Summer 7-14-2020

Maximum likelihood estimation of species trees and anomaly zone detection using ranked gene trees

Anastasiia Kim

Anastasiia Kim
Candidate

Mathematics and Statistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

James Degnan , Chairperson

Erik Erhardt

Helen Wearing

Jeffrey Long

Maximum likelihood estimation of species trees and anomaly zone detection using ranked gene trees

by

Anastasiia Kim

B.S., Applied Mathematics, Taras Shevchenko National University of Kyiv, 2012

M.S., Applied Mathematics, Taras Shevchenko National University of Kyiv, 2014

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2020

Dedication

To my boyfriend, sister, grandmother, and parents.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor James Degnan, for all his guidance and support over the past three years. Our numerous insightful discussions about the research kept me motivated. I appreciate all the time that he spent discussing various research ideas with me. His valuable suggestions encouraged me to constantly learn more about statistics and phylogenetics.

I would like to thank my committee members Professor Erik Erhardt, Professor Helen Wearing, and Professor Jeffrey Long for their time and patience. I am also grateful to Professor Noah Rosenberg for providing great suggestions and detailed comments in the process of writing the paper.

I would also like to extend my thanks to the faculty, staff, and my fellow students at the department of the Mathematics and Statistics. Most of this research was supported by National Institute of Health R01 grant GM117590.

Thanks to my parents and grandmother for their love and endless support.

My most heartfelt thanks to my boyfriend Oleksii and my sister Anna. Thanks for everything that helped me get to this day.

Maximum likelihood estimation of species trees and anomaly zone detection using ranked gene trees

by

Anastasiia Kim

B.S., Applied Mathematics, Taras Shevchenko National University of Kyiv, 2012

M.S., Applied Mathematics, Taras Shevchenko National University of Kyiv, 2014

Ph.D., Statistics, University of New Mexico, 2020

ABSTRACT

A phylogenetic tree represents the evolutionary relationships among a set of organisms. Due to several biological processes, the evolutionary histories of the parts of the genome, called gene trees, might not agree with each other, or with the evolution of the species from which the genes were sampled. Such gene trees can be used to reconstruct phylogenetic trees. The multispecies coalescent is a stochastic process that often used to model sources of gene-species tree discordance.

The methods in this dissertation focus on the gene tree topologies with emphasis on ranked gene tree topologies. A ranked tree depicts the order in which nodes appear in the tree together with topological relationships among gene lineages. One challenge that arises during phylogenetic inference is the existence of the *anomaly zones*, the regions of branch-length space in the species tree that can produce *anomalous gene trees*, gene trees that have topologies differing from the species tree topology but are more probable than the gene tree matching the species tree. The recognition of the

possibility of anomalous gene trees motivated development of inference methods that are not affected by anomaly zones. In spite of the many analytic results known about the various types of anomalous gene trees, less is known about how often they arise in practice. In this work, we show how the parameters of a constant-rate birth-death process used to simulate species trees affect the probability that the species tree lies in the anomaly zone. We prove that the probability that a species tree is in an anomaly zone approaches 1 as the number of species and the birth rate go to infinity in a pure birth process. We propose a heuristic approach to infer whether species trees have different types of anomalous gene trees when it is intractable to compute the entire distribution of gene tree topologies. The method proposed cannot have false positives and serves as a quick test to check the presence of anomaly zones.

In this dissertation, we develop the first maximum likelihood method that infers a species tree from the collection of ranked gene tree topologies. We introduce the software *PRANC*, which was initially designed to compute the probabilities of ranked gene tree topologies under the coalescent process and extended to infer a maximum likelihood species tree. *PRANC* has other useful options to work with phylogenetic trees. We propose methods to estimate a starting tree to be able to locate the maximum likelihood species tree quickly. We evaluate the computational and accuracy performance of *PRANC* under different settings. To illustrate the methods proposed, we analyze two experimental studies of skinks and gibbons.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Evolution is a historical natural process by which organisms change over time as a result of changes in heritable physical or behavioral traits (Darwin, 1872). It is likely that all life on Earth evolved from a single common ancestor approximately a billion years ago (Weiss *et al.*, 2016). Evolution plays a central role in all fields of biology. Variation in physical traits can occur among organisms of different species or among organisms of the same species. For example, long necked saddle-backed tortoises lived on an island with a lot of tall plants. On the other hand, short-necked domed tortoises lived in areas with a lot of short plants (Schafer and Krekorian, 1983). Species adapted to their enviroments that led to a genetic change in a population in future generations. Therefore, these tortoises have different evolutionary histories which can be studied by phylogenetic trees.

The main goal of phylogenetics is to reconstruct the evolutionary history of living organisms, and describe this history via *species trees*. Of course, the fossil record can be used but it is incomplete and fragmentary. Since all organisms can be uniquely identified by deoxyribonucleic acid (DNA) (ribonucleic acid (RNA) for some viruses), one can study the evolutionary relationships of organisms by comparing their DNA

(Felsenstein, 1981; Nei, 1987). DNA consists of four types of nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). The main cause of evolution is the mutational changes in these characters due to substitution, insertion, inversion, or deletion of nucleotides (Wakeley, 2009). A phylogenetic tree represents the evolutionary history of a set of species given DNA sequences. Therefore, the phylogeny can be reconstructed from the various sets of genes, segments of the DNA sequence in the genome. Prior to the reconstruction, the species samples are gathered and samples are sequenced. Then, sequences are extracted for each gene and aligned to match each other as closely as possible.

