

M.Sc. Engg. (CSE) Thesis

**Estimating species trees from multi-locus data in the presence of incomplete lineage sorting by maximizing quartet consistency and minimizing deep coalescence**

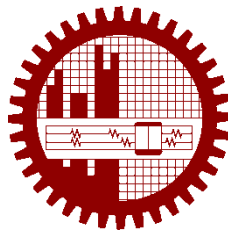
Submitted by

Ishrat Tanzila Farah

1017052063

Supervised by

Dr. Md. Shamsuzzoha Bayzid



Submitted to

**Department of Computer Science and Engineering**  
**Bangladesh University of Engineering and Technology**  
Dhaka, Bangladesh

in partial fulfillment of the requirements for the degree of  
Master of Science in Computer Science and Engineering

March 2022

## Candidate's Declaration

I, do, hereby, certify that the work presented in this thesis, titled, "Estimating species trees from multi-locus data in the presence of incomplete lineage sorting by maximizing quartet consistency and minimizing deep coalescence", is the outcome of the investigation and research carried out by me under the supervision of Dr. Md. Shamsuzzoha Bayzid, Associate Professor, Department of CSE, BUET.

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

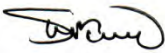


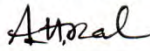
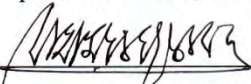
Ishrat Tanzila Farah

Ishrat Tanzila Farah

1017052063

The thesis titled “**Estimating species trees from multi-locus data in the presence of incomplete lineage sorting by maximizing quartet consistency and minimizing deep coalescence**”, submitted by Ishrat Tanzila Farah, Student ID 1017052063, Session October 2017, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfilment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents on March 1, 2022.

### Board of Examiners

1.   
\_\_\_\_\_
- Dr. Md. Shamsuzzoha Bayzid  
Associate Professor  
Department of CSE, BUET, Dhaka
- Chairman  
(Supervisor)
2.   
\_\_\_\_\_
- Dr. A.K.M. Ashikur Rahman  
Professor and Head  
Department of CSE, BUET, Dhaka
- Member  
(Ex-Officio)
3.   
\_\_\_\_\_
- Dr. Md. Saidur Rahman  
Professor  
Department of CSE, BUET, Dhaka
- Member
4.   
\_\_\_\_\_
- Dr. Atif Hasan Rahman  
Assistant Professor  
Department of CSE, BUET, Dhaka
- Member
5.   
\_\_\_\_\_
- Dr. Mohammad Kaykobad  
Distinguished Professor  
Department of CSE  
Brac University
- Member  
(External)

## Acknowledgement

All praise to the Almighty for His countless blessings upon me. Apart from that, I would like to thank the people who were always there for me throughout this erratic as well as illuminating journey.

Firstly, I would like to convey my sincere gratitude to my supervisor Dr. Md. Shamsuzzoha Bayzid Sir. When I started working on this research, this was merely a course project which gradually extended towards being an extensive experimental study. As I didn't have any prior experience in doing research work, I made a lot of silly mistakes and asked for his guidance in almost each and every step. Every time he showed me the correct path with utmost patience, and guided me in all possible ways he could from the very beginning till the end. It took quite a long time for me to complete this thesis due to my personal and professional reasons, and often I had to take a break from this research. I am truly grateful to my supervisor for always understanding my circumstances and encouraging me to continue my research despite all the ups and downs. I really feel blessed to have such a humble person as my supervisor!

I cannot deny the continuous support and encouragement of my husband Md. Muktadirul Islam throughout this journey. Whenever I fell down, he gave me the strength to stand up again; sometimes being a friend, and sometimes being my research partner. Another person to whom my gratitude cannot be expressed in words is my mother. She has always created such a world for me where I never had to worry about anything despite my studies, and could provide my complete concentration on my career and studies only. I believe without my husband and my mother by my side, I wouldn't have been where I am today.

Lastly, I would like to thank my undergraduate course teacher, Shantanu Shipan Sarker Sir. He is the person who insisted me to pursue a Masters at BUET, and I still remember the day when I got the chance at BUET and he was the first person whom I informed. You were more than happy for me Sir, and I know you are still smiling (as always) at me from the heaven and feeling proud of me. No matter whatever I achieve in my life, you will always be there in my heart as a mentor, an idol, and an encouragement who considered me more like his sister than his student.

Dhaka  
March 1, 2022

Ishrat Tanzila Farah  
1017052063

# Contents

<b>Candidate's Declaration</b>	<b>i</b>
<b>Board of Examiners</b>	<b>ii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>viii</b>
<b>Abstract</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
Phylogenetic tree .....	1
Species trees and gene trees .....	1
Species tree estimation from gene trees: state of the art and knowledge gaps .	3
1.4 Our contributions.....	4
1.5 Thesis organization.....	5
<b>2 Preliminaries</b>	<b>6</b>
Phylogenies .....	6
Gene tree and species tree .....	8
Gene tree-species tree discordance .....	8
Incomplete Lineage Sorting (ILS) .....	8
Gene tree reconciliation.....	9
Statistical consistency .....	11
Phylogenomic analysis pipeline .....	11
Concatenation .....	11
Summary methods .....	12
Evaluation of species tree estimation methods.....	13
Evaluation on simulated datasets.....	13
Error metrics .....	13

Evaluation on real biological datasets .....	15
<b>3 Related work</b>	<b>16</b>
Species tree estimation from multi-locus data under ILS.....	16
Species tree estimation under MDC and MQC .....	17
Minimize Deep Coalescence (MDC).....	17
Maximize Quartet Consistency (MQC).....	18
Tools Used.....	19
Phylonet-MDC .....	19
ASTRAL-III .....	19
Phylogenetic terraces.....	20
<b>4 Pseudo terraces in species tree estimation from gene trees</b>	<b>21</b>
Pseudo EL terrace.....	22
Characterization of the trees in a pseudo EL terrace .....	23
Pseudo quartet terrace .....	27
Characterization of the trees in a pseudo quartet terrace .....	27
Quartet scores of the trees in a neighborhood of a species tree .....	28
Additional Remarks .....	30
<b>5 Experimental Studies</b>	<b>31</b>
Datasets .....	31
Materials and Methods .....	32
Measurements.....	33
Results on datasets simulated from a biological example (37-taxon mam-	
malian dataset).....	33
Results on 11-taxon dataset .....	39
Consensus trees of the trees in a pseudo terrace.....	43
Results on biological dataset .....	48
Amniota dataset .....	48
Mammalian dataset.....	50
<b>6 Conclusions</b>	<b>52</b>
<b>References</b>	<b>55</b>

# List of Figures

1.1 A Phylogenetic tree of coronavirus strains.....	2
A phylogenetic tree.....	7
Unrooted tree and rooted tree.....	8
Gene tree-species tree discordance.....	9
Reconciliation of a discordant gene tree-species tree pair under incomplete lineage sorting.....	10
Optimal and non-optimal reconciliations under the deep coalescence model.....	11
Two approaches for constructing species trees from gene trees.....	12
A simulation protocol for evaluating species tree estimation techniques.....	14
False negative rate or, missing branch rate.....	15
Extra lineages inside a pectinate and a balanced symmetric tree for the cases when gene lineages do not coalesce with each other on the internal branches until they reach the root of the tree.....	25
Computing the extra lineage score.....	25
NNI move on an internal edge.....	29
Comparison of ASTRAL and Phylonet on 37-taxon simulated mammalian dataset.....	35
EL scores for ASTRAL, Phylonet and true species tree on 37-taxon simulated mammalian dataset.....	37
Quartet and EL terraces on 37-taxon dataset.....	39
Demonstration of quartet and EL terraces in 37-taxon dataset.....	40
Quartet and EL terraces in 37-taxon dataset.....	40
Average FN rates of ASTRAL and Phylonet on 11-taxon dataset.....	41
Extra lineage scores for ASTRAL and Phylonet on 11-taxon dataset under various model conditions.....	43
Quartet and EL terraces on 11-taxon dataset.....	44
Quartet and EL terrace in 11-taxon dataset.....	46
Demonstration of quartet and EL terrace in 11-taxon dataset.....	47

Analysiss of the Amniota AA dataset by maximizing quartet score (ASTRAL) and minimizing extra lineage score (Phylonet) .....	49
Analysis of the Amniota DNA dataset by maximizing quartet score (ASTRAL) and minimizing extra lineage score (Phylonet) .....	49
Analyses of the mammalian dataset by maximizing quartet score (ASTRAL) and minimizing deep coalescence (Phylonet) .....	51



# List of Tables

4.1 Pseudo EL terraces in the tree space of four taxa with respect to a set $G$ of rooted and binary gene trees.....	24
Properties of the simulated datasets .....	32
Quartet scores for ASTRAL, Phylonet and true species tree on 37-taxon simulated mammalian dataset. ....	36
Quartet and EL scores of the ASTRAL- and Phylonet-estimated trees and the model species tree for the model condition analyzed in Figure 5.3.....	38
Quartet scores of ASTRAL- and Phylonet-estimated trees and the model species tree on 11-taxon datasets under various model conditions. ....	42
EL scores of ASTRAL and Phylonet estimated trees and the model species tree on 11-taxon datasets under various model conditions.....	45
Quartet terraces on 11-taxon dataset .....	45
EL terraces on 11-taxon dataset .....	46
Pseudo EL and quartet terraces on around 2.2 million candidate species trees for the 11-taxon dataset. ....	47
Comparison of the optimality scores of the consensus trees and the corresponding pseudo terraces. ....	48
Quartet and EL scores of ASTRAL- and Phylonet-estimated trees on the amniota dataset (both DNA and AA). ....	50
Quartet and EL scores of ASTRAL- and Phylonet-estimated trees on the biological mammalian dataset. ....	50

## Abstract

Species tree estimation from multi-locus datasets is extremely challenging, especially in the presence of gene tree heterogeneity across the genome due to incomplete lineage sorting (ILS). *Summary methods* have been developed which estimate gene trees and then combine the gene trees to estimate a species tree by optimizing various optimization scores. In this study, we have extended and adapted the concept of phylogenetic terraces to species tree estimation by “summarizing” a set of gene trees, where multiple species trees with distinct topologies may have exactly the same optimality score (i.e., *quartet score*, *extra lineage score*, etc.). We particularly investigated the presence and impacts of equally optimal trees in species tree estimation from multi-locus data using summary methods by taking ILS into account. We analyzed two of the most popular ILS-aware optimization criteria: *maximize quartet consistency* (MQC) and *minimize deep coalescence* (MDC). Methods based on MQC are provably statistically consistent, whereas MDC is not a consistent criterion for species tree estimation. We present a comprehensive comparative study of these two optimality criteria. Our experiments, on a collection of datasets simulated under ILS, indicate that MDC may result in competitive or identical quartet consistency score as MQC, but could be significantly worse than MQC in terms of tree accuracy – demonstrating the presence and impacts of equally optimal species trees. This is the first known study that provides the conditions for the datasets to have equally optimal trees in the context of phylogenomic inference using summary methods.

# Chapter 1

## Introduction

### Phylogenetic tree

Phylogenetic trees (also known as evolutionary trees) is a representation of the evolutionary relationships of a set of entities (species, genes, etc.). Phylogenetic trees provide insights into basic biology, including how life evolved, the mechanisms of evolution and how it modifies function and structure, biodiversity, medical diagnosis, drug design, and criminal investigation. Various organisms on earth are genetically related, and the relationships of living things can be represented by a vast evolutionary tree – the “Tree of Life.” The Tree of Life is one of the most ambitious goals and grand challenges of modern science [1]. The ability to efficiently analyze the vast amount of genomic data available these days due to significant advancements in sequencing techniques, is critical to assembling this tree of life. Figure 1.1 shows a phylogenetic tree of various coronavirus strains.

### Species trees and gene trees

A *species tree* represents the evolutionary history of a group of organisms, while a *gene tree* represents the evolution of a particular *gene* within a group of species. Species tree estimation from multiple genes has become an emerging field of study in comparative and evolutionary biology. Also, it has drawn significant attention from the systematists due to the high availability of molecular data. Estimation of species trees give us insights on how different species have evolved from their common ancestors. Because a species contains many genes, one may expect the species tree to match the evolutionary histories of the genes present inside the species. However, various biological processes can result in different genes having different evolutionary histories. As a result, a gene tree could be discordant to the species tree. This incongruence between a gene tree and its containing species tree is known as *gene*

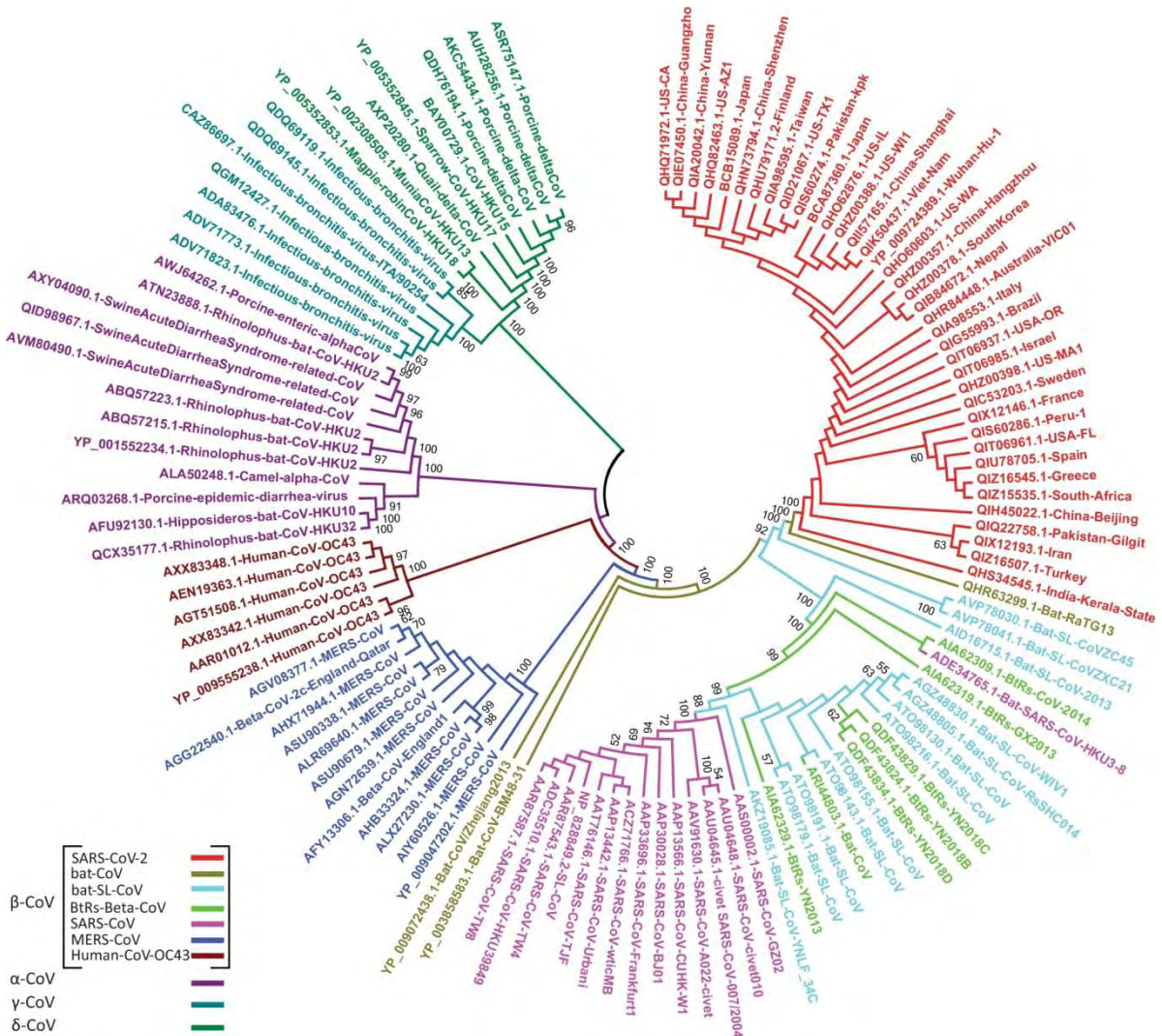


Figure 1.1: A Phylogenetic tree representing the evolution of different coronavirus strains (SARS-CoV-2). This figure has been taken from [2]

*tree - species tree discordance*. This discordance can be caused by a variety of biological reasons such as, incomplete lineage sorting (ILS), horizontal gene transfer, gene duplication and loss etc. Among these reasons, ILS is considered to be a dominant cause for gene tree heterogeneity.

## **Species tree estimation from gene trees: state of the art and knowledge gaps**

Reconstructing species trees from a set of gene trees is not an easy task, especially in the presence of gene tree - species tree discordance. In the presence of gene tree - species tree discordance, species tree estimation involves the estimation of trees and alignments on many different genes, so that the species tree can be based upon many different parts of the genome. Species trees can be estimated from a collection of gene sequences in two fundamental ways: i) combined analysis, and ii) summary method. In combined analysis, the individual gene sequences are concatenated with each other to create a super gene sequence (which is called the “supermatrix”), and the species tree is estimated from that concatenated gene sequence. Combined analysis does not take the gene tree discordance into account, and hence, it can be statistically inconsistent [3, 4], and produce incorrect trees with high support [5]. In contrast, in summary methods, individual gene trees are constructed from the corresponding individual gene sequences and the species tree is finally estimated by reconciling (“summarizing”) all the gene trees. Summary methods can explicitly take the reason for gene tree discordance into account during the species tree estimation process and thus some of them are provably statistically consistent, meaning that they will converge in probability to the true species tree given sufficiently large numbers of genes and sites per gene.

Fundamental to developing summary methods is to find appropriate optimization criteria to summarize the given collection of gene trees. Maximizing quartet consistency (MQC) is one of the leading optimization criteria for estimating statistically consistent species trees from gene trees in the presence of ILS [6–9]. MQC seeks a species tree that is consistent with the largest number of quartets induced by the set of gene trees. ASTRAL [6, 10], which is one of the most accurate and popular coalescent based summary methods, solves this optimization problem.

Another optimization problem that takes ILS into account is ‘Minimize Deep Coalescence’ (MDC), which was first introduced by Maddison in his seminal paper [11] and was further investigated in [12, 13]. In the presence of gene tree discordance due to ILS, the MDC criterion seeks the species tree that minimizes the number of “extra-lineages” (resulting from “deep” coalescence events) required for a given collection of gene trees [12, 14, 15]. This

is computed by embedding a gene tree  $gt$  into the species tree  $T$ , and then counting the number of lineages on each edge of the species tree. The number of *extra lineages* on an edge is one less than the total number of lineages on that edge [16]. Thus, it is a natural candidate for species tree inference, under the parsimony setting, when discordance among gene trees is caused by incomplete lineage sorting. Several exact algorithms and heuristics for implementing this criterion have been developed [12, 17]. Phylonet [18] is a popular tool for estimating species trees under MDC, which has both exact and heuristic versions similar to ASTRAL. Although it has been shown to be statistically inconsistent under multi-species coalescent model [19], this is not agnostic to the gene tree heterogeneity as it takes into account the specific nature of the way incomplete lineage sorting occurs. Simulation studies have suggested a high degree of accuracy of species tree estimates obtained by this criterion [12, 20]. Yet, to our knowledge, the accuracy of species trees estimated under MDC has not been explored in comparison to highly accurate ILS-aware and statistically consistent summary methods like ASTRAL. Although statistically consistent methods like ASTRAL are expected to perform better than statistically inconsistent methods like Phylonet, it is important to evaluate the relative performance under various realistic model conditions as even the best coalescent-based summary methods are sensitive to gene tree estimation error and can be worse than concatenation in some cases (mostly when the gene trees have poor phylogenetic signal or when the level of ILS is low) [21–23]. There have been a few studies [6, 21, 24, 25], which present comparisons among concatenation and various summary methods, including ASTRAL, MP-EST, BUCKy, NJst, and SVDquartets. Huang *et al.* [26] evaluated STEM and Phylonet in the presence of mutational and coalescent variance. However, to the best of our knowledge, there is no study evaluating MQC and MDC.

## Our contributions

This study shows that MQC is in general a better optimization criterion compared to MDC. However, MDC achieves competitive results on some of the model conditions that we analyzed in this study. Interestingly, this study reveals that the search under MDC criterion may result in trees that have competitive *quartet score* ( $QS$ ) (number of quartets induced by the gene trees that a tree is consistent with) compared to the trees estimated under MQC criterion. However, trees estimated under the MDC criterion are generally significantly worse than the trees identified by MQC criterion. As we will show in the following sections it is, in fact, expected that even for small numbers of species and genes there will be sets of trees with identical scores. In previous studies, sets of trees having optimality scores greater than a threshold were described as “islands” [27, 28], and those with identical scores were termed phylogenetic terraces [29–32], which occur in the presence of missing data. Here, for the first

time, we present evidence for equally optimal species trees in the context of summary methods using MDC and MQC optimization criteria. The phylogenetic terraces refer to regions of tree space having identical optimality scores purely due to certain patterns of missing data among the taxa sampled. We extend this concept to the species tree - gene tree context, and show that specific combinatorial properties of species trees with respect to gene trees imply “equally good trees” – which we call *pseudo species tree terrace*. In particular, this study entails the following contributions.

