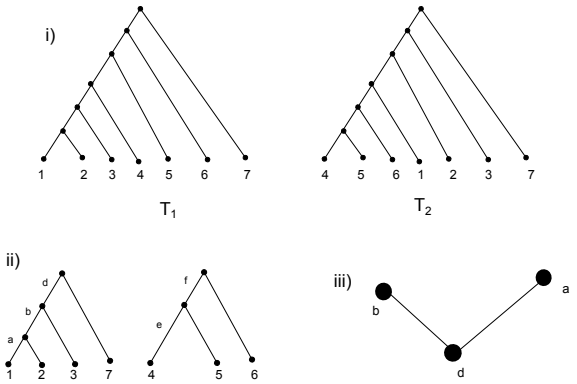

Figure 3.10: i) 2 phylogeny trees T1 and T2. ii) Components in a forest which are topologically compatible in both T1 and T2. iii) For Separability check the components are in conflict with T1. So the graph has been formed with conflicted edges. The vertex cover of this graph will give d. So remove d from ii to get MAF.

T_i under consideration. Chataigner's algorithm proceeds as follows:

- build the graph G whose nodes represent the edges of F and an edge $(a, b) \in E(\mathrm{G})$ represent a collision, i.e., some i such that $P_i(a) \cap P_i(b) \neq \phi$.
- compute a vertex cover of G.
- for each node a of the vertex cover, remove one edge from $P_1(a)$.

This has been illustrated in figure 3.10.

If we consider the minimum number of components required to ensure P_{sep} is bounded by $m*$ then at most $2m*$ edges can be removed. Since the minimum vertex cover problem can be approximated within a ratio 2, the step needs at most $4m*$ edges to be removed. Hence, altogether $8m*$ edges are removed from T_1 which gives an approximation ratio 8.

Chapter 4

Our Contribution

4.1 Approximation Algorithms

The work in this chapter has been motivated by the work of Bordewich et al. [15]. To begin with, we introduce some terms and definitions that are relevant to our work.

4.1.1 Terms and Definitions

1. **Partial Order [15]:** In mathematics, Partial order means the concept of ordering the arrangement of certain pairs of elements in a set. In the forest F of a phylogenetic tree T, for two elements x and y (elements are the union of the vertex and edge sets of F) which are in the same component of F, we write $x < y$ if $x \neq y$ and y lies on the path from x to the root of this component.

2. **Incompatible triple [15]:** A triple is a rooted binary phylogenetic tree which has 3 leaves. A triple with leaf set $\{a, b, c\}$ can be denoted by $ab|c$ if c does not lie on the path from a to b in that triple.

If T and T' be two rooted binary phylogenetic trees and $\{a, b, c\}$ be the leaf set in both T and T', then we say the triple $ab|c$ is an incompatible triple of T with respect to T' if $ab|c$ is a triple in T only. The concept of partial order can also be implied in

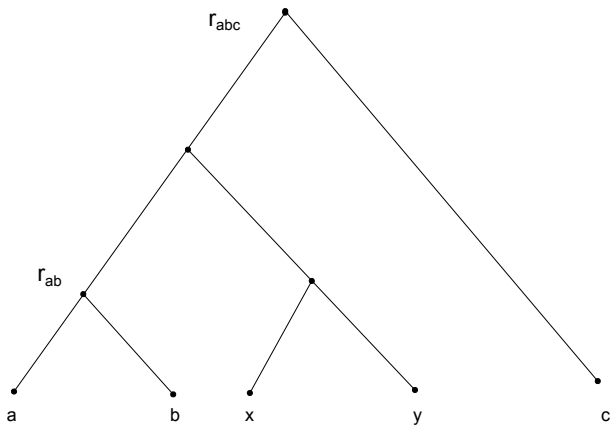

Figure 4.1: Minimum incompatible triple is $ab|c$ as $ab|c < xy|c$

the incompatible triple. Let $ab|c$ be an incompatible triple in T and let r_{abc} represent the most recent common ancestor of a and c in T and r_{ab} represent the most recent common ancestor of a and b in T. We say $ab|c < xy|z$ if either i) r_{xyz} lies on the path from r_{abc} to the root of T or ii) r_{xy} is on the path from r_{ab} to the root of T if r_{abc} and r_{xyz} are equal. We say an incompatible triple of T with respect to T' is *minimal* if it is minimal with respect to this partial order.

3. **Inseparable Components [15]:** Let F be the forest obtained from the two rooted binary phylogenetic trees T and T' after all the incompatible triples have been taken care of. Let T_s and T_t be the two components of F with leaf sets $L(T_s)$ and $L(T_t)$ respectively such that no two edges or vertices of T_s and T_t are common in T but they share at least a common element in T', then T_s and T_t are said to be *inseparable components* with respect to T'.

The following sections give details about our contribution in the thesis.

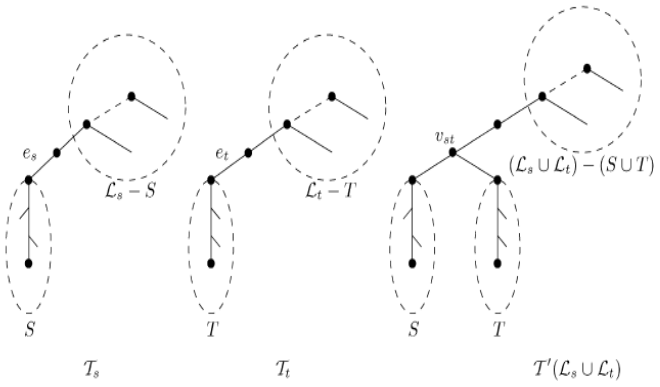

Figure 4.2: Overlap Components [[15], p:6]

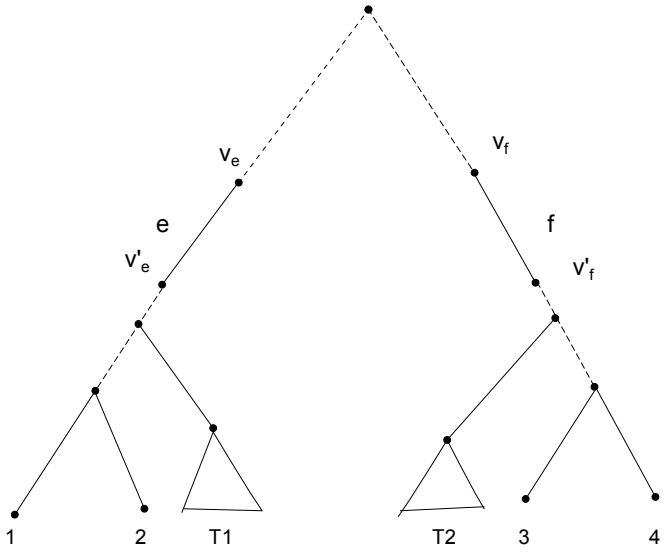

Figure 4.3: Central Lemma

4.1.2 3-Approximation for MAF on k binary trees (rooted)

The following lemma is central to our result.