The most intuitive approach to reconstruct the evolutionary relationships among species is to concatenate all DNA sequences contained in genes in a single alignment (Philippe *et al.*, 2005) and thus reduce the variation among genes. In a such way individual gene data might be concatenated into single supermatrix, which is then used for the phylogeny reconstruction (e.g., by maximum likelihood). Due to several reasons, such concatenation of sequences from multiple genes can lead to poor estimates of the true relationships of species (Kubatko and Degnan, 2007; Leaché and Rannala, 2011; Jiang *et al.*, 2019). Moreover, the maximum likelihood estimation of concatenated alignments is statistically inconsistent under the multispecies coalescent model (Roch and Steel, 2015) due to the incorrect assumption that all genes share the same tree. Another approach is to estimate a tree for each gene, called a *gene tree*, and then estimate a phylogeny from the sample of gene trees assuming the latter were estimated accurately.

In this dissertation, most of the methods we will discuss either estimate a phylogeny and the gene trees at the same time, or just estimate a species tree from a given sample of gene trees (Liu *et al.*, 2009a).

1.1 Gene trees and species trees

A phylogeny or phylogenetic tree is a tree that represents the evolutionary relationships of a set of organisms. The tree itself can be viewed as a connected undirected graph without cycles. Each phylogenetic tree has nodes that usually represent populations of species (called taxa), where a single population is all organisms of the same species. The leaves of the tree represent current species. The branches of the tree (i.e. edges) represent the evolutionary time between the species and its offspring.

The branch lengths of the tree can represent various things. They can show the amount of evolutionary time between two consecutive nodes. Branch lengths could also show the amount of mutation that occurred in the evolutionary time between lineages.

A tree can be rooted or unrooted (Figure 1.1). A rooted tree is a tree with the root specified. The root of the tree corresponds to the most recent common ancestor (MRCA) of all the organisms at the leaves of the tree. An unrooted tree has no known or inferred root. Such a tree just describes the relatedness of the leaves. We can observe an ancestral hierarchy in the rooted tree but not in the unrooted tree. We can always obtain an unrooted tree from the rooted tree by omitting the root. However, to infer the root for the unrooted tree is difficult. A common approach is to add an outgroup, a distantly related species, and assume that the root is between the outgroup and rest of the tree. Another approach is to assume a molecular clock, i.e., that the evolutionary rate is constant over time. To safely use a molecular clock rooting, one needs to test for the validity of this assumption. This typically involves calculating a maximum likelihood of two trees, one with the molecular clock enforced and one without the molecular clock enforced and then conducting a likelihood ratio test (Felsenstein, 1983).

Both rooted and unrooted phylogenetic trees can be either bifurcating or mul-

tifurcating. In a bifurcating (binary) tree each species gives rise to exactly two descendants, whereas a multifurcating tree may have more than two descendants at some nodes. The trees can be labeled or unlabeled. In labeled trees all the leaves are considered distinct. An unlabeled tree is equivalent to a labeled tree where every leaf has the same label. In this work, we primarly work with rooted and unrooted labeled ultrametric binary trees where ultrametric means that the leaves are equally distant from the root. Ultrametric trees satisfy the molecular clock assumption.

In phylogenetic studies, gene trees are often used to reconstruct a species tree that describes evolutionary relationships among species. Gene trees that are contained within the branches of the species phylogeny represent the evolutionary histories of the sampled genes. The species tree is treated as a parameter, and gene trees are considered as random variables whose distributions depend on the species tree (Yang, 2014).

The methods in this dissertation involve only topologies of gene trees; however, the species tree is always a rooted binary ultrametric tree with branch lengths specified. The methods use unrooted and rooted gene trees. There are two types of rooted gene trees: ranked and unranked. A ranked gene tree not only accounts for the topology, as an unranked tree does, but also for the order in which lineages join. We write a ranked tree topology as a modified unranked tree topology using the Newick format, in which each clade (a group of organisms that includes a single ancestor and all of its descendants) is represented by a pair of parentheses, and we add a number after each clade to indicate its ranking. For example, the unranked gene tree $((A, B), (C, D))$ produces two ranked gene trees because the most recent ancestral gene of the A and B lineages could be either more or less recent than the most recent ancestral gene of the C and D lineages. Therefore, we count the following ranked gene trees as distinct $((A, B)_2, (C, D)_3)$ and $((A, B)_3, (C, D)_2)$, where the subscript indicates the ranking of the nodes. In the first of these two ranked gene

trees, the (C, D) coalescence, indicated by the largest subscript, is the most recent. In the Newick format, we suppress the labeling of the root node, which has rank 1. The term "gene tree" will be used to refer to the topology of the gene tree (without branch lengths) unless otherwise noted.