- We have introduced the concept of pseudo species tree terraces in the context of constructing species trees from a collection of gene trees using summary methods.
- We showed analytical results showing that pseudo species tree terraces are obvious especially when we have a large number of taxa. We also proved combinatorial characteristics, for both MDC and MQC criteria, about which tree topologies will share the same score.
- We investigated MDC (Phylonet) and MQC (ASTRAL) criteria in terms of tree accuracy under different model conditions with varying numbers of genes, amounts of ILS, and gene sequence lengths. We also investigated their performance on real biological datasets.
- We systematically analyzed, through simulation studies, the presence and impacts of pseudo species tree terraces in species tree inference using summary methods under MDC and MQC criteria.
- We investigated and demonstrated the applicability of consensus trees to handle the ambiguity—in finding an optimal tree under a particular optimization criterion—resulting from the presence of pseudo species tree terraces.

## Thesis organization

This thesis is organized as follows. Chapter 2 describes the background materials related to the the problem of species tree estimation from gene trees in the presence of gene tree discordance. Chapter 3 discusses the two optimization criteria that we have analyzed in this experimental study and relevant prior studies. In Chapter 4, we formally introduce the concept of pseudo terraces in the context of species tree estimation. Chapter 5 presents the experimental results on a collection of simulated and biological datasets. Finally, we conclude in Chapter 6 with a brief discussion of our contributions and several future research directions.



# Chapter 2

## Preliminaries

This chapter describes the basic definitions and concepts that will be used throughout this thesis <sup>1</sup>. We begin with a discussion on the concept of phylogeny. Next, we have discussed the concepts of gene trees and species trees, and some concepts associated with these trees such as the discordance between gene trees and species tree, Incomplete Lineage Sorting (ILS) as a reason of gene tree-species tree discordance etc. Later, the traditional pipelines for phylogenomic analysis have been discussed along with the evaluation of species tree estimations methods on simulated and real biological datasets. Lastly, we have discussed the error metric that has been used in this study to evaluate species tree reconstruction methods. Terminologies that are not included in this section have been introduced later as they are needed.

### Phylogenies

A phylogeny represents the evolutionary relationships among a set of entities such as species, genes, languages etc.). Phylogenetic entities are in general known as *taxa*. A *tree*  $T$  is a connected acyclic graph with a set of vertices  $V$  and a set of edges  $E$ . The evolutionary history of phylogenetic entities can be best represented using trees, which are particularly called *phylogenetic trees*. The *leaf* nodes of a phylogenetic tree typically represent the existing taxa, whereas the internal nodes represent the hypothetical ancestral taxa from which the descendant taxa are considered to have evolved. Which means, the internal nodes represent extinct species that existed in the past, but do not exist at present. A *branch*, or an *edge*  $e = (u, v) \in E$  of a phylogenetic tree represents an evolutionary relationship between the two taxa represented by the nodes  $u$  and  $v$ . Throughout this thesis, we denote the set of nodes of a phylogenetic tree  $T$  by  $V(T)$ , the set of internal nodes by  $V_{int}(T)$ , the set of edges by

---

<sup>1</sup>Most of the material in this chapter has been taken/adapted from [33] and [34].



$E(T)$ , and the set of leaf nodes by  $L(T)$ .

An example of a phylogenetic tree is shown in Figure 2.1 that displays the evolutionary history of primates e.g. orangutans, gorillas, chimpanzees and humans. According to this phylogenetic tree, humans are more closely related to chimpanzees than they are to gorillas and orangutans, as humans and chimpanzees share a common nearest ancestor.



Figure 2.1: **A phylogenetic tree.** A phylogenetic tree representing the evolutionary history of primates: orangutans, gorillas, chimpanzees and humans.

The length of the branches in an evolutionary tree is known as the *branch length*, which is a non-negative real number that represents various quantities measured on a branch such as the rate or amount of evolutionary change, or the amount of time between two taxa. Trees that are not provided with any branch lengths are usually referred as *topologies*.

A phylogenetic tree can be rooted, or unrooted. Phylogenetic trees can be rooted by designating a single vertex  $r \in V$  as the root of the tree. A rooted tree best represents the true evolutionary histories among a set of species, but to locate the actual root of the tree is a complex problem that requires specific knowledge of the set of taxa being studied or the assumption of a “*molecular clock*” to accurately root that phylogenetic tree. Often, phylogenetic trees are being rooted using an *outgroup*. An outgroup is a taxon that is known to have branched off before all the other taxa under consideration. Conversely, a phylogenetic tree can be rooted based on the estimated time between speciation. In that case, the molecular data used to reconstruct that phylogeny are assumed to have a constant rate of evolution overtime. However, this assumption is often violated in real datasets. Figure 2.2 shows an example of an unrooted tree and its corresponding rooted tree. We denote the root of a tree  $T$  by  $root(T)$ .

A group of taxa that are more closely related to each other than they are to any other taxon in

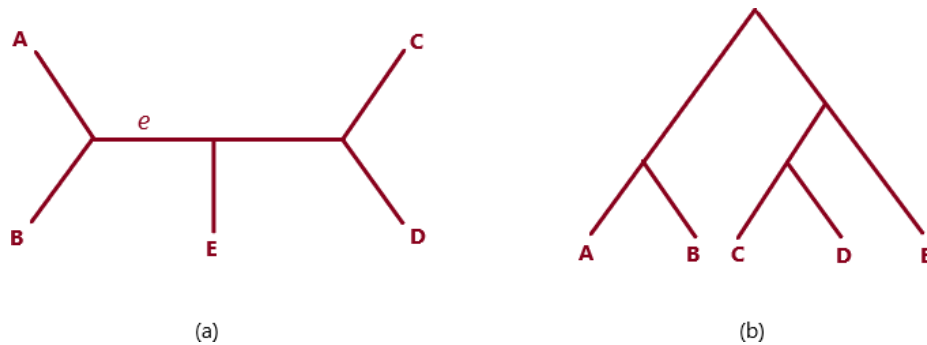


Figure 2.2: **Unrooted tree and rooted tree.** (a) An unrooted tree, and (b) the rooted tree resulting from rooting the unrooted tree on the edge  $e$  shown in (a).

that rooted tree is called a *clade*. If  $T$  is a phylogenetic tree, then a clade  $clade_T(v)$  is actually a rooted subtree of  $T$ , where the clade is rooted at node  $v$ . A *cluster*  $c_T(v)$  is a set of leaves of a clade at node  $v$ .

## Gene tree and species tree

A species tree represents the pattern of branching of species lineages through speciation, whereas, a gene tree shows how a particular “gene” evolves within a group of species. The gene trees are contained within the species trees [16].

### Gene tree-species tree discordance

The gene trees contained within the species trees may have discordant evolutionary histories due to various biological processes. An example of discordance between a species tree and a gene tree is shown in figure 2.3. From the figure it is evident that species  $A$  and species  $B$  are “sister” species in the species history, although  $B$  is closer to  $C$  than  $A$  in the gene history.

### Incomplete Lineage Sorting (ILS)

Although there are various reasons that might result in discordance between gene trees and species tree, **Incomplete Lineage Sorting (ILS)** is considered to be a dominant cause for gene tree heterogeneity, which is best understood under the coalescent model [35–42]. According to the coalescent model, the evolutionary process is considered to be operated backwards

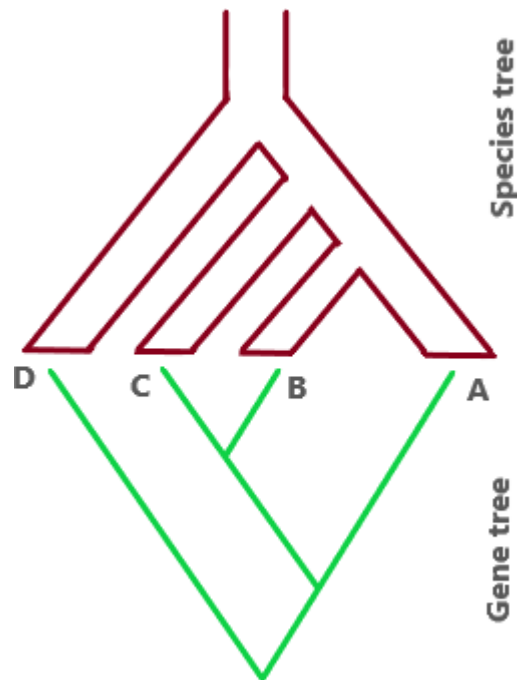


Figure 2.3: **Gene tree-species tree discordance.** A species tree (on top) and a gene tree (on bottom) on the same set  $\{A, B, C, D\}$  of taxa with different topologies.

in time, and the gene lineages are connected to a common ancestor through the process of “coalescence”.

ILS or deep coalescence refers to the case in which two lineages fail to coalesce at their speciation point, hence the gene copies at a single locus extends deeper towards their ancestors. Figure 2.4 shows an example of incongruence due to incomplete lineage sorting. Considering the lineages to go back in time, from the figure we can see that the gene copies within species *A* and *B* at first meet at their corresponding speciation point which is the recent-most common ancestor of species *A* and *B*), but they fail to coalesce at the speciation point. Hence, both of these gene copies extends backward in time, and thus we have two gene lineages (dashed and solid green lines in Fig. 2.4) on deeper ancestral branch. Then the gene from *B* at first coalesces with the gene from species *C*, and subsequently with the gene from *A*. In this particular example, we have two gene lineages instead of one on a branch of the tree. Therefore, this coalescence event has resulted in one extra lineage.

### Gene tree reconciliation

Reconstruction of species tree from a set of gene trees sampled from throughout the whole genome is not an easy task in the presence of gene tree-species tree discordance. Developing mathematical models to explain (or reconcile) gene tree-species tree incongruence

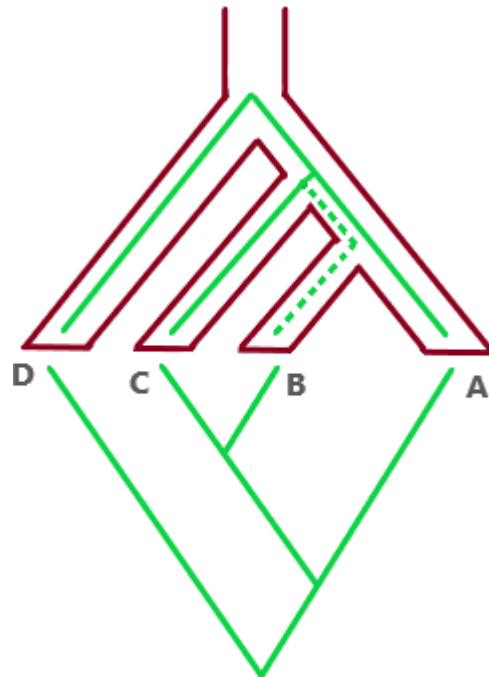


Figure 2.4: **Reconciliation of a discordant gene tree-species tree pair under incomplete lineage sorting.** Here, the tree with fat branches represents the species tree, and the reconciled gene tree is embedded inside the species tree. Going backwards in terms of time, the gene copies within species *A* and *B* first meet at their corresponding speciation point (i.e., the recent-most common ancestor of the species *A* and *B*). They fail to coalesce at their speciation point according to the species tree topology. Rather, both of the gene copies extend deeper towards their ancestor, and hence we have two gene lineages (dashed and solid green lines) on the ancestral branch. As a result, we have one extra lineage on the ancestral branch. The gene from species *B* at first coalesces with the gene from species *C*, and later with the gene from *A*.

assuming specific reasons for discordance is central to addressing the challenge in species tree estimation from a collection of gene trees. For example, in order to explain the difference between a gene tree and a species tree assuming that the discordance is due to ILS, we have to embed/map the gene tree inside the species tree using a number of deep coalescence events. Fundamental to this reconciliation problem is to find an optimal embedding (i.e., most parsimonious embedding in terms of the number of confounding evolutionary events) of the gene tree inside a species tree. Figure 2.5 shows an example of both the optimal as well as non-optimal reconciliation of a rooted, binary gene tree *gt* with respect to a rooted, binary species tree *ST* under the deep coalescence model.

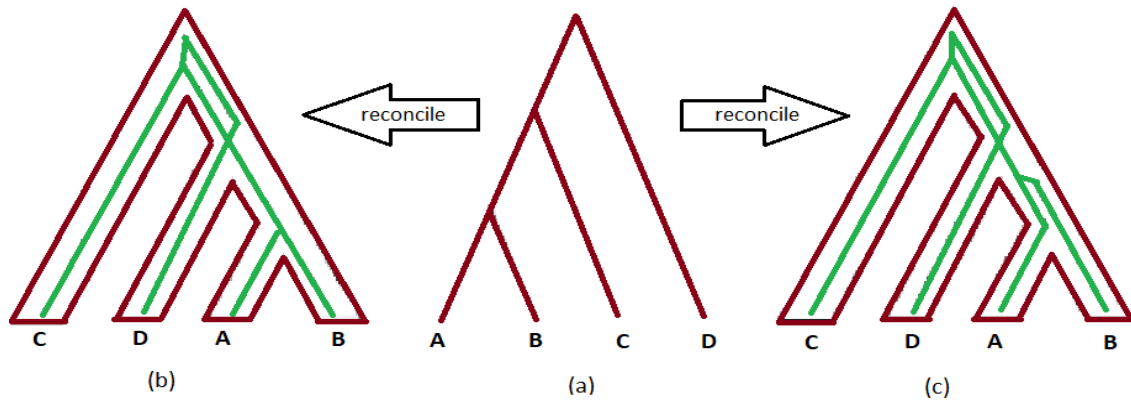


Figure 2.5: **Optimal and non-optimal reconciliations under the deep coalescence model.** (a) A rooted, binary gene tree  $gt$ , (b) an optimal reconciliation of  $gt$  with a rooted, binary species tree that yields 1 extra lineage, and (c) a non-optimal reconciliation of  $gt$  using 2 extra lineages [33].

### Statistical consistency

A statistically consistent species tree reconstruction method is such a method whose probability of returning the true species tree converges to **one** (under a particular model of evolution) as the amount of data increases. Certainly, statistically consistent methods are preferred over the methods that are not statistically consistent [22,43–47]. Various statistically consistent methods have already been developed in the last decade to estimate species tree from a set of gene trees in the presence of discordance between gene trees and species tree. Some of the leading statistically consistent methods under the multi-species coalescent model are BEAST [48], BUCKy-pop [49], MP-EST [47] and ASTRAL [6, 10] and STELAR [50].

## Phylogenomic analysis pipeline

There are several approaches for estimating species trees from a collection of gene trees. Two of the most popular approaches are **concatenation** and **summary methods** (See Figure 2.6).

### Concatenation

Concatenation, which is also known as *combined analyses*, is the most elementary pipeline for species tree estimation. In this pipeline, alignments are estimated for each gene and then they are being concatenated into a *supermatrix*. This supermatrix is subsequently used to estimate the species tree. Since concatenation combines all the gene alignments into a supermatrix, it does not consider the discordance between gene trees and their corresponding species tree. Hence, the assumption that all the genes have the same evolutionary history

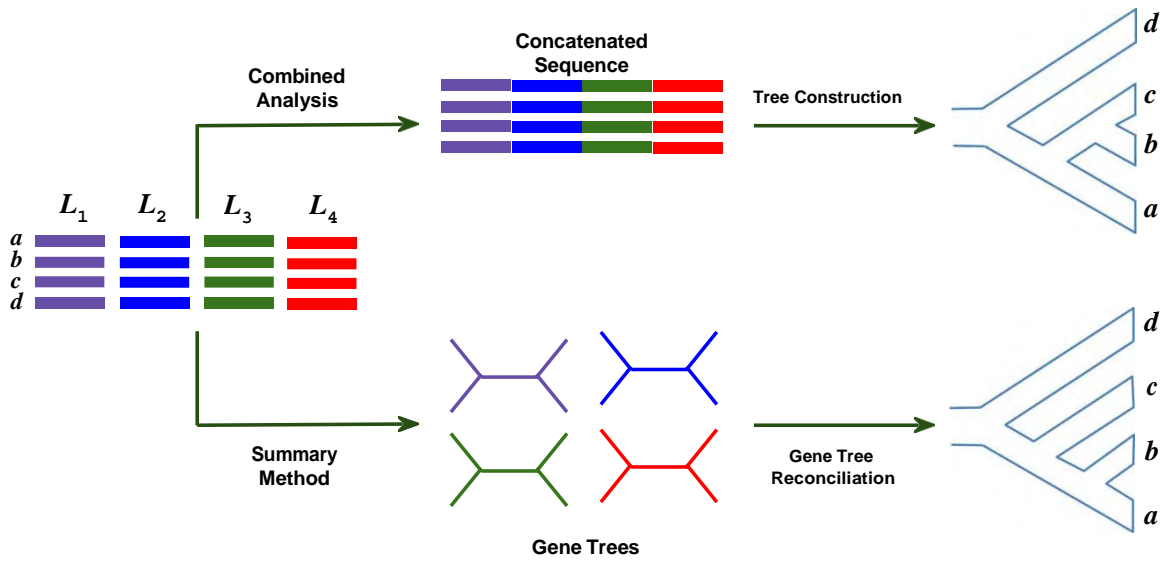


Figure 2.6: Two approaches for constructing species trees from gene trees. The figure is taken without alteration from [56].

is implicit in this approach. However, it has been proved lately that concatenation using an unpartitioned maximum likelihood analysis can be statistically inconsistent under the multi-species coalescent model [3, 4]. Also, empirical studies suggest that incorrect trees can be returned with a high confidence with this approach [22, 44, 46–49]. Nevertheless, the statistical consistency of concatenation using partitioned analyses is still not known.

### Summary methods

Summary method refers to the reconstruction of species tree by summarizing a collection of gene trees. Some examples of summary methods can be gene tree parsimony methods such as estimating species trees by *minimizing deep coalescence (MDC)* and *minimizing duplication and loss (MGDL)*. A good thing about summary methods is that they are not necessarily sceptic about the reason for gene tree-species tree discordance. Also, they can be statistically consistent. As a result, summary methods have become more popular these days as a species tree reconstruction method [6, 47, 48, 51–55].

Formally, summary methods can be defined as follows considering  $C(gt, ST)$  as the cost (e.g., the number of extra lineages, the number of consistent quartets) associated with reconciling  $gt$  with  $ST$ :

**Input:** A set  $G = gt_1, gt_2, \dots, gt_k$  of gene trees, and a reason for discordance (e.g. ILS).

**Output:** A species tree  $ST$  that minimizes  $\sum_{gt \in G} C(gt, ST)$  assuming the presence of the given reason for discordance.

## Evaluation of species tree estimation methods

Throughout this thesis, we have performed extensive experimental studies upon two species tree estimation methods: MDC (Minimizing Deep Coalescence) and MQC (Maximizing Quartet Consistency). We have used both simulated and real biological datasets for this purpose.

### Evaluation on simulated datasets

Typically, the techniques of species tree estimation from simulated datasets can be evaluated with the protocol shown in Figure 2.7, which illustrates the different steps in this simulation protocol.

- **Step 1:** The simulation study typically begins with a **model species tree**, which is also known as a *true species tree*. A model species tree can be generated using a birth-death process. Moreover, a biologically-based species tree estimated on real biological datasets from existing literature can also be chosen as a model tree.
- **Step 2:** In the next step, a set of gene trees are simulated from the model species tree under a particular model such as gene duplication and loss, ILS etc. These simulated gene trees are considered as the **true gene trees**.
- **Step 3:** Simulation of the set of gene sequences are performed next, by evolving nucleotide sequence down the true gene trees under a particular sequence evolution model.
- **Step 4:** Afterwards, the gene trees are being estimated from the gene sequence alignments, and these gene trees are called the **estimated gene trees**.
- **Step 5:** Finally, a species tree is estimated from the set of estimated gene trees using a particular method, and the **estimated species tree** is compared to the model species tree using an appropriate error metric.

### Error metrics

Since the ground truth (which we call the *model tree* or *true tree*) is known in simulation studies, the estimated species trees by the methods of consideration can be compared with the true species tree. There are several standard ways of measuring estimation errors. We now describe the **false negative rate** as the error metric as we have used this metric in our study being a widely used error metric to quantify the reconstruction errors.

