B.Sc. in Computer Science and Engineering Thesis

Fast and Accurate Species Tree Estimation using Consensus Methods

Submitted by

Kazi Abdun Noor

201605061

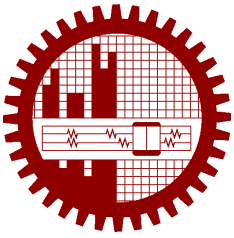
Adiba Shaira

201605097

Supervised by

Dr Rezwana Reaz

Dr. Md. Shamsuzzoha Bayzid



Department of Computer Science and Engineering

Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

February 2022

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis, titled, "Fast and Accurate Species Tree Estimation using Consensus Methods", is the outcome of the investigation and research carried out by us under the supervision of Dr Rezwana ReazDr. Md. Shamsuzzoha Bayzid.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Kazi Abdun Noor

201605061

Adiba Shaira

201605097

CERTIFICATION

This thesis titled, **"Fast and Accurate Species Tree Estimation using Consensus Methods"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in February 2022.

Group Members:

Kazi Abdun Noor

Adiba Shaira

Supervisor:

Dr Rezwana Reaz

Assistant Professor

Department of Computer Science and Engineering

Bangladesh University of Engineering and Technology

Dr. Md. Shamsuzzoha Bayzid

Associate Professor

Department of Computer Science and Engineering

Bangladesh University of Engineering and Technology

ACKNOWLEDGEMENT

First, we want to express our eternal appreciation to Almighty Allah. For everything, particularly for allowing us to survive the COVID-19 epidemic and keeping us both psychologically and physically healthy.

We are indebted to our Supervisors Dr Rezwana Reaz and Dr. Md. Shamsuzzoha Bayzid.We only have this fantastic possibility to undertake research in this intriguing field because of them. Their invaluable comments, thoughts, insights, and guidance transformed our mindset for the better, allowing us to develop new techniques for this work while also enhancing our skill set for the future.As this was our first research project, we are truly fortunate to have such persons as knowledgeable, resourceful, and fantastic as our supervisors to guide us along the way. Thank you Dr Rezwana Reaz and Dr. Md. Shamsuzzoha Bayzid sir for your outstanding supervision and for making our undergraduate final year a pleasant and welcoming environment. This thesis would not have been possible without your guidance.

Last but not least, we want to express our heartfelt gratitude to both of our parents for their unwavering love and support not only during our undergraduate years, but throughout our lives. Their unwavering support was the enormous wall we could lean against whenever things got tough.

Finally, we'd like to express our gratitude to every single person we met during our undergraduate years. All of them made a great positive difference in our lives.

Dhaka

February 2022

Kazi Abdun Noor

Adiba Shaira

Contents

List of Figures

List of Algorithms

ABSTRACT

Gene trees are used to estimate a phylogenetic tree. Gene trees differ from species trees due to incomplete lineage sorting (ILS), gene duplication, horizontal gene transfer, and so forth. One of the most common sources among this discordance is incomplete lineage sorting (ILS), in which alleles coexist in populations for durations that may span numerous speciation event. The consensus methods for estimating species trees from gene trees are becoming more popular due to having less complexity and theoretically proven (some) to be consistent.

We propose $\frac{2}{3}$ Majority Method, a highly accurate method for species tree estimation from gene trees, by extending upon the Majority method.$\frac{2}{3}$ Majority Method was accessed on a collection of simulated datasets. We compared $\frac{2}{3}$ Majority Method with Strict Consensus Method, which is the best alternate method for estimating species trees with zero False Positive rate (FP) and with many other consensus methods. We also compared BUCKy method with our proposed $\frac{2}{3}$ Majority Method. Our result suggest that $\frac{2}{3}$ Majority Method can estimate species tree with False Positive rate (FP) zero and has better performance than Strict Consensus Method.

Chapter 1

Introduction

The evolutionary links between organisms are represented by phylogenetic trees (evolutionary trees).A phylogenetic tree's branching pattern indicates how species or other groups emerged from a common ancestor. This reveals information about basic biology, such as how Orthology, how life evolved, the mechanisms of evolution, and how they modify function and structure detection, illness progression, and so on [1–4]. One of contemporary science's most ambitious goals and huge problems is to construct the phylogenetic tree [5]. The trees that have been built are just that: hypotheses. As new data becomes available, trees are edited and updated on a regular basis.

A consensus tree is a phylogenetic tree that summarizes a given collection of phylogenetic trees having the same leaf labels but different branching structures. Consensus trees are used to resolve structural differences between two or more existing phylogenetic trees arising from conflicts in the raw data, to find strongly supported groupings, and to reconcile large sets of candidate trees obtained by bootstrapping when trying to infer a new phylogenetic tree accurately [6–9]

For validating the estimated species tree, various measures of error are used like false negative number/rate (FP rate) or false positive number/rate(FN rate). False positive edges are defined as those edges which are present in the estimated tree but not in the true one.False neg-

ative edges are defined as those edges that are present in the true tree but not in the estimated tree.Consensus methods like majority rule(+), frequency difference, greedy methods are used frequently to estimate a species tree. But a few of them guarantee us with zero number of false positive edge.

1.1 Our Goal

A consensus tree is a phylogenetic tree that summarizes a given collection of phylogenetic trees having the same leaf labels but different branching structures. The purpose of consensus methods is to produce a consensus tree. A phylogenetic tree can be divided into a set of clades. Any subtree of a tree will be considered as a clade. In another words, a clade can be produced from each node of a tree comprising of the leaves under that node. Among them, the root node and the leaf nodes are defined as trivial clades.Our goal is to make the false positive rate equals to zero when estimating species trees using consensus methods.

1.2 Motivation

The strict consensus will miss a lot of splits if any outlier gene are present so the FN rate will be high.On the other hand,The majority consensus method keeps all the edges that are common to 50 $\%$ or more of the gene trees which may add some edges that are not in the true tree . For this reason, the FP rate is not zero in majority consensus method.Greedy method gives a fully binary tree so most of the time it may have some edges that are not present in the true tree.We are trying to find a method that will give a estimated species tree which won't have any false positive edge.

1.3 Our Contribution

We have proposed a novel consensus method to estimate species tree in faster and more accurate way.The evaluation metric we have used for this is false positive rate and false negative rate.The name of our consensus method is $\frac{2}{3}$ rd Majority consensus method.The clades which are present in 66.67 $\%$ or more than 66.67 $\%$ are selected for the final output consensus tree. By running the method in several simulated datasets we have seen that the false positive rate remains zero in all cases.

1.4 Organization of our dissertation

The rest of the dissertation is organized as follows.

Chapter 2 provides necessary background for understanding the problem and methods used in this book.

In chapter 3 we discuss all/(some of) the works done on estimating species trees based on consensus methods such as Strict Consensus method, Majority Consensus method, Greedy Consensus method, BUCKy etc.

Chapter 4 discuss about the methodology of our work.

Chapter 2

Background

2.1 Tree

2.1.1 Tree

Mathematically,a graph is a pair G = (V, E),where V is the vertex set and E is the edge set.A connected and acyclic graph is called a tree. So,a tree is a graph so that for every pair of vertices (v, w) in the graph there is a unique path between v and w.(So, There is always a unique path between every pair of vertices (v,w) in a tree.)

2.1.2 Binary and non binary trees

Binary and non-binary phylogenetic trees exist.If all the internal nodes of a tree have degree at most three, it is called a binary (also known as fully-resolved) tree. Otherwise, the tree is a non-binary tree.If a tree has at least one node with degree greater than three, it is known as a polytomy tree.

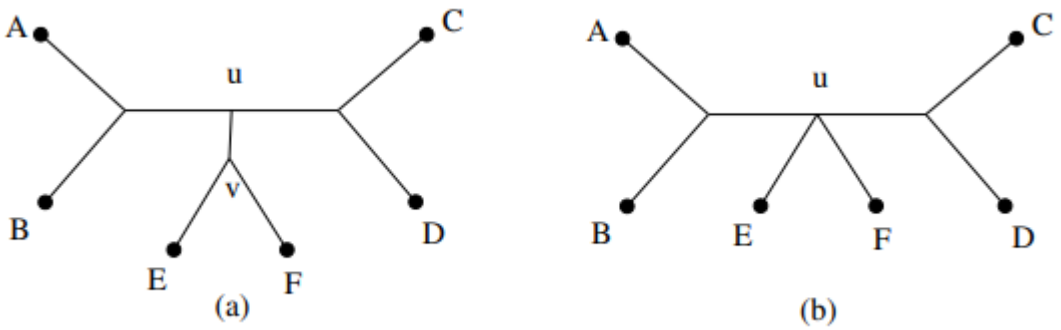


Figure 2.1: (a) A binary tree and (b) a non-binary tree representing the evolutionary history of 6 species: A, B, C, D, E, F, where u is a polytomy in (b)

2.2 Phylogenies

2.2.1 Phylogenetic Tree

A representation of the evolutionary relationship of a set of entities is known as a phylogeny.Taxa is commonly used to describe phylogenetic entities. Generally a tree is used to represent the evolutionary relationships,which is called a phylogenetic tree. A tree T is a connected acyclic graph with a set of vertices V and a set of edges E.A taxon that typically exists in the present day is represented by a leaf in a phylogenetic tree.The internal nodes represent the hypothetical ancestral taxa from which the descendant taxa evolved.

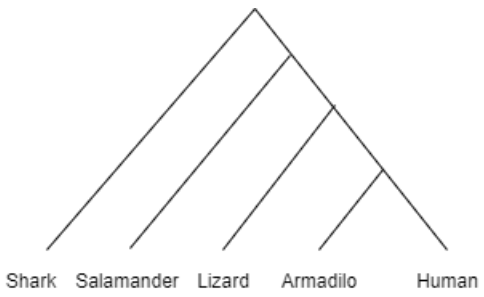


Figure 2.2: Phylogenetic Tree

2.2.2 Basic Units of a Phylogenetic Tree

A quartet is an unrooted tree for a quadruple of taxa. By studying quartets from multiple gene trees, we can infer rich information about the evolutionary history of species and genes. Gene trees generated from different gene families may have different topologies. A tree is complicated and hard to be analyzed and compared

Triplets

A triplet is a rooted tree with three leaves. This is the most basic piece of information of phylogenetic tree.

Figure 2.3: $AB|C, AC|B$ and $BC|A$ are triplet with three topologies of taxa A, B and C

Quartlets

A quartet is the minimal informative element in the tree. So it is simpler to use it as unit to be compared; it only has three kinds of topologies.

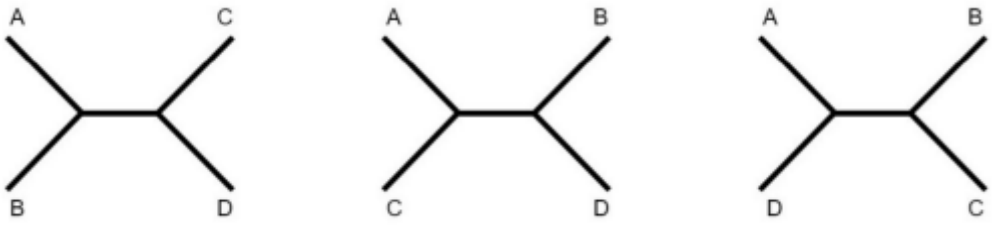


Figure 2.4: $AB|CD, AC|BD$ and $AD|BC$ are Quartet with three topologies of taxa A, B, C and D

2.2.3 Unrooted and Rooted Phylogenetic Trees

A phylogenetic tree T = (V, E) can be rooted by designating a single vertex r in V as the root of the tree, and where the root is undefined, it is called an unrooted phylogenectic tree. Although true evolutionary histories are often best represented by a rooted tree, locating the root of the tree is usually hard to achieve. Accurately rooting a phylogenetic tree is a complex problem requiring specific knowledge of the set of taxa being studied or the assumption of a "molecular clock".