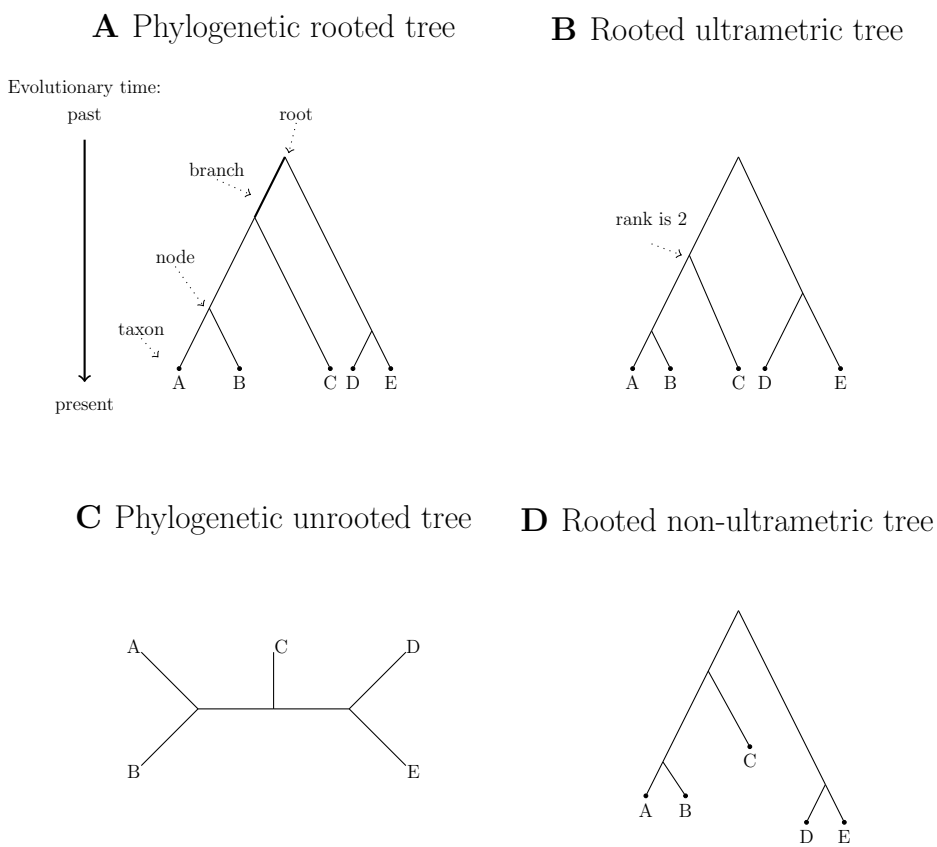


Figure 1.1: (A)–(B) Rooted phylogenetic trees that have the same unranked topology $(((A, B), C), (D, E))$ but different ranked topologies. (A) Tree with the ranked topology $(((A, B)_3, C)_2, (D, E)_4)$. (B) Tree with the ranked topology $(((A, B)_4, C)_2, (D, E)_3)$. (C) An unrooted tree. (D) A tree that does not satisfy the molecular clock assumption that tips are equidistant from the root. The tree in (D) is also called non-ultrametric.

1.2 The coalescent process

It is impossible to trace the history of all gene variants, called *alleles*, because some of them didn't produce descendants in the past and have no trace in the current generation. Because we don't have high quality fossils to study such ghost lineages, we can only study the alleles that have descendants in the current population. A phylogeny represents the history of the lineages that reproduced and have descendants in the present.

We also don't take every allele in the whole population to analyze. Instead, we aim to infer a true evolutionary history about the population from a representative sample of alleles from a set of populations. The history of the sample can be represented as a tree embedded in the larger tree of all the populations. Under the Fisher-Wright model (Fisher, 1930; Wright, 1931), where genetic differences between individuals have no influence on their probability of reproducing (all mutations are selectively neutral), each allele is equally likely to have been passed from one generation to the next. We need to make a clear distinction between census size (actual number of individuals) and effective population size, which is based on an ideal population where all parents have an equal expectation of being the parents of any offspring. Therefore, populations with the same effective population size have similar patterns of genetic variation and genetic drift (random fluctuations in allele frequency in a population over time) as randomly mating populations regardless of census size (Hamilton, 2011).

When DNA is replicated, two gene copies descend from a single common ancestor. Viewed backwards in time from present to past, the process of merging of alleles into a single ancestor called *coalescence*. The theory of coalescence was developed independently by several researchers but mainly attributed to John Kingman (Kingman, 1982a,b; Tajima, 1983; Hudson, 1983).

Let's consider the case of coalescence of any two alleles from a sample of alleles. The question then arises: going backwards in time, how many generations it will take to find their most recent common ancestor (MRCA)? Given a constant effective population size N ($2N$ copies in diploid organisms), the probability that two alleles descended from the same ancestral parent in the previous generation is $1/(2N)$. Therefore, the probability that two alleles came from different parental alleles is $1 - 1/(2N)$. Figure 1.2 illustrates the coalescence process of two alleles. The probability that two alleles coalesce and find their MRCA t generations back in time is

$$P(\text{coalescence occurs at time } t) = P(T_1 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}, \tag{1.1}$$

where T_1 is the coalescent waiting time. Because the first successful coalescence doesn't occur until t generations back, T_1 follows a geometric distribution with mean $2N$. Thus, on average, it will take $2N$ generations for two alleles to coalesce. For example, on average, two randomly choosen alleles from a population of 50 individuals (i.e., 100 gene copies) find their MRCA 100 generations back in time.