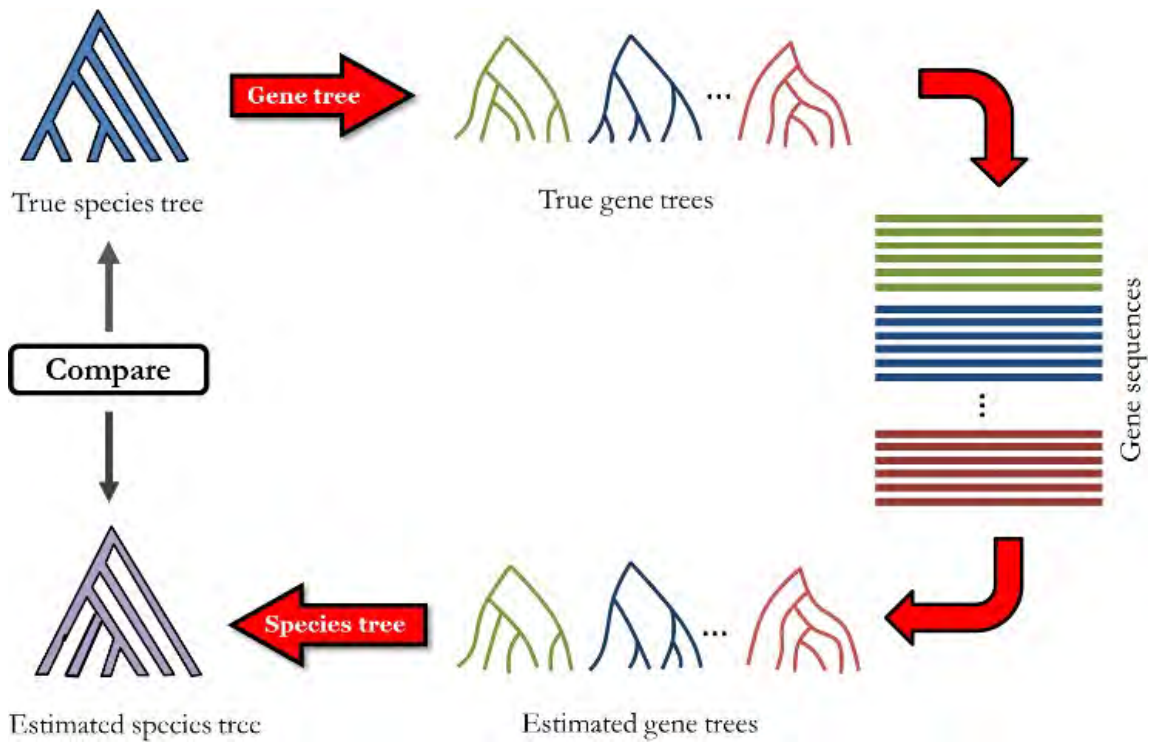


Figure 2.7: **A simulation protocol for evaluating species tree estimation techniques.** The protocol starts with a model species tree. At first, a collection of true gene trees are inferred from the model species tree, and gene sequences are also estimated from the collection of true gene trees. Next, gene trees are estimated from the gene sequence alignments. Finally, a species tree is estimated from the estimated gene trees, and then it is compared to the true species tree. This figure has been taken without any alternation from [33].



**False negative (FN) rate :** The false negative (FN) rate (also known as missing branch rate) is the proportion of the edges that are present in the true tree, but are absent from the estimated tree. Figure 2.8 shows an example of a true tree and an estimated tree where one true branch is not present in the estimated tree.

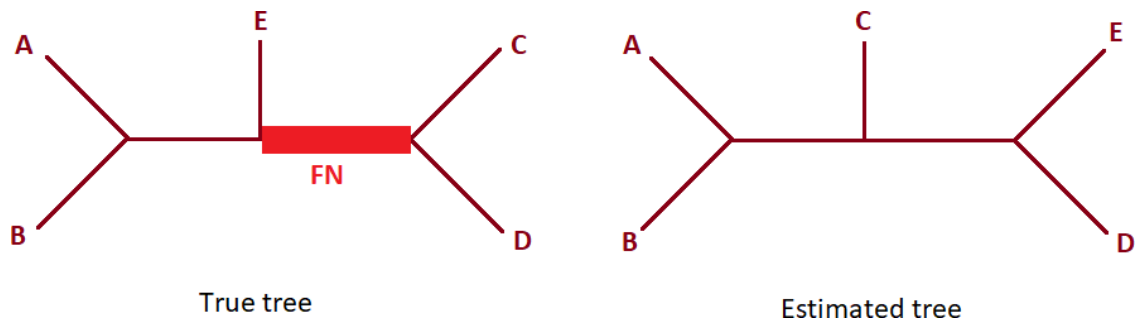


Figure 2.8: **False negative rate or, missing branch rate.** The branch separating  $u, v, w$  and  $x, y$  is not reconstructed in the estimated tree, although it is present in the true tree.

### Evaluation on real biological datasets

Since the ground truth is not known in real biological datasets, we cannot use any error metrics to evaluate the estimation techniques. Rather, we have to depend on the existing literature, biological beliefs and evidence regarding the evolutionary history of the species concerned. For instance, humans are considered to be more closely related to chimpanzees than they are to gorillas or orangutans according to existing literature. Therefore, a method is expected to reconstruct this relationship using genome-scale data from these primates e.g. humans, chimpanzees, gorillas and orangutans.

# Chapter 3

## Related work

In this chapter, we discuss various concepts and prior works that are related to this thesis. Because of its relevance to the rest of this thesis, we elaborate on species tree estimation techniques from a collection of gene trees in the presence of ILS. We elaborate on two optimization criteria for estimating species trees under ILS that we have considered in this thesis. We also briefly discuss another relevant concept called phylogenetic terraces, and the previous studies thereof.

### **Species tree estimation from multi-locus data under ILS**

Biological processes can result in different loci having different evolutionary histories [14], and therefore species tree estimation involves the estimation of trees and alignments on many different genes, so that the species tree can be based upon many different parts of the genome. While many processes can result in discordance between gene trees and species tree such as gene duplication and loss [57, 58], incomplete lineage sorting (ILS) is considered to be a dominant cause for gene tree heterogeneity, which is best understood under the coalescent model [35–42]. ILS or deep coalescence refers to the case in which two lineages fail to coalesce at their speciation point. Under the coalescent model, deep coalescence can be a source of discordance, because the common ancestry of gene copies at a single locus can extend deeper than speciation events.

Many scientific problems are more focused on the evolutionary history of organisms (i.e., species trees) than on the evolutionary history of a particular gene. Species tree estimations are complicated in the presence of gene tree discordance. Therefore, species tree estimation from a collection of gene trees in the presence of ILS is gaining substantial attention from the scientific community, and many summary methods have been developed over the last decade. Examples of statistically consistent coalescent based summary methods include MP-

EST [47], ASTRAL [6, 10], BUCKy [49], GLASS [59], STEM [60], SVDquartets [61], STEAC [62], NJst [63], ASTRID [64], STELAR [50], wQFM [65]. Other statistically consistent species-tree estimation methods include BEST [53] and \*BEAST [48], which co-estimate gene trees and species tree from input sequence alignments. These methods can produce substantially more accurate trees than other methods; however, these methods are extremely computationally intensive and do not scale to large numbers of genes [21, 66, 67]. Therefore, summary methods are comparatively more feasible for use on genome-scale datasets.

## Species tree estimation under MDC and MQC

We have considered two optimization criteria in this study for estimating species tree from gene trees in the presence of gene tree-species tree discordance due to ILS, namely **MDC** and **MQC**. The corresponding tools used for estimating species trees from a collection of gene trees under these criteria will be discussed in subsequent sections.

### Minimize Deep Coalescence (MDC)

This approach takes a set of rooted binary gene trees as input, where each of the gene trees are on the same set of taxa. It seeks the species tree with the minimum amount of deep coalescence. Although this method of species tree estimation is not statistically consistent [15], it is one of the most popular techniques for estimating species trees when gene trees can differ from the species tree due to ILS (incomplete lineage sorting).

It is elementary to know what are **Extra Lineages** in order to understand this approach.

### Extra Lineages

Let us denote the number of extra lineages on an edge  $e \in E(ST)$  by  $XL(gt, e)$ . This number is always one less than the total number of lineages. The total number of extra lineages is denoted by  $XL(gt, ST)$  within an optimal reconciliation of  $gt$  and  $ST$ . We refer to this total number of extra lineages as the **Extra Lineage Score**. Thus,  $XL(gt, ST) = \sum_{e \in E(ST)} XL(gt, e)$ .

### Problem Definition

The MDC problem can be defined as follows:

- **Input:** A set  $T = t_1, t_2, \dots, t_k$  of rooted, binary gene trees with each tree  $t_i$  on the same set  $S$  of taxa.

- **Output:** a rooted, binary species tree  $T$  that minimizes the number of extra lineages with respect to  $\mathbb{T}$ , denoted by  $XL(T, \mathbb{T}) = \sum_i XL(T, t_i)$

Thus, we need to define  $XL(T, t_i)$  which represents the number of extra lineages of a species tree  $T$  with respect to a gene tree  $t_i$  in order to define the MDC problem. This is visually defined by embedding the gene tree  $t_i$  into the species tree  $T$ , and then counting how many lineages are there on each edge of the species tree. As stated earlier, the number of extra lineages is one less than the total number of lineages on the edge for a given edge [14].

### Maximize Quartet Consistency (MQC)

Maximum Quartet Consistency (MQC) takes a quartet set as the input and finds a phylogenetic tree that satisfies the maximum number of quartets [68]. In this study, we refer to this number of consistent quartets as the **Quartet Score**.

We need to understand what are **Quartets** and **Quartet Consistency** prior to describing MQC.

#### Quartets and Quartet Consistency

Quartet is an unrooted tree with four taxa. If we denote a quartet by  $q = ab|cd$ , then it indicates that the internal edge in  $q$  separates  $a$  and  $b$  from  $c$  and  $d$ . In other words,  $ab|cd$  is the bipartition defined by the internal edge in  $q$ .

A quartet  $q$  is considered to be consistent with a tree when the tree has an internal edge that separates the same pairs of taxa as in  $q$ . It may not always be possible to find a tree which is consistent with all the  $\binom{n}{4}$  quartets [69]. In that case, the aim is to find out a tree such that maximum number of quartets are consistent with it.

#### Problem Definition

MQC is a well-studied optimization criterion for species tree estimation, which provides a statistically consistent approach for estimating species trees despite gene tree discordance [70]. The MQC problem can be formally defined as follows.

**Input:** A multiset of quartets  $Q$  on a taxa set  $P$ .

**Output:** A phylogenetic tree  $T$  on  $P$  such that  $T$  satisfies the maximum number of quartets of  $Q$ .

MQC is an NP-hard optimization problem [71]. MQC has both the exact and heuristic approaches available for species tree estimation. The exact version is computationally expensive as the run time of the exact version grows exponentially with the increase of number of taxa [72]. Therefore, we need to take the heuristic approach when the datasets are large enough.

## Tools Used

In this study, we have used Phylonet-MDC for species tree estimation by maximizing Deep Coalescence (MDC) and ASTRAL-III for estimating species tree by Minimizing Quartet Consistency (MQC). The following sections provide a brief explanation of these two tools used in this study.

### Phylonet-MDC

Phylonet [12, 18] is a tool which has been developed to analyze, reconstruct, and evaluate evolutionary relationships of phylogenetic networks. Although species tree estimation with MDC is not statistically consistent [15], Phylonet-MDC can infer species trees with high accuracy [73] especially when it is applied to gene trees in which all the low support branches are collapsed. Phylonet also has both exact and heuristic versions. The exact version can be applied even for unrooted gene trees [67, 74], which can be used on datasets with a small number of taxa. On the other hand, the heuristic version can be used for datasets with large number of taxa.

### ASTRAL-III

ASTRAL [75] (Accurate Species TRee ALgorithm) is a coalescent-based species tree estimation tool. This tool provides a statistically consistent estimation of the true species tree from unrooted gene trees under the multi-species coalescent model. Since MQC is an NP-hard problem [6, 71, 76], ASTRAL has two versions: the exact version is guaranteed to return the globally optimal solution, but runs in exponential time, and the heuristic version returns the optimal tree for a constrained search space and runs in polynomial time. However, the relative accuracy of the species trees estimated by ASTRAL depends on the amount of ILS. Typically, ASTRAL has an advantage when ILS levels are at least moderate.

ASTRAL takes a set of unrooted gene trees as input, and finds the species tree that agrees with the largest number of quartet trees induced by the set of gene trees as output. Since ASTRAL uses dynamic programming (DP) approach, it does not need to explicitly enumerate the set of all possible quartet trees.

## Phylogenetic terraces

Missing data is one of the major challenges in phylogenetic analyses. Terraces [29–32] are collections of equally optimal trees that may arise in tree space as a result of taxon coverage patterns (presence/absence pattern of the taxa) in alignments, resulting from missing data pattern which are commonly found in multi-locus data matrices. Terraces can both slow down tree search and mislead heuristic search algorithms [29, 30]. The presence of multiple equally good trees adds ambiguity to the tree inference as the tree search algorithms may need to pick one optimal tree among many. The underlying structure of terraces provides significant insights that can improve the accuracy and speed of searching for optimal trees, as well as the analysis and visualization of findings [77–79].

## Chapter 4

# Pseudo terraces in species tree estimation from gene trees

Species tree inference from multiple loci using summary methods begins by inferring gene trees from individual gene sequence alignments and then summarizes them to get a coherent species tree. Summarizing is typically done by optimizing various optimization criteria, such as minimizing deep coalescence [11,12,80], maximizing pseudo-likelihood [47], maximizing quartet consistency [6, 7], etc. Since the tree space grows exponentially as the number of taxa increases, navigating through the tree space to find an optimal tree is challenging. Sanderson *et al.* [29] showed that when phylogenetic trees are estimated from sequence alignments using maximum likelihood (ML), multiple distinct trees can have exactly the same likelihood score due to missing data (i.e., missing genes) – a condition which was referred to as terraces and was further investigated in subsequent studies [30–32]. A similar concept is “islands” of trees under ML and MP (maximum parsimony) criteria for estimating trees from sequence data [27,28]. Sanderson *et al.* [29,30] formalized the concept of phylogenetic terraces arising from missing data in multi-locus datasets, where they considered the supermatrix resulting from concatenating the gene alignments, and characterized terraces for ML score. The trees in a phylogenetic terrace are “close” to one another and have particular topological relationships (e.g., all trees on a terrace of rooted trees can be reached by a series of nearest neighbor interchanges (NNIs) between trees of the same optimality score) [29, 30]. The trees in a phylogenetic terrace are indistinguishable in two important ways: they “display” the same set of subtrees, and they have the same optimality score [29]. Similar phenomenon can arise in phylogenomic analyses where species trees are estimated from a set of gene trees using summary methods under various optimization criteria.

We have extended the concept of terraces to phylogenomic analyses in the context of species tree inference using summary methods. We show that if two species trees display the same set

---

Much of the material in this chapter is taken without alteration from Farah *et al.* (2021) [81]

of *maximal clusters* with respect to a set of gene trees, then they have the same extra lineage score. We also show an analogous property for quartet scores. Note that, in phylogenetic terraces, it can happen that two trees have the same optimality score but do not belong to the same terrace because they do not induce/display an identical set of locus-specific subtrees due to a lack of topological closeness [29,30,82]. However, in our context, questions regarding the topological proximity of the equally good species trees remain open. Since their topological “closeness” is yet to be explored, we refer to these sets of equally optimal species trees as the “pseudo species tree terrace”, where potentially large numbers of distinct species trees may have exactly the same optimality score with respect to a set of input gene trees. Such equally optimal trees – while inferring a species tree by summarizing gene trees – can arise even without the presence of missing data (i.e., missing leaves in the gene trees). We refer to this set of equally optimal species trees as “pseudo species tree terrace”. For an optimization criterion  $C$  and a set  $G$  of gene trees, we define *pseudo C terrace* to be a set of trees with distinct tree topologies, but with exactly the same  $C$  score with respect to  $G$ .

**Definition 1.** For a set  $G$  of gene trees and an optimization criterion  $C$ , the *pseudo C terrace*  $T_{G,C}(s)$  represents a set of species trees having exactly the same  $C$  score  $s$  with respect to the input set  $G$  of gene trees.

In this thesis, we particularly investigate pseudo extra lineage (EL) terrace and quartet terrace. We prove theoretical results showing the possibility of the existence of pseudo EL and quartet terraces, especially when the number of taxa is high. We have provided combinatorial characteristics and conditions for datasets to have equally optimal trees with respect to extra lineage (EL) and quartet consistency scores.

## Pseudo EL terrace

**Theorem 2.** For a set  $G$  of  $k$  gene trees on  $n$  taxa, the species tree space will have at least one pseudo EL terrace if  $\prod_{i=2}^n (2i - 3) > k(n - 2)(n - 1)/2 + 1$ .

*Proof.* For a gene tree  $gt$  and a species tree  $ST$  on  $n$  taxa, the number of extra lineages increases as the number of deep coalescence increases, that means the lineages do not coalesce at their common ancestors and go deeper in time. Therefore, the number of extra lineages is maximized when the gene lineages do not coalesce until they reach the root of the phylogeny, resulting in extra lineages on all the internal branches. Assuming that gene lineages do not coalesce with each other until they reach the root node, the number of extra lineages will be maximized for a pectinate tree (also known as caterpillar tree) since all the gene lineages (except for one) will go through all the internal branches “above” its most immediate ancestor.



See Figure 4.1 for an example, which shows how a pectinate species tree results in more extra lineages than relatively more balanced trees. Therefore, the internal branch incident on the root of the phylogenetic tree will contain  $n - 1$  lineages and hence  $n - 2$  extra lineages. The more recent internal branches (from ancient to recent) will contain  $n - 3, n - 4, \dots, 1$  extra lineages, respectively. Thus, the maximum number of extra lineages that may occur for a gene tree  $gt$  and a species tree  $ST$  is

$$(n - 2) + (n - 3) + \dots + 1 = (n - 2)(n - 1)/2.$$

For a set  $G$  of  $k$  gene trees, the maximum number of extra lineages with respect to a species tree  $ST$  will be  $k(n - 2)(n - 1)/2$ . The number of rooted species trees in the tree space with  $n$  taxa ( $n \geq 2$ ) is

$$1.3.5. \dots (2n - 3) = \prod_{i=2}^{n-1} (2i - 3)$$

[83, 84]. The EL scores of these trees in the species tree space will be within the range  $0 \sim k(n - 2)(n - 1)/2$ . That means there are  $k(n - 2)(n - 1)/2 + 1$  possible distinct EL scores

for  $\prod_{i=2}^{n-1} (2i - 3)$  trees in the tree space with respect to  $G$ . Therefore, using the pigeonhole principle, there will be at least one pseudo terrace  $T_{G,EL(s)}$  with more than one tree having the same EL score  $s$  provided that  $\prod_{i=2}^{n-1} (2i - 3) > k(n - 2)(n - 1)/2 + 1$ .

□

Table 4.1 shows an example demonstrating the pseudo EL terraces in the tree space for four taxa with respect to a set  $G$  of four rooted gene trees. There are 15 possible rooted species tree topologies with four taxa and we examined the EL scores of all of them with respect to  $G$ . We identified six pseudo terraces, containing more than one tree with EL scores of 4, 6, 7, 8, 9, and 11:  $T_{G,EL(4)}$ ,  $T_{G,EL(6)}$ ,  $T_{G,EL(7)}$ ,  $T_{G,EL(8)}$ ,  $T_{G,EL(9)}$ , and  $T_{G,EL(11)}$ . Here, two candidate species trees ( $((B, D), C), A$ ) and  $((C, D), B), A$ ) belong to  $T_{G,EL(4)}$  and are optimal (most parsimonious) under the MDC criteria.

### Characterization of the trees in a pseudo EL terrace

Let  $G = \{gt_1, gt_2, \dots, gt_k\}$  be a set of  $k$  gene trees, and  $ST$  be a species tree (both on a set  $X$  of  $n$  taxa). Let  $L(T)$  and  $E(T)$  be the set of leaves and the set of edges in a tree  $T$ , respectively. A clade in  $T$  is a subtree of  $T$  rooted at a node in  $T$ , and the set of leaves of the clade is called a cluster. Let  $CL(G)$  be the set of clusters in an input set  $G$  of gene trees, and  $CL(ST)$  be the set of clusters in a species tree  $ST$ . Given a binary, rooted gene tree  $gt$  and a species tree  $ST$  on a set  $X$  of  $n$  taxa, the deep coalescence cost (i.e., extra lineage cost) can be computed by using the most recent common ancestor (MRCA) mapping of the nodes

Table 4.1: **Pseudo EL terraces in the tree space of four taxa with respect to a set  $\mathcal{G}$  of rooted and binary gene trees.**  $\mathcal{G}$  contains four gene trees, where two of the genes have the gene tree topology of  $gt_1 = (((B, D), C), A)$ , one gene tree has the topology of  $gt_2 = ((A, B), (C, D))$ , and one gene has the topology of  $gt_3 = (((C, D), A), B)$ . Two species trees  $(((B, D), C), A)$  and  $(((C, D), B), A)$  are both optimal under MDC with 4 extra lineages.

Species tree	MDC-score			Total
	$gt_1$ $(((B, D), C), A)$	$gt_2$ $((A, B), (C, D))$	$gt_3$ $((((C, D), A), B)$	
$(((A, B), C), D)$	3	1	3	10
$(((A, B), D), C)$	2	1	3	8
$(((A, C), B), D)$	3	2	3	11
$(((A, C), D), B)$	3	2	1	9
$(((A, D), B), C)$	2	2	3	9
$(((A, D), C), B)$	3	2	1	9
$(((B, C), A), D)$	3	2	3	11
$(((B, C), D), A)$	1	2	2	6
$(((B, D), A), C)$	1	2	3	7
$(((\mathbf{B}, \mathbf{D}), \mathbf{C}), \mathbf{A})$	0	2	2	<b>4</b>
$(((C, D), A), B)$	3	1	0	7
$(((\mathbf{C}, \mathbf{D}), \mathbf{B}), \mathbf{A})$	1	1	1	<b>4</b>
$((A, B), (C, D))$	2	0	1	5
$((A, C), (B, D))$	1	2	2	6
$((B, C), (A, D))$	2	2	2	8

in  $gt$  to the nodes in  $ST$  [16, 55]. Than and Nakhleh [55] and Yu *et al.* [80] showed that it is possible to compute the number of extra lineages in an edge  $e$  in  $ST$  without using an MRCA mapping. They introduced a concept of “maximal cluster” defined as follows.