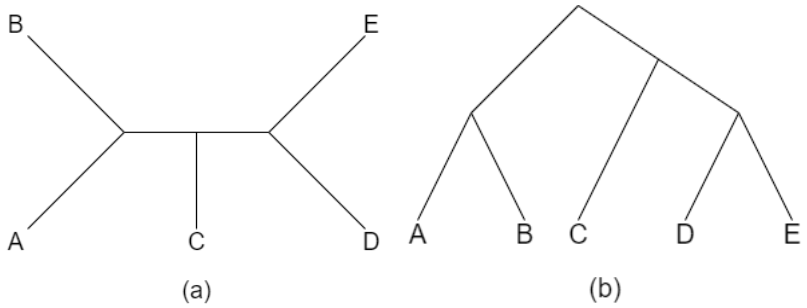


Figure 2.5: (a) Unrooted Tree & (b) Rooted Tree

Figure 2.5 shows (a) an unrooted tree (b) a rooted tree of five taxa A,B,C,D and E.

2.2.4 Clade

Each vertex in a rooted tree defines a group of taxa that are more closely related to each other than they are to any other taxon in the tree; such a group is called a clade. Formally, a clade in a phylogenetic tree T is a rooted subtree of T, which can be identified by a node v in T rooting the clade (represented by $clade_T(v)$). The set of leaves of a clade $clade_T(v)$ is called a cluster . We denote the cluster at v by $c_T(v)$; however, when the tree T is understood, we may also write c(v). The set of all clades in a tree T is denoted by C(T).

2.2.5 Bipartition

Unrooted trees have bipartitions of the taxon set, which are defined by edges rather than vertices, and are bipartitions of the taxon set.Every edge e in a phylogenetic tree T defines a bipartition

π_e. Deleting the edge e from T creates two subtrees T1 and T2, resulting into a bipartition of leaves L(T1)|L(T2).Because the bipartitions corresponding to the edges incident on the leaves convey no information about the topology of the tree, they are referred to as trivial bipartitions; nevertheless, bipartitions belonging to internal edges are referred to as non-trivial bipartitions.

2.3 Representation of Tree (Newick Format)

Newick notation is the standard way that trees,both rooted and unrooted,are represented in phylogenetic software. The Newick notation for a rooted binary tree with subtrees A and B is given by (A',B'), where A' is the Newick notation for the subtree A and B' is the Newick notation for the subtree B. Since we do not care about the left-to-right ordering of subtrees,it follows that (A', B') is the same thing as (B', A'). Also,the Newick notation for a leaf is the leaf itself.

2.4 Gene Tree and Species Tree

Most often, the goal of a phylogenetic reconstruction is to infer an evolutionary tree depicting the history of speciation events that lead to a currently extant set of taxa. Species tree is a phylogenetic tree which depicts the evolutionary relationships of a set of species..On the other hand, a gene tree is the representation of the evolution of a particular "gene" (we interpret gene as a particular part of the whole genome) within a group of species.

Gene trees and species tree often show different relationship between taxa.This is due to discordant evolutionary histories among different genes which can appear for various biological reasons and this phenomenon is known as gene tree discordance.

We now briefly describe some biological reasons for gene tree discordance.

2.4.1 Incomplete Lineage Sorting

Incomplete lineage sorting (ILS) is also known as deep coalescence.The coalescent model describes the evolutionary process as if it operates backwards in time, and connects gene lineages

to a common ancestor through a process of “coalescence“ of lineage pairs. The multi-species coalescent (MSC) model is the extension of this general coalescent framework where multiple randomly mating populations corresponding to multiple species are present. Thus, a gene tree inside a species tree is represented by multispecies coalescent. In the case in which two lineages fail to coalesce at their speciation point is refered to as incomplete lineage sorting.

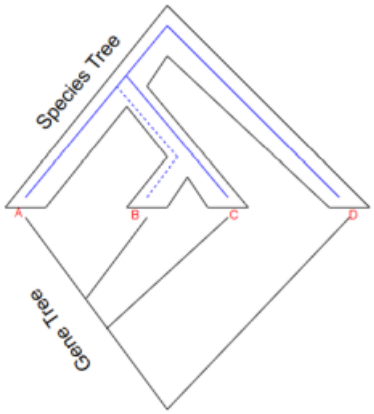


Figure 2.6: Incomplete Lineage Sorting

Figure 2.6 shows Incomplete Lineage Sorting where B and C reach their speciation point but don’t coalesce at that point and go further up.

2.4.2 Gene Duplication

The process of generating multiple gene lineages in coexisting in a species lineage is called gene duplication [10]. A gene duplication event causes a second “locus“, and these duplicated loci evolve independent of each other resulting in incongruence between gene tree and the containing species tree [11].

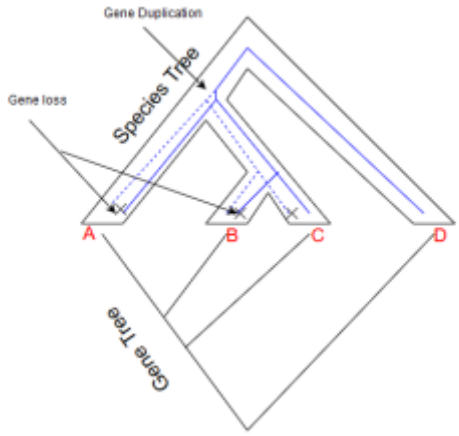


Figure 2.7: Gene Duplication and Gene Loss

Figure 2.7 shows duplication of gene and loss of gene and therefore discordance in gene tree and species tree.

2.4.3 Horizontal Gene Transfer

When genetic information is passed “sideways“ to a relatively unrelated organism is called Horizontal (or lateral) gene transfer. This causes the discordance in gene tree and species tree.

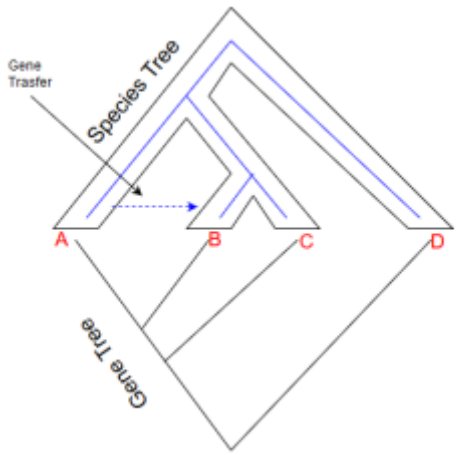


Figure 2.8: Horizontal Gene Transfer

Figure 2.8 show horizontal gene transfer changing gene tree from its species tree.

2.5 Evaluation on simulated datasets

The following is a typical simulation protocol for assessing species tree estimate algorithms.

- **Step 1:** A model species tree (sometimes known as a true-species tree) is the first step in a simulation study. A biologically-based species tree (a tree estimated on real biological datasets) from existing literature can be chosen as a model tree, or a model species tree can be constructed (usually via a birth-death process).
- **Step 2:** Under a specific model, a set of gene trees is simulated down the model species tree (e.g., gene duplication and loss, ILS etc.). True gene trees are what they're called.
- **Step 3:** A set of gene sequences are simulated by evolving nucleotide sequence down the true gene trees under a particular sequence evolution model.
- **Step 4:**The gene trees are generated using gene sequence alignments. Estimated gene trees are what they're termed.
- **Step 5:**We generate set of quartets with appropriate weights from estimated gene trees, where weights correspond to the relative importance of quartets.
- **Step 6:** Finally, we use the method of our consideration to estimate a species tree from the set of quartets, and we compare this estimated species tree to the model species tree using an appropriate error metric specified below.

2.6 Error Metric

In simulation studies, since the true result (which is known as the model tree or true tree) is known, we compare the species trees estimated by the methods of consideration with the true tree.

2.6.1 False Positive (FP) Rate

The proportion of the edges present in the estimated tree but not present in the true tree is known as The false positive (FP) rate. The false positive rate has a maximum value of 1.

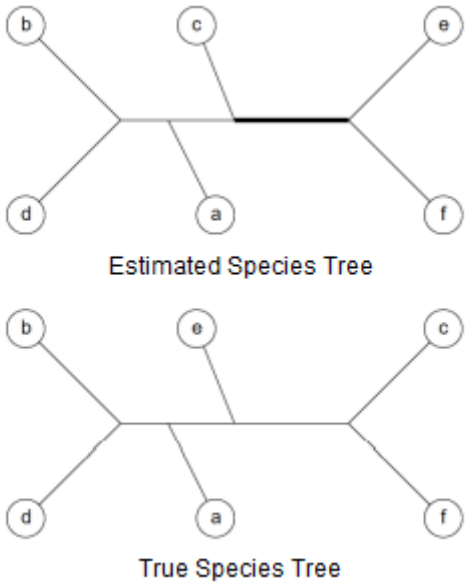


Figure 2.9: False Positive Rate

Figure 2.9 shows an estimated tree and a true tree where one estimated branch is not in the true tree, which is marked by bold line, the branch separating {a,b,c,d} {e,f}.

2.6.2 False Negative (FN) Rate

The proportion of the edges present in the true tree but not present in the estimated tree is known as The false negative (FN) rate. The false negative rate has a maximum value of 1.

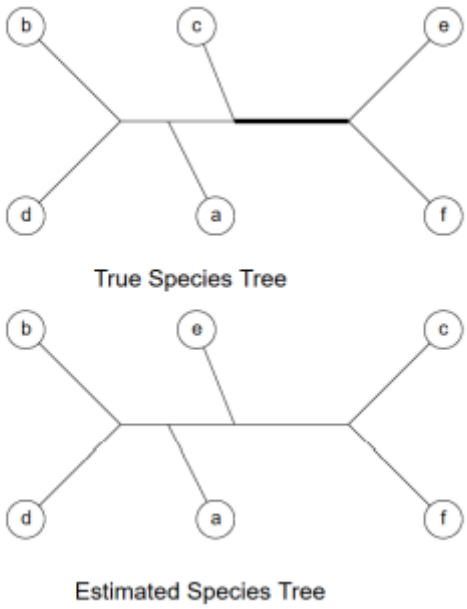


Figure 2.10: False Negative Rate

Figure 2.10 shows a true tree and an estimated tree where one true branch is not reconstructed in the estimated tree. In figure 2.10, the branch is marked missing in the true tree, the branch separating {a,b,c,d} {e,f}.

2.6.3 Robinson–Foulds metric (RF distance)

The Robinson–Foulds (RF distance) metric also known as symmetric difference metric, is a simple way to calculate distance between phylogenetic trees [12]. It is defined as (X + Y) where X is the number of partitions of data implied by the first tree but not the second tree and Y is the number of partitions of data implied by the second tree but not the first tree then scaled to maximum value of 1 [13]. Robinson-Foulds metric represents a simple and intuitive way to quantify the distances between between phylogenetic trees and therefore remain widely used.

2.7 Different Types of Summary Consensus Methods

A consensus tree is a summary of the agreement among a set of fundamental trees.There are many consensus methods that differ in:

- the kind of agreement
- the level of agreement

Consensus methods can be used with multiple trees from a single analysis or from multiple analyses.

2.7.1 Strict Consensus method

The strict consensus tree is the most basic of all consensus approaches.Given a collection of unrooted trees, the strict consensus tree contains exactly those splits common to all the trees in the collection [14].Other relationships (those in which the fundamental trees disagree) are shown as unresolved polytomies .It can be less optimal than any of the optimal trees. When the collection consists of rooted trees the strict consensus tree contains those clusters common to all the input trees.