Lemma 4.1 [15]: *Let T be an X-tree, F a forest of T, e and f edges in the same component of F, and E a subset of edges of F such that $f \in E$ and $e \notin E$. Let v_f be the end-vertex of f closest to e, and v_e an end-vertex of e.*
(i) If $v_f \sim v_e$ in F - E and
(ii) $\neg(x \sim v_f)$ in F-(E+e) for all $x \in$ leaf-set in F,
then F-E and F-(E-f+e) yield the same forest.

Figure 4.3 illustrates the main idea of the lemma. In order for the lemma to hold good, the set of edges from e to f should be linear, that is, there should not be any branching of subtree like $T1$ or $T2$ between the path v'_e and v'_f. Then we can take any edge e $\notin E$ and exchanged it with f $\in E$ so as to have both F-E and F-(E-f+e) as isomorphic forests.

The algorithm we present here is a 3-approximation algorithm for calculating MAF on k rooted binary trees. The algorithm is inspired from the one described in [15] for estimating the approximation ratio for the rooted SPR distance that is for rooted MAF on 2 trees. We have extended the idea and have shown that for k trees it holds a 3-approximation ratio for calculating the MAF. The algorithm MAF-Approx takes k rooted binary phylogenetic trees $\{T_1, T_2, ..., T_k\}$ as input. It initially singles out a tree, say T_1, and proceeds to make edge cuts from the forest F of T_1 until the agreement forest of k trees is obtained. The algorithm recursively computes the set of minimal incompatible triples $ab|c$ of F with respect to any one of the remaining k-1 trees T_i (i$\neq$1) and deletes some associated edges from F. When all the incompatible triples have been taken care of in the forest, the algorithm searches for the inseparable components t_x and t_y of F which overlaps in any one of the T_i where $[i = 2, 3, ..., k]$ and

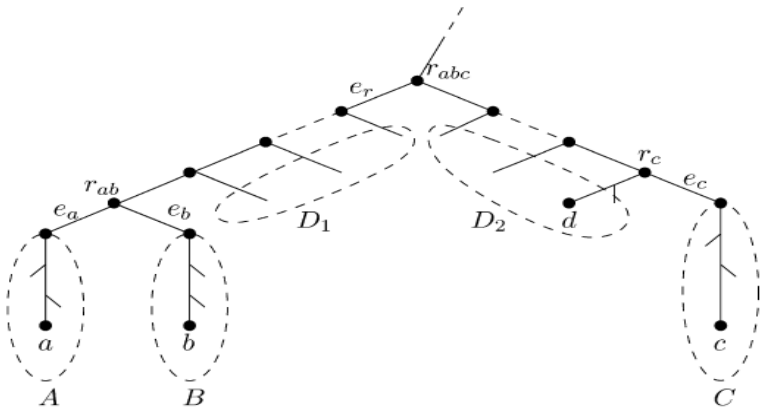

Figure 4.4: Layout of a minimal incompatible triple[[15],p:5]

accordingly edges are cut from forest F . The edges being cut have been introduced in the paper [15]. In the next paragraph we give our readers a brief description of the edges.

For a minimum incompatible triple $ab|c$ in T_1 with respect to any T_{i}, we consider r_{abc} as the most recent common ancestor of a and c in T and r_{ab} represent the most recent common ancestor of a and b in T. The child edge of r_{ab} leading to a is denoted by e_a and the child edge of r_{ab} leading to b is denoted by e_b. We denote e_r as the child edge of r_{abc} leading to r_{ab}. Finally we represent e_c as the first edge on the path from r_{abc} to c such that for elements c' in the leaf-set of T-c below e_c, there exists triples of the form $cc'|a$ and $cc'|b$ in all the trees. For a pair of inseparable components t_{x} and t_y in any T_{i} $[i = 2, 3, ..., k]$ with respect to T_1, we define a minimal common vertex v_{xy} in $t_{\mathrm{x}} \cup t_{\mathrm{y}}$ with respect to the partial order on the vertices in T_{i}. $\mathrm{e_x}$ denotes the minimal edge in T_1 whose set of descendants in the leaf-set is also the descendants of v_{xy} in t_{x}. Similarly e_y denotes the minimal edge in T_1 whose set of descendants in

the leaf-set is also the descendants of v_{xy} in t_y.

MAF-Approx(T_1,T_2,...,T_k)

1. $F \leftarrow T_1$;

2. while there exists an incompatible triple in T_1 with respect to any T_i (i≠1)

do

2.1. Consider the minimal incompatible triple ab|c in T_1 with respect to that particular T_i

2.2. $E \leftarrow$ {e_a, e_c, e_r} in $ab|c$

2.3. $F \leftarrow F$ - E

end;

3. while there exists a pair of inseparable components in any T_i (i≠1) with respect to T_1

do

3.1. consider the inseparable components t_x and t_y in that particular T_i with respect to T_1

3.2. $E \leftarrow$ {e_x, e_y} in t_x and t_y

3.3. $F \leftarrow F$ - E

end;

4. return F;

Lemma 3.1.2: *Let there be k rooted binary phylogenetic trees and let F be a forest of T_1.*

i)If there exists a minimal incompatible triple ab|c of F with respect to any of the (k-1) trees, then

$e(F-\{e_a, e_c, e_r\}, \{T_2, T_3,...,T_k\}) \leq e(F,\{T_2, T_3,...,T_k\}) - 1$, where $e(F,\{T_2, T_3,...,T_k\})$ denotes the size of a minimum set E of edges of F such that F - E forms an agreement forest of k trees.