**Definition 3.** (From [80]) For  $B \subseteq X$  and gene tree  $gt$ , we set  $k_B(gt)$  to be the number of  $B$ -maximal clusters in  $gt$ , where a  $B$ -maximal cluster is a cluster  $Y \subseteq L(gt)$  in  $\text{CL}(G)$  such that  $Y \subseteq B$  but no other cluster of  $gt$  containing  $Y$  is a subset of  $B$ .

Yu *et al.* [80] showed that for any edge  $e$  in  $ST$ , where  $B$  is the cluster below  $e$ ,  $k_B(gt)$  is the number of lineages going through the edge  $e$ , and so  $k_B(gt) - 1$  is the number of extra lineages going through  $e$ . Therefore, for a gene tree  $gt$  and for any edge  $e \in E(ST)$ , the number of extra lineages in  $e$  can be defined as follows.

$$XL_{ST}(gt, e) = k_B(gt) - 1 \quad (4.1)$$

Figure 4.2 illustrates an embedding of a gene tree  $gt$  within a species tree  $ST$  which results in one extra lineage in the branch  $(u, v)$ . This can also be obtained by analyzing the maximal clusters as defined in Def. 3 and Eqn. 4.1. Note that  $(u, v) \in E(ST)$  induces the cluster

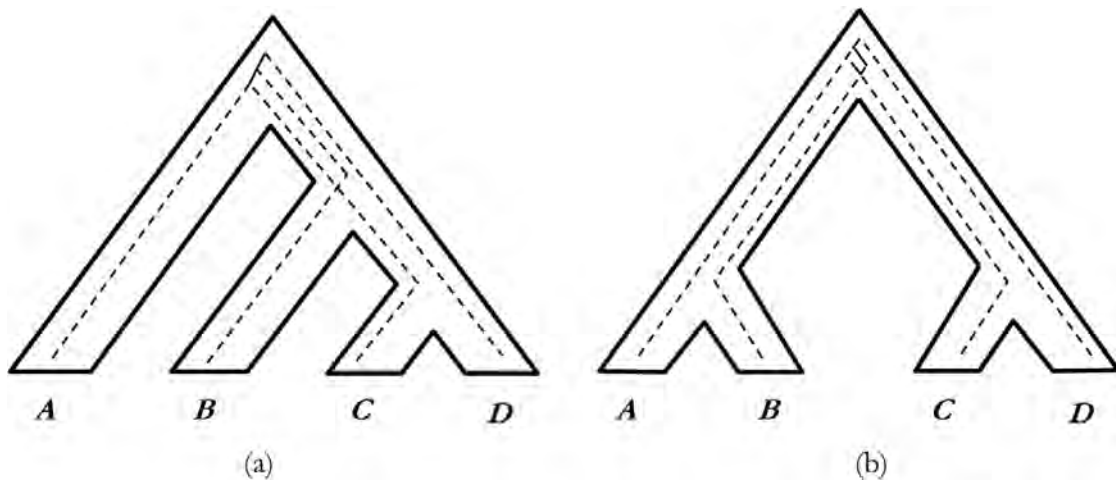


Figure 4.1: **Extra lineages inside a pectinate and a balanced symmetric tree for the cases when gene lineages do not coalesce with each other on the internal branches until they reach the root of the tree.** Species boundaries are shown in solid lines while gene lineages are represented by dashed lines. (a) Reconciliation of  $gt = ((A, C), B), D$  with a pectinate species tree  $ST = (((C, D), B), A)$ , which results in three extra lineages. (b) Reconciliation of  $gt = (((A, C), B), D$  with a balanced species tree  $ST = ((A, B), (C, D))$ , which results in two extra lineages. In both cases, the gene lineages do not coalesce with each other until they go backwards to the root node.

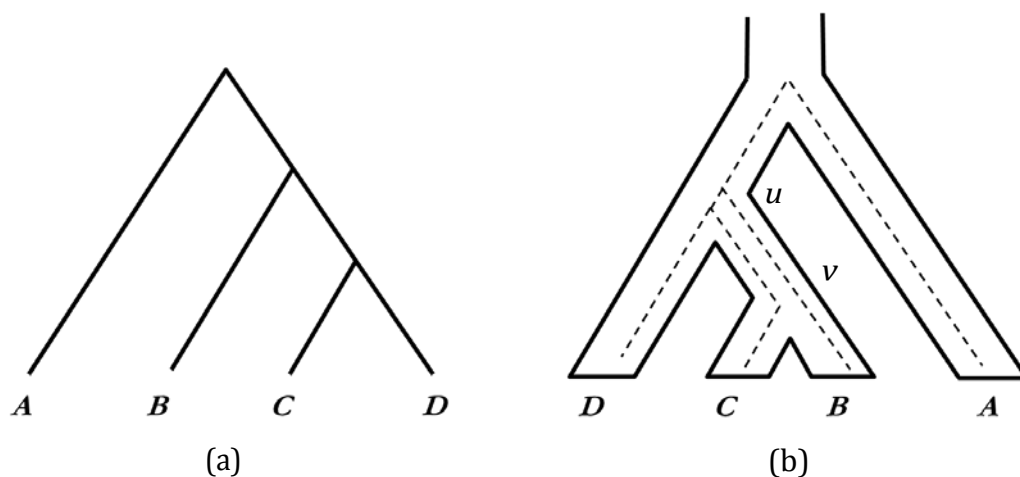


Figure 4.2: **Computing the extra lineage score.** (a) A gene tree  $gt = (A, (B, (C, D)))$  and (b) the embedding/reconciliation of  $gt$  within a species tree  $ST = (A, (D, (B, C)))$ . This optimal reconciliation results in one extra lineage on the  $(u, v)$  edge in  $ST$ .

$cl \in \text{CL}(ST) = \{B, C\}$ . There are two clusters  $\{B\}$  and  $\{C\}$  in  $\text{CL}(gt)$ , that are subsets of  $cl$  and are  $cl$ -maximal. Therefore, from Eqn. 4.1, the number of extra lineages in  $(u, v)$  is  $2 - 1 = 1$ . The deep coalescence cost  $DC(\mathbf{G}, ST)$  of a candidate species tree  $ST$  with respect to a set  $\mathbf{G}$  of gene trees can be computed as follows.

$$\begin{aligned} DC(\mathbf{G}, ST) &= \sum_{gt \in \mathbf{G}} DC(gt, ST) \\ &= \sum_{gt \in \mathbf{G}} \sum_{e \in E(ST)} XL_{ST}(gt, e). \end{aligned} \quad (4.2)$$

Note that each edge  $e \in E(ST)$  corresponds to a cluster in  $\text{CL}(ST)$ , and  $XL_{ST}(gt, e)$  can be computed independently for each edge in  $ST$ . Let  $\text{MC}(gt, ST)$  be the set of clusters in  $\text{CL}(gt)$  that are maximal with respect to the clusters in  $\text{CL}(ST)$ . Then, from Eqn. 4.1, it is easy to see that the deep coalescence cost  $DC(gt, ST)$  can be defined as follows.

$$\begin{aligned} DC(gt, ST) &= \sum_{e \in E(ST)} XL_{ST}(gt, e) \\ &= |\text{MC}(gt, ST)| - |E(ST)| \\ &= |\text{MC}(gt, ST)| - (2n - 2) \end{aligned} \quad (4.3)$$

Let  $\text{MC}(\mathbf{G}, ST)$  be the multiset of maximal clusters in  $\mathbf{G}$  with respect to the clusters in  $\text{CL}(ST)$ . Thus, Eqn. 4.2 can be written as follows.

$$\begin{aligned} DC(\mathbf{G}, ST) &= \sum_{gt \in \mathbf{G}} DC(gt, ST) \\ &= \sum_{gt \in \mathbf{G}} (|\text{MC}(gt, ST)| - (2n - 2)) \\ &= |\text{MC}(\mathbf{G}, ST)| - k(2n - 2) \end{aligned} \quad (4.4)$$

Therefore, it is easy to see that two candidate species trees  $ST_1$  and  $ST_2$  on the same set of taxa  $X$  will incur identical deep coalescence cost and subsequently will be members of a particular pseudo EL terrace if and only if  $\text{CL}(ST_1)$  and  $\text{CL}(ST_2)$  induce equal number of maximal clusters in  $\text{CL}(\mathbf{G})$ , i.e.,  $|\text{MC}(\mathbf{G}, ST_1)| = |\text{MC}(\mathbf{G}, ST_2)|$ . Thus, the following theorem follows.

**Theorem 4.** *Given a pseudo EL terrace  $T_{\mathbf{G}, \text{EL}}(s)$ , any two trees  $ST_i$  and  $ST_j$  will belong to  $T_{\mathbf{G}, \text{EL}}(s)$  if and only if  $|\text{MC}(\mathbf{G}, ST_i)| = |\text{MC}(\mathbf{G}, ST_j)|$ . Also,  $s = |\text{MC}(\mathbf{G}, ST_i)| - k(2n -$*

$$2) = |\text{MC}(\mathbf{G}, ST_j)| - k(2n - 2).$$

## Pseudo quartet terrace

**Theorem 5.** For a set  $\mathbf{G}$  of  $k$  gene trees on  $n$  taxa, the species tree space will have at least one pseudo quartet terrace if  $\prod_{i=2}^n (2i - 3) > k \binom{n}{4} + 1$ .

*Proof.* There are  $\binom{n}{4}$  quartets in a tree with  $n$  taxa. Therefore, for a gene tree  $gt \in \mathbf{G}$  and a species tree  $ST$ , both on the same set of  $n$  taxa,  $ST$  can satisfy at most  $\binom{n}{4}$  quartets (this is when  $gt$  and  $ST$  have an identical topology). Hence, for a set  $\mathbf{G}$  of  $k$  gene trees, the maximum number of consistent quartets with respect to a species tree  $ST$  is  $k \binom{n}{4}$ . Following the same argument as described in Thm. 5, there will be at least one pseudo quartet terrace  $T_{\mathbf{G}, \text{QS}(s)}$  with an identical quartet score (QS)  $s$  if  $\prod_{i=2}^n (2i - 3) > k \binom{n}{4} + 1$ . □

## Characterization of the trees in a pseudo quartet terrace

Fundamental to characterizing or identifying the trees in a pseudo quartet terrace is the ability to *efficiently* compute the quartet score of a candidate species tree with respect to a given set of input gene trees. Siavash *et al.* [6] leveraged a dynamic programming based solution in ASTRAL, which can efficiently find the quartet score without needing to explicitly enumerate the set of quartets in the gene trees and species trees. They showed that the quartet score of a species tree can be computed using the “tripartitions” in the input gene trees and the species tree.

Let  $T$  be an unrooted binary tree on a set  $X$  of  $n$  taxa, and  $q = ab|cd$  be a quartet where  $\{a, b, c, d\} \in X$ . The subtree of  $T$  induced by  $\{a, b, c, d\}$  will have exactly two nodes ( $x$  and  $y$ ) with degree three. We say that  $q = ab|cd$  is mapped (or associated) to  $x$  and  $y$ . An internal node  $u$  in an unrooted tree  $T$  defines a *tripartition*  $X_1|X_2|X_3$ , where  $X_1, X_2$ , and  $X_3$  are three partitions of the leaves in  $T$ . It is easy to see that the number of quartets mapped to  $u$ , where  $|X_1| = n_1, |X_2| = n_2$ , and  $|X_3| = n_3$  is:

$$\begin{aligned} \text{NQ}(n, n_1, n_2, n_3) &= \binom{n_1}{2} \binom{n_2}{1} \binom{n_3}{1} + \binom{n_2}{2} \binom{n_1}{1} \binom{n_3}{1} + \binom{n_3}{2} \binom{n_1}{1} \binom{n_2}{1} \\ &= \frac{n_1 n_2 n_3 (n_1 + n_2 + n_3 - 3)}{2} \end{aligned} \quad (4.5)$$

**Lemma 6.** (From [6, 85]) Given two tripartitions  $x = X_1|X_2|X_3$  and  $y = Y_1|Y_2|Y_3$ , the number of quartets common between these two tripartitions is given by the following equation.

Here,  $n_{ij} = |X_i \cap Y_j|$ .

$$\begin{aligned} \text{MQ}(x, y) = & \text{NQ}(n_{11}, n_{22}, n_{33}) + \text{NQ}(n_{11}, n_{23}, n_{32}) + \\ & \text{NQ}(n_{12}, n_{21}, n_{33}) + \text{NQ}(n_{12}, n_{23}, n_{31}) + \\ & \text{NQ}(n_{13}, n_{21}, n_{32}) + \text{NQ}(n_{13}, n_{22}, n_{31}) \end{aligned} \quad (4.6)$$

Let the set of tripartitions in a tree  $T$  be  $\text{tpt}(T)$ . Then, we can compute the number of quartets that a tripartition  $x$  in  $\text{tpt}(ST)$  shares with  $G$  according to Eqn. 4.7. Finally, we can compute the quartet score of  $ST$  using Eqn. 4.8. The division by two in this equation is due to the fact that a quartet is mapped to two internal nodes in a tree.

$$QS_G(x) = \sum_{gt \in G} \sum_{y \in \text{tpt}(gt)} \text{MQ}(x, y). \quad (4.7)$$

$$QS_G(ST) = \frac{1}{2} \sum_{x \in \text{tpt}(ST)} QS_G(x). \quad (4.8)$$

Therefore, two candidate species trees will have identical quartet scores if and only if the total numbers of quartets – in the gene trees – mapped to their induced set of tripartitions are equal (Thm. 7). Note that Eqn. 4.7 allows us to score a tripartition in  $ST$  independently from the other tripartitions in  $ST$ . Hence, a preprocessing step – which involves enumerating all possible tripartitions on  $X$  and computing  $QS_G(x)$  for each  $x \in \cup_{gt \in G} \text{tpt}(gt)$  – will facilitate exploring and enumerating pseudo quartet terraces for a given input set  $G$  of gene trees.

**Theorem 7.** *Given a pseudo quartet terrace  $T_{G, QS}(s)$ , any two trees  $ST_i$  and  $ST_j$  will belong to  $T_{G, QS}(s)$  if and only if  $QS_G(ST_i) = QS_G(ST_j) = s$ .*

### Quartet scores of the trees in a neighborhood of a species tree

We now briefly discuss how the quartet scores of the neighboring trees of a particular tree can be computed efficiently. We consider the Nearest Neighbor Interchange (NNI) operation for exploring the neighborhood of a tree. Let  $ST$  be a species tree, and  $e = (u_1, u_2)$  be an internal edge in  $T$ . Let  $A$ ,  $B$ ,  $C$ , and  $D$  be the sets of taxa in the four subtrees on the four branches incident on the two endpoints  $u_1$  and  $u_2$  of  $e$  (see Fig. 4.3a). An NNI move on edge  $e$  will result in two neighboring trees  $ST_1$  and  $ST_2$  (see Figs. 4.3b and 4.3c). Note that the sets of tripartitions in the four subtrees around edge  $e$  remain the same after an NNI move. Only the tripartitions on the endpoints ( $u_1$  and  $u_2$ ) of  $e$  are changed. The tripartitions on  $u_1$  and  $u_2$  in

$ST$  are  $A|B|(C \cup D)$  and  $C|D|(A \cup B)$ , respectively, whereas, the tripartitions on  $u_1$  and  $u_2$  in  $ST_1$  are  $A|C|(B \cup D)$  and  $B|D|(A \cup C)$ , respectively. Therefore, it is easy to see that the set of tripartitions in  $T_1$  and  $T_2$  can be computed as follows.

$$tpt(T_1) = tpt(T) - \{A|B|(C \cup D), C|D|(A \cup B)\} \cup \{A|C|(B \cup D), B|D|(A \cup C)\}$$

$$tpt(T_2) = tpt(T) - \{A|B|(C \cup D), C|D|(A \cup B)\} \cup \{A|D|(B \cup C), B|C|(A \cup D)\}$$

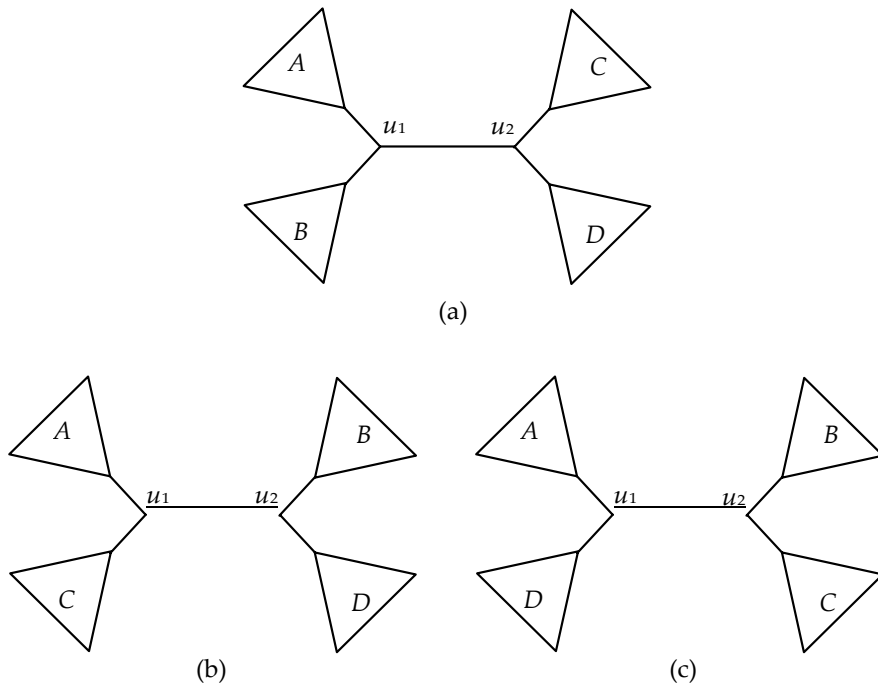


Figure 4.3: **NNI move on an internal edge.** (a) A species tree  $ST$ , and (b)-(c) the neighbors of  $ST$  resulting from one NNI move on edge  $e = (u_1, u_2)$ .  $A$ ,  $B$ ,  $C$ , and  $D$  are the sets of taxa in the four subtrees around edge  $e$ .

Subsequently,  $QSG(T_1)$  can be obtained from  $QSG(T)$  using Eqn. 4.9. Note that once we have calculated the quartet score of  $ST$ , we can obtain the quartet score of  $ST_1$  by additionally calculating the scores of only two new tripartitions. These results can be leveraged to efficiently explore the neighborhood of a tree and identify pseudo QS terraces.

$$\begin{aligned}
QS_G(ST_1) &= \frac{1}{2} \sum_{x \in \text{tpt}(ST_1)} QS_G(x). \\
&= \frac{1}{2} \sum_{x \in \text{tpt}(ST)} QS_G(x) \\
&\quad - \frac{1}{2} (QS_G(\{A|B|(C \cup D)\}) + QS_G(\{C|D|(A \cup B)\})) \\
&\quad + \frac{1}{2} (QS_G(\{A|C|(B \cup D)\}) + QS_G(\{B|D|(A \cup C)\}))
\end{aligned} \tag{4.9}$$

**Corollary 8.** *Given a set  $G$  of gene trees, a species tree  $ST$  and its one NNI move away neighbor  $ST_1$  as shown in Figure 4.3,  $ST$  and  $ST_1$  will belong to the same pseudo quartet terrace if and only if*

$$QS_G(\{A|B|(C \cup D)\}) + QS_G(\{C|D|(A \cup B)\}) = QS_G(\{A|C|(B \cup D)\}) + QS_G(\{B|D|(A \cup C)\}).$$

### Additional Remarks

Note that the numbers of distinct EL score and quartet score grow at a polynomial rate, whereas the number of unique species trees grows exponentially as we increase the number of taxa. This is also valid for most other optimization criteria (e.g., minimize gene duplication and loss (MGDL) score [52, 57, 58], triplet consistency [50]), that are commonly used in phylogenomic analyses. Thus, search for an optimal species tree under a particular optimization criterion may result in a tree which is a member of a pseudo terrace with a potentially large number of species trees with different tree topologies. Therefore, we can have multiple trees with the optimal (or near-optimal) score, but with different “closeness” to the true tree. Our experimental results also support this as we will show that, in some cases, the trees estimated by Phylonet achieve competitive or identical quartet scores, but are not as accurate as the trees estimated by ASTRAL.



# Chapter 5

## Experimental Studies

In this chapter, we report, on an extensive experimental studies, the performance of MDC and MQC for species tree estimation in the presence of ILS, as well as the presence and implications of pseudo species tree terraces.

### Datasets

We used previously studied simulated and biological datasets to evaluate the performance of MQC and MDC. We studied two collections of simulated datasets: one based on a biological dataset (37-taxon mammalian dataset) that was generated in a prior study [86], and another simulated dataset from [86]. Table 5.1 presents a summary of these datasets, generated under various model conditions with varying ILS levels (reflected in the average topological distance between true gene trees and the true species tree) and gene sequence lengths. All the estimated and true gene trees and the true species trees used in this study are taken from previous studies.