Example:

Let T be the collection of rooted trees {((((a,b),c),d),(e,(f,g))),((((a,b),c),e),(d,(f,g)))} .The clusters {a, b, c},{d,e}, {e, f, g} appear in both trees, so the strict consensus tree is (((a,b),c),(d,e),(f,g)).[3]

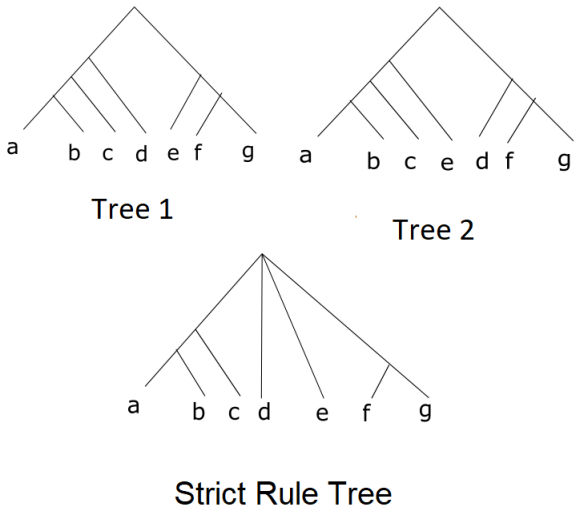


Figure 2.11: Strict Consensus Method

2.7.2 Majority Consensus Method

The majority rule tree contains exactly those clusters or splits that appear in more than half of the input trees.Other relationships are shown as unresolved polytomies .Thus every cluster (split) of the strict consensus tree will also be a cluster (split) of the majority rule tree, and the majority rule tree refines the strict consensus tree [14]. It is particularly used in bootstrapping and Bayesian Inference (best not to use for single searches) .It is implemented in PAUP* and MrBayes.

Example:

Let T be the collection of three rooted trees {((a,(b, c)), d),(((a, b), c), d),(((a, b), d), c) } The clusters {a, b}, {a, b, c} and {a, b, c, d} appear in two out of three trees, so the majority rule tree is (((a, b), c), d). Note that the strict consensus tree for this collection is (a, b, c, d).

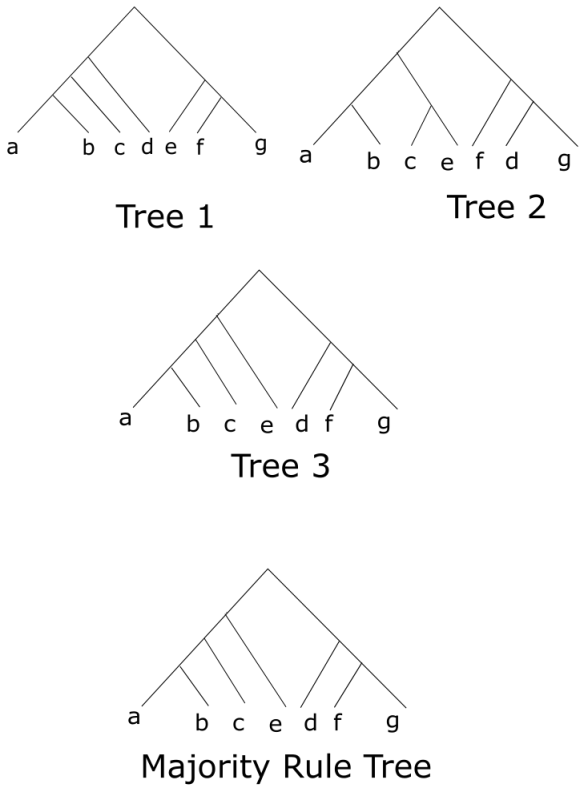


Figure 2.12: Majority Consensus Method

2.7.3 Greedy Consensus Method

This method allows additional splits or clusters to be included in the majority rule tree. The splits (clusters) are selected using what is called a greedy strategy. Suppose that T is a collection of unrooted trees. We write out the set of splits appearing in trees in T in order of frequency, with those splits appearing in the largest number of trees coming first in the order and ties broken arbitrarily. A collection of compatible splits S is built up step by step. The first split in the ordering is put in S. The remaining splits are considered in order: if a split is compatible with all of the splits already in S then it is included in S. At the end of the process we will obtain a collection of splits corresponding to some phylogenetic tree. This gives the greedy consensus tree for T . The same process works for clusters. One problem with the greedy approach is that if two clusters or splits appear the same number of times, then they could be put in either order: this decision made arbitrarily by the program. For example, in Phylip, the consensus of rooted

trees ((a, b), c) and ((a, c), b)) is either ((a, b), c) or ((a, c), b), depending on the order of the trees in the input file. [14]

Example:

Let T be the collection of three rooted trees {(((a,b),c),d),(a,(b,c),d),(a,(b,d),c)}. The greedy consensus tree for these trees is(((a,b),c),d) . The strict consensus tree is (a,b,c,d) and the majority rule tree is (a,(b,c),d).

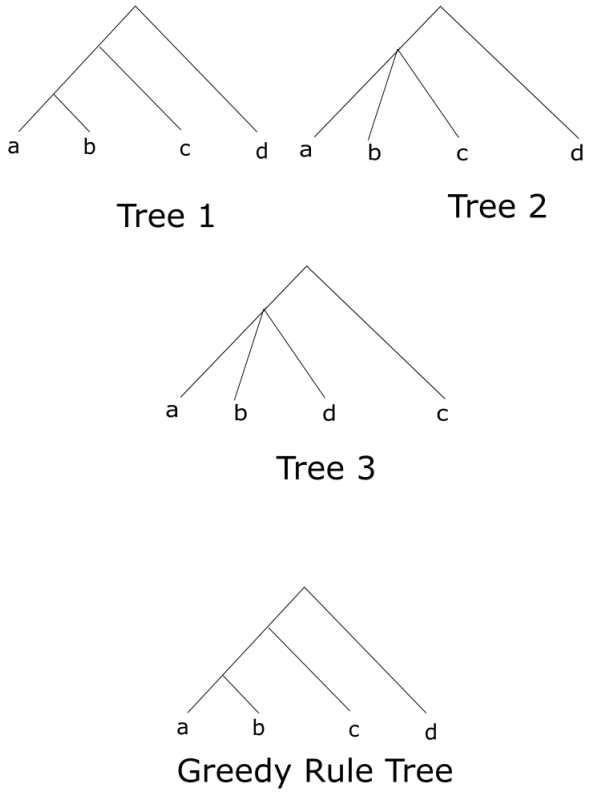


Figure 2.13: Greedy Method

2.7.4 Bucky

BUCKy is a C++ program that implements Bayesian concordance analysis. The method uses a non-parametric clustering of genes with compatible trees, and reconstructs the primary concordance tree from clades supported by the largest proportions of genes. A population tree with branch lengths in coalescent units is estimated from quartet concordance factors [15].

Chapter 3

Literature Review

Estimation of a species tree from a given set of gene trees or gene alignments is a very hard task. The challenge of this task is mainly due to the discordance of the gene trees with the species tree. Incomplete Lineage Sorting(ILS) or Deep Coalescence, Hybridization, Horizontal Gene Transfer(HGT), gene duplication/loss are the leading causes for disagreement of the gene tree with the species tree. Deep Coalescence occurs when the gene trees show change in topology or branch lengths with the species tree.

Methods concentrating on the presence of ILS infer species trees are called coalescent-based methods. Many such methods have already been explored. Concatenation-based techniques takes multiple sequence alignments(MSA) for different genes as input and then concatenates them to construct super gene alignment. Then it is used to estimate tree. The simplest version of concatenation analysis using maximum likelihood (CA-ML) seeks the maximum likelihood model tree. However, concatenation based methods can return incorrect species tree with high confidence and many of them are not statistically consistent.On the other hand, co-estimation methods develops species trees concurrently estimating gene trees. This method can estimate trees with improved accuracy but it has to compromise with the running time. Some notable methods like BEST and BEAST are statistically consistent but computationally expensive.Distance Metrics based methods like ASTRID, IDXL are among the most notable methods that construct distance matrix being aware of ILS. But, these methods suffer from inability to efficiently use information about local high-variation regions that appear across multiple sub-

trees.

Consensus method comes into play as it summarises a collection of trees without boiling everything down to one measure or number[].Fundamentally, consensus is nothing more than an algorithm that takes in a collection of gene trees (on the same set of taxa) and returns a single species tree (on the same set of taxa).

Some of the most notable consensus methods are Strict consensus method, Majority consensus method, Greedy consensus method , BUCKy.

Strict Consensus Method

The strict consensus tree of a set $T = \{T1; T2; \cdots; T_k\}$ of trees is the most resolved common contraction of the trees in T . Hence, T is the strict consensus of T if and only if every tree $T_i \in T$ refines T, and every other tree satisfying this property is a refinement of T.

Majority Consensus Method

To construct the majority consensus, the bipartitions that appear in more than half the trees in the profile are considered. The tree that has exactly those bipartitions is called the "majority consensus". The majority consensus minimizes the total Robinson-Foulds distance to the input trees, i.e., $\sum_{i=1}^{k} RF(T, T_i)$.

Greedy Consensus Method

To construct the greedy consensus, it is necessary to order the bipartitions by the frequency with which they appear in the profile. Then, starting with the majority consensus,it is needed to "add" bipartitions (if it can be), one by one, to the tree computed so far. When it is attempted to add a bipartition A|B to a tree T, it has to be checked whether it can be refined T so that it contains the bipartition A|B. If T already contains this bipartition the answer is "Yes", and T

doesn't need to change at all. If T does not already have the bipartition, it is needed to look for a polytomy in T so that this polytomy can be refined (by adding just one edge) to create the desired bipartition. It will be stopped either when a fully resolved tree (because in that case no additional bipartitions can be added) is constructed, or the examination of the entire list is done. Note that the order of listing the bipartitions will determine the greedy consensus – so that this particular consensus is not uniquely defined for a given profile of trees . On the other hand, the strict consensus and majority consensus do not depend upon the ordering, and are uniquely defined by the profile of trees.

BUCKy

BUCKy which is a program that implements BCA(The Bayesian concordance approach). BCA is a method that integrates over gene tree uncertainty. It does not make any particular assumption regarding the reason for discordance. It assumes no recombination within loci and free recombination between loci [15]. A non-parametric clustering of genes with information sharing across compatible genes is used by the Bayesian concordance approach.Its primary target is to estimate each clade's concordance factor (CF). The clades with the largest CFs are used to reconstruct the primary concordance tree. clades with moderately low CFs display relationships that are not in the primary concordance tree but are true for a minority of genome.

BUCKy takes as input the file or files consisting of gene tree(which is generated by mbsum). BUCKy's output from the Bayesian analysis includes a sample of gene trees from their joint distribution, from which CFs are estimated with credibility intervals. Finally, these CFs are used to produce the main output: a concordance tree and a population tree.

The primary concordance tree features relationships inferred to be true for a large proportion of genes. Clades are ranked by their estimated CFs. Clades are included in the concordance tree one by one as long as they do not contradict with a clade with a higher CF already in the tree [15].

In case of the population tree, . CFs are estimated from the full taxon set alignment and quartets are considered only afterwards. The posterior mean CF of each quartet is computed and transformed to an integer weight as follows, to ensure consistency. For each set of four taxa, the quartet with the largest estimated CF is favored and given weight 1, while the other two conflicting quartets are given weight 0. All quartets with weight 1 must be compatible and identify the true population tree. In practice, incompatibilities between the favored quartets are resolved with the quartet-joining algorithm described by Xin [16], which starts from the star tree and progressively joins pairs of nodes.
There have been many previous works around these topics.Consensus Methods have always been on the top interest for the researchers.

In [7],to investigate the theoretical properties of consensus trees that would be obtained from large numbers of loci evolving according to a basic evolutionary model, they construct consensus trees from rooted gene trees that occur in proportion to gene-tree probabilities derived from coalescent theory.They consider majority-rule, rooted triple (R$\star$), and greedy consensus trees obtained from known, rooted gene trees, both in the asymptotic case as numbers of gene trees approach infinity and for finite numbers of genes.when there is sufficient gene tree discordance due to incomplete lineage sorting,majority-rule consensus trees can have a high probability of being at least partially unresolved.The fact that under the multispecies coalescent model R trees are asymptotically guaranteed to be fully resolved and to match the species-tree topology.greedy consensus trees can be increasingly likely (as the number of gene trees increases) to have a topology that differs from that of the species tree. Thus,greedy consensus trees can be misleading if used as estimators of species trees.