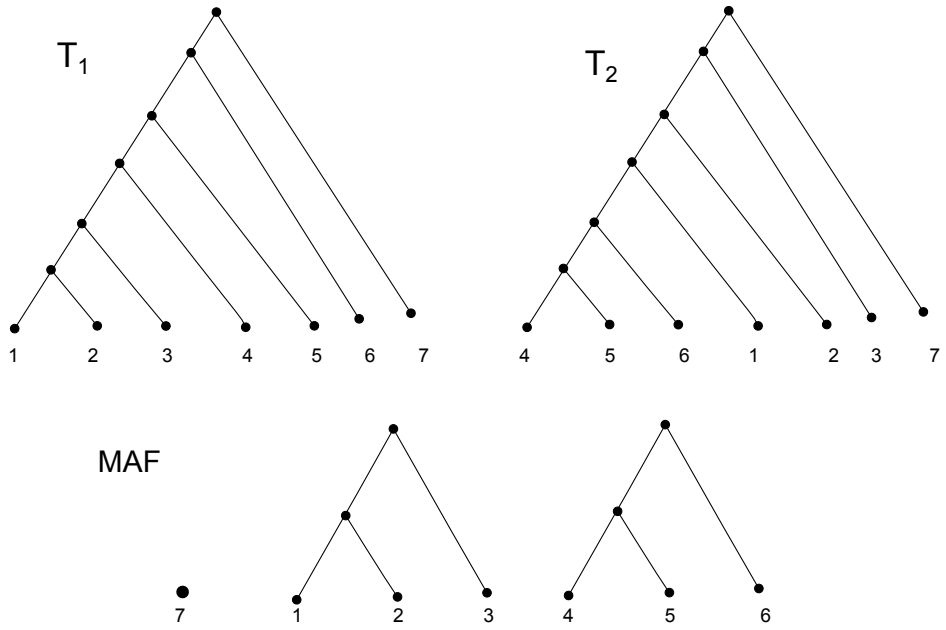

Figure 4.5: MAF of T_1 and T_2

ii) If there is no incompatible triple of F with respect to any other tree, but there exist two components t_x and t_y of F that overlap in any one of the other (k-1) trees, then for some j ϵ {x, y},

e(F - e_j, $\{T_2, T_3,...,T_k\}$) = $e(F,\{T_2, T_3,...,T_k\})$ - *1.*

Proof: Though we have k trees, we have fixed one tree T_1 and cut the edges of the forest F of T_1. While checking the incompatible triples at a time we are considering two trees in our algorithm F and T_i where i $\neq$ 1. For this, we use Lemma 4.1 [15] as a subroutine. Similarly for overlap components we consider two trees F and T_i at a time.

Claim: $\alpha \leq e(T_1,T_2,T_3,...,T_k) \leq 3\alpha$, $\alpha = \alpha_1 + \alpha_2$

Proof: Let us suppose there are α_1 iterations of the algorithm in the 1st while

loop and α_2 iterations of the algorithm in the 2nd while loop over all k trees.
We are going to show that $\alpha \leq e(T_1,T_2,T_3,...,T_k) \leq 3\alpha$ where $\alpha = \alpha_1 + \alpha_2$.
Let us assume that after i iterations, the minimum set of edge-cuts is represented by $e(F_i, T_2,T_3,...,T_k)$.

- For $i \leq \alpha_1$, (1st while loop)

According to Lemma 3.1.2. (i)
$e(F_i, T_2,T_3,...,T_k) \leq e(F_{i-1}, T_2,T_3,...,T_k)$ - 1
Hence, $e(F_i, T_2,T_3,...,T_k) + i \leq e(F_0, T_2,T_3,...,T_k)$
$\Rightarrow e(F_i, T_2,T_3,...,T_k) + i \leq e(T_1, T_2,T_3,...,T_k)$ $[F_0 = T_1]$
Moreover, since F_i has 3 fewer edges than F_{i-1} we can say
$e(F_{i-1}, T_2,T_3,...,T_k) \leq e(F_i, T_2,T_3,...,T_k) + 3$
$\Rightarrow e(F_0, T_2,T_3,...,T_k) \leq e(F_i, T_2,T_3,...,T_k) + 3i$
$\Rightarrow e(T_1, T_2,T_3,...,T_k) \leq e(F_i, T_2,T_3,...,T_k) + 3i$

- For $i > \alpha_1$, (2nd while loop)

According to Lemma 3.1.2. (ii)
$e(F_i, T_2,T_3,...,T_k) \leq e(F_{i-1}, T_2,T_3,...,T_k)$ - 1
Hence, $e(F_i, T_2,T_3,...,T_k) + i \leq e(F_0, T_2,T_3,...,T_k)$
$\Rightarrow e(F_i, T_2,T_3,...,T_k) + i \leq e(T_1, T_2,T_3,...,T_k)$
Moreover, since F_i has 2 fewer edges than F_{i-1}we can say
$e(F_{i-1}, T_2,T_3,...,T_k) \leq e(F_i, T_2,T_3,...,T_k) + 2$
Thus, $e(T_i, T_2,T_3,...,T_k) \leq e(F_i, T_2,T_3,...,T_k) + 3\alpha_1 + 2(i - \alpha_1)$

At the end of all the while loops,
$e(F, T_2,T_3,...,T_k) + \alpha_1 + \alpha_2 \leq e(T_1, T_2,T_3,...,T_k) \leq e(F, T_2,T_3,...,T_k) + 3\alpha_1 + 2\alpha_2$
Now, when the MAF is generated, no further edge-cut is necessary. So, $e(F, T_2,T_3,...,T_k) = 0$.

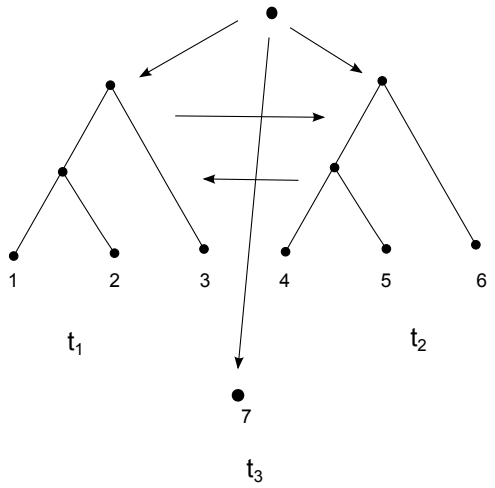

Figure 4.6: Directed Graph obtained from the MAF components in Figure 4.5. There is a cycle between t_1 and t_2.

$\alpha_1 + \alpha_2 \leq \mathrm{e}(T_1, T_2, T_3, ..., T_\mathrm{k}) \leq 3\alpha_1 + 2\alpha_2$

$\Rightarrow \alpha \leq \mathrm{e}(T_1, T_2, T_3, ..., T_\mathrm{k}) \leq 3\alpha_1 + 3\alpha_2$

$\Rightarrow \alpha \leq \mathrm{e}(T_1, T_2, T_3, ..., T_\mathrm{k}) \leq 3\alpha$

Hence, the claim of 3-approximation ratio.

The best known approximation algorithm for computing MAF on k trees is the 8-approximation ratio [18]. Our MAF-Approx algorithm gives a better approximation ratio.

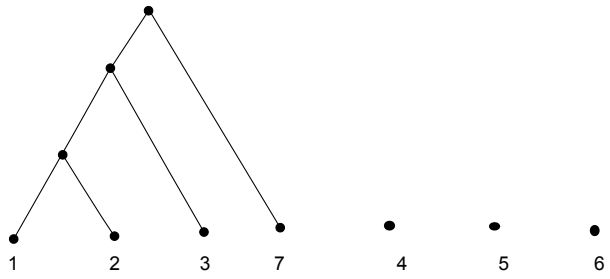

Figure 4.7: MAAF obtained from the Directed Graph in Figure 5.6

4.1.3 2-approximation for MAAF on k rooted binary trees

For the definition of MAAF please refer to Section 2.4 of Chapter 2.