In the mammalian simulation, we explored the impact of varying numbers of genes and the impact of phylogenetic signal by varying the sequence length (250bp, 500bp, 1000bp and 1500bp) of the genes. In both cases, three levels of ILS are simulated by multiplying or dividing all internal branch lengths in the model species tree by two, and we also explore various numbers of genes. These datasets have been generated under a multi-stage simulation process: a species tree was estimated on a mammalian dataset [87] using MP-EST, gene trees were simulated down the species tree under the multi-species coalescent model and then gene sequence alignments were simulated down the gene trees under the GTRGAMMA model.

We used another simulated dataset: 11-taxon dataset (generated by [88] and subsequently studied by [21, 67]). The 11-taxon dataset vary in the level of ILS (low and high amount of ILS; see Table 5.1). Thus, the simulated datasets provide a range of conditions in which we

---

Much of the material in this chapter is taken without alteration from Farah *et al.* (2021) [81]

Table 5.1: **Properties of the simulated datasets.** Level of ILS is presented in terms of the average topological distance between true gene trees and true species tree.

Data set	ILS level	No. of genes	No. of sites	No. of replicates	Ref.
11-taxon	38%, 85%	5 - 100	2000	20	[88]
37-taxon	18%, 32%, 54%	25 - 800	250 -1000	20	[86]

explored the performance of MQC and MDC and investigated the impacts of pseudo species tree terrace.

We used two biological datasets: the 37-taxon mammalian datasets studied by Song *et al.* [87] with 424 genes, and the the amniota dataset from [89] containing 16 species and 248 genes.

## Materials and Methods

We compared ASTRAL-III [75] (which solves the MQC problem) with Phylonet [12, 18], which is based on the MDC problem. We ran the exact versions of ASTRAL and Phylonet on dataset with 11 taxa, and the heuristic versions for larger datasets (37-taxon).

The following commands were used in this study to run various methods.

### Estimation of species tree by minimizing deep coalescence using Phylonet

The following command was executed in Phylonet to infer species trees from set of gene trees:

```
infer_st -m MDC -i <input gene tree files> -x -o <output species tree>
```

The -x option was used to specify that the exact version will be run. For the 37-taxon dataset, the default heuristic version was run (without the -x option).

### Estimation of species tree by maximizing quartet consistency using ASTRAL-III

The following command was executed in Astral to infer species trees from set of gene trees by maximizing the number of consistent quartets:

```
-x -i <gene trees> -o <output species tree>
```

The -x option was used to specify that the exact version will be run. For the 37-taxon dataset, the default heuristic version was run (without the -x option).

### Estimation of Extra Lineage Score

The following command was executed in Phylonet to count the total number of extra lineages required to reconcile a set of gene trees in a species tree.

```
deep_coal_count <species-tree-file> <gene-trees-file>
```

### Estimation of Quartet Score

The following command was executed in ASTRAL to count the total number of quartets induced by the gene trees that are consistent with a species tree.

```
-q <species tree to be scored> -i <gene trees> -o <output-file>
```

## Measurements

We compared the quartet support scores, EL scores (number of extra lineages required to reconcile the input set of gene trees with a species tree [11, 12]), and topological accuracy of the trees estimated by ASTRAL and Phylonet. We used false negative (FN) rate to measure the topological error. All the trees estimated by ASTRAL and Phylonet in this study are binary and so False Positive (FP) rates and False Negative (FN) rates are identical. For the biological dataset, we compared the estimated species trees with the existing literature and biological beliefs. Since EL score is defined for a pair of rooted gene tree and species tree, and ASTRAL does not provide rooted trees, we made the ASTRAL-estimated species trees rooted using an appropriate outgroup before computing the EL scores. We performed Wilcoxon signed-rank test (with  $\alpha = 0.05$ ) to measure the statistical significance of the differences between two methods.

### Results on datasets simulated from a biological example(37-taxon mammalian dataset)

**Missing branch rate:** Substantial differences were observed between ASTRAL and Phylonet on all the model conditions that we analyzed. Figure 5.1a shows the average FN rates on three model conditions with varying amounts of ILS. Both ASTRAL and Phylonet incurred the highest missing branch rates (around 5% and 20% respectively) for the 0.5X model condition, which has the highest amount of gene tree discordance. ASTRAL obtained the lowest FN rate (2.5%) for the 2X model condition, which is expected as it has the lowest level of ILS. However, Phylonet was slightly better on the 1X model condition than on the 2X model

condition. Figure 5.1b shows the error rates on various model conditions with varying lengths of gene sequence alignments and hence varying amounts of gene tree estimation error. We also analyzed the true gene trees. Both ASTRAL and Phylonet incurred the highest FN rate on the model condition with the shortest sequences (i.e., highest amount of gene tree estimation error), and the difference between ASTRAL and Phylonet was substantial (6.32% vs. 21.47%). As we increase the sequence lengths and hence decrease the gene tree estimation errors, both methods produce more accurate trees, but the differences between ASTRAL and Phylonet are still substantial. Additionally, Phylonet incurred slightly higher FN rate on 1000bp (10.29%) than on 500bp (9.26%). Even on the true gene trees, Phylonet was much worse than ASTRAL. Figure 5.1c shows the FN rates for varying numbers of genes. ASTRAL, being a statistically consistent method, showed improved accuracy as we increased the number of genes. However, a similar trend was not observed for Phylonet as increasing genes from 100 to 800 did not improve the tree accuracy. On the model conditions with 50, 100, 400, and 800 genes, Phylonet achieved similar FN rates (12.35% ~12.65%). Our overall observation is that ASTRAL is much better in terms of FN rate than Phylonet and the differences are statistically significant ( $P \ll 0.05$ ).

**Quartet score:** Table 5.2 shows the comparison between the quartet scores of the species trees estimated by ASTRAL and Phylonet. We also show the quartet score of the true species tree (denoted by “true quartet score”). As expected, ASTRAL achieved higher quartet scores than Phylonet under all model conditions as ASTRAL estimates species trees by maximizing the quartet score. The quartet scores of the ASTRAL-estimated trees are usually closer to the quartet score of the true species tree than those of the Phylonet-estimated trees. However, we observed that ASTRAL generally overestimates the quartet scores (compared to the true quartet score). Phylonet, in general, underestimates the quartet score since it does not take quartet consistency into account. Interestingly, Phylonet achieved closer quartet scores (with respect to the true quartet scores) than ASTRAL on 25- and 50-gene model conditions. ASTRAL’s average quartet scores are higher than true quartet scores by 4308 and 4415 on 25 and 50-gene model conditions, whereas the average quartet scores of Phylonet are less than the true quartet scores by only 122 and 516 on these two model conditions. However, Phylonet is significantly worse than ASTRAL on these two model conditions (see Fig. 5.1c) – suggesting that overestimating the quartet score may not hurt the tree accuracy as much as underestimating it. These results suggest that MDC criteria may sometimes achieve reasonably high QS, but the search under MDC may lead to incorrect trees by underestimating the amount of extra lineages.

**EL score:** We show the comparison between ASTRAL and Phylonet in terms of the number of extra lineages in Figure 5.2. We also show the EL scores of the true species trees (which we refer to as the “true EL score”). As expected, Phylonet obtained the lowest EL scores under

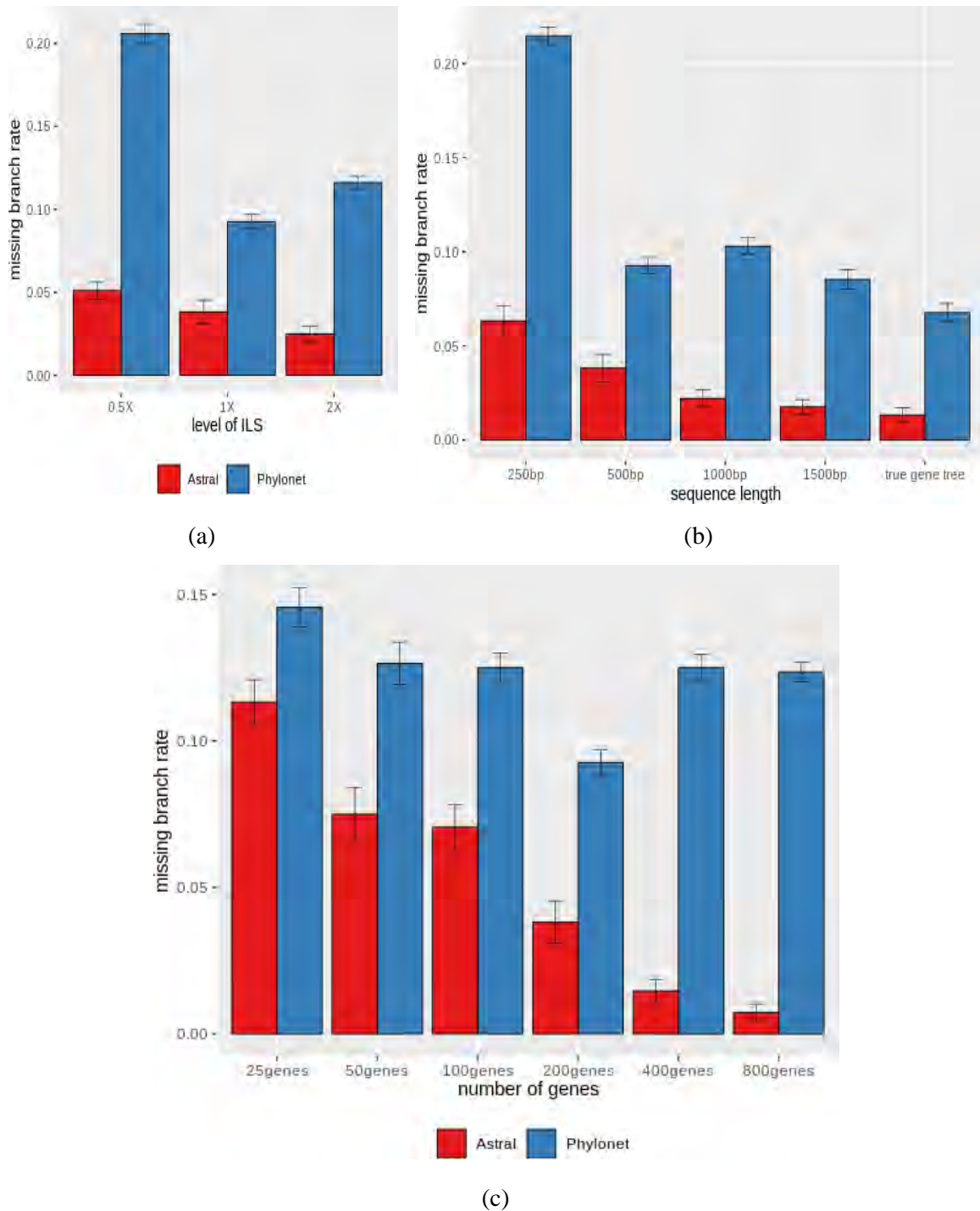


Figure 5.1: **Comparison of ASTRAL and Phylonet on 37-taxon simulated mammalian dataset.** We show the average FN rates with standard error bars over 20 replicates. (a) We fixed the sequence length to 500 bp and number of genes to 200, and varied the amounts of ILS. The 2X model condition contains the lowest amount of ILS while the 0.5X refers to the model condition with the highest amount of ILS. (b) We varied the amount of gene tree estimation error by varying the sequence lengths from 250 to 1000 bp, while keeping the ILS level (moderate) and the number of genes (200) fixed. (c) We fixed the sequence length to 500 bp and amount of ILS to 1X, and varied the numbers of genes from 25 ~ 800.

Table 5.2: **Quartet scores for ASTRAL, Phylonet and true species tree on 37-taxon simulated mammalian dataset.** We show average quartet scores (number of quartets in the gene trees that are consistent with the species tree) over 20 replicates for various model conditions by controlling the levels of ILS, gene tree estimation error, and numbers of genes.

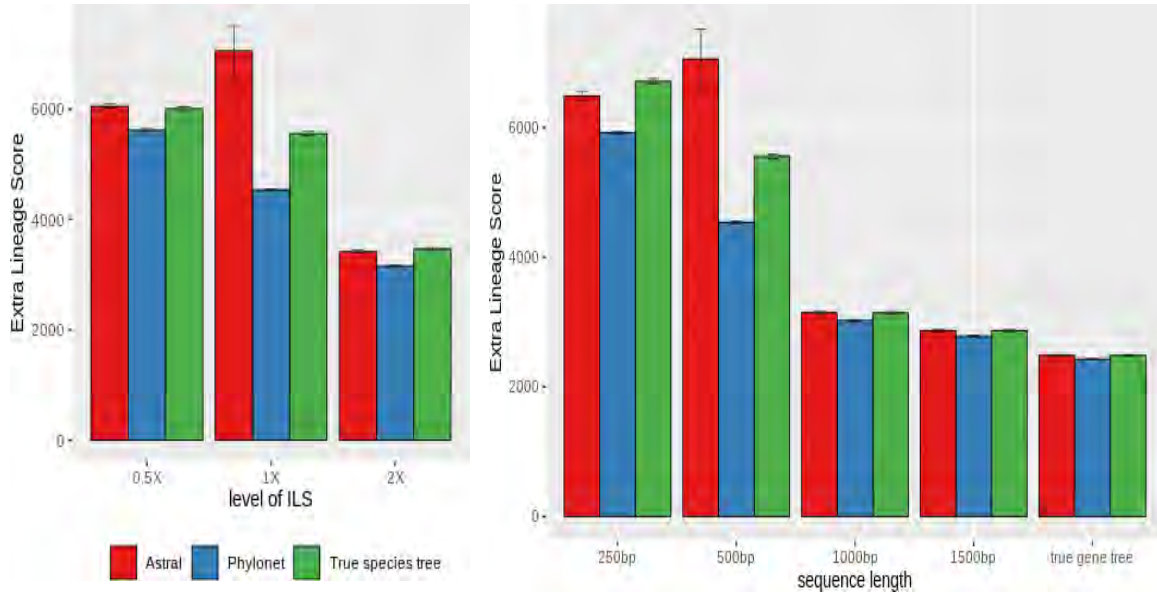
Model condition	Quartet score		
	ASTRAL	Phylonet	Model tree
0.5X, 200gt, 500bp	10,007,425.55	9,970,770.00	10,003,929.55
1X, 200gt, 500bp	11,270,657.55	11,249,179.6	11,266,716.15
2X, 200gt, 500bp	11,944,654.40	11,919,007.15	11,943,759.05
1X, 200gt, 250bp	10,561,825.45	10,491,273.00	10,557,321.60
1X, 200gt, 500bp	11,270,657.55	11,249,179.60	11,266,716.15
1X, 200gt, 1000bp	11,586,641.45	11,568,183.15	11,568,183.15
1X, 200gt, 1500bp	11,658,445.25	11,645,163.15	11,657,249.90
1X, 200gt, true-tree	11,746,203.50	11,731,108.60	11,744,078.75
1X, 25gt, 500bp	1,414,559.25	1,410,127.95	1,410,250.70
1X, 50gt, 500bp	2,821,123.85	2,816,191.40	2,816,708.35
1X, 100gt, 500bp	5,636,032.85	5,622,636.40	5,630,260.55
1X, 200gt, 500bp	11,270,657.55	11,249,179.60	11,266,716.15
1X, 400gt, 500bp	22,534,424.55	22,487,163.00	22,531,906.20
1X, 800gt, 500bp	45,096,874.45	44,996,519.20	45,096,639.80

all model conditions since it estimates species trees by minimizing extra lineages (resulting from deep coalescence). However, the true species trees may have higher amounts of extra lineages as we can see in Figure 5.2. ASTRAL, on these datasets, overestimated the EL scores. Overall, these results indicate that Phylonet underestimates the amount of ILS by estimating trees that minimize the number of extra lineages.

The EL scores of ASTRAL estimated trees are much closer to the true EL scores compared to Phylonet except for one model condition containing 1X ILS, 200 genes, and 500 bp sequence length (see Fig. 5.2a 1X, Fig. 5.2b 500bp, and Fig. 5.2c 200 genes). This is due to the fact that topological accuracy of ASTRAL is much higher than Phylonet and hence embedding/reconciling the gene trees inside the ASTRAL estimated species trees results in closer EL scores (with respect to the true EL scores).

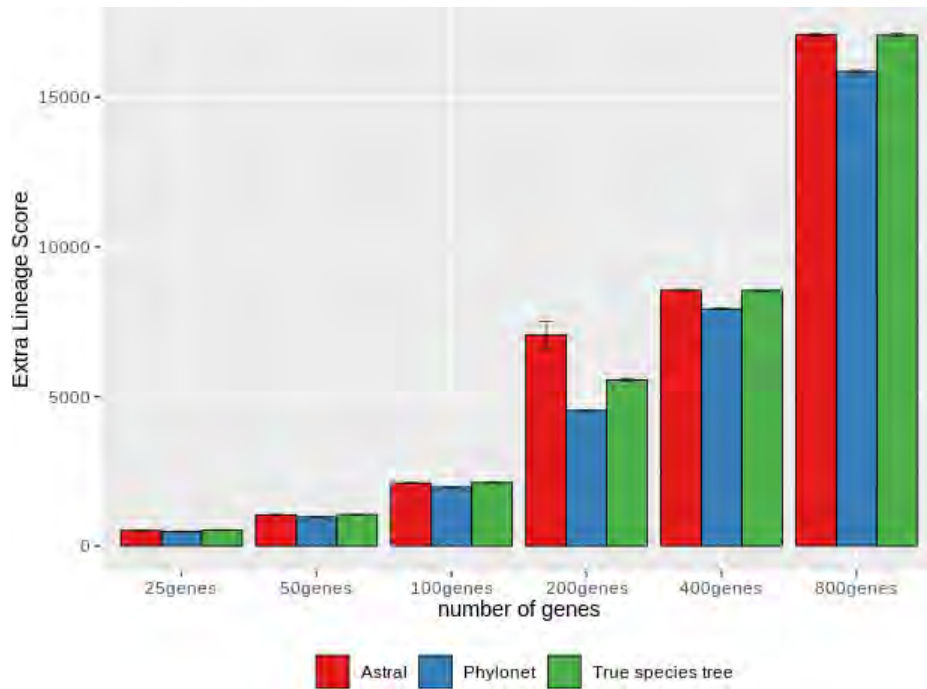
**Pseudo species tree terraces:** We have observed that Phylonet may achieve as high quartet score as ASTRAL, but may not be as accurate as ASTRAL. That means we may have a collection of trees with the same quartet score but with different topologies – indicating the presence of quartet terraces. To further investigate this, we generated about 9500 neighboring species trees of the species trees estimated by ASTRAL and Phylonet (for one replicate in 1X, 200gt, 500bp model condition) using subtree prune-and-regraft (SPR) operations. We plotted the QS scores and EL scores of these trees against their corresponding FN rates (see





(a)

(b)



(c)

Figure 5.2: EL scores for ASTRAL, Phylonet and true species tree on 37-taxon simulated mammalian dataset. We show average EL scores with standard error bars over 20 replicates for various model conditions (controlling the levels of ILS, gene tree estimation error and numbers of genes).

Fig. 5.3). We identified 29 pseudo quartet terraces each containing two trees, and 1678 EL terraces of different sizes (from 2 to 7 trees). In fact, these sets of equally scoring trees can be much larger if we analyze more species trees. Surprisingly, the topological accuracy of the trees in a particular pseudo terrace may vary substantially. For example, the FN rates of the seven trees in an EL terrace (out of the 1678 EL terraces) range from 20% to 44%. Similarly, we observed substantial differences in quartet scores and EL scores between the trees in a set of trees with the same FN rate. We identified a set of 376 trees (out of around 9000 trees we analyzed) with an identical FN rate of 0.294, but the quartet scores of which vary from 9,129,372 to 10,192,450, and the EL scores vary from 12,824 to 24,366. These results support previous findings reported in [25], where they observed some cases where wQMC [9] produced trees with better quartet support scores than ASTRAL, but ASTRAL matched the accuracy of wQMC.

To further investigate the relationships between FN rate, quartet score and EL score, we color the data points (corresponding to the 9500 trees) plotted in Figure 5.3 with a color gradient which varies continuously from dark red to dark blue with increasing EL scores (in Figure 5.3a), and with increasing quartet scores (in Figure 5.3b). Trees with higher quartet scores are expected to have relatively lower ILS scores. This is more evident in Figure 5.8 than in this particular figure on a subset of around 9500 trees for the 37-taxon dataset. From the scores of the ASTRAL- and Phylonet-estimated trees and the true species tree (see the yellow, green and black dots, Tables 5.2, 5.3, and Figure 5.2), it is evident that both ASTRAL and Phylonet “overshoot” respective optimization criteria – ASTRAL *overestimates* the quartet score, and Phylonet *underestimates* the EL score. In doing so, ASTRAL also tends to overestimate the EL score and Phylonet tends to underestimate the quartet score. This underscores the need to apply multi-objective optimization [90–93] in species tree estimation, where multiple optimization criteria (e.g, MQC and MDC) would be simultaneously optimized to reduce the tendency of overshooting a particular criterion. Similar results are shown separately for the neighborhoods of the ASTRAL- and Phylonet-estimated trees in Figures 5.4 and 5.5.