A new consensus method Majority Rule(+) was discussed in [17].Some of the findings discussed are: 1.A majority-rule (+) consensus tree is one of the optimal candidate trees of a profile, while a majority-rule consensus tree is one of the median trees. 2.A majority-rule (+) consensus tree is, by definition, the intersection of all optimal candidate trees, while a majority-

rule consensus tree is the intersection of all median trees. 3.An optimal candidate tree is the union of a majority-rule (+) consensus tree and a set of balanced clusters, while a median tree is the union of a majority-rule consensus tree and a set of clusters that are in exactly half of the input trees.

This [18] article show that the current used software packages for consensus methods(including PAUP*) may have a worst-case unbounded running time.In this article , simple deterministic algorithms for constructing (1)majority rule consensus tree, (2) the loose consensus tree and (3) the greedy consensus tree were develpoed .

Chapter 4

Methodology

4.1 Definitions and Notations

In our research, we seek to solve the problem of inferring species tree from gene trees in a faster and more accurate way by obtaining zero false positive rate.Now, there are many biological processes like gene duplication and losses, deep coalescence, horizontal gene transfer etc. which causes genes to be evolved differently from species. So, gene trees are different in structure from species trees.For this reason there remain clades in the estimated species tree which are not present in the true gene tree.So we try to find consensus methods that will guarantee of having false positive zero rate in every case.We adopt a modified version of majority method which is $\frac{2}{3}$ rd majority. $\frac{2}{3}$ rd majority is also a majority rule consensus method where a clade(cluster) is get selected if the percentage of that clade is strictly more than $\frac{2}{3}$ of the total number of trees(66.67 percent).

Now we formally define our problem:
Given a collection of k gene trees $G = \{g1, g2, ..., gk\}$, where g_i is ith gene tree, we seek to infer the species tree T by using $\frac{2}{3}$ rd majority method which will guarantee of giving a species tree with false positive value zero.

4.1.1 Formulation

Majority consensus method selects a clade (cluster) if the percentage of that clade is strictly more than $\frac{2}{3}$ of the total number of trees (66.67 percent). $\frac{2}{3}$ majority consensus method is a modified version of majority consensus method. In $\frac{2}{3}$ majority consensus method, only those clades are selected for formation of the final species tree which are present in more than 2/3 of the total gene trees which means it's more strict than majority consensus methob but lenient that strict consensus method.

Example:

Let T be the collection of three rooted trees

- (((a,e),c),(f,(b,d)))
- (((a,e),c),(b,(f,d)))
- (((a,c),e),(d,(f,b)));

The clusters {a},{b,d,f},{c},{e} appear in three out of three trees, so the two third majority rule tree is (a,(b,d,f),c,e). Note that the strict consensus tree for this collection is also (a,(b,d,f),c,e).

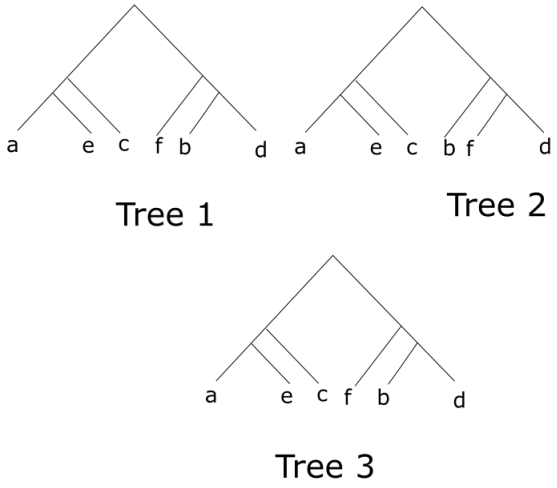


Figure 4.1: $\frac{2}{3}$ rd Majority Rule Method

4.2 Theoretical Results

It can also be theoretically proven that the $\frac{2}{3}$ rd majority method will have FP value zero.Let us see the theoretical proof now. Let (T,θ) be a model species tree, i.e., a rooted binary tree with θ its associated model parameters (branch lengths and population size). Under the coalescent model, this defines a probability distribution on rooted binary gene trees.As a result, the model species tree (topology and numeric parameters) also defines a probability on clades (subsets of taxa that appear as the leaves below some node) of the gene trees, by summing up the probabilities of all rooted gene trees that have this clade. Similarly, the model species tree defines a probability distribution on bipartitions, computed by summing up the probabilities of all rooted gene trees such that their unrooted forms have the bipartition. (Note, however, that these probabilities do not sum to 1.)

In [19], Allman, Degnan, and Rhodes proved the following:

Theorem 1: (Theorem 3 in [3]) Let (T,θ) be a binary species tree on taxon set S with edge lengths greater than $\epsilon \geq 0. If\ C \subseteq$ S satisfies Pr(C $\in clade(t)) \geq \frac{1}{3}$ exp(-ϵ) , then C is a clade in T with probability 1.

Here we show the following:

Theorem 2: Let (T,θ) be a model species tree on species set S. Let π be a bipartition on S that has probability at least $\frac{2}{3}$ exp(-ϵ). Then $Pr(\pi \in C(T)) = 1$, where C(T) denotes the set of bipartitions of T.

Proof: Let bipartition π = A|B have probability of appearing in a gene tree that is greater than $\frac{2}{3}$ exp(-ϵ). Note that for any rooted gene tree $T\prime$,$\pi \in C(T\prime)$ if and only if $A \in Clades(T\prime)$ or $B \in Clades(T\prime)$. Without loss of generality, we can assume that clade A has probability of being a clade in a gene tree that is at least that of B. Hence, it follows that the probability that clade A is a clade in a gene tree is at least half that of the probability that π = A|B is a bipartition in a gene tree. Therefore, clade A has probability at least $\frac{1}{3}$ exp(-ϵ) of appearing in a gene tree. By Theorem 1, this means that A is a clade in the model tree with probability 1. The result then follows.

Theorem 2 states that given a large enough number of real gene trees, only those bipartitions that present in strictly more than 2/3 of the gene trees are true of the true species tree.Putting this observation into practice poses various difficulties. First, gene trees are estimated.As a result, no assurances can be given for bipartitions based on estimated gene trees.

However, the following theorem follows easily:

Theorem 3: Let $t_1, t_2, \cdots t_k$ be gene trees that evolve within a species tree under ILS, and let S_i be a set of sequences that evolved down gene tree $ti, i = 1, 2, \cdots, k$ under the GTR +Γ +I model. If a statistically consistent estimation technique is used to estimate the gene trees and the $greedy_{\frac{2}{3}}$ method is applied to the estimated gene trees, then given sufficiently long sequences and sufficiently many genes, the estimated species tree will contain no false positive branches.

Note that the theorem holds even if the genes evolve under different GTR + Γ+ I parameters, since the gene trees are estimated independently. Theorem 3 implies that under optimal conditions (enough genes and long enough sequences for each gene), the species tree estimated by $greedy\_\frac{2}{3}$ will have no false positive edges. However, it does not follow that the $greedy\_\frac{2}{3}$ is statistically consistent, because it is possible for the true species tree to have bipartitions which will appear in fewer than $\frac{2}{3}$ of the true gene trees [19]. Even so, the result shows that this tree represents a valid constraint on the topology of the true species tree under these optimal conditions.

However, estimated gene trees may not be completely accurate, and even if correctly estimated, there may not be enough gene trees for the bipartitions that appear in at least $\frac{2}{3}$ to be true of the true species tree. Therefore, interpretations of this theorem in practice (and indeed any theorem about statistical consistency) requires some care. algorithm arevmath algpseudocode [margin=3cm]geometry algorithm2e

Chapter 5

Implementation Details

PAUP* (Phylogenetic Analysis Using Parsimony *and other methods) is a computational phylogenetics program for inferring evolutionary trees (phylogenies).PAUP only implemented parsimony, but from version 4.0 (when the program became known as PAUP*) it also supports distance matrix and likelihood methods.PAUP(Phylogenetic Analysis Using Parsimony *and other methods) will help us in getting expected results in $\frac{2}{3}$ rd majority consensus method.To check the performance of this modified majority method,other methods like BUCKy is also run to check the performances. For running BUCKy, we implemented converter that converts Newick formatted trees to the required format. We also implemented scripts to convert BUCKy result to Newick format for evaluations. Now to check the FP rate another tool is used which will take true tree and estimated tree as input and will give the False positive (FP rate), false negative(FN rate).

5.1 Algorithmic Pipeline of $\frac{2}{3}$ rd Majority Method

Algorithm 1 $\frac{2}{3}$ rd Majority Consensus Method

Input : A file containing a set of gene trees in Newick form
Output : An estimated species tree st

1: Write paup file
2: $67majortree \leftarrow ConvertNewickToNexus(gt)$
3: GetTaxaList(gt)
4: Begin Paup:
5: contree all / strict = no majrule = yes le50 = no percent = 67 treefile
6: $67majortree \leftarrow 67treefile$
7: End Paup:
8: Run paup
9: $st \leftarrow Convertnexustonewick(67majortree)$

5.2 Algorithmic Pipeline of $\frac{3}{4}$ th Majority Method

In order to find consensus methods that will guarantee of having false positive zero rate in every case,we adopt another modified version of majority method which is $\frac{3}{4}$ th majority consensus method. $\frac{3}{4}$ th majority is also a majority rule consensus method where a clade(cluster) is get selected if the percentage of that clade is strictly more than $\frac{3}{4}$ th of the total number of trees (75 percent).

Algorithm 2 $\frac{3}{4}$ th Majority Consensus Method

Input : A file containing a set of gene trees in Newick form
Output : An estimated species tree st

1: Write paup file
2: $75majortree \leftarrow ConvertNewickToNexus(gt)$
3: GetTaxaList(gt)
4: Begin Paup:
5: contree all / strict = no majrule = yes le50 = no percent = 75 treefile
6: $75majortree \leftarrow 75treefile$
7: End Paup:
8: Run paup
9: $st \leftarrow Convertnexustonewick(75majortree)$

5.3 Pipeline of BUCKy Method

BUCKy is built as a greedy consensus method as clades are ranked by their estimated CFs and included in the concordance tree one by one as long as they do not contradict a clade with a

higher CF already in the tree. For our work we used the current latest available version 1.4.4 (June 22, 2015).

Our dataset was gene trees represented in Newick format. BUCKy is not compatible with Newick format. The input gene tree leafs are needed to be represented as natural numbers, also header part needed to be added before the gene trees and each tree needed to be given a unique variable name.

Algorithm 3 Converting Newick format to BUCKy compatible .t file

Input : A file containing a set of gene trees in Newick form
Output : A .t file in BUCKy compatible format

```
1:  if file is not empty then
2:      read first line
3:      remove " ", "(", ")"
4:      split the sting based on "," {this will give us each gene names}
5:      store number of gene and store the gene names(string representing the gene) in an array
6:      generate the header part using array with gene names
7:      tree_serial = 0
8:      move to the start of the file
9:      while file is not empty do
10:         read line
11:         for i : 0 to numberofgene − 1 do
12:             replace the occurrence of array[i] in the line with i+1
13:         end for
14:         tree_serial++
15:         add "tree" and tree_serial to generate the variable name then add " = " and add the line
            and store in outline
16:         write outline to output
17:     end while
18:     add the footer part
19: end if
```

First we use 3 to covert the input files consisting of Newick formatted gene trees to .t files. Then we use mbsum method summarize each .t file to .in file, which consist of each unique trees present in .t tree file and the number of time they are repeated. BUCKy takes in .in file as input and outputs five output files namely, a **.input** which lists input files (loci) with their assigned ID , a **.out** file which lists parameters values, reports average SD of mean sample-wide CF (to assess convergence) and acceptance probabilities when swapping between cold/heated chains , a **.cluster** which lists posterior distribution on the number of clusters, i.e. the number of groups of genes that share the same tree, a **.gene** which lists its support for each tree from the

individual analysis (input), and from the combined analysis and a **.concordance** file which contains estimated population & concordance trees and more information. We extract the estimated population tree & concordance tree form the .concordance file using 4.