Whidden and Zeh(2009) in [58] derived a 3-approximation ratio algorithm for computing the MAAF on 2 phylogenetic trees. To the best of our knowledge, an approximation algorithm for computing the MAAF on k trees had not been obtained before.

Lemma 3.2.1. *Let F' be a rooted tree in the Maximum Agreement Forest of k trees (F_A). Let e and f edges in the same component of F', and E a subset of edges of F_A such that $f \epsilon E$ and $e \notin E$. Let v_f be the end-vertex of f closest to e, and v_e an end-vertex of e. If (i) $v_f \sim v_e$ in F_A-E and (ii) $\neg (x \sim v_f)$ in F_A-$(E+e)$ for all $x \epsilon$ leaf-set in F', then F_A-E and F_A-$(E$-$f+e)$ yield the same forest.*

Proof: This proof is similar to the proof given in Lemma 5.1 of the paper of Bordewich et al [15]. We need to show that if $x, y \in \chi$, (χ is the leaf-set in F'), then $x \sim y$ in F_{A} - E if and only if $x \sim y$ in F_{A} - (E-f+e). We will prove it by the method of contradiction. We assume that $x \sim y$ in F_{A} - E but no path between x and y in F_{A} - (E-f+e). The path from x to y in F_{A} - E contains e but not f. From (i) we can conclude that either $x \sim \mathrm{v_f}$ or $y \sim \mathrm{v_f}$ in F - (E+e) which is a contradiction to the statement in (ii). So $x \sim y$ in F_{A} - (E-f+e).

Now, we assume that $x \sim y$ in F_{A} - (E-f+e) but $\neg$ ($x \sim y$) in F_{A} - E. The path from x to y in F_{A} - (E-f+e) contains e but not f. From (i) we can conclude that either $x \sim \mathrm{v_f}$ or $y \sim \mathrm{v_f}$ in F - (E+e) which is a contradiction to the statement in (ii). So $x \sim y$ in F_{A} - E. This completes the proof.

In the previous section we have described an algorithm to obtain an approximation on the MAF for k rooted binary trees. In section 2.4 of Chapter 2 we define a directed graph G_{F} with the trees derived from MAF, $F_{\mathrm{M}} = \{t_1, t_2, ..., t_{\mathrm{n}}\}$. Let us consider 2 trees t_{a} and t_{b} in F_{M} which form a cycle in G_{F} . Such a tree-pair in G_{F} is said to be infeasible. In order to obtain the Maximum Acyclic Agreement Forest F_{A} from F_{M}, at least one of the trees of this infeasible tree-pair cannot be realized and the leaves of this tree form isolated vertices in F_{A}.

The algorithm MAAF-Approx initially colors all the root vertices of the trees in F_{M} white (not processed). It selects one white tree and changes the color to gray (in process). It is checked against all other white trees to see whether there is any cycle. Once processed, the color is changed to black (complete).

MAAF-Approx(T_1,T_2,...,T_{k})

1. $F_{\mathrm{A}} \leftarrow \{t_1, t_2, ..., t_{\mathrm{n}}\}$, the MAF obtained from k trees
2. color all the root nodes of the trees in F_{A} as white
3. while there exists a white root

do

3.1. pick one white root (r_p) and change the color to gray

3.2. check whether r_p forms a cycle with any other white root in any one of the trees T_i.

3.3. if there is no cycle change the color of r_p to black.

3.4. if it forms a cycle with r_q(white root)

3.4.1. $C_A \leftarrow \{r_p, r_q\}$

3.5. while there is a cycle in C_A with respect to r_q

do

3.5.1. $r'_p \leftarrow$ {the subtree forming cycle with r_q}

3.5.2. $E \leftarrow$ {the two incident edges to the root of r'_p}

3.5.3. $C_A \leftarrow C_A$ - E

3.5.4. color the isolated vertices black

3.5.5. $F_A \leftarrow C_A$ {the set of isolated black vertices and r_q(white) }

end;

end;

4. return F_A;

Lemma 3.2.2: *Let F' be a rooted tree of an infeasible tree-pair in G_F obtained from a Maximum Agreement Forest of k trees F_M. Then there exists an edge $f \epsilon E$ in F' such that if e_l and e_r are the set of two incident edges to the root of F' then*

i) the forest F_A - (E - f + e_i) is isomorphic to F_A - E for some $i \epsilon \{l, r\}$, where E is the minimum set of edge-cuts required to produce a MAAF from a MAF of k rooted binary trees.

ii) $e(F_A - e_i) \leq e(F_A) - 1$, where $e(F_A)$ is the minimum set of edge-cuts required to produce MAAF from the MAF of k rooted binary trees and e_i is an edge in F'.

Proof: i) Let $r_{F'}$ be the root and let L and R be the left and right subtree of the

root of F'. Assume that for all nodes $l'\epsilon L$, we have $\neg(l'\sim r_F)$ in F_{A} - E. Let f be the first edge in E on the path from $\mathrm{r_F}$ to l' in F'. According to Lemma 4.1 [15], F_{A} - E is isomorphic to F_{A} - (E - f + e_{l}).

Similarly if for all nodes $r'\epsilon R$, we have $\neg(r'\sim r_F)$ in F_{A} - E, then taking f to be the first edge in E on the path from $\mathrm{r_F}$ to r' in F', we have F_{A} - E is isomorphic to F_{A} - (E - f + e_{r}). This completes the proof of part (i).

ii) We have assumed that F' is a tree in the MAF which forms an infeasible tree-pair with another tree in the directed graph and F' is not realised in MAAF. From Lemma 3.2.2.(i) we can say there exists an $f \epsilon E$ and $i \epsilon$\{a,b\} in F' such that F_{A} - E is isomorphic to F_{A} - (E - f + $\mathrm{e_l}$). Note that $E = \mathrm{e}(F_{\mathrm{A}})$. Hence F_{A} - (E - f + e_{l}) yields MAAF of F_{A} - e_{l}. Thus $\mathrm{e}(F_{\mathrm{A}}$ - $e_{\mathrm{l}}) \leq |E$ - $f| = \mathrm{e}(F_{\mathrm{A}})$ - 1. This completes the proof.