Because the first-order Taylor expansion for e^x is $1 + x + o(x^2)$, the geometric distribution can be rewriten as its continuos analog, the exponential distribution. In particular, the cumulative distribution function for a geometric random variable T_1 is written in equation (1.2). As $N \to \infty$, we can expand $e^{-\frac{1}{2N}}$ as $(1 - \frac{1}{2N})$ which is the exponential cumulative distribution function for $t/(2N)$ (eq. (1.3)). Therefore as $N \to \infty$, it follows that for two alleles $T_1/(2N) \xrightarrow{d} T_2$, where T_2 is exponential with mean $2N$.

$$P(T_1 < t) = 1 - \left(1 - \frac{1}{2N}\right)^t, \tag{1.2}$$

$$P(T_2 < t) = 1 - e^{-\frac{t}{2N}}, \tag{1.3}$$

More generally, one can calculate the probability that u lineages coalesce into v lineages. Given that $T = t/(2N)$ is a branch length, the probability that u lineages

coalesce into v lineages ($1 \leq v \leq n$) in the length of time T is (Tavaré, 1984)

$$P_{uv}(T) = \sum_{k=v}^{u} e^{-\binom{k}{2}T} \frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)} \prod_{y=0}^{k-1} \frac{(v+y)(u-y)}{u+y}. \tag{1.4}$$

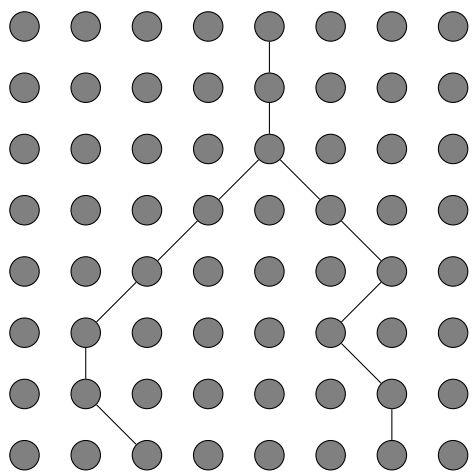


Figure 1.2: Illustrations of a Kingman's coalescent process. It starts from the current generation (bottom) tracing backward in time to the most recent common ancestral (MRCA). Two alleles coalesced at the sixth generation backward time. The distributions of allele frequencies in successive generations follow a Fisher-Wright model.

The coalescent is a stochastic process that considers all genealogies of all possible rooted labeled binary trees with internal nodes ordered in time, with coalescence times $0 < T_i < \infty$ for $2 \leq i \leq n$ lineages. Any particular genealogy will specify the branching pattern of relationships among the members of the sample and the coalescence times (Wakeley, 2009). For n lineages, Kingman (1982a,b) showed that as $N \to \infty$, the coalescence times T_i are independent and exponentially distributed as

$$f_{T_i}(t_i) = \binom{i}{2} e^{-\binom{i}{2}t_i}, \qquad t_i \geq 0,\ i = 2, ..., n \tag{1.5}$$

The coalesent model for a single population was generalized to the Multi-Species Coalescent (MSC) (Rannala and Yang, 2003) model for multiple populations. The MSC

framework allows tracing an evolutionary history of multiple species. Each branch is modeled using one instance of the Kingman coalescent process with a fixed population size. A coalescent history describes where on the species tree the coalescent events occurred (Degnan and Salter, 2005). The probability of a coalescent history is the product of the probabilities of the events on each branch. The probability of a gene tree is the sum of the probabilities of the coalescent histories. For the rest of this dissertation, branch lengths in the species tree are in coalescent units $t/(2N)$, where t is the number of generations and N is the effective population size. Here, 1.0 coalescent unit represents $2N$ generations, where $2N$ is the effective number of gene copies (Felsenstein, 2004).

1.3 Incomplete lineage sorting (ILS)

For a variety of reasons, gene trees may fail to reflect the relationships of the species from which the genes were sampled. This discord can arise from horizontal transfer (including hybridization), lineage sorting, and gene duplication and extinction (Maddison, 1997). Moving backward in time, when two lineages from closely related species enter the parental population, they might or might not coalesce. The probability of coalescence largely depends on the population size and branch length in the species tree. If the lineages don't coalesce in a more recent population, they coalesce at the more ancestral population where other lineages from other species that are less related to these two are also present. Sometimes lineages from these other species have a chance of coalescing with lineages from one of the two closely related species before those two lineages coalesce with each other. This phenomenon of the failure of ancestral copies to coalesce into a common ancestral copy until deeper than previous speciation events is called *deep coalescence*. The *Incomplete Lineage Sorting (ILS)* phenomenon is characterized by regions of DNA that produce gene trees with

different topologies compared to the species tree.

The multispecies coalescent has emerged as a powerful framework that allows modeling sources of gene-species tree discordance due to ILS and deep coalescence (Figure 1.3). Both deep coalescence and ILS more frequently occur when effective population sizes are large and/or branch lengths are short. Many researches do not distinguish between ILS and deep coalescence. Both terms are used to describe the discordance of the gene and species tree topologies.