Algorithm 4 Extracting output form .concordance file

Input : A .concordance file from output of BUCKy

Output : two .tree files, one containing concordance tree another containing population tree

1: locate the estimated population tree and concordance tree
2: read and store estimated population tree in population_tree and estimated concordance tree in concordance_tree
3: **for** i : $numberofgene - 1$ to 0 **do**
4: replace the occurrence of i+1 in the concordance_tree with array[i]
5: replace the occurrence of i+1 in the population_tree with array[i]
6: **end for**
7: write concordance_tree and population_tree to the respective output files

The concordance tree and population tree are then evaluated to find FP , FN and RF values in respect of their corresponding true tree.

Chapter 6

Results and Discussions

We calculated the metrics, Average FP rate and Average RF rate, and compared our $\frac{2}{3}$ rd majority consensus technique to the species-tree estimation method Strict Consensus, Majority Consensus, Greedy Consensus, and another competitive, widely used, method Bucky on simulated datasets.

6.1 Datasets

6.1.1 Simulated Dataset

We examined the performance of various consensus methods using the previously used simulated datasets.We examined two biologically based simulated datasets (37 taxon (Mammalia) and 48 taxon (Avian)) studied in [20].We have also analyzed some more simulated datasets 11 taxon,15 taxon,17 taxon,101 taxon,500 taxon,1000 taxon and some of them were studied in [20] [21] [22].

The species tree calculated by MP-EST on the biological dataset examined by Song et al. [23] was used to replicate this mammalian dataset. We utilized the coalescent model to generate a series of gene trees from this species tree, which had branch lengths in coalescent units. As a re-

sult, the model tree has an ILS level based on a coalescent analysis of the biological mammalian dataset, as well as other simulation aspects tailored to the biological sequences they investigated. We looked at the effects of different numbers of genes (25 -800), different quantities of gene tree estimate error (i.e., the amount of phylogenetic signal), and different marker sequence lengths (250bp-1500bp). The levels of ILS were different in both cases (shorter branches).

By doubling or dividing all internal branch lengths in the model species tree by two, the degrees of ILS were altered in both situations (shorter branches increase ILS). As a result, we have three model conditions: 1X (moderate ILS), 0.5X (high ILS), and 2X (extreme ILS) (low ILS). The 48-taxon avian simulated dataset was created using the same approach as the mammalian dataset and is based on the species tree calculated using MP-EST on the avian dataset of [23]. It has three different ILS levels (1X, 0.5x, and 2X), similar to the mammalian dataset, except the ILS levels are larger (i.e., more discordance between the genuine gene trees and the species tree). [24]

For the study in [22], 11-taxon datasets were constructed and simulated using a complicated approach to guarantee high variation between genes and to diverge from the molecular clock. Model trees with long branches (LB) produced low levels of ILS, while those with short branches (SB) produced significant levels of ILS. These two model conditions are referred to as weakILS and strongILS, respectively. Our 11-taxon datasets (both strongILS and weakILS) contain 100 replicates each containing 100 genes. To evaluate the impact of the number of genes on the performance of different methods, it was subsampled different number of genes (5, 10, 25, and 50 genes) from the available set of 100 genes in [25]. It was randomly subsampled a particular number of genes (5, 10, 25 etc.) from a replicate that contains 100 genes. 20 set of such subsamples from each replicate was generated.

17 taxon dataset prepared in [26],two collections of gene trees were simulated in this model.one with only 8 gene trees and one with 32 gene trees; however, the 8-gene dataset is not a subset of the 32-gene dataset. These gene trees were simulated within the species trees using the “Coa-

lescence Contained Within Current Tree" module within Mesquite, with an effective population size of Ne = 100, 000. Then sequences were evolved down the gene trees under the Jukes-Cantor model (without any rates-across-sites), using Seq-gen (Rambaut and Grassly (1997)), with each sequence having length 2000 [25].

The 15-taxon datasets have a high level of ILS and different sequence lengths and gene counts. The 500 and 1000 taxon dataset was generated using the probabilistic gene evolution model.50 replicatants of both the dataset were used in our experiment.

6.1.2 Biological Dataset

We have also worked with the biological dataset.The 37-taxon mammalian dataset from Song et al. [23], the avian phylogenomic dataset containing 48 species and 14,446 loci (including exons, introns and UCEs) [27]

6.2 Result on Simulated Datasets

In this chapter,we will see the changes of FP rate in different model conditions of different dataset.

6.2.1 11-Taxon Simulated Dataset

The performance of various consensus methods on 11-taxon strong-ILS dataset with varying numbers of estimated and true gene trees is shown in following figure.On this dataset, in small gene numbers we have observed that the FP rate is zero in case of $\frac{2}{3}$ rd Majority method but in case of Majority method the FP rate is not zero.In this model condition,Strict,$\frac{3}{4}$ th Majority method,$\frac{2}{3}$ rd Majority method gives FP value of zero.

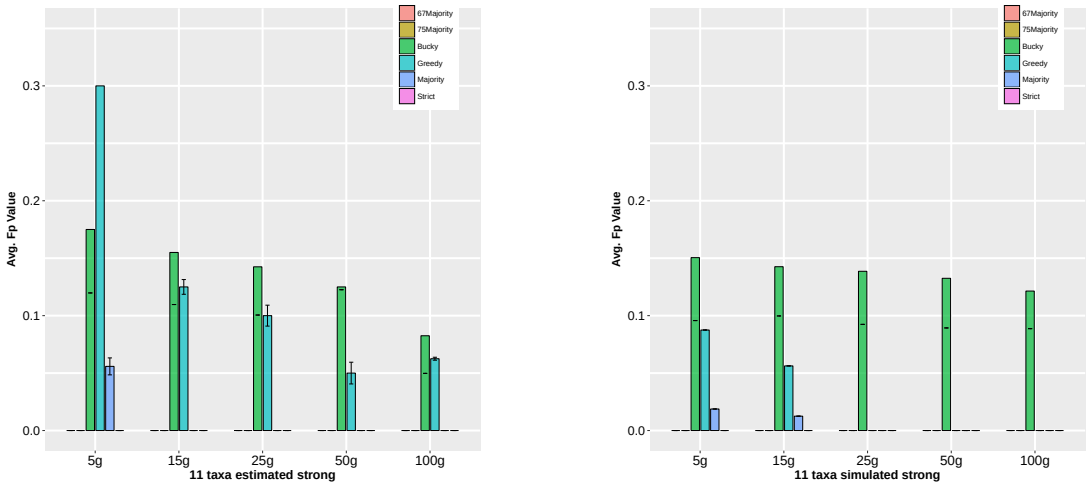


Figure 6.1: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

The performance of various consensus method on 11-taxon strong-ILS dataset with varying numbers of estimated and true gene trees with respect to RF value is shown in the figure 6.2. Here, we can see that $\frac{2}{3}$ th Majority and $\frac{3}{4}$ th Majority method has lower RF rate than Strict method. We can observe that very little changes occurs with increase in genes (gene tree). This indicates that $\frac{2}{3}$ th Majority and $\frac{3}{4}$ th Majority method are less susceptible to outliers.

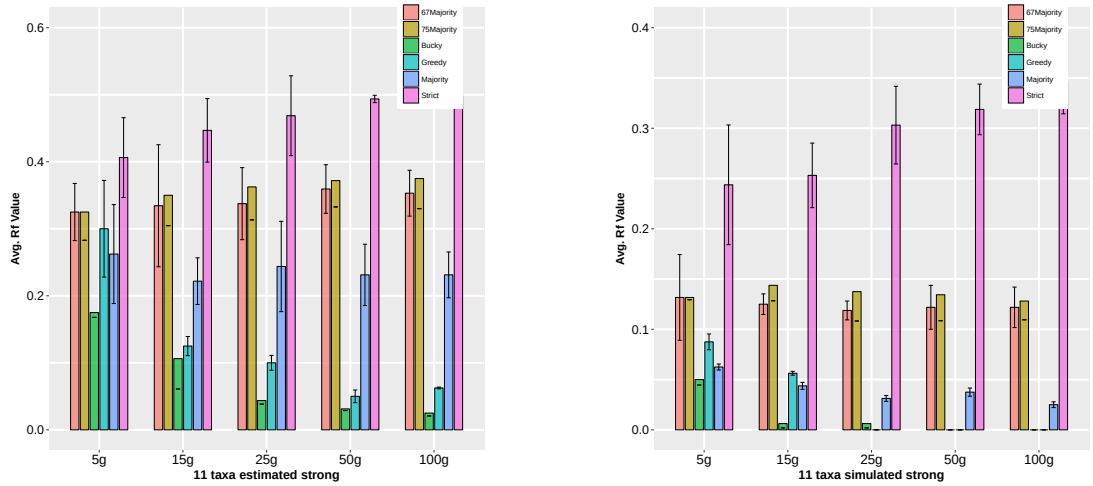


Figure 6.2: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

On the contrary,in the weak ILS for the estimated one,we have discerned the above mentioned pattern where the $\frac{2}{3}$ method gives FP value zero but majority doesn't in smaller gene num-

bers.On the contrary,in the simulated weak version only bucky gives some false positive edge which eventually diminishes in the higher gene numbers.

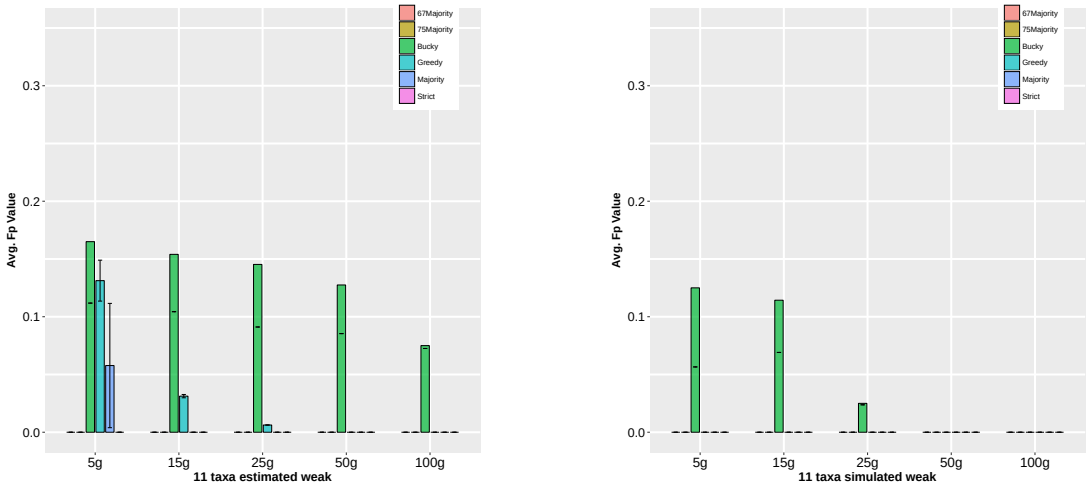


Figure 6.3: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

The performance of various consensus method on 11-taxon weak-ILS dataset with varying numbers of estimated and true gene trees with respect to RF value is shown in the figure 6.4.We can see that for simulated weak RF value start with small value and with increase in genes(gene trees) becomes zero 6.4, Which means that for 25g 50g & 100g the result is the true tree, as both FP and RF equals to zero.