Claim: Assume there are β iterations of the algorithm, β' of which form the cycle in C_A where $\beta \geq \beta'$. We claim that

$$\beta' \leq \mathrm{e}(F_{\mathrm{A}}) \leq 2\beta'$$

Proof: Let $e(F)_\theta$ be the forest generated after θ iterations

According to Lemma 3.2.2.(ii) $e(F)_\theta \leq e(F)_{\theta\text{-}1}$ - 1

Hence after β' iterations, $e(F)_{\beta'} + \beta' \leq \mathrm{e}(F_{\mathrm{A}})$ [as $e(F)_0 = \mathrm{e}(F_{\mathrm{A}})$]

So, $\mathrm{e}(F_{\mathrm{A}}) \geq \beta'$

Conversely, the algorithm MAAF-Approx(T_1,T_2,...,T_{k}) accounts for at most 2 edge-cuts for every iteration.

Hence, $e(F)_{\theta\text{-}1} \leq e(F)_\theta + 2$.

$\Rightarrow \mathrm{e}(F_{\mathrm{A}}) \leq e(F)_{\beta'} + 2\beta'$

Now, after β' iterations the Acyclic-MAF is generated and we do not require any further edge-cut. So, we can say $e(F)_{\beta'} = 0$

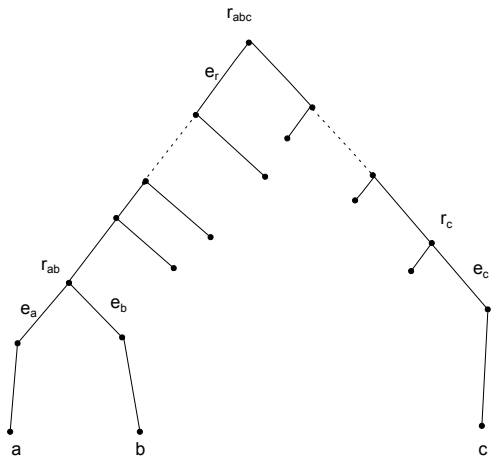

Figure 4.8: An incompatible triple $ab|c$

This proves that $\beta' \leq \mathrm{e}(F_{\mathrm{A}}) \leq 2\beta'$ and hence our claim of 2-approximation ratio.

4.1.4 Heuristic

This section gives an approximation algorithm for finding the rSPR distance between 2 trees, whose approximation ratio we conjecture to be 2.

The work has been motivated from Lemma 5.4. [15] which states that in order to remove the incompatibility of a triple, $ab|c$ the edges e_a, e_c, e_r are removed. See Figure 5.8. But here we have proposed a Lemma where we have proved that we will have the same result by removing the edges e_c and e_r.

Lemma 3.2.3:*Let $ab|c$ be an incompatible triple of F with respect to T'. Then there exists an edge $f \in E$ such that F-(E-f+$\{e_c, e_r\}$) is isomorphic to a subforest of F-E.*

Proof: The proof is in accordance with Figure:4.4.
Let us consider for all $c' \epsilon C$, we have $\neg(c' \sim r_c)$ in F-E. Further let f to be the first edge in E on the path from r_c to c in F. According to Lemma 5.1 [15] F-E is

isomorphic to F - (E-f+e_{c}).

Let us assume that for some $y \epsilon B + D_1$ the relation $y \sim r_{\mathrm{ab}} \sim a'$ in F-E holds ($a' \epsilon A$). According to Lemma 5.4 [15] under this assumption F-E is isomorphic to F - (E-f+e_{r}).

Now, let us suppose that there is no y ϵ B + D_1 such that $y \sim r_{ab}$ in F-E. Then in particular, $\neg(\mathrm{b}' \sim r_{ab})$ for all $b' \epsilon B$. Assume f to be the first edge in E on the path from r_{ab} to b in F. In order to show that F-(E-f+{$\mathrm{e_c}$, $\mathrm{e_r}$}) is isomorphic to a subforest of F-E it is enough to show that for all $x, y \epsilon \chi$ such that $x \sim y$ in F-(E-f+{$\mathrm{e_c}$, $\mathrm{e_r}$}) we have x $\sim$ y in F-E. In order to prove this by the method of contradiction, consider $i, j \epsilon \chi$ such that $i \sim j$ in F-(E-f+{$\mathrm{e_c}$, $\mathrm{e_r}$}) but i, j have no path between them in F-E. Under this assumption, the path from i to j in F-(E-f+{$\mathrm{e_c}$, $\mathrm{e_r}$}) contains f but none of the elements in {$\mathrm{e_c}$, $\mathrm{e_r}$}. It now follows that if we consider i ϵ B then it is true that j ϵ A. Moreover, by Lemma 5.1, F-E is isomorphic to F-(E-f+e_{b}) which concludes that $j \notin B$. Again, since e_r is not in the path from i to j in F-(E-f+{$\mathrm{e_c}$, $\mathrm{e_r}$}), we can say $j \epsilon D_1$ implying that $y \sim r_{ab}$ which is a contradiction. This completes the proof of the Lemma.

4.2 Implementation

We have implemented the algorithm on 3-approximation ratio in [15] in Java and tested with a few pair of phylogenetic trees to justify the validity.

In Figure 4.9, we have tested on two rooted binary phylogenetic trees having the structures ((A,B),C) in one tree and ((A,C),B) in the other tree. It can be observed that these two trees are incompatible. So the MAF is [(A,C) and B].

Figure 4.10 shows two rooted binary phylogenetic trees having the structures ((((A,B),3),4),(D,C)) in one tree and ((((A,C),3),4),(D,B)) in the second tree. The MAF of these trees is