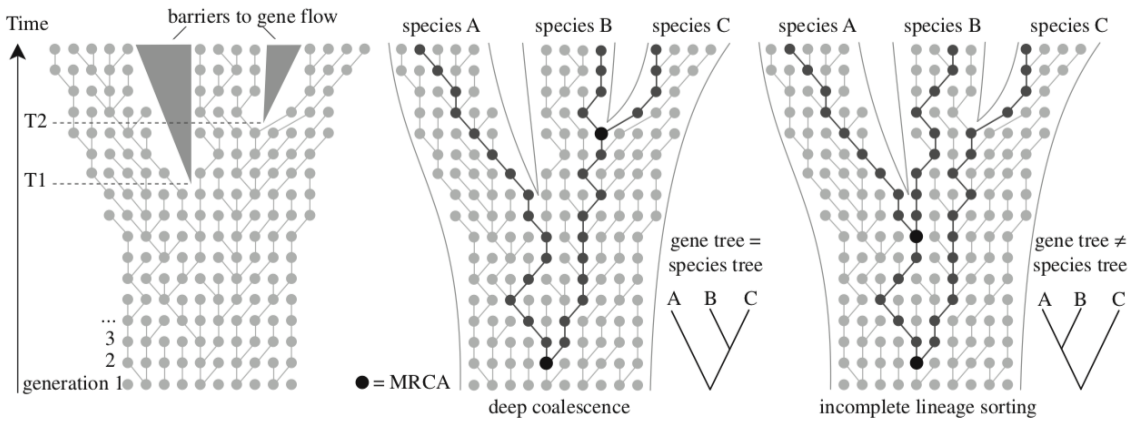


Figure 1.3: Example of a Fisher-Wright coalescent process. The middle and right diagrams show a gene tree is embedded in the species tree. Each dot represents a gene copy and each line connects a gene copy to its ancestor in the previous generation. Starting with a set of individuals in the present generation, each subsequent generation is created by randomly selecting a parent from the previous generation. $T1$ and $T2$ in the leftmost figure represent speciation events. The central figure depicts the event called *deep coalescence* that occurs when the most recent common ancestor (MRCA) of two gene copy samples from species A, B, and C is much older than the speciation event T1. The figure to the right shows a deep coalescence in combination with ILS. ILS causes the discordance in gene and species tree topologies. For example, the gene tree illustrates that among the three species, the lineage sampled from species B is more closely related to the lineage sampled from species A than to the one from species C, even though B is more closely related to C at the species level. The illustration is from Leliaert *et al.* (2014).

If a rooted gene tree and species tree have the same unranked topology, then

we describe the unranked topologies as identical and refer to the unranked gene tree as *matching* the unranked species tree; otherwise, the gene tree topology is *nonmatching*. Similarly, we say the ranked gene tree matches the ranked species tree if, and only if, they have the same ranked topology. At times we will also be interested in cases where a ranked gene tree has the same unranked topology as the species tree, meaning that if the ranks are ignored, the two trees are matching.

1.4 Constant-rate birth-death process

The birth-death model is widely used in the field of biology to model speciation processes and disease transmission (Tanaka *et al.*, 2006; Volkov *et al.*, 2003; Ford *et al.*, 2009). It is a continuous time Markov process with the state transitions of birth or death. The model is characterized by the two parameters, the birth rate $\lambda \geq 0$ and the death rate $\mu \geq 0$. In the birth-death model transitions are allowed only between neighboring states, and it assumes that the birth and death events are independent. The transition from ith to the $(i+1)$th state, indicating that a birth occurs before death, happens with the probability $P_{i,i+1} = \frac{\lambda_i}{\lambda_i+\mu_i}$, whereas the transition from the ith to the $(i-1)$th state, indicating that a death occurs before birth, happens with the probability $P_{i,i-1} = \frac{\mu_i}{\lambda_i+\mu_i}$. The constant birth-death process (Kendall *et al.*, 1948) assumes that at any time point, birth and death can happen at constant rates. In the case when $\mu = 0$, the process is called the pure birth process or Yule process. It is often used to describe the reproduction process of living organisms. In the Yule process, at any time point the rate of birth is proportional to the population size.

In the context of species tree terminology, each tree branch gives birth to a new branch at an exponential rate λ. Lineages can also go extinct at an exponential rate μ. In this model, each species is equally likely to be the next to speciate. In the

Yule model, each branch only gives a birth and never go extinct. When $\lambda = \mu$, the branching process is critical (Popovic, 2004), and we expect each tree branch to have one new branch before going extinct.

In this dissertation, we limited our choices for λ and μ such that the turnover $\frac{\mu}{\lambda} = 0.25, 0.5, 0.75, 1$ with λ primarily in the range of values $0 < \lambda \leq 1$. Cases for the extreme values of λ (up to 30) were also considered. Because the length of a randomly selected interior branch in a Yule tree on n leaves is exponentially distributed with rate 2λ (Stadler and Steel, 2012), for $\lambda = 0.1$ and $\lambda = 1$ a species tree has a mean branch length of $1/(2 \cdot 0.1) = 5$ and $1/(2 \cdot 1) = 0.5$, respectively. Values of λ near 0.5 are chosen to be reasonably plausible for hominid evolution (Stadler *et al.*, 2016). The range of $\lambda = 0.1$ to $\lambda = 1$ thus gives a range of low to moderate levels of incomplete lineage sorting that are plausibly consistent with empirical studies.

1.5 Species tree inference

Numerous methods were developed during the last two decades to estimate a phylogeny. The multispecies coalescent process (Rannala and Yang, 2003; Degnan *et al.*, 2009a) is widely used to infer species trees directly from sequence data or from gene trees, assuming that gene trees were estimated correctly. There are several inference methods that have been proven to be statistically consistent under this model. The majority of the developed methods fall into one of the following categories: Bayesian inference methods, full or pseudo maximum likelihood methods, and heuristic approaches.

Bayesian coalescent-based methods are typically slowest. They perform multiple MCMC runs to estimate a posterior distribution for a species tree. SNAPP is a Bayesian method that infers the species tree directly from sequences (Bryant *et al.*, 2012). However, a few methods were developed that co-estimate gene trees and

species tree simultaneously from the alignments: MrBayes (Huelsenbeck and Ronquist, 2001), BEST (Liu, 2008), *BEAST (Heled and Drummond, 2009), and BPP (Yang, 2015).