Table 5.3: **Quartet and EL scores of the ASTRAL- and Phylonet-estimated trees and the model species tree for the model condition analyzed in Figure 5.3.** We show the scores of the three trees corresponding to the yellow, green and black dots in Figure 5.3.

Optimality criterion	ASTRAL	Phylonet	Model tree
Quartet score	10,239,226	10,198,005	10,237,882
EL score	5,830	5,414	5,751



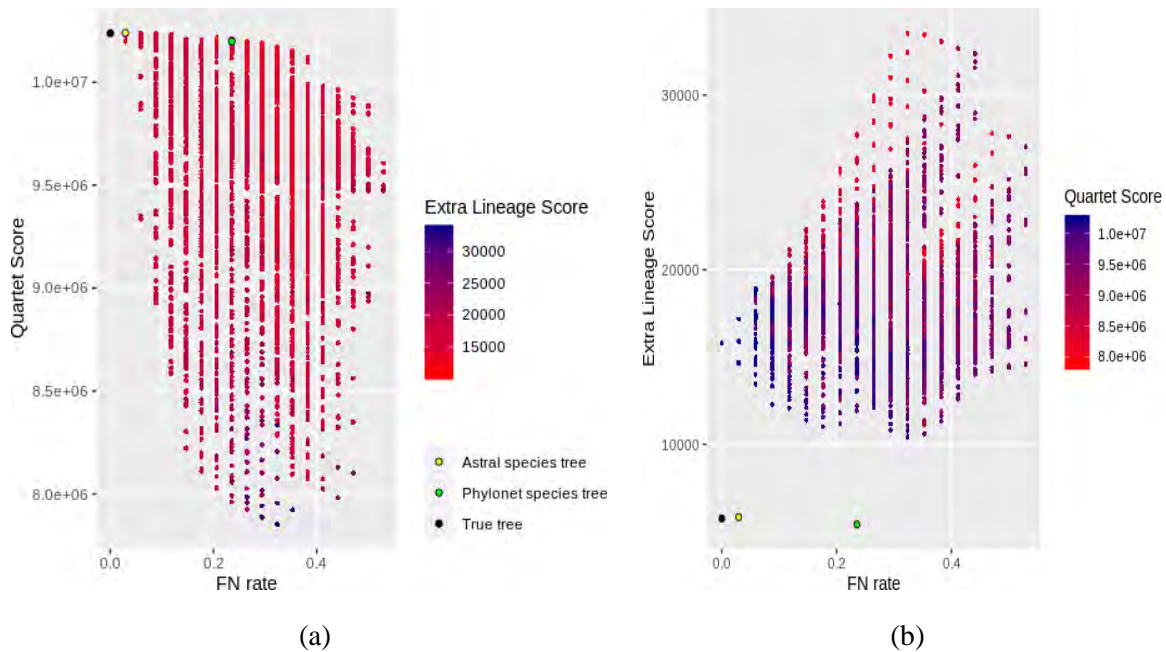


Figure 5.3: **Quartet and EL terraces on 37-taxon dataset.** We show the results for around 9500 trees: around 4700 neighbors of both ASTRAL- and Phylonet-estimated trees were generated using subtree prune-and-regraft (SPR) operations. We also show the scores of the true species tree. (a) Species tree estimation error vs. quartet score for ASTRAL- and Phylonet-estimated trees and their neighboring trees. (b) species tree estimation error vs. EL score for ASTRAL- and Phylonet-estimated trees and their neighboring trees. We color the data points with a color gradient which varies continuously from dark red to dark blue with increasing EL or quartet scores.

## Results on 11-taxon dataset

The 11-taxon dataset was simulated under a complex process in order to have substantial heterogeneity between genes and to deviate from the molecular clock [88]. It contains two model conditions: one with long branches that produce low levels of ILS (weak ILS), and the other one with short branches that result in high amounts of ILS (strong ILS). We analyzed the estimated maximum likelihood gene trees as well as the true gene trees. We also varied the number of genes from 5 to 100.

**Missing branch rate:** Figure 5.6 shows the average FN rates of ASTRAL and Phylonet on various model conditions. Both these methods improved as we decreased the amounts of ILS and increased the numbers of genes. Unlike the 37-taxon dataset, under most of the model conditions (strong ILS and weak ILS true gene trees, and weak ILS estimated gene trees), Phylonet matched the accuracy of ASTRAL. ASTRAL improved on Phylonet only on the strong ILS data with estimated gene trees. Both ASTRAL and Phylonet produced highly accurate species trees on true gene trees, especially for weak ILS where they recovered the true species tree even with only 5 genes. Therefore, MDC seems to be more sensitive to the

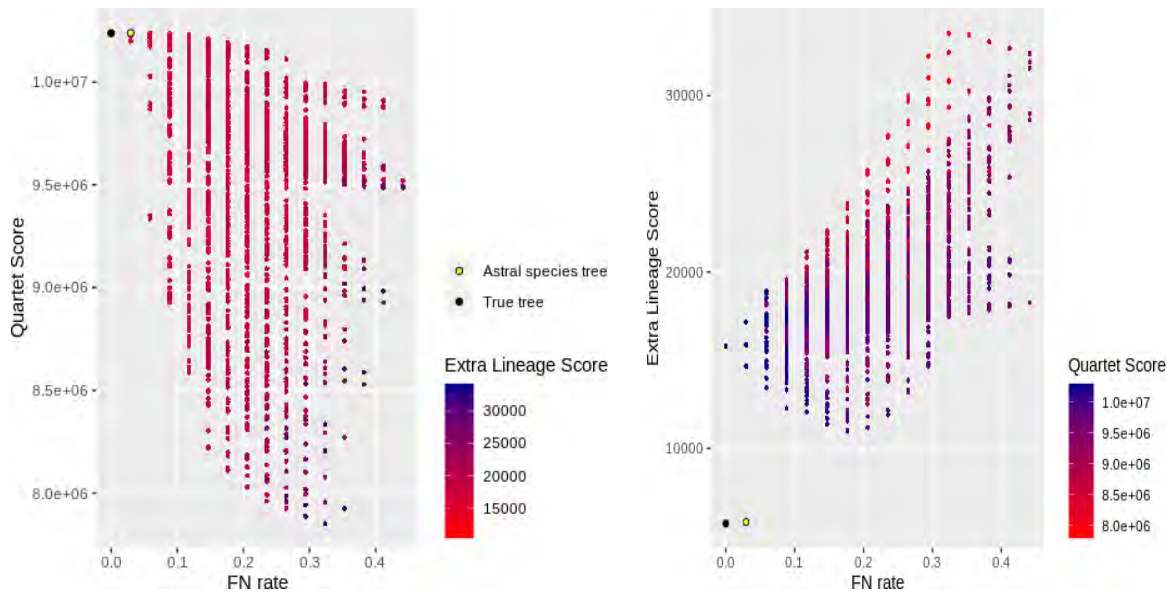


Figure 5.4: **Demonstration of quartet and EL terraces in 37-taxon dataset.** We show the results for around 4700 neighboring trees, generated using subtree prune-and-regraft (SPR) operations, of ASTRAL-estimated trees. We also show the scores of the true species tree. (a) Species tree estimation error vs. quartet score for ASTRAL-estimated tree and its neighboring trees. (b) species tree estimation error vs. EL score for ASTRAL-estimated trees and its neighboring trees. We color the data points with a color gradient which varies continuously from dark red to dark blue with increasing EL or quartet scores.

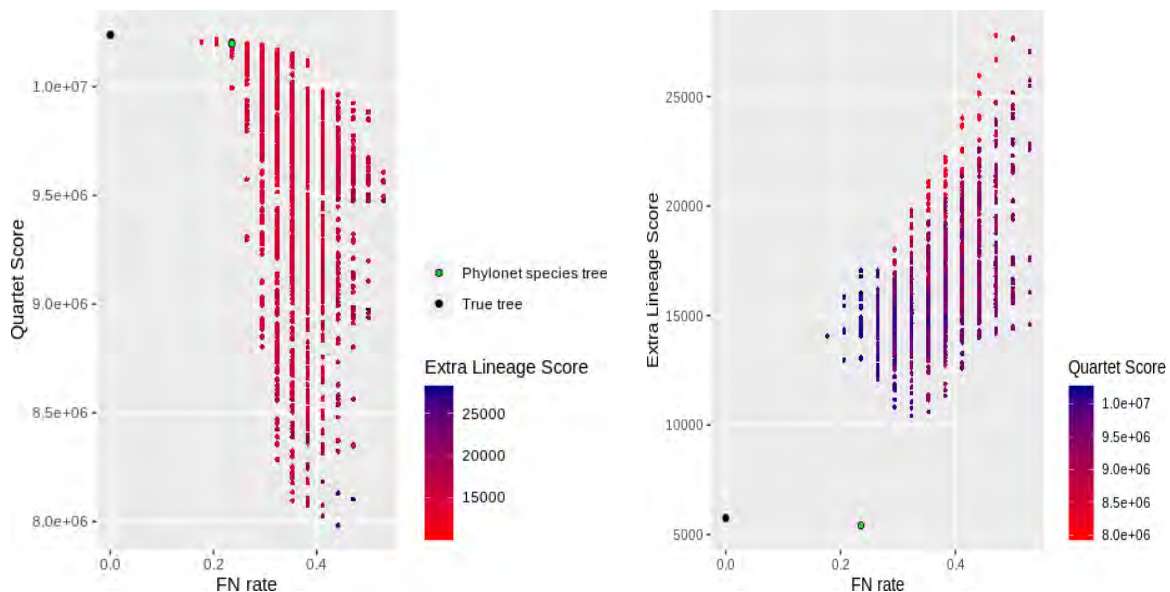


Figure 5.5: **Quartet and EL terraces in 37-taxon dataset.** Similar to Fig. 5.4, We show the results for around 4700 neighboring trees of the Phylonet-estimated tree.

gene tree estimation error than MQC.

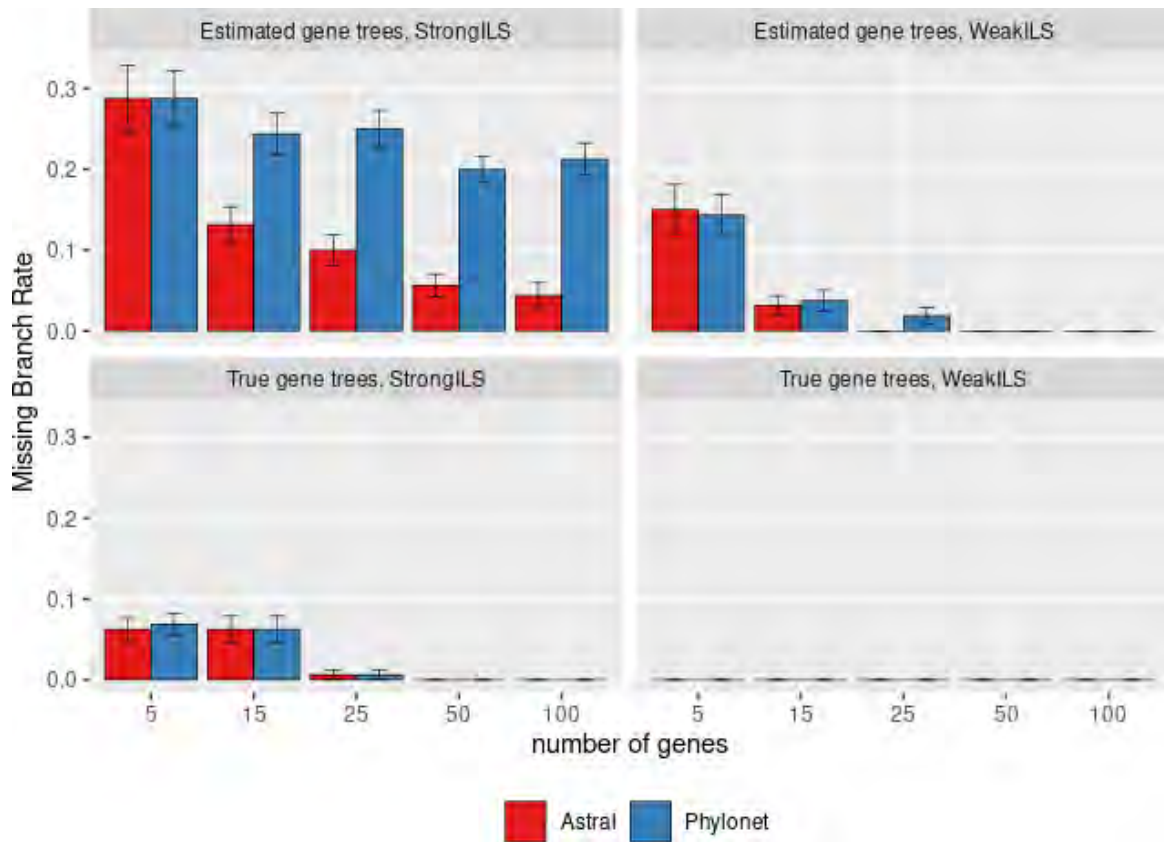


Figure 5.6: **Average FN rates of ASTRAL and Phylonet on 11-taxon dataset.** We analyzed two model conditions with varying amounts of ILS (strong ILS and weak ILS), and considered both estimated and true trees. We varied the numbers of genes from 5 to 100. Average FN rates are shown with standard error bars over 20 replicates.

**Quartet score:** Since both ASTRAL and Phylonet recovered the true species tree on many of the model conditions, the quartet scores are equal to the true QS on those model conditions (see Table 5.4). On the model conditions where they had differences (e.g., all the strong ILS model conditions with estimated gene trees and a few model conditions on weak ILS estimated gene trees), ASTRAL obtained higher quartet scores than Phylonet.

**EL score:** Figure 5.7 and Table 5.5 show the EL scores of various trees on the 11-taxon dataset. On true gene trees, both Phylonet and ASTRAL recovered the true species tree in most of the cases, and therefore, the EL scores of Phylonet- and ASTRAL-estimated trees and the model species tree are identical for those model conditions. On estimated gene trees, Phylonet achieved lower EL scores than the true EL scores, and ASTRAL’s EL scores were closer to the true EL scores than those of Phylonet.

**Pseudo species tree terraces:** Similar to the 37-taxon dataset, we observed the presence of species tree terraces in the 11-taxon dataset. We investigated the FN rates, quartet scores,

Table 5.4: **Quartet scores of ASTRAL- and Phylonet-estimated trees and the model species tree on 11-taxon datasets under various model conditions.** We show the average number of consistent quartets over 20 replicates.

Model Condition	# genes	ASTRAL	Phylonet	Model tree
True gene trees, Strong ILS	5	1,536.80	1,536.80	1,528.00
	15	4,637.90	4,637.90	4,619.50
	25	7,708.45	7,708.45	7,708.25
	50	15,412.55	15,412.55	15,412.55
	100	30,844.25	30,844.25	30,844.25
True gene trees, Weak ILS	5	1,630.40	1,630.40	1,630.40
	15	4,871.30	4,871.30	4,871.30
	25	8,138.80	8,138.80	8,138.80
	50	16,269.90	16,269.90	16,269.90
	100	32,536.90	32,536.90	32,536.90
Estimated gene trees, Strong ILS	5	1,430.55	1,415.90	1,390.40
	15	4,068.25	4,024.30	4,042.90
	25	6,695.40	6,578.25	6,679.45
	50	13,383.10	13,112.25	13,362.85
	100	26,705.40	26,148.85	26,686.55
Estimated gene trees, Weak ILS	5	1,438.70	1,433.40	1,414.15
	15	4,311.40	4,311.05	4,302.90
	25	7,152.60	7,135.70	7,152.60
	50	14,241.20	14,241.20	14,241.20
	100	28,547.90	28,547.90	28,547.90

and EL scores of the trees in the neighborhood of ASTRAL- and Phylonet-estimated trees on a single replicate in strong ILS, 25 gene model condition, and observed MDC and quartet terraces (Fig. 5.8). We identified 81 EL terraces each containing two trees, and 23 quartet terraces of different sizes (from 2 to 4 trees). Moreover, we identified a few replicates (among the 20 replicates) under different model conditions, where Phylonet and ASTRAL obtained identical quartet scores but differed in tree topologies – indicating the presence of pseudo quartet terraces (see Table 5.6). Similarly, we identified some replicates where ASTRAL- and Phylonet-estimated trees are topologically different, but have identical EL scores – indicating the presence of pseudo EL terraces (see Table 5.7). This figure suggests a negative correlation between quartet score and EL score – higher quartet scores result in lower EL scores, and higher EL scores correspond to lower quartet scores. Figs. S11 and S12 in the Supplementary Material present the same analysis but was performed separately on the neighborhoods of ASTRAL and Phylonet tree.



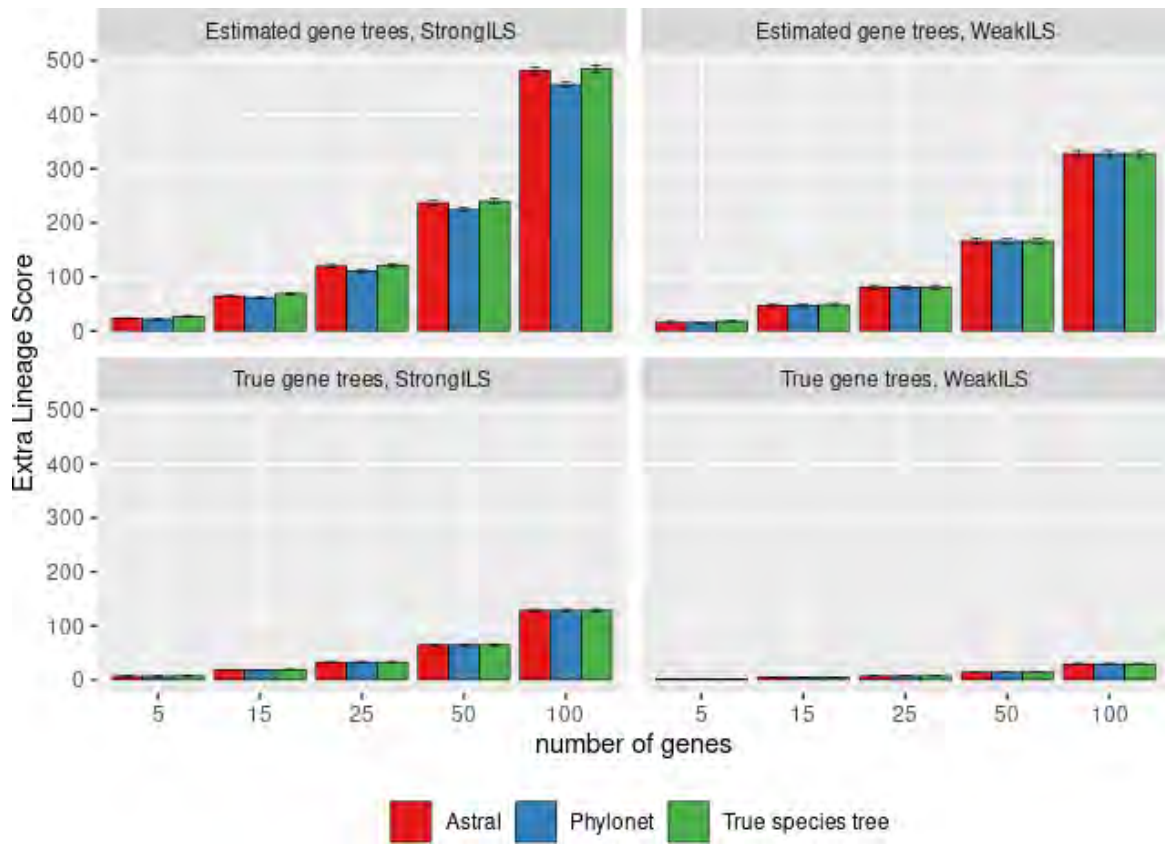


Figure 5.7: **Extra lineage scores for ASTRAL and Phylonet on 11-taxon dataset under various model conditions.** We show average EL scores with standard error bars over 20 replicates.

## Consensus trees of the trees in a pseudo terrace

The multiplicity of equally good trees in a species tree space introduces ambiguity as it imposes the challenge of identifying relatively more reliable trees within the pseudo terraces and their neighborhood. However, leveraging the trees in a pseudo terrace by finding various consensus trees (e.g., greedy consensus, maximum clade credibility, etc.) may help address the uncertainty and may lead to trees with higher optimality scores than the scores of the corresponding pseudo terraces.