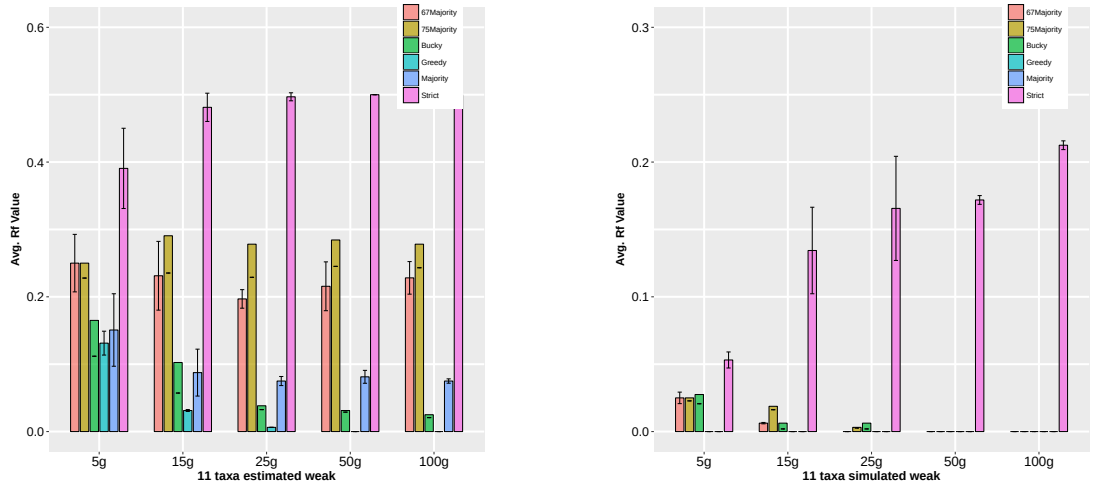


Figure 6.4: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.2.2 15-Taxon Simulated Dataset

As shown in following figure, we investigated the performance of different gene tree estimate errors using 100bp and 1000bp sequence lengths and different numbers of genes (100 and 1000).Greedy and BUcky have false positve values because both of the method provide binary tree.Other three four methods in these dataset produce same results (FP = 0).

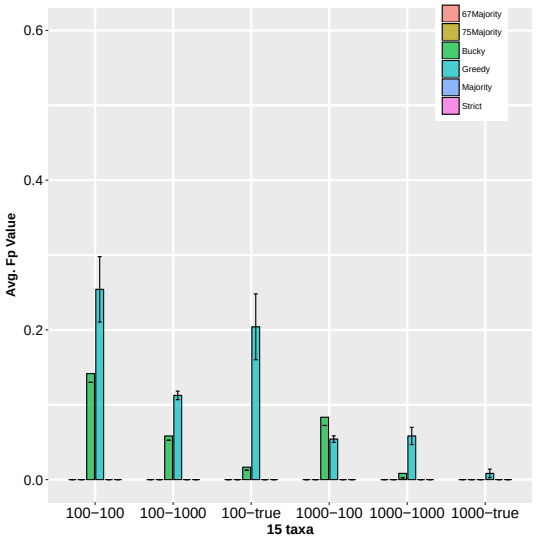


Figure 6.5: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

Figure at 6.6 shows the RF values of various consensus methods on 15-taxon dataset. Here we can see that $\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method Rf values does not change with change in bp sequence and number of genes. We can also see that Strict and Greedy method also producing the same result.

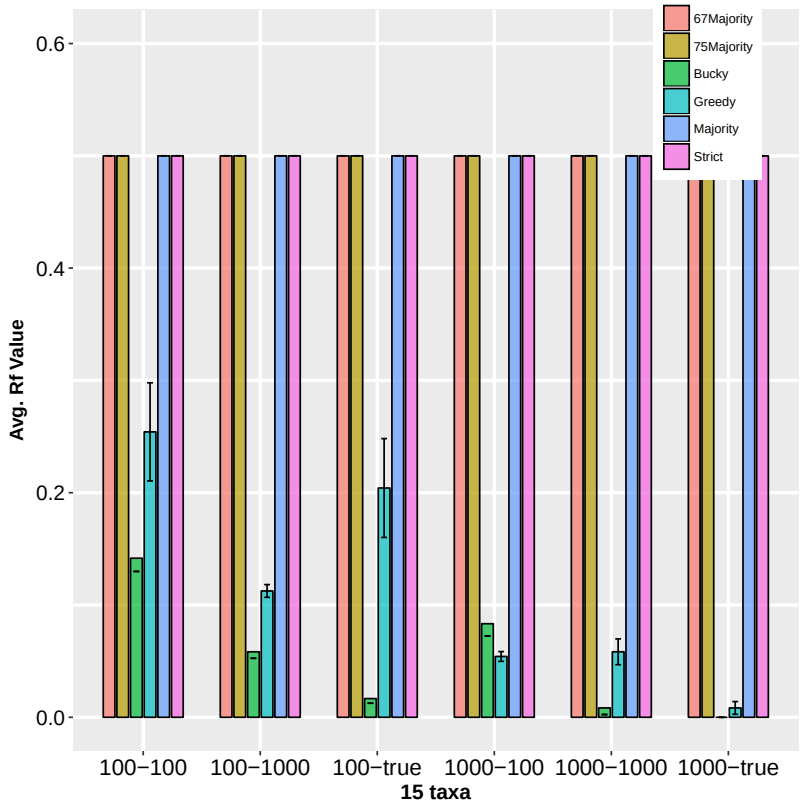


Figure 6.6: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.2.3 17-Taxon Simulated Dataset

There are two collections of gene trees in 17 taxon dataset.In both of the variation,we can observe that the Strict,$\frac{2}{3}$ rd Majority method,$\frac{3}{4}$ th Majority method give zero FP value whereas majority method estimated trees have false positive edge.

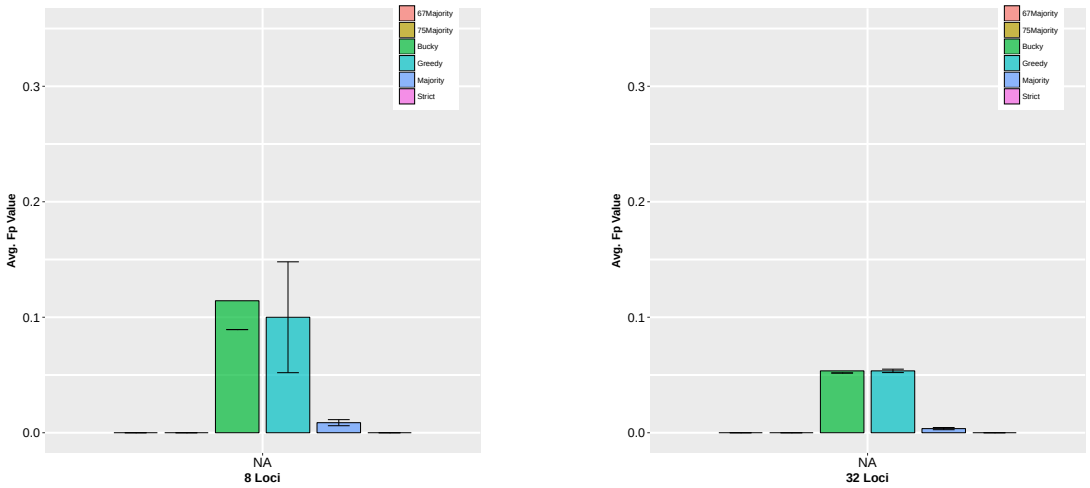


Figure 6.7: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

In following figure 6.8, we showed the RF values of various consensus methods on 17-taxon dataset below. Here, we can see that the $\frac{3}{4}$ th Majority Method's RF value increase with increase in loci. $\frac{2}{3}$ rd Majority method's RF value stays almost the same while change in loci.

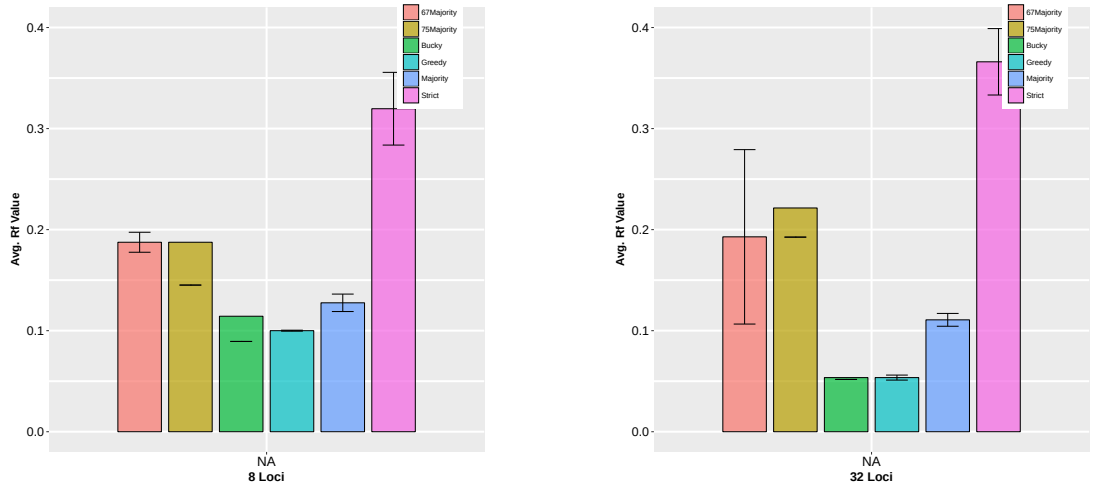


Figure 6.8: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.2.4 37-Taxon Simulated Dataset

The following figure shows the average FP rates of different consensus methods under various model conditions in a 37-taxon dataset.The FP value decreases with the increase of num-

ber of genes in first figure,also only BUCKy and Greedy show FP value.Species tree FP error rates doesn't change signaficantly with the ILS level.In case for both of the variation,estimated species trees by Strict,Majority, $\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority don't have any false positve edge so the average FP rate is zero.

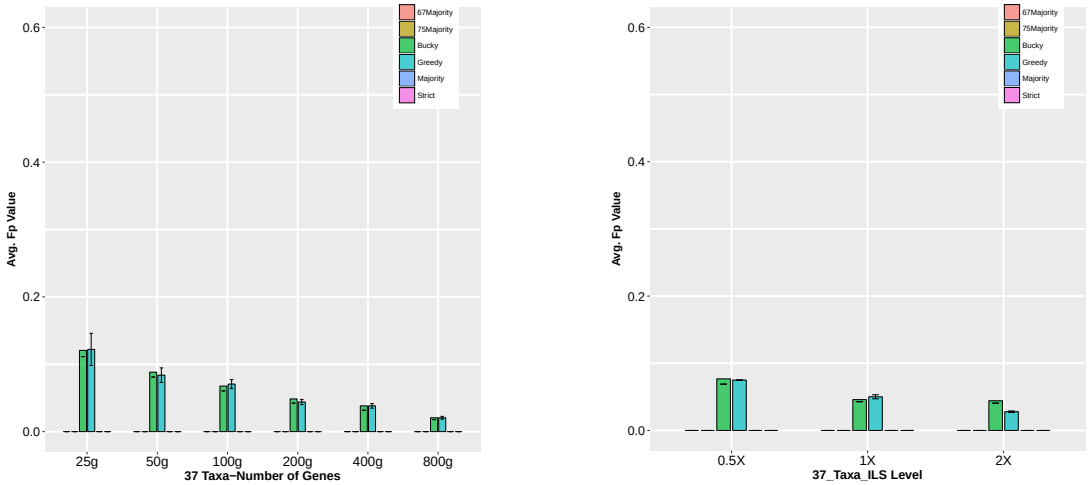


Figure 6.9: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

In the following figure 6.10, we showed the average RF values of various consensus methods on 37-taxon dataset. Here, we can see that RF value of Strict method increases with increse in genes, but $\frac{2}{3}$ rd Majority and $\frac{3}{4}$ th Majority method's RF values almost stays the same (very little fluctuation) with increase in genes (gene trees). We can see that $\frac{2}{3}$ rd Majority and $\frac{3}{4}$ th Majority method's RF values decrease with increase in ILS level.