Maximum likelihood methods typically search in the tree space to infer the species tree topology and the set of branch lengths by maximizing the likelihood of the gene trees. There are several methods that infer the ML species tree from the set of unrooted or unranked gene tree topologies. For example, STELLS and STELLS2 (Wu, 2012; Pei and Wu, 2017) infer a species tree from the unranked gene tree topologies. $\text{InferNetwork}_{\text{ML}}$ (Yu *et al.*, 2014) can infer the phylogenetic network or a species tree from a sample of unrooted gene trees. These methods use only gene tree topologies for inference ignoring the additional information contained in branch lengths. Several methods were developed that estimate the species tree from the gene trees with branch lengths. For example, STEM (Kubatko *et al.*, 2009) uses the gene trees with branch lengths for the estimation. However, the accuracy of STEM applied to estimated ML gene trees can decrease with the number of gene trees (McCormack *et al.*, 2009; Leaché and Rannala, 2011). Since full ML methods are computationally extensive, some pseudo-likelihood methods were developed. The most popular is MP-EST (Liu *et al.*, 2010) that maximizes the pseudo-likelihood of the triplets in the unranked gene tree topologies to estimate the species tree.

Several methods were developed to further decrease the computational time. For example, NJst (Liu and Yu, 2011) reconstructs the species tree from the distance matrix between pairs of species in the unrooted gene tree topologies. The recent method ASTRAL (Mirarab *et al.*, 2014) became very popular. It minimizes the quartet distance between unrooted gene tree topologies and the species tree. Other methods include STAR (Liu *et al.*, 2009b), which uses only topologies, and STEAC (Liu *et al.*, 2009b), which uses average coalescence times of the unranked gene trees to estimate the species tree.

There is an uncertainty in both the estimated gene tree topologies and their branch lengths. DeGiorgio and Degnan (2014) showed that topology-based methods have similar or better performance as some current methods using coalescence times. Therefore, it might be better to rely not on the exactly estimated branch lengths but on the temporal order of nodes in the gene tree (ranked gene tree topology). Although branch lengths in gene trees are not estimated very accurately, the relative hierarchical order of the nodes might still be estimated more correctly than the branch lengths. Therefore, in this dissertation, we propose the maximum likelihood method that infers a species tree from the collection of ranked gene tree topologies.

1.6 Ranked gene tree probabilities

Let $\mathcal{T}$ be a rooted species tree with branch lengths in coalescent units. Assuming that we have observed a collection of N ranked gene trees $\mathcal{G}_i$s, the maximum likelihood (ML) species tree is (Rannala and Yang, 2003)

$$\mathcal{T}_{ML} = \underset{\mathcal{T}}{\operatorname{argmax}}\, P[\mathcal{G}_1, ..., \mathcal{G}_N|\mathcal{T}] = \underset{\mathcal{T}}{\operatorname{argmax}} \prod_{i=1}^{N} P[\mathcal{G}_i|\mathcal{T}]. \tag{1.6}$$

Here the same set of labels is used for both species and genes, and all gene trees have one gene sampled per species. As shown in eq. (4.1), to get a maximum likelihood estimate, the probabilities of individual gene trees should be calculated first.

For a species tree with n labeled leaves, we assign ranks to the nodes according to their speciation order. Denote the time of the interior node of rank i (ith speciation) by s_i, $i = 1, 2, ..., n-1$. Time is zero for the leaves and increases going backwards in time: $s_1 > s_2 > ... > s_{n-1}$, where s_1 is the time of the root (Figure 1.4). For $i = 2, 3, ..., n-1$, denote the interval between the $(i-1)$th and ith speciation events by τ_i and its length by $t_i = s_{i-1} - s_i$. As discussed in Section 1.1, we write a ranked tree topology as a modified unranked tree topology using the Newick format, in

which each clade is represented by a pair of parentheses, and we add a number after each clade to indicate its ranking. For example, the species tree in Figure 1.4A can be written as $(((A, B)_3, C)_2, (D, E)_4)$. In the Newick format, we supress the labeling of the root node, which has rank 1.

1.6.1 Ranked histories

Let $\mathcal{G}$ be a ranked gene tree topology with the same labels for the leaves as species tree $\mathcal{T}$. Given a gene tree that evolves on a species tree $\mathcal{T}$, a *ranked history* can be defined as a non-decreasing sequence $x = (x_1, x_2, ..., x_{n-1})$, where for $i = 1, 2, ..., n-1$, $x_i = j$ if the ith coalescence occurs in species tree interval $\tau_j = [s_{j-1}, s_j)$ (Degnan *et al.*, 2012b). For example, in Figure 1.4B, the ranked history of the gene tree is $(1, 2, 3, 3, 3)$. One coalescence occurs in the species tree interval τ_1, one in τ_2, and three in τ_3. We denote the probability under the coalescent model of a ranked gene tree topology with the particular ranked history x by $P(\mathcal{G}, x|\mathcal{T})$.