In order to investigate this, we generated all possible rooted trees with 11 taxa, resulting in a set of 65,47,29,075 candidate species tree. We considered one set of gene trees in the 25-gene strong ILS model condition, and identified the pseudo terraces. Since analyzing this large number of trees is prohibitively computationally expensive, we selected a subset of around million trees (around 1.1 million trees before and after the true species tree in the sequence of 65,47,29,075 generated trees). We computed the EL and quartet scores of these 2.2 million trees and identified 733 and 4,320 EL and quartet terraces, respectively (see Table 5.8). The

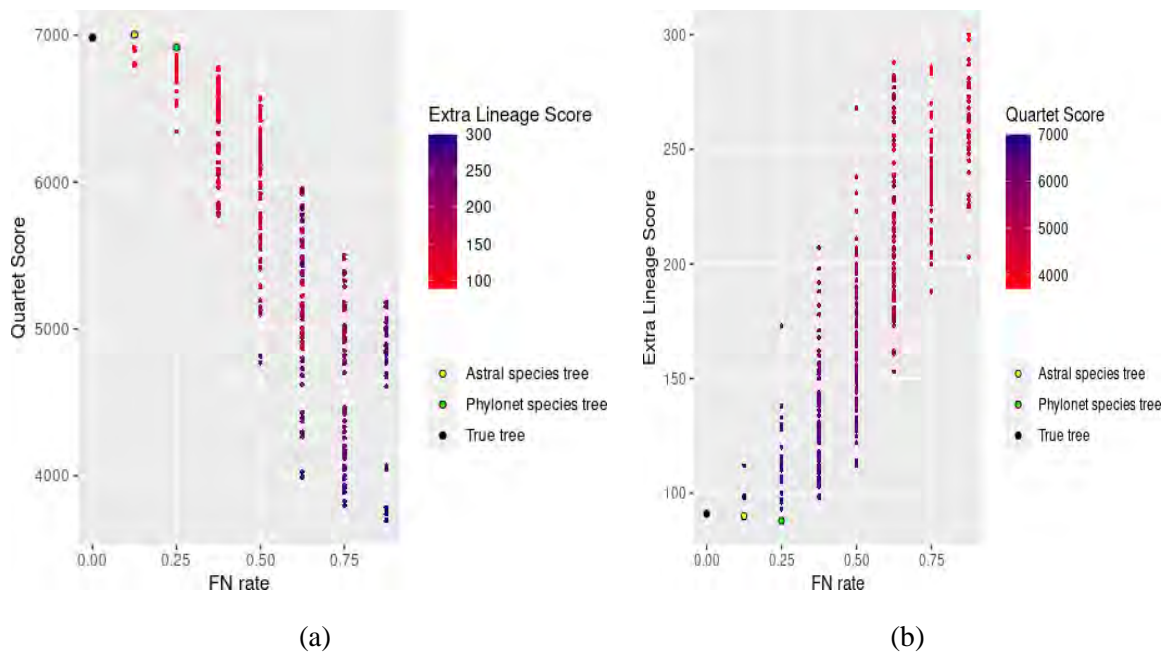


Figure 5.8: **Quartet and EL terraces on 11-taxon dataset.** We show the results for around 500 trees: around 250 neighbors of both ASTRAL- and Phylonet-estimated trees were generated using subtree prune-and-regraft (SPR) operations. We also show the scores of the true species tree. (a) Species tree estimation error vs. quartet score for ASTRAL- and Phylonet-estimated trees and their neighboring trees. (b) species tree estimation error vs. EL score for ASTRAL- and Phylonet-estimated trees and their neighboring trees. We color the data points with a color gradient which varies continuously from dark red to dark blue with increasing EL or quartet scores.

Table 5.5: **EL scores of ASTRAL and Phylonet estimated trees and the model species tree on 11-taxon datasets under various model conditions.** We show the average EL scores over 20 replicates.

Model Condition	# genes	ASTRAL	Phylonet	Model tree
True gene trees, Strong ILS	5	6.65	6.65	7.15
	15	18.20	18.20	18.95
	25	32.70	32.70	32.70
	50	64.90	64.90	64.90
	100	128.50	128.50	128.50
True gene trees, Weak ILS	5	1.67	1.67	1.67
	15	4.80	4.80	4.80
	25	7.10	7.10	7.10
	50	14.50	14.50	14.50
	100	29.25	29.25	29.25
Estimated gene trees, Strong ILS	5	23.20	21.45	26.70
	15	64.45	61.55	68.85
	25	119.75	110.75	121.50
	50	236.80	224.85	240.00
	100	481.50	455.80	485.40
Estimated gene trees, Weak ILS	5	15.95	15.15	17.70
	15	47.40	47.40	48.25
	25	81.20	80.60	81.20
	50	166.25	166.25	166.25
	100	327.70	327.70	327.70

Table 5.6: **Quartet terraces on 11-taxon dataset.** We show different sets of gene trees (among the 20 replicates of data that we analyzed for each of the model conditions), where Phylonet and ASTRAL achieved identical quartet score but differed in tree topologies.

Model Condition	Gene tree set	Quartet Score		FN rate	
		Phylonet	ASTRAL	Phylonet	ASTRAL
True gene trees, Strong ILS, 5 genes	1	1528	1528	0.125	0.000
	2	1536	1536	0.125	0.000
	3	1548	1548	0.000	0.125
	4	1502	1502	0.000	0.125
	5	1548	1548	0.125	0.000
	6	1518	1518	0.000	0.125
	7	1488	1488	0.125	0.000
Estimated gene trees, Strong ILS, 5 genes	8	1408	1408	0.125	0.000
Estimated gene trees, Weak ILS, 5 genes	9	1534	1534	0.125	0.250

sizes of the EL and quartet terraces vary from 2 to 8,527 trees, and 2 to 1,981 trees, respectively. The EL scores of 733 pseudo EL terraces range from 118 to 862, and the quartet scores of the

Table 5.7: **EL terraces on 11-taxon dataset.** We show different sets of gene trees, where Phylonet and ASTRAL achieved identical EL scores but differed in tree topologies.

Model Condition	Gene tree set	EL Score		FN rate	
		Phylonet	ASTRAL	Phylonet	ASTRAL
True gene trees, Strong ILS, 5 genes	1	7	7	0.000	0.125
	2	10	10	0.000	0.125
	3	6	6	0.125	0.000
	4	7	7	0.000	0.125
	5	8	8	0.125	0.000

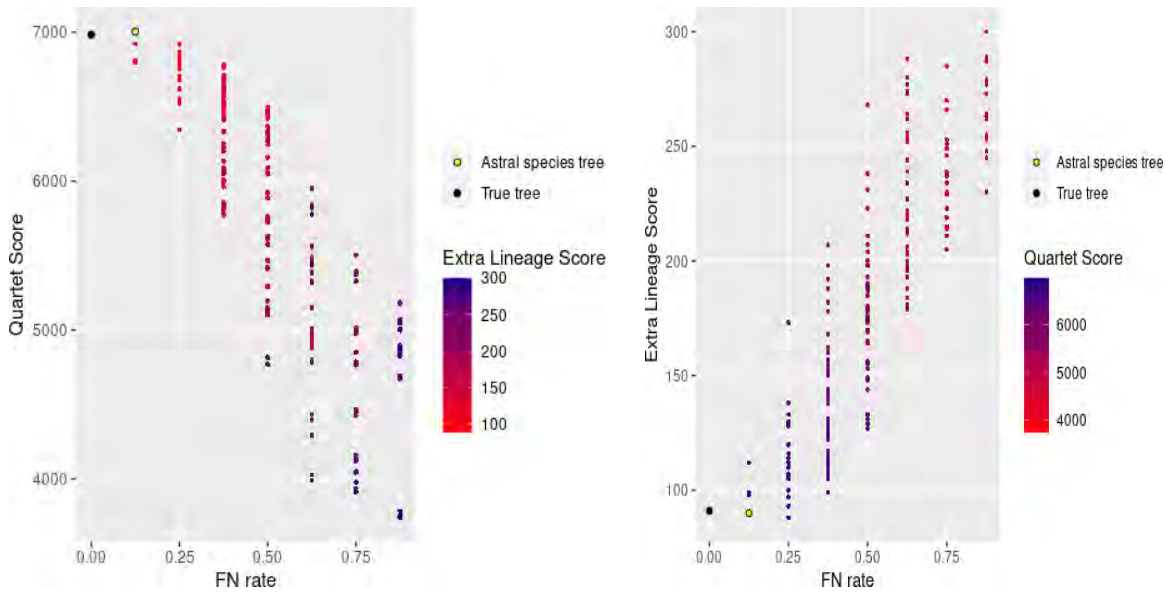


Figure 5.9: **Quartet and EL terrace in 11-taxon dataset.** We show the results for around 250 neighboring trees, generated using subtree prune-and-regraft (SPR) operations, of ASTRAL-estimated trees. We also show the scores of the true species tree. (a) Species tree estimation error vs. quartet score for ASTRAL-estimated tree and its neighboring trees. (b) species tree estimation error vs. EL score for ASTRAL-estimated trees and its neighboring trees. We color the data points with a color gradient which varies continuously from dark red to dark blue with increasing EL or quartet scores.

4,320 pseudo quartet terraces range from 1,768 to 6,702. For each of these pseudo terraces, we computed greedy consensus trees and compared the optimality scores of the consensus trees with the optimality scores of the corresponding pseudo terraces. We observed that, in most of the cases, the optimality score of the greedy consensus tree is better than the score of the corresponding psuedo terrace (see Table 5.9). Computing the greedy consensus trees led to better (lower) EL scores for around 99% of the 733 EL terraces, and better (higher) quartet scores for around 61% of the quartet terraces.

Note that consensus trees can be non-binary (i.e., not fully resolved), having fewer internal edges than bifurcating topologies. Therefore, consensus trees may have relatively lower EL



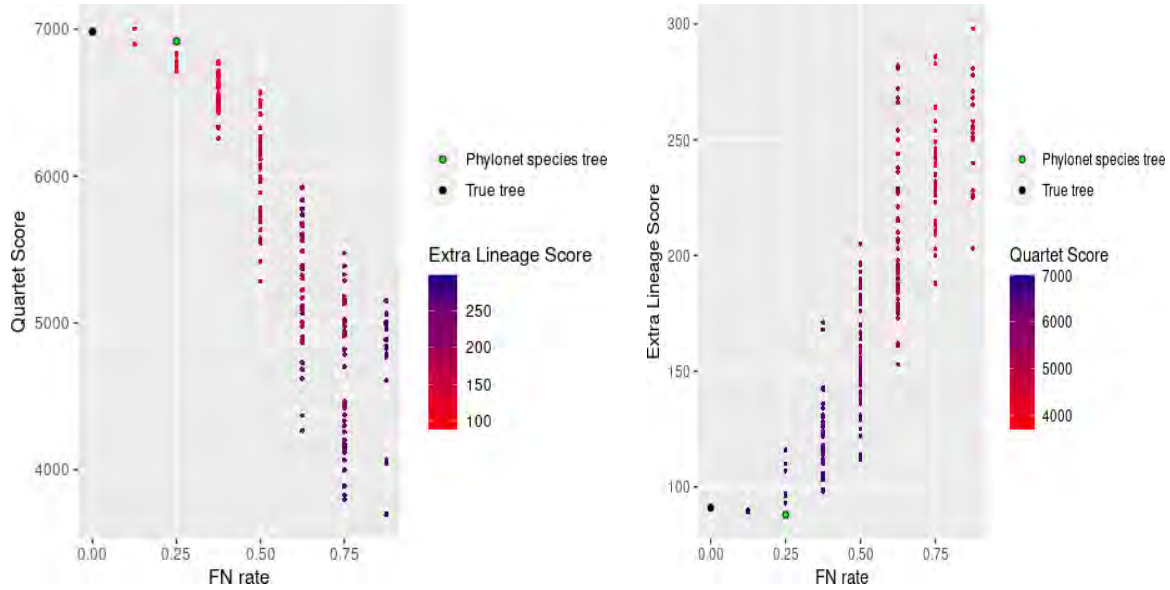


Figure 5.10: **Demonstration of quartet and EL terrace in 11-taxon dataset.** Similar to Fig. 5.9, We show the results for around 250 neighboring trees of the Phylonet-estimated tree.

scores simply due to fewer numbers of internal lineages. However, this is not the case for this dataset where computing greedy consensus trees resulted in non-binary trees only on two EL terraces (out of 733) and 161 quartet terraces (out of 4,320). All of these non-binary trees contain seven internal edges (whereas the unrooted topology of a bifurcating tree with 11 taxa contains eight internal edges). Thus, the improved (lower) EL scores for 722 consensus trees, only two of which were non-binary, indicates that the improvement in EL score is not solely due to the decreased resolution. Similarly, unresolved trees may result in relatively lower (i.e., worse) quartet scores than fully resolved trees. On this dataset, 21 non-binary consensus trees (out of 161) obtained lower quartet scores than the corresponding pseudo terraces.

Table 5.8: **Pseudo EL and quartet terraces on around 2.2 million candidate species trees for the 11-taxon dataset.** We show the total number of pseudo terraces and minimum, maximum, and average numbers of trees in those pseudo terraces.

Optimality criterion	No. of terraces	Minimum # trees	Maximum # trees	Average # trees
EL	733	2	8,527	2,946.78
Quartet	4,320	2	1,981	499.98

We also empirically investigated the degree of difference between the trees in a pseudo terrace by computing the SPR distance between all pairs of trees in a pseudo terrace. Due to the huge computational burden for considering all pairs of trees, we could not analyze all the terraces that we identified on a collection of 2.2 million trees. We analyzed 20 quartet and 20 EL terraces, sizes of which range from 10 to 204 trees. The minimum, maximum, and average

Table 5.9: **Comparison of the optimality scores of the consensus trees and the corresponding pseudo terraces.** We show the number of pseudo terraces, where computing the greedy consensus trees resulted in better, worse, and identical EL and quartet scores.

Optimality criterion	Change in score	# terraces	% terraces
EL	Better (lower)	722	98.50
	Worse (higher)	4	0.55
	Identical	7	0.95
Quartet	Better (higher)	2,646	61.25
	Worse (lower)	866	20.05
	Identical	808	18.70

SPR distances between the trees in the EL terraces are 1, 6, and 3.45, respectively. Similar values (1, 6, and 3.95) were observed for the quartet terraces. Future studies need to perform more exhaustive empirical studies on relatively larger terraces for different datasets, as well as provide analytical results showing the degree of similarity between the trees in a pseudo terrace in terms of tree rearrangement operations.

## Results on biological dataset

### Amniota dataset

We analyzed the Amniota dataset from Chiari *et al.* [89] containing 248 genes across 16 amniota taxa in order to resolve the position of turtles relative to birds and crocodiles. This dataset contains both amino acid (AA) and nucleotide (DNA) gene trees. We ran exact versions for both ASTRAL and Phylonet.

Previous studies suggest that placing turtles as a sister group to Archosaurs (birds and crocodiles) is more reliable [23, 89]. ASTRAL on both DNA and AA data produced (turtles,(birds,crocodiles)), and thus recovered the Archosaurs (see Figure 5.12 and Figure 5.12).

Phylonet estimated trees on both AA and DNA gene trees are same except for the resolution within the turtles (*Phrynops hilarii*, *Caretta caretta*, *Chelonoidis nigra*, and *Emys orbicularis*). Phylonet, on both AA and DNA data, produced the Archosaurs, but it placed turtles as sister to Squamates (snakes and lizards) and placed Archosaurs as sister to the clade containing turtles and Squamates. Thus, Phylonet did not produce the (turtles,(birds,crocodiles)) relationship. The quartet and extra lineage scores of ASTRAL and Phylonet estimated trees are given in Table 5.10.

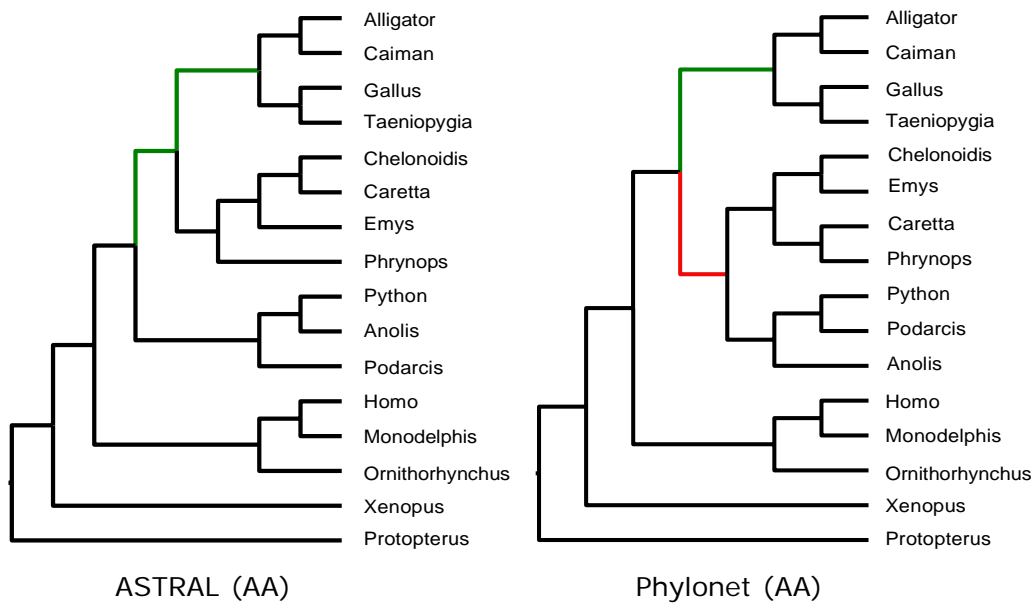


Figure 5.11: Analysis of the Amniota AA dataset by maximizing quartet score (ASTRAL) and minimizing extra lineage score (Phylonet). We show the rooted versions of the ASTRAL-estimated trees using the outgroup (*Protopterus annectens*).

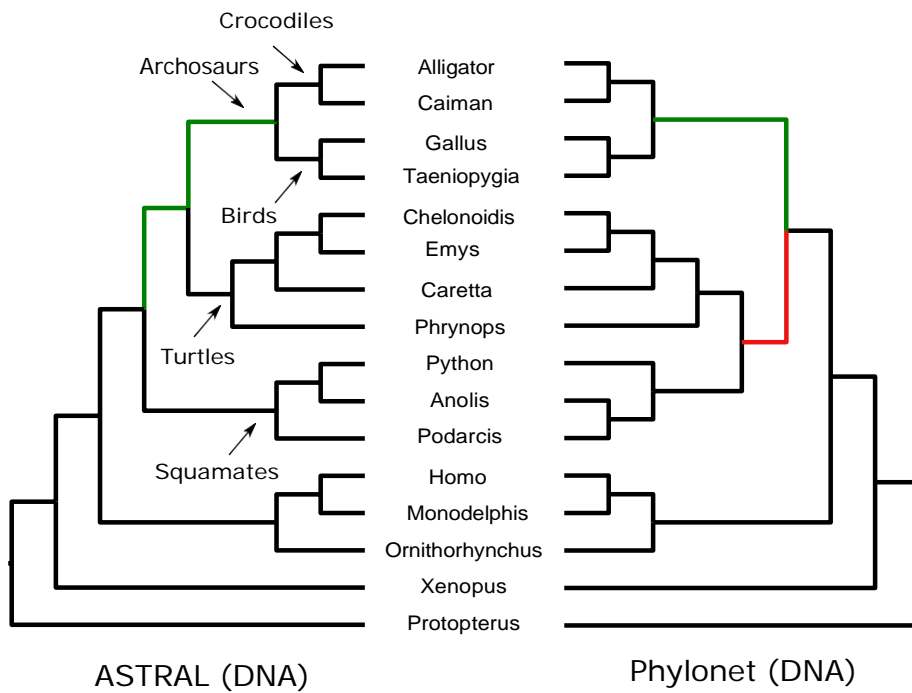


Figure 5.12: Analysis of the Amniota DNA dataset by maximizing quartet score (ASTRAL) and minimizing extra lineage score (Phylonet). We show the rooted version of the ASTRAL-estimated trees using the outgroup (*Protopterus annectens*).

Table 5.10: **Quartet and EL scores of ASTRAL- and Phylonet-estimated trees on the amniota dataset (both DNA and AA).**

Tool	EL score	Quartet score
ASTRAL (DNA)	2,620	97,890
Phylonet (DNA)	1,170	93,018
ASTRAL (AA)	3,293	83,604
Phylonet (AA)	1,916	80,507

### Mammalian dataset

Song *et al.* analyzed a dataset containing 447 genes across 37 mammals using MP-EST [47] and concatenation using maximum likelihood [87]. We reanalyzed this dataset after removing 21 mislabeled genes (confirmed by the authors), and two other outlier genes using ASTRAL and Phylonet. The placement of bats (*Myotis lucifugus* and *Pteropus vampyrus*) and tree shrews (*Tupaia belangeri*) were two of the questions of greatest interest, and alternative relationships have previously been reported [94–97]. ASTRAL and Phylonet estimated trees also exhibit similar differences with respect to the placement of bats and tree shrews (see Fig. 5.13). ASTRAL placed tree shrews as sister to Glires (Rodentia, Lagomorpha) which is consistent to the tree estimated by concatenation using maximum likelihood reported in [87]. Phylonet recovered a tree that placed tree shrews as sister to the Primates, which is consistent to the tree estimated by MP-EST using multi-locus bootstrapping [98] (reported in [6,23,87]). Both trees put Perissodactyla (*Equus caballus*) as a sister to Carnivora (*Canis familiaris*, *Felis catus*). With respect to the position of bats, ASTRAL agrees with MP-EST which placed bats as sister to the (Cetartiodactyla, (Perissodactyla, Carnivora)) clade. Phylonet placed bats as sister to Cetartiodactyla, and put (Perissodactyla, Carnivora) as the sister clade of (bats, Cetartiodactyla), and thus agrees with concatenation [23]. The extra lineage and quartet scores of these two trees are reported in Table 5.11.