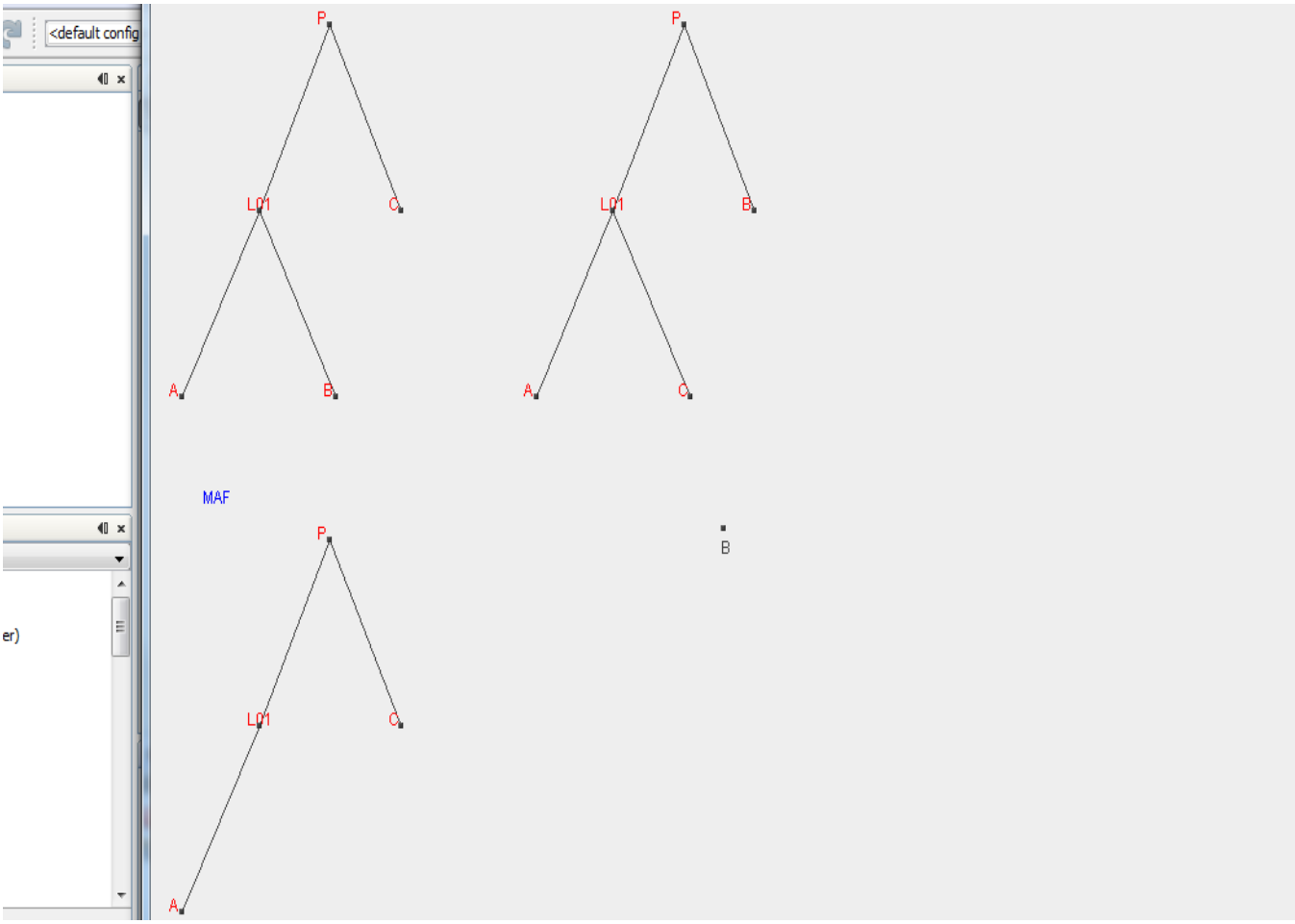

Figure 4.9: Trees with leaf-set [A,B,C] and the MAF

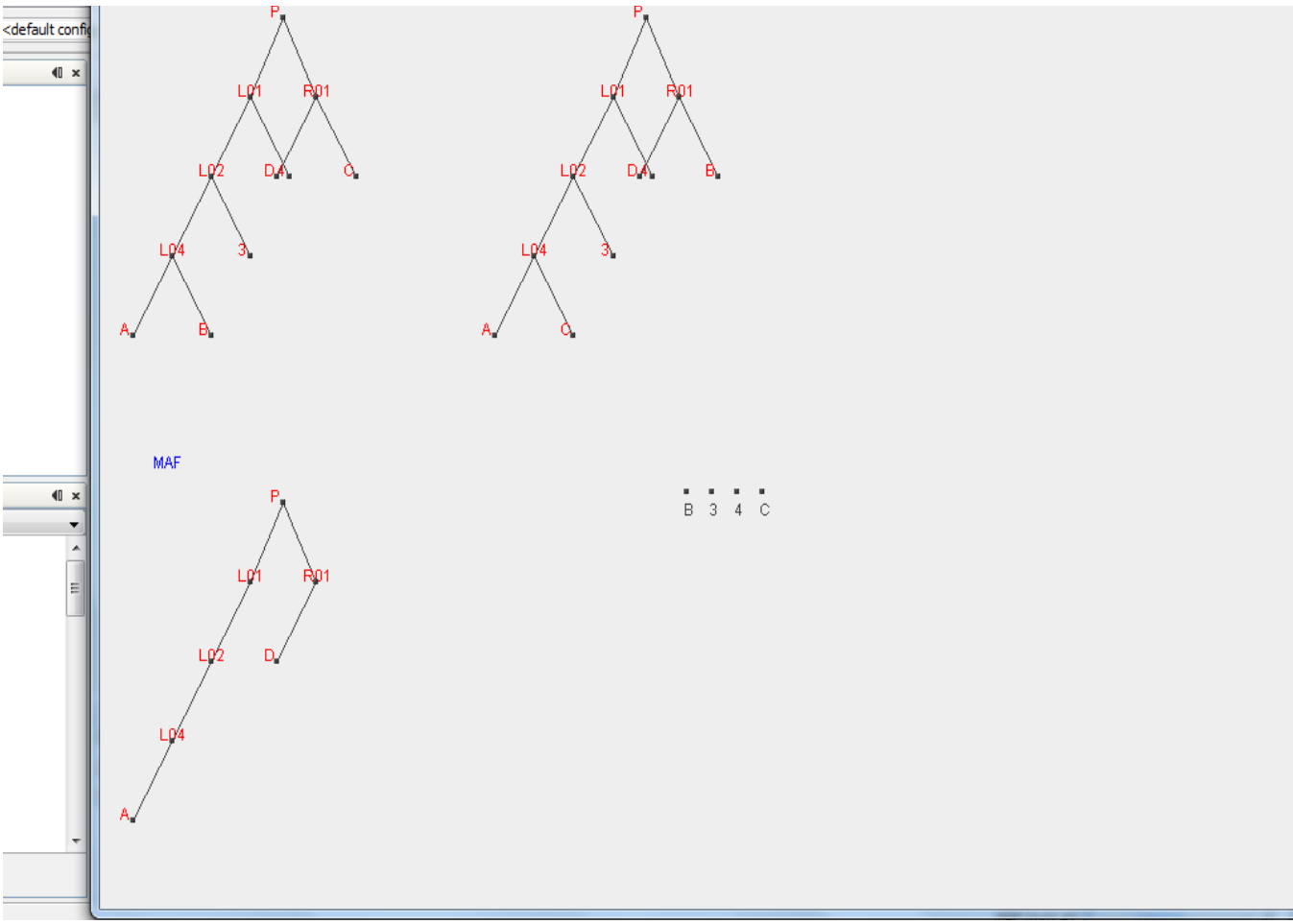

Figure 4.10: Trees with leaf-set [A,B,3,4,D,C] and the MAF

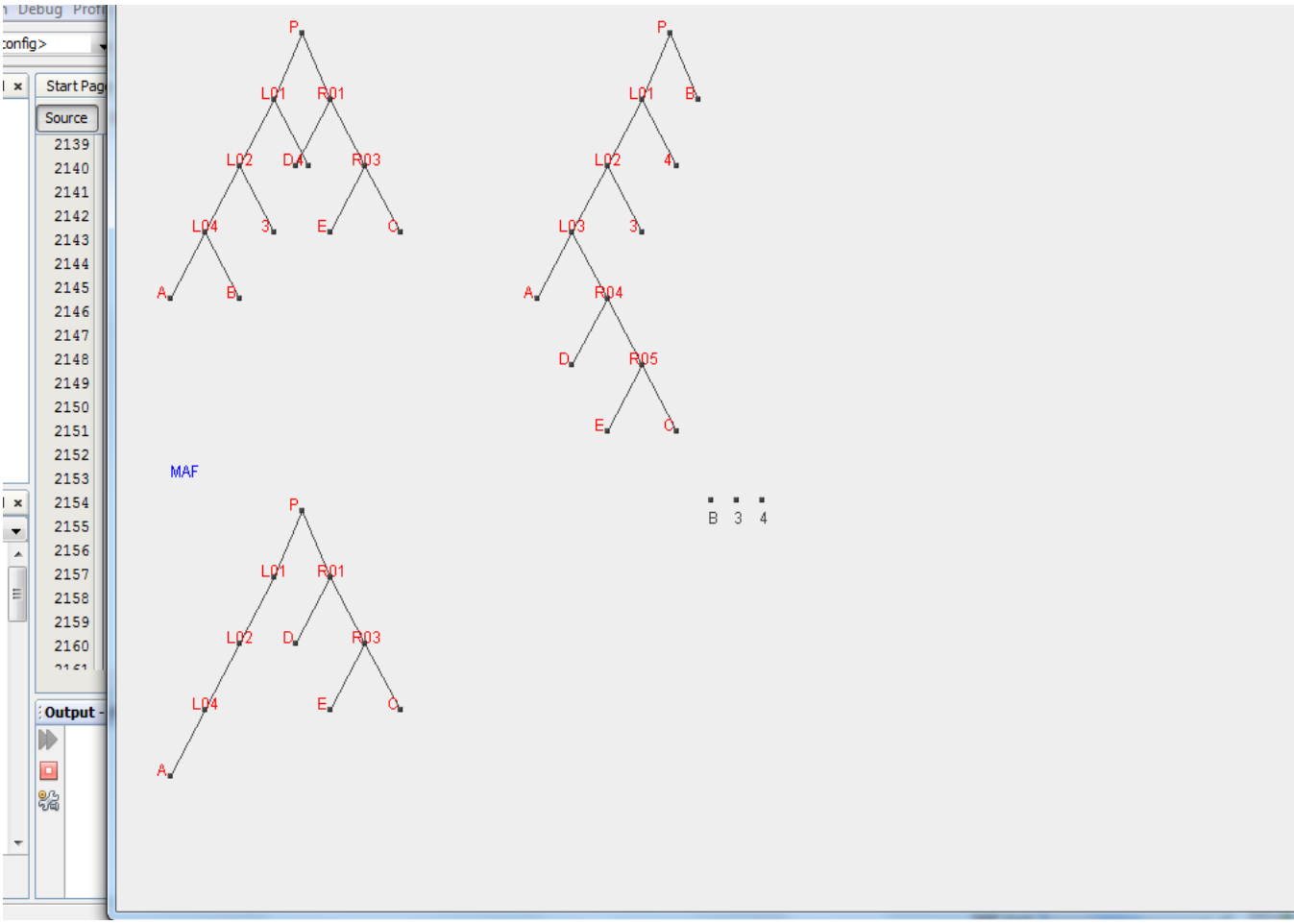

Figure 4.11: Trees with leaf-set [A,B,3,4,D,E,C] and the MAF

[(A,D), B, 3, 4, C].

Figure 4.11 consists of two rooted binary phylogenetic trees having the structures ((((A,B),3),4),(D,(E,C))) in one tree and ((((A,(D,(E,C))),3),4),B) in the second tree. The MAF of these trees is [(A,(D,(E,C))), B, 3, 4].

Chapter 5

Conclusion

In this thesis we have provided an approximation ratio for finding the Maximum Agreement Forest and the Maximum Acyclic Agreement Forest on k rooted phylogenetic trees. We have implemented the algorithm for $k = 2$ and tested with some random pair of trees. As future work, we should look into larger random tests and other biological datasets and make a thorough and rigorous test on these datasets. Besides one can look into the aspect of extending the idea of finding the approximation ratio of MAF and MAAF on k unrooted trees. Whatever results we have are based on the binary trees. It will be interesting to know the results if we have k trees of degree d ($d \geq 2$). So this can be another future course of research. In this thesis we have proposed an approximation algorithm for finding the rSPR distance between 2 phylogenetic trees. We conjecture that the approximation ratio of this algorithm is 2. Efforts can be given to see whether this heuristic can be established with a definite proof. Finally, it can be explored if the Fixed-Parameter Tractability approach can be applied to the problems of calculating the MAF and MAAF on k rooted binary phylogenetic trees.

References