Eventually, we will need to compute the probability of a ranked gene tree for every possible ranked history. To find the set of all possible ranked histories, we need to determine the *maximal ranked history*. The maximal ranked history depicts the sequence of coalescence events, where each coalescence of gene lineages occurs in the most recent possible species tree interval. For the matching ranked gene tree topology with n leaves, the maximal ranked history is always $(1, 2, 3, ..., n-1)$. After determining the maximal ranked history, each next ranked history can be obtained by decreasing the rightmost element by 1 to have a non-descreasing sequence and resetting each element to the right of this element to its possible maximum depicted in the maximal ranked history. For example, the maximal ranked history of the matching gene tree depicted in Figure 1.4A is $(1, 2, 3, 4)$. The maximal ranked history of the nonmatching gene tree depicted in Figure 1.5A is $(1, 2, 3, 3)$. There are eight other possible ranked histories of the nonmatching gene tree depicted in Figure 1.5A.

Three of them are depicted in Figures 1.5B–D, and the five others are $(1,1,2,3)$, $(1,1,2,2)$, $(1,1,1,3)$, $(1,1,1,2)$, and $(1,1,1,1)$. The "last" ranked history in this enumeration is always $(1,1,1,1)$ for any five-taxon gene tree representing that all four coalescences occur in the species tree interval τ_1.

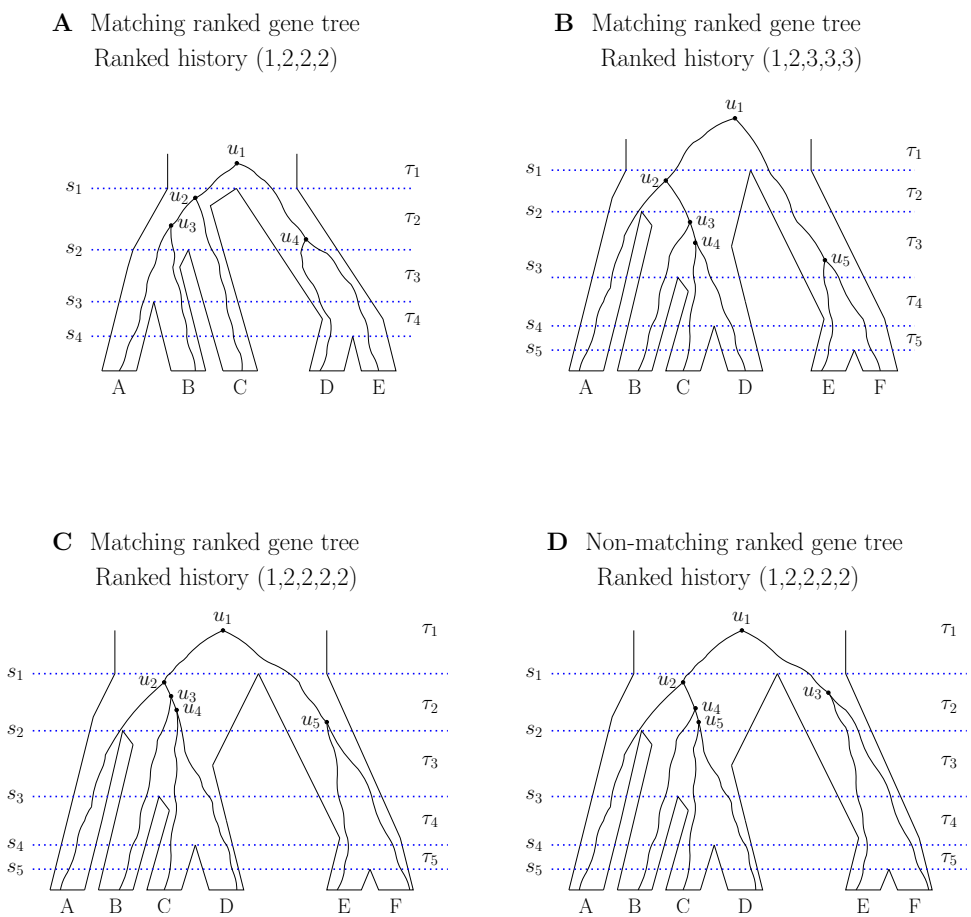


Figure 1.4: Gene trees evolving on five-taxon and six-taxon species trees. (A) Five taxa. (B)–(D) Six taxa. The gene trees in (B)–(D) have the same unranked topology ((A,(B,(C,D))),(E,F)). Only the ranked gene tree topology in (D) does not match the ranked species tree topology. For each $i = 1, 2, ..., n-1$, $s_i \geq 0$ denotes the time of the ith speciation, τ_i represents the interval between the $(i-1)$th and ith speciation events, and u_i represents the ith coalescence (node with rank i) in the gene tree. Interval τ_1 has infinite length.