Table 5.11: **Quartet and EL scores of ASTRAL- and Phylonet-estimated trees on the biological mammalian dataset [87].**

Tool	EL score	Quartet score
ASTRAL	5,909	25,526,915
Phylonet	5,675	25,479,405

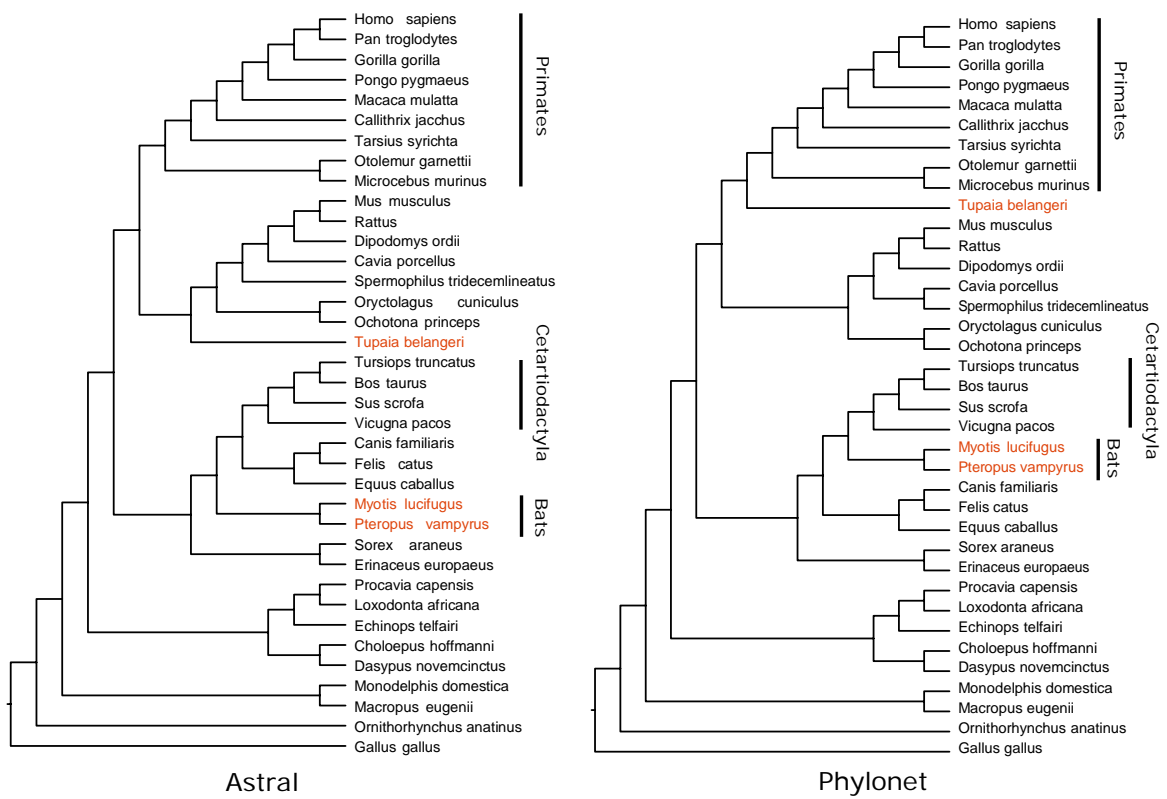


Figure 5.13: Analyses of the mammalian dataset by maximizing quartet score (ASTRAL) and minimizing deep coalescence (Phylonet). These two trees differ in the placement of tree shrews and bats.

# Chapter 6

## Conclusions

We report, based on an extensive evaluation study, the comparison between MQC and MDC criteria for estimating species trees in the presence of incomplete lineage sorting. While the superiority of MQC over MDC is expected (since MQC is a statistically consistent estimator of species tree under ILS), and the observations that MDC criteria underestimates the amount of deep coalescence is not novel [15], this study is the first to evaluate MQC and MDC and confirms these results on various simulated and real biological datasets, and hence provide additional support for the consistency properties of MQC and MDC. Although the presence of multiple equally good species trees with respect to a particular optimization criteria is not unexpected [11], we – for the first time – provided the conditions for the datasets to have equally optimal trees in the context of phylogenomic inference using summary methods under MDC and MQC criteria.

Our experimental study suggests that species trees estimated under MQC and MDC may belong to sets of equally optimal trees and their neighborhoods, but ASTRAL’s search strategy usually leads to trees that are closer to the true species trees than the trees estimated under MDC criterion. This study reveals various interesting trends regarding the FN rates, quartet scores and EL scores of the trees estimated by MQC and MDC under various model conditions.

Pseudo species tree terraces have implications in the search strategies under various optimization criteria. Species tree space grows exponentially with the number of taxa. Thus, algorithms to find optimal species trees under various optimization criteria requires navigating a large tree space. Considering the presence of large set of equally optimal trees, efficient algorithms to strategically explore the terraces and their neighborhood is crucial. Fundamental to most of the summary methods is the ability to efficiently explore and score (with respect to a particular optimization criterion) the trees inside the tree space. Since all the trees inside a particular pseudo terrace have the same optimization score, identifying a terrace may help reduce the computational efforts by avoiding the computation time that might be

---

unnecessarily spent to evaluate many trees with identical score. Thus, efficient identification of a pseudo terrace and directing the tree search from one terrace to the other ones with higher optimization scores may result in faster convergence. Exact solutions to various optimization criteria, such as MQC [6], MDC [12] and MGDL [57, 58] are available that are guaranteed to find the globally optimal solution under respective optimization criteria. However, the application of exact versions to large datasets has been limited by the prohibitive amount of time required by the available algorithms to explore the tree space exhaustively. To the best of our knowledge, these algorithms cannot detect multiple equally good solutions. Therefore, utilizing the knowledge of terraces may help prune the search space. However, it could also be possible that a particular tree in a terrace is topologically more correct than the other ones, and hence navigating off from a terrace may lead us to miss more reliable (in terms of topological accuracy) trees. Therefore, the presence of potentially large sets of equally optimal trees imposes the challenge of identifying relatively more reliable trees within the terraces and their neighborhoods. Thus, the multiplicity of equally good trees in a species tree terrace introduces ambiguity. One plausible option for reducing the ambiguity is to estimate consensus trees (greedy consensus, majority consensus, maximum agreement subtree, maximum clade credibility tree, etc.) of the trees in a terrace. Indeed, our experimental results suggest that computing the greedy consensus trees results in better optimization scores. This can also be used to draw branch supports on the species tree without having to rely on multi-locus bootstrapping [98]. Thus, future studies need to investigate the properties of the consensus trees of the trees in a pseudo species tree terrace. Another important direction would be optimizing multiple optimization criteria [90,92] simultaneously instead of a single one.

Navigating trees within a terrace could be easier due to their similarity with respect to a particular optimization criterion. *Terrace-aware* data structures led to substantial speedup of RAxML [79, 99] and IQ-tree [100] for estimating ML trees from alignments [31]. Thus, efficient *terrace-aware* algorithms and data structures for strategically navigating trees both inside a pseudo species tree terrace and its neighborhood would contribute to the improvement of the summary methods both in terms of accuracy and scalability. Efficiently characterizing a pseudo species tree terrace – by (empirically and analytically) quantifying the difference between the trees using various distance measures such as, average Robinson-Foulds (RF), nearest neighbor interchange (NNI), subtree prune-and-regraft (SPR), and tree bisection and re-connection (TBR) distances – would be another interesting research direction. Thus, the discovery of terraces poses various challenges as well as opens up several important research avenues.

This study is limited in scope and can be extended in several directions. We analyzed complete gene trees with full set of taxa. Future studies need to investigate the impact of missing data

(i.e., incomplete gene trees) in pseudo species tree terraces. This study analyzed small to moderate sized datasets. Small datasets enabled us to run the exact versions of ASTRAL and Phylonet. However, impacts of equally optimal trees in larger datasets with hundreds of taxa need to be investigated as the possibility of the presence of potentially large terraces is relatively higher for larger numbers of taxa. This study investigated relatively long sequences (250 ~2000 bp); subsequent studies should investigate the relative performance of MQC and MDC on very short sequences, since recombination-free loci can be very short [101]. Finally, investigating further combinatorial properties of species tree terraces, the problems they induce, and strategies for overcoming them is crucial for formalizing this concept as well as for developing terrace-aware data structures and tree search algorithms.



## References

- [1] Kenneth M Halanych and Leslie R Goertzen. Grand challenges in organismal biology: the need to develop both theory and resources. *Integrative and comparative biology*, 49(5):475–479, 2009.
- [2] Irfan Hussain, Nashaiman Pervaiz, Abbas Khan, Shoaib Saleem, Huma Shireen, Dong-Qing Wei, Viviane Labrie, Yiming Bao, and Amir Ali Abbasi. Evolutionary and structural analysis of sars-cov-2 specific evasion of host immunity. *Genes & Immunity*, 21(6):409–419, 2020.
- [3] Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, 2015.
- [4] J H Degnan, M DeGiorgio, D Bryant, and N A Rosenberg. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*, 58:35–54, 2009.
- [5] L S Kubatko and J H Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56:17, 2007.
- [6] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.
- [7] Rezwana Reaz, Md Shamsuzzoha Bayzid, and M Sohel Rahman. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS One*, 9(8):e104008, 2014.
- [8] Sagi Snir and Satish Rao. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, 62(1):1–8, 2012.
- [9] Eliran Avni, Reuven Cohen, and Sagi Snir. Weighted quartets phylogenetics. *Systematic Biology*, 64(2):233–242, 2014.

- [10] Siavash Mirarab and Tandy Warnow. Astral-ii: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.
- [11] W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
- [12] C. V. Than and L. Nakhleh. Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, 5(9), 2009.
- [13] Jimmy Yang and Tandy Warnow. Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics*, 12(9):1–12, 2011.
- [14] W. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
- [15] C. V. Than and N. A. Rosenberg. Consistency properties of species tree inference by minimizing deep coalescences. *Journal of Computational Biology*, 18:1–15, 2011.
- [16] W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
- [17] Mukul S. Bansal, J. Gordon Burleigh, and Oliver Eulenstein. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC bioinformatics*, 11(1):S42, 2010.
- [18] C. V. Than, D. Ruths, and L. Nakhleh. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9:322, 2008.
- [19] Cuong V. Than and Noah A. Rosenberg. Consistency properties of species tree inference by minimizing deep coalescences. *Journal of Computational Biology*, 18(1):1–15, 2011.
- [20] Wayne P. Maddison and L. Lacey Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30, 2006.
- [21] Md. Shamsuzzoha Bayzid and Tandy Warnow. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–2284, 2013.
- [22] Michael DeGiorgio and James H. Degnan. Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular Biology and Evolution*, 27(3):552–569, 2009.
- [23] Siavash Mirarab, Md. Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–380, 2014.

- [24] Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16(10):S2, 2015.
- [25] Ruth Davidson, Pranjal Vachaspati, Siavash Mirarab, and Tandy Warnow. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC genomics*, 16(10):S1, 2015.
- [26] Huateng Huang, Qixin He, Laura S Kubatko, and L Lacey Knowles. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, 59(5):573–583, 2010.
- [27] David R Maddison. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Biology*, 40(3):315–328, 1991.
- [28] Laura A Salter. Complexity of the likelihood surface for a large dna dataset. *Systematic Biology*, 50(6):970–978, 2001.
- [29] Michael J Sanderson, Michelle M McMahon, and Mike Steel. Terraces in phylogenetic tree space. *Science*, 333(6041):448–450, 2011.
- [30] Michael J Sanderson, Michelle M McMahon, Alexandros Stamatakis, Derrick J Zwickl, and Mike Steel. Impacts of terraces on phylogenetic inference. *Systematic Biology*, 64(5):709–726, 2015.
- [31] Olga Chernomor, Arndt Von Haeseler, and Bui Quang Minh. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, 65(6):997–1008, 2016.
- [32] Katherine St. John. The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83–e94, 2017.
- [33] Md Bayzid et al. *Estimating species trees from gene trees despite gene tree incongruence under realistic model conditions*. PhD thesis, The University of Texas at Austin, 2016.
- [34] Tandy Warnow. *Computational phylogenetics: an introduction to designing methods for phylogeny estimation*. Cambridge University Press, 2017.
- [35] J H Degnan and N A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2:762 – 768, 2006.

- [36] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37, January 2005.
- [37] R. R. Hudson. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37:203 – 217, 1983.
- [38] M. Nei. Stochastic errors in dna evolution and molecular phylogeny. In *In H. Gershowitz, D. L. Rucknagel, and R. E. Tashian, editors, Evolutionary Perspectives and the New Genetics*, pages 133 – 147, 1986.
- [39] M. Nei. *Molecular evolutionary genetics*. New York, 1987. Columbia University Press.
- [40] N. Rosenberg. The Probability of Topological Concordance of Gene Trees and Species Trees. *Theoretical Population Biology*, 61(2):225–247, March 2002.
- [41] Fumio Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, October 1983.
- [42] N. Takahata. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.
- [43] A.J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7:214, 2007.
- [44] Scott V Edwards, Liang Liu, and Dennis K Pearl. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941, 2007.
- [45] Bret R Larget, Satish K Kotha, Colin N Dewey, and Cécile Ané. Bucky: gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.
- [46] A D Leaché and B Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol*, 60(2):126–137, 2011.
- [47] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10:302, 2010.
- [48] J Heled and A J Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27:570–580, 2010.

- [49] B Larget, S K Kotha, C N Dewey, and C Ané. BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.
- [50] Mazharul Islam, Kowshika Sarker, Trisha Das, Rezwana Reaz, and Md Shamsuzzoha Bayzid. Stelar: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. *BMC Genomics*, 21(1):1–13, 2020.
- [51] C Ané, B Larget, D A Baum, S D Smith, and A Rokas. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, 24:412–426, 2007.
- [52] R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca, and O Eulenstein. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics*, pages 574–574, 2010.
- [53] L Liu. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24:2542–2543, 2008.
- [54] Nam Nguyen, Siavash Mirarab, and Tandy Warnow. Mrl and superfine+ mrl: new supertree methods. *Algorithms for Molecular Biology*, 7(1):1–13, 2012.
- [55] C. V. Than and L. Nakhleh. Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, 5(9), 2009.
- [56] Zaineb Chelly Dagdia, Pavel Avdeyev, and Md Shamsuzzoha Bayzid. Biological computation and computational biology: survey, challenges, and discussion. *Artificial Intelligence Review*, pages 1–67, 2021.
- [57] M. S. Bayzid and T. Warnow. Gene tree parsimony for incomplete gene trees: addressing true biological loss. *Algorithms for Molecular Biology*, 13:1, 2018.
- [58] M. S. Bayzid, S. Mirarab, and T. Warnow. Inferring optimal species trees under gene duplication and loss. In *Proc. of Pacific Symposium on Biocomputing (PSB)*, volume 18, pages 250–261, 2013.
- [59] E Mossel and S Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *EEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):166–171, 2011.
- [60] L S Kubatko, B C Carstens, and L L Knowles. Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25:971–973, 2009.

- [61] Julia Chifman and Laura Kubatko. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, 374:35–47, 2015.
- [62] Liang Liu, Lili Yu, Dennis K Pearl, and Scott V Edwards. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477, 2009.
- [63] Liang Liu and Lili Yu. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5):661–667, 2011.
- [64] Pranjal Vachaspati and Tandy Warnow. Astrid: accurate species trees from internode distances. *BMC Genomics*, 16(10):S3, 2015.
- [65] Mahim Mahbub, Zahin Wahab, Rezwana Reaz, M Saifur Rahman, and Md Shamsuzzoha Bayzid. wqfm: highly accurate genome-scale species tree estimation from weighted quartets. *Bioinformatics*, 37(21):3734–3743, 2021.
- [66] Brian Tilston Smith, Michael G Harvey, Brant C Faircloth, Travis C Glenn, and Robb T Brumfield. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, 63(1):83–95, 2013.
- [67] Md Shamsuzzoha Bayzid and Tandy Warnow. Estimating optimal species trees from incomplete gene trees under deep coalescence. *Journal of Computational Biology*, 19(6):591–605, 2012.
- [68] Rezwana Reaz, Md Shamsuzzoha Bayzid, and M Sohel Rahman. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PloS one*, 9(8):e104008, 2014.
- [69] Nazifa Ahmed Mousi, Badhan Das, Zarin Tasnim Promi, Nishat Anjum Bristy, and Md Shamsuzzoha Bayzid. Quartet-based inference of cell differentiation trees from chip-seq histone modification data. *Plos one*, 14(9):e0221270, 2019.
- [70] J. H. Degnan. Anomalous unrooted gene trees. *Systematic Biology*, 62(4):574–590, 2013.
- [71] Michael Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9(1):91–116, 1992.
- [72] Trevor R Hodkinson and John AN Parnell. *Reconstructing the tree of life: taxonomy and systematics of species rich taxa*. CRC Press, 2006.

- [73] Jimmy Yang and Tandy Warnow. Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics*, 12(9):1–12, 2011.
- [74] Yun Yu, Tandy Warnow, and Luay Nakhleh. Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11):1543–1559, 2011.
- [75] Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, 2018.
- [76] T Jiang, P Kearney, and M Li. A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its applications. *SIAM Journal on Computing*, 30(6):1924–1961, 2001.
- [77] SJ Katherine. Review paper: the shape of phylogenetic treespace. *Syst. Biol.*, 66(1):e83–e94, 2017.
- [78] Rudolf Biczok, Peter Bozsoky, Peter Eisenmann, Johannes Ernst, Tobias Ribizel, Fedor Scholz, Axel Trefzer, Florian Weber, Michael Hamann, and Alexandros Stamatakis. Two c++ libraries for counting trees on a phylogenetic terrace. *Bioinformatics*, 34(19):3399–3401, 2018.
- [79] Alexandros Stamatakis and Nikolaos Alachiotis. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics*, 26(12):i132–i139, 2010.
- [80] 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*, 18(11):1543–1559, 2011.
- [81] Ishrat Tanzila Farah, Muktadirul Islam, Kazi Tasnim Zinat, Atif Hasan Rahman, and Shamsuzzoha Bayzid. Species Tree Estimation from Gene Trees by Minimizing Deep Coalescence and Maximizing Quartet Consistency: A Comparative Study and the Presence of Pseudo Species Tree Terraces. *Systematic Biology*, 70(6):1213–1231, 04 2021.
- [82] Barbara H Dobrin, Derrick J Zwickl, and Michael J Sanderson. The prevalence of terraced treescapes in analyses of phylogenetic data sets. *BMC Evolutionary Biology*, 18(1):46, 2018.
- [83] Luigi L Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570, 1967.

- [84] Joseph Felsenstein. The number of evolutionary trees. *Systematic Zoology*, 27(1):27–33, 1978.
- [85] Siavash Mirarab. *Novel scalable approaches for multiple sequence alignment and phylogenomic reconstruction*. PhD thesis, University of Texas at Austin, 2015.
- [86] Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):1250463, 2014.
- [87] Sen Song, Liang Liu, Scott V Edwards, and Shaoyuan Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947, 2012.
- [88] Y Chung and C Ané. Comparing two Bayesian methods for gene tree/species tree reconstruction: A simulation with incomplete lineage sorting and horizontal gene transfer. *Syst Biol*, 60(3):261–275, 2011.
- [89] Ylenia Chiari, Vincent Cahais, Nicolas Galtier, and Frédéric Delsuc. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). *BMC Biology*, 10(1):65, 2012.
- [90] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*. John Wiley & Sons, 2001.
- [91] M. A. Nayeem, M. S. Bayzid, A. H. Rahman, R. Shahriyar, and M. S. Rahman. Multiobjective formulation of multiple sequence alignment for phylogeny inference. *IEEE Transactions on Cybernetics*, pages 1–12, 2020.
- [92] Muhammad Ali Nayeem, Md Shamsuzzoha Bayzid, Atif Hasan Rahman, Rifat Shahriyar, and M Sohel Rahman. A ‘phylogeny-aware’ multi-objective optimization approach for computing MSA. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 577–585, 2019.
- [93] Muhammad Ali Nayeem, Md. Shamsuzzoha Bayzid, Sakshar Chakravarty, M. Saifur Rahman, and M. Sohel Rahman. A multi-objective metaheuristic approach for accurate species tree estimation. In *IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020. (to appear).
- [94] HU Jingyang, Yaping ZHANG, and YU Li. Summary of laurasiatheria (mammalia) phylogeny. *Zoological Research*, 33.



- [95] Jan E Janečka, Webb Miller, Thomas H Pringle, Frank Wiens, Annette Zitzmann, Kristofer M Helgen, Mark S Springer, and William J Murphy. Molecular and genomic data identify the closest living relative of primates. *Science*, 318(5851):792–794, 2007.
- [96] Vikas Kumar, Björn M Hallström, and Axel Janke. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PLoS One*, 8(4):e60019, 2013.
- [97] Bastien Boussau, Gergely J Szöllösi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, 2013.
- [98] Tae-Kun Seo. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, 25(5):960–971, 2008.
- [99] Alexandros Stamatakis and Michael Ott. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):3977–3984, 2008.
- [100] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.
- [101] John Gatesy and Mark S Springer. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80:231–266, 2014.

Generated using Postgraduate Thesis L<sup>A</sup>T<sub>E</sub>X Template, Version 1.02. Department of  
Computer Science and Engineering, Bangladesh University of Engineering and  
Technology, Dhaka, Bangladesh.

This thesis was generated on Monday 21<sup>st</sup> March, 2022 at 5:33am.