[1] L. Addario-Berry, M. Hallett, and J. Lagergren. Towards identifying lateral gene transfer events. In *Pacific Symposium on Biocomputing*, volume 290, pages 279–290, 2003. [51]

[2] B. L. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:2001, 2000. [2, 3, 14, 18, 22, 24]

[3] N. Amenta and J. Klingner. Case study: Visualizing sets of evolutionary trees. In *IEEE Symposium on Information Visualization*, pages 71–74, 2002. [51]

[4] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. Complexity and approximation. *Springer, Berlin*, 1999. [51]

[5] M. Baroni, S. Grunewald, V. Moulton, and C. Semple. Bounding the number of hybridisation events for a consistent evolutionary history. *Journal of Mathematical Biology*, 51:171–82, 2005. [2, 3, 7, 9, 12, 21]

[6] M. Baroni, C. Semple, and M. Steel. A framework for representing reticulate evolution. *Annals of Combinatorics*, 8:391–408, 2005. [21]

[7] B. R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10, 1992. [51]

[8] B.DasGupta, X.He, T.Jiang, M.Li, J.Tromp, and L.Zhang. On distances between phylogenetic trees. In *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, pages 427–436, Philadelphia, PA, USA, 1997. Society for Industrial and Applied Mathematics. [51]

[9] R. G. Beiko and N. Hamilton. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*, 6(15), 2006. [51]

[10] R. G. Beiko, T. J. Harlow, and M. A. Ragan. Highways of gene sharing in prokaryotes. In *Proceedings of the National Academy of Sciences*, volume 102(40), pages 14332–14337, 2005. [51]

[11] M. A. Bender and M. Farach-colton. The LCA problem revisited. In *Latin American Theoretical Informatics*, pages 88–94. Springer, 2000. [51]

[12] M. Bonet, K. S. John, R. Mahindru, and N. Amenta. Approximating subtree distances between phylogenies. *Journal of Computational Biology*, 13:1419–1434, 2006. [51]