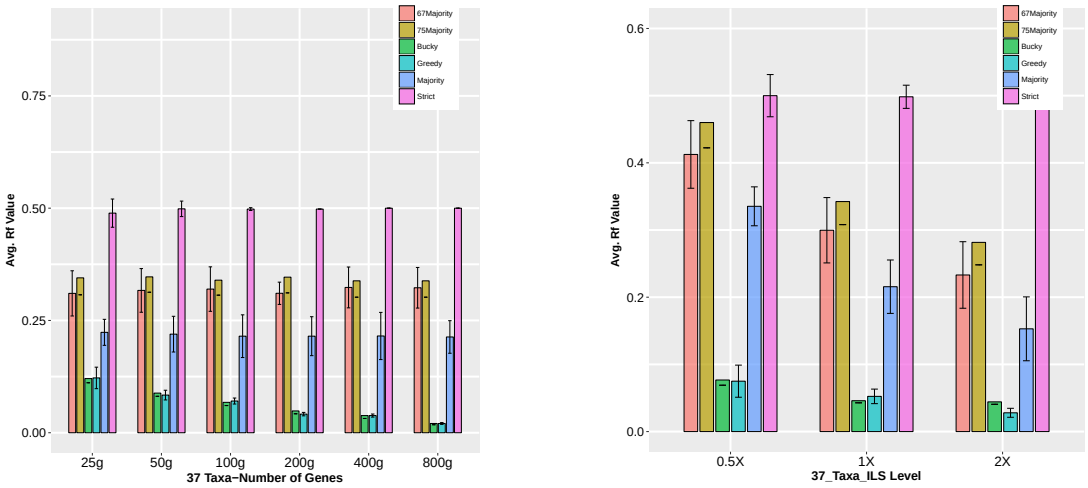


Figure 6.10: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.2.5 48-Taxon Simulated Dataset

48 Taxon shows a different pattern in FP rates in regarding of BUCKy than the previous ones.Only the 1X ILS level gene trees's estimated species tree gets false positive value when we used BUCKy .The other three consensus methods except Greedy as usual have zero FP value.Following figure shows the performance on varying the ILS levels (0.5X, 1X, 2X) with 1000 genes and fixed default sequence length (500 bp).

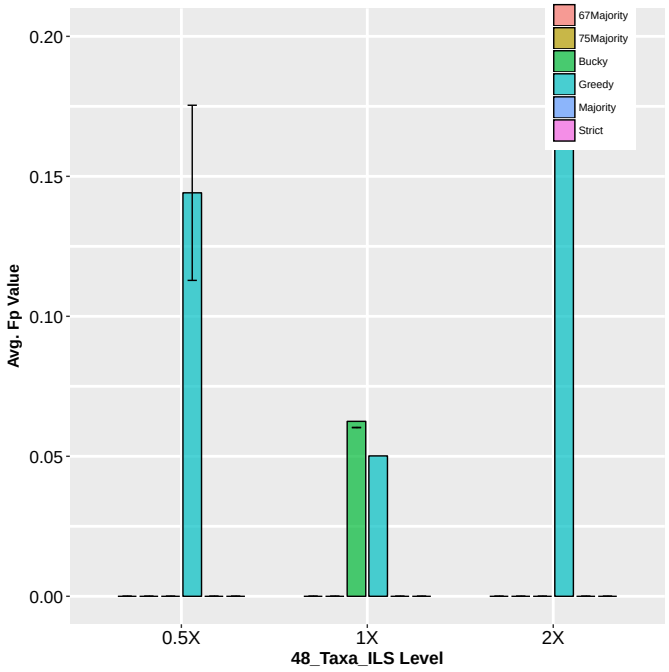


Figure 6.11: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

The figure 6.12 below shows the RF values of different consensus method on 48-taxon simulated dataset. Here, we can that $\frac{2}{3}$ rd Majority and $\frac{3}{4}$ th Majority method's RF value remains almost the same with increase in ILS level but Majority method's RF value increase with ILS value, which means $\frac{2}{3}$ rd Majority and $\frac{3}{4}$ th Majority method works better as ILS level increases for this dataset than Majority method as it's RF value increase at higher rate.

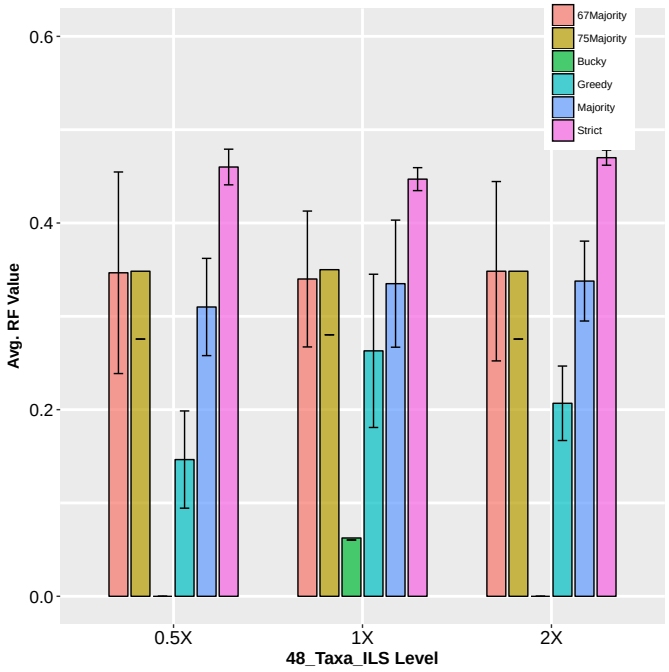


Figure 6.12: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.2.6 101-Taxon Simulated Dataset

We show average FP rates of consensus methods on 50 replicates of 101-taxon dataset with 1000 true gene trees.Only Greedy method and BUCKy have some average FP rate value in their estimated species tree.The greedy method produce much lower FP rate error than BUCKy in case of 101 taxon.

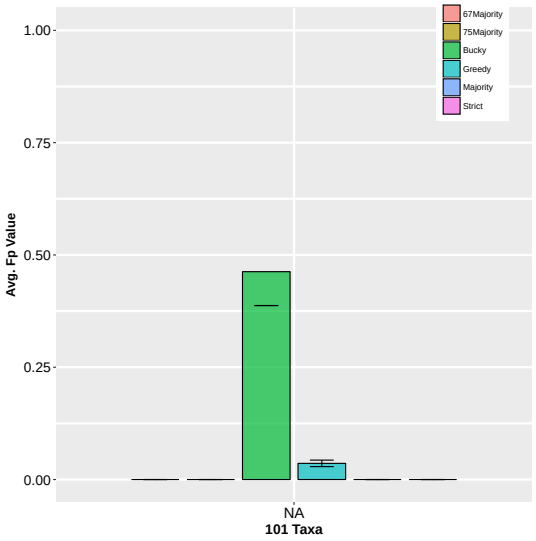


Figure 6.13: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCky in respect of Average FP value

Figure 6.14 shows the RF values of various consensus method on 101-taxon simulated dataset. Here, we see that $\frac{2}{3}$ rd Majority and $\frac{3}{4}$ th Majority method has lower RF values than Strict and BUCKy methods but higher values than Majority and Greedy methods.

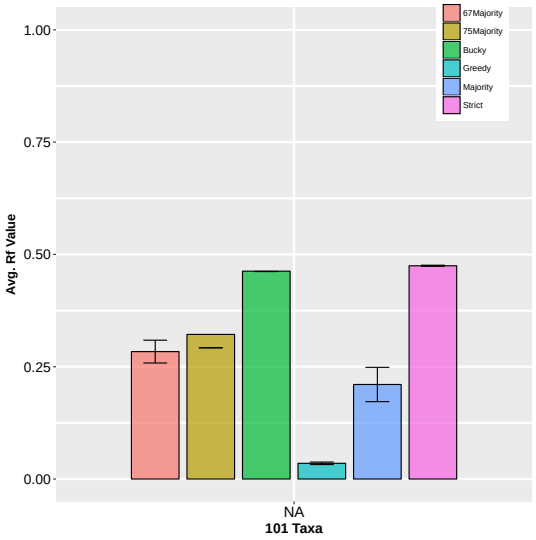


Figure 6.14: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCky in respect of Average RF value

6.2.7 500-1000 Taxon Simulated Dataset

In the large taxon,we faced a lot of difficulties to run the methods.It took a lot of time to run the methods.BUCKy wasn't compatible for large files.Greedy gave pretty good result in large taxons in respect of FP rates.

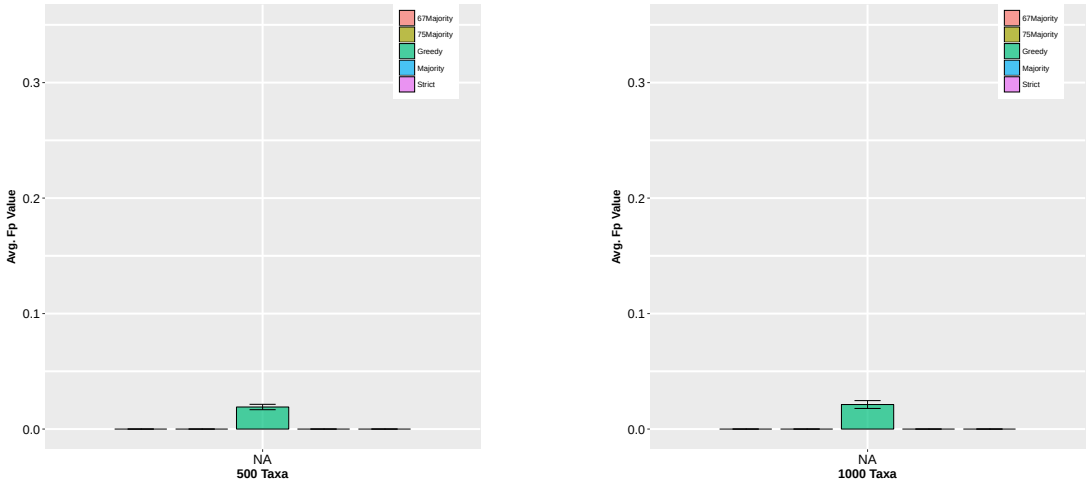


Figure 6.15: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average FP value

In following figure 6.16 the RF values of various consensus methods on 500-taxon and 1000-taxon are shown.Here, we can see that result remained almost same for both datasets. $\frac{2}{3}$ rd Majority method performed better than Strict and $\frac{3}{4}$ rd Majority method. Greedy method has the lowest RF value among these method.