[13] M. L. Bonet and K. S. John. On the complexity of USPR distance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7:572–576, 2010. [24]

[14] M. Bordewich, S. Linz, K. S. John, and C. Semple. A reduction algorithm for computing the hybridization number of two trees. *Evolutionary Bioinformatics*, 3:86–98, 2007. [51]

[15] M. Bordewich, C. McCartin, and C. Semple. A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*, 6:458–471, September 2008. [x, 3, 4, 13, 23, 32, 33, 34, 35, 36, 38, 42, 44, 45, 46]

[16] M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8:409–423, 2004. [3, 7, 22]

[17] M. Bordewich and C. Semple. Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:458–466, July 2007. [22, 23]

[18] F. Chataigner. Approximating the maximum agreement forest on k trees. *Information Processing Letters*, 93:239–244, March 2005. [x, 3, 25, 26, 28, 29, 30, 40]

[19] M. Chlebik and J. Chlebikova. Inapproximability results for bounded variants of optimization problems. *Fundamentals of Computation Theory*, LNCS 2751:123–145, 2003. [51]

[20] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On computing the nearest neighbor interchange distance. In *Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications*, volume 55, pages 125–143, 1999. [15]

[21] W. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2:7–28, 1985. [51]

[22] R. Downey and M. Fellows. Parameterized complexity. *Springer, New York,*, 1998. [51]

[23] N. C. Ellstrand, R. Whitkus, and L. H. Rieseberg. Distribution of spontaneous plant hybrids. In *Proceedings of the National Academy of Science, USA*, volume 93, pages 5090–5093, 1996. [51]

[24] J. Felsenstein. Inferring phylogenies. *Sinauer, Sunderland, MA*, 2004. [51]

[25] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990. [51]

[26] N. Garg, V. V. Vazirani, and M. Yannakakis. Approximate max-floor min-(multi)cut theorems and their applications,. *SIAM Journal on Computing*, 25:235–251, 1996. [51]

[27] D. Gusfield. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In *Proceedings of Research in Computational Molecular Biology*, pages 217–232, 2005. [51]

[28] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2:173–213, 2003. [51]

[29] M. Hallett and C. Mccartin. A faster FPT algorithm for the maximum agreement forest problem. *Theory of Computing Systems*, 41:539–550, October 2007. [24]

[30] M. T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 149–156, New York, NY, USA, 2001. ACM. [51]

[31] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36:396–405, 1993. [ix, 16, 17]

[32] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71:153–169, 1996. [3, 17, 18, 23]

[33] J. Hein and Y. S. Song. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98:185–200, 1990. [51]

[34] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin. Spr distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4:17–27, 2008. [24]

[35] D. M. Hillis, T. A. Heath, and K. S. John. Analysis and visualization of tree space. *Systematic Biology*, 54:471–482, 2005. [51]

[36] D. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–267, 2006. [51]

[37] D. H. Huson, P. J. Lockhart, and M. A. Steel. Reconstruction of reticulate networks from gene trees. In *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology*, pages 233–249. Springer, 2005. [51]

[38] V. Kann. Maximum bounded 3-dimensional matching is MAX SNP-complete. *Information Processing Letters*, 37:27–35, January 1991. [51]

[39] M. Li, J. Tromp, and L. Zhang. On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, 182:463–467, 1996. [51]

[40] S. Linz and C. Semple. Hybridization in nonbinary trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:30–45, 2009. [51]

[41] D. MacLeod, R. L. Charlebois, F. Doolittle, and E. Bapteste. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evolutionary Biology*, 27, 2005. [51]

[42] W. Maddison. Gene trees in species trees. *Systematic Biology*, 46:523–536, 1997. [51]

[43] J. Mallet. Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, 20:229–237, 2005. [2]

[44] G. W. Moore, M. Goodman, and J. Barnabas. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Theoretical Biology*, 38:423–457, 1973. [15]

[45] L. Nakhleh. Evolutionary phylogenetic networks: models and issues. 2010. [51]

[46] L. Nakhleh, C. R. Linder, T. Warnow, and K. S. John. Reconstructing reticulate evolution in species: theory and practice. In *Proceedings of 8th Annual International Conference on Computational Molecular Biology*, pages 337–346, 2004. [51]

[47] L. Nakhleh, D. Ruths, and L. Wang. RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer. In *Proceedings of the Eleventh International Computing and Combinatorics Conference. LNCS 3595*, pages 84–93. Springer, 2005. [51]

[48] L. Nakhleh, T. Warnow, C. R. Linder, and K. S. John. Reconstructing reticulate evolution in speciestheory and practice. *Journal of Computational Biology*, 12:796–811, 2005. [51]

[49] G. J. Olsen, H. Matsuda, R. Hagstrom, and R. A. Overbeek. FastDNAmL: a tool for construction of phylogenetic trees of dna sequences using maximum likelihood. *Computer Applications in the Biosciences*, pages 41–48, 1994. [51]

[50] C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 229–234, New York, NY, USA, 1988. ACM. [51]

[51] L. H. Rieseberg, O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexar.

Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301:1211–1216, 2003. [51]

[52] D. F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory*, 11:105–119, 1971. [15]

[53] E. M. Rodrigues, M. F. Sagot, and Y. Wakabayashi. The maximum agreement forest problem: Approximation algorithms and computational experiments. *Theoretical Computer Science*, 374:91–110, 2007. [2, 3, 23]

[54] C. Semple and M. Steel. Phylogenetics. *Oxford University Press*, 2003. [51]

[55] Y. S. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events. In *Workshop on Algorithms in Bioinformatics*, volume 2812, pages 287–302, 2003. [51]

[56] Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, 12:147–169, 2005. [51]

[57] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. In *Proceedings of the 2001 ACM symposium on Applied computing*, pages 46–50, New York, NY, USA, 2001. ACM. [51]

[58] C. Whidden and N. Zeh. A unifying view on approximation and fpt of agreement forests. In *Proceedings of the 9th international conference on Algorithms in bioinformatics*, pages 390–402, Berlin, Heidelberg, 2009. Springer-Verlag. [3, 15, 41]

[59] Y. Wu. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, pages 190–196, 2009. [51]

[60] Y. Wu and J. Wang. Fast computation of the exact hybridization number of two phylogenetic trees. *International Symposium on Bioinformatics Research and Applications*, pages 203–214, 2010. [9, 24]

Vita Auctoris

NAME	:	Puspal Bhabak
BIRTH YEAR	:	1983
BIRTH PLACE	:	INDIA
EDUCATION		
2009–2011	:	**Master of Science**
		Computer Science
		University of Windsor, Windsor, Ontario, Canada
2001–2005	:	**Bachelors of Technology**
		Computer Science and Engineering
		West Bengal University of Technology, West Bengal, India