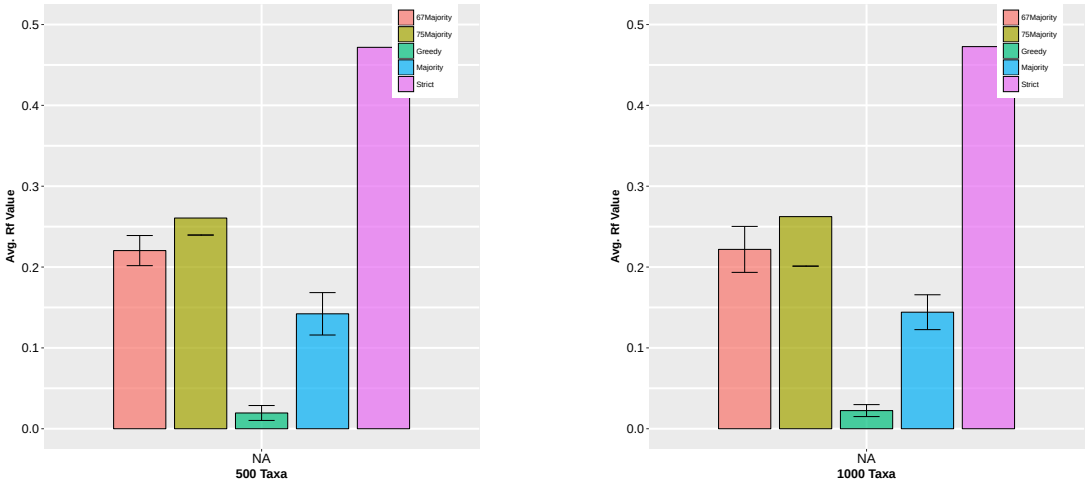


Figure 6.16: Comparison among Strict,Majority,$\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority Method,BUCKy in respect of Average RF value

6.3 Results on Biological Datasets

6.3.1 Mammalian Dataset

We analyzed the mammalian dataset from [23] containing 447 genes across 37 mammals.Majority method, $\frac{2}{3}$ rd Majority,$\frac{3}{4}$ th Majority give same tree as output.In these trees Red Junglefowl (Chicken) and Turkey are sisters which is consistent with the analysis of MP-EST [28].Most of the leaves were unresolved in the above mentioned methods.Greedy method gives a resolved tree but there are multiple false negative edge if we consider the MP-EST analysis as true.

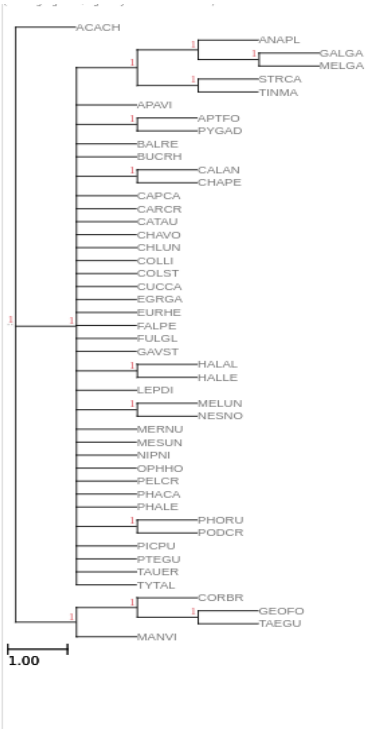


Figure 6.17: $\frac{2}{3}$ rd Majority Method on Mammalian Dataset

6.3.2 Avian Dataset

The avian biological dataset, which contains 14 446 loci from 48 taxa, was re-analyzed (including exons, introns and UCEs). Separately, we looked at 8251 exon, 2516 intron, and 3679 UCE gene trees.This is a difficult dataset to analyze because it has a lot of gene tree disagreement, maybe due to their ancestors' rapid radiation. [29]. In Majority method American Crow,Medium Ground-finch and Zebra Finch have been grouped together with Golden-collared Manakin which is present in the MP-EST analysis. [28]. On the other hand ,this grouping is missing in $\frac{2}{3}$ rd Majority methods' analysis.Both of the methods couldn't reconstruct the well-established Australaves clade (passeriformes, parrot, falcon and seriema) and Columbea(flamingo, grebe, pigeon, mesite and sandgrouse).The Strict method gives a completely unresolved tree.Greedy gives a resolved one where many of the groups are reconstructed.

Figure 6.18: MP-EST analysis on Avian Dataset

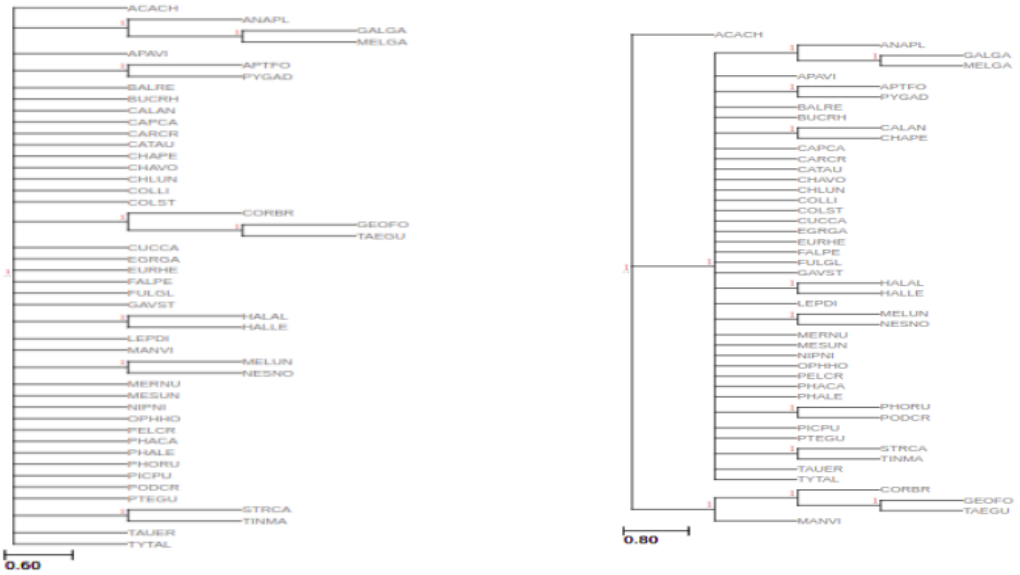


Figure 6.19: $\frac{2}{3}$ rd Majority and Majority analysis on Avian dataset

Chapter 7

Conclusions

One of the most difficult aspects of evolution theory is reconstructing the tree of Life which we refer to as species tree.Due to gene tree discordance, reconstructing a species tree from a set of gene trees has proven to be difficult.This thesis contributes to the challenge of fast and accurate estimating species trees from gene trees quickly and accurately in the presence of gene tree discordance. This thesis is mainly focused on consensus methods.

In this thesis we proposed a novel consensus method to estimate species tree in faster and more accurate way. We proved theoretically that our $\frac{2}{3}$ Majority model will have zero False Positive rate (FP). We simulated our methods and compared with other existing consensus methods and got some interesting results. Our model zero False positive rate (FP) for all tested datasets just as our theoretical proved suggested. We have seen our model operate better than Strict method and have way lesser RF value changes than Strict method with increase in number of gene trees, which indicates that our method was less susceptible to outliers.However, this study can be expanded in numerous ways. This study is limited to small to moderate size datasets. Also the species tree produced by our model are non-binary. Future studies need to investigate a way to resolve the non-binary node to have a way more accurate result. On an ending note, our paper shows that the idea of estimating species trees with consensus condition between Strict and Majority method has merit and should be pursued and used in the future phylogenomic studies.

References

[1] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch, "Predicting the evolution of human influenza a," *Science*, vol. 286, no. 5446, pp. 1921–1925, 1999.

[2] J. S. Brown, "Brooks, d. r., and d. a. mclennan. 1991. phylogeny, ecology, and behavior: A research program in comparative biology. the university of chicago press, chicago, 434 pp. isbn 0-226-07572-9.," *Journal of Mammalogy*, vol. 75, pp. 243–246, 02 1994.

[3] P. H. Harvey, M. D. Pagel, *et al.*, *The comparative method in evolutionary biology*, vol. 239. Oxford university press Oxford, 1991.

[4] C. R. Linder and T. Warnow, "An overview of phylogeny reconstruction," 2001.

[5] K. M. Halanych and L. R. Goertzen, "Grand challenges in organismal biology: the need to develop both theory and resources," *Integrative and comparative biology*, vol. 49, no. 5, pp. 475–479, 2009.

[6] N. Amenta, F. Clarke, and K. St John, "A linear-time majority tree algorithm," in *International Workshop on Algorithms in Bioinformatics*, pp. 216–227, Springer, 2003.

[7] J. H. Degnan, M. DeGiorgio, D. Bryant, and N. A. Rosenberg, "Properties of consensus methods for inferring species trees from gene trees," *Systematic Biology*, vol. 58, no. 1, pp. 35–54, 2009.

[8] J. Felsenstein and J. Felenstein, *Inferring phylogenies*, vol. 2. Sinauer associates Sunderland, MA, 2004.

[9] W.-K. Sung, *Algorithms in bioinformatics: A practical introduction*. Chapman and Hall/CRC, 2009.

[10] R. D. M. Page, "Genes, organisms, and areas: The problem of multiple lineages," *Systematic Biology*, vol. 42, no. 1, pp. 77–84, 1993.

[11] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda, "Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences," *Systematic Biology*, vol. 28, no. 2, pp. 132–163, 1979.

[12] D. Robinson and L. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1, pp. 131–147, 1981.

[13] M. K. Kuhner and J. Yamato, "Practical Performance of Tree Comparison Metrics," *Systematic Biology*, vol. 64, pp. 205–214, 12 2014.

[14] D. Bryant, "A classifica of co sensus methods for phylogenetics," in *Bioconsensus: DIMACS Working Group Meetings on Bioconsensus: October 25-26, 2000 and October 2-5, 2001, DIMACS Center*, vol. 61, p. 163, American Mathematical Soc., 2003.

[15] B. R. Larget, S. K. Kotha, C. N. Dewey, and C. Ané, "Bucky: gene tree/species tree reconciliation with bayesian concordance analysis," *Bioinformatics*, vol. 26, no. 22, pp. 2910–2911, 2010.

[16] L. Xin, B. Ma, and K. Zhang, "A new quartet approach for reconstructing phylogenetic trees: Quartet joining method," in *International Computing and Combinatorics Conference*.

[17] J. Dong, D. Fernández-Baca, F. McMorris, and R. C. Powers, "Majority-rule (+) consensus trees," *Mathematical Biosciences*, vol. 228, no. 1, pp. 10–15, 2010.

[18] J. Jansson, C. Shen, and W.-K. Sung, "Improved algorithms for constructing consensus trees," *Journal of the ACM (JACM)*, vol. 63, no. 3, pp. 1–24, 2016.

[19] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Determining species tree topologies from clade probabilities under the coalescent," *Journal of theoretical biology*, vol. 289, pp. 96–106, 2011.

[20] S. Mirarab, M. S. Bayzid, B. Boussau, and T. Warnow, "Statistical binning enables an accurate coalescent-based estimation of the avian tree," *Science*, vol. 346, no. 6215, p. 1250463, 2014.

[21] S. Mirarab and T. Warnow, "Astral-ii: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes," *Bioinformatics*, vol. 31, no. 12, pp. i44–i52, 2015.

[22] Y. Chung and C. Ané, "Comparing two bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer," *Systematic biology*, vol. 60, no. 3, pp. 261–275, 2011.

[23] S. Song, L. Liu, S. V. Edwards, and S. Wu, "Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model," *Proceedings of the National Academy of Sciences*, vol. 109, no. 37, pp. 14942–14947, 2012.

[24] M. Mahbub, Z. Wahab, R. Reaz, M. S. Rahman, and M. S. Bayzid, "wqfm: Statistically consistent genome-scale species tree estimation from weighted quartets," *bioRxiv*, 2020.

[25] M. S. Bayzid and T. Warnow, "Supplementary materials, naive binning improves phylogenomic analyses,"

[26] Y. Yu, T. Warnow, and L. Nakhleh, "Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles," *Journal of Computational Biology*, vol. 18, no. 11, pp. 1543–1559, 2011.

[27] Z. Xi, L. Liu, J. S. Rest, and C. C. Davis, "Coalescent versus concatenation methods and the placement of amborella as sister to water lilies," *Systematic biology*, vol. 63, no. 6, pp. 919–932, 2014.

[28] S. Mirarab, M. S. Bayzid, and T. Warnow, "Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting," *Systematic Biology*, vol. 65, no. 3, pp. 366–380, 2016.

[29] E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, *et al.*, "Whole-genome analyses resolve early branches in the tree of life of modern birds," *Science*, vol. 346, no. 6215, pp. 1320–1331, 2014.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Tuesday 11th October, 2022 at 5:04pm.