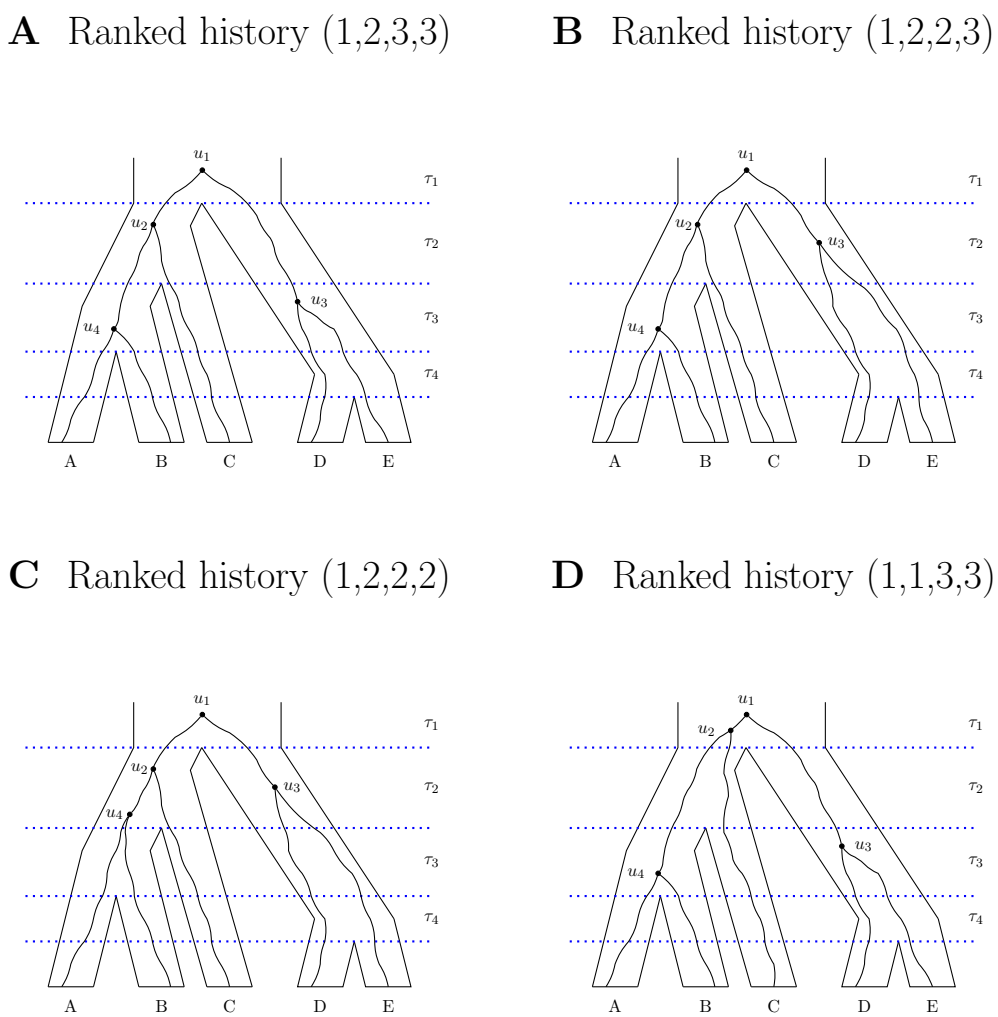


Figure 1.5: (A)–(D) Illustration of the different ranked histories. The same non-matching gene tree $(((A,B)_4,C)_2,(D,E)_3)$ is evolving on the five-taxon species tree $(((A,B)_3,C)_2,(D,E)_4)$. The maximal ranked history is depicted in (A). For each $i = 1, 2, ..., n-1$, τ_i represents the interval between the $(i-1)$th and ith speciation events, and u_i represents the ith coalescence (node with rank i) in the gene tree. Interval τ_1 has infinite length.

1.6.2 Calculating ranked gene tree probabilities

The probability of the ranked gene tree $P(\mathcal{G}|\mathcal{T})$ can be computed as a sum over all ranked histories. Denote the probability in interval τ_i for a particular ranked history

x by $P(\mathcal{G}_{\tau_i}, x|T)$. The probability of a ranked gene tree topology $\mathcal{G}$ with ranked history set $\mathcal{Y}$ given a species tree $\mathcal{T}$ can be written as

$$P\left(\mathcal{G}|\mathcal{T}\right) = \sum_{x \in \mathcal{Y}} H_{\ell_1}(x) \prod_{i=2}^{n-1} P(\mathcal{G}_{\tau_i}, x|\mathcal{T}), \tag{1.7}$$

where $H_{\ell_1}(x)$ is the probability that the coalescences above the root appear in the order that follows the ranked gene tree (Stadler and Degnan, 2012). If the number of lineages above the root is ℓ_1, then (Rosenberg, 2006)

$$H_{\ell_1}(x) = \frac{2^{\ell_1 - 1}}{\ell_1!(\ell_1 - 1)!}. \tag{1.8}$$

Denote the number of lineages available for coalescence in population z just after (going forward in time) the jth coalescence in interval τ_i by $k_{i,j,z}$. For example, there are always i populations in the interval τ_i. The probability that ℓ lineages fail to coalesce in a time interval of length t_i is $e^{-\binom{\ell}{2}t_i}$. Hence, the waiting time until the next coalescent event (going backward in time) has rate $\lambda_{i,j} = \sum_{z=1}^{i} \binom{k_{i,j,z}}{2}$. The density for the coalescent events in the interval τ_i is (Degnan *et al.*, 2012b)

$$f_i(v_0, v_1, ..., v_{m_i}) = \exp\left(-\sum_{j=0}^{m_i} \lambda_{i,j} v_j\right), \tag{1.9}$$

where v_j is the time between the jth and $(j+1)$st coalescent events, with v_0 being the time between s_{i-1} and the least recent coalescent event in τ_i and with v_{m_i} being the time between s_i and coalescent event m_i.

For example, consider the second speciation interval τ_2 for the species tree in Figure 1.4A. Here, v_0 is the time between s_1 and the least recent coalescent event u_2 in interval τ_2. Similarly, v_1 is the time between u_2 and u_3, v_2 is the time between u_3 and u_4, and $v_{m_i} = v_3$ is the time between u_4 and s_2. Using the fact that the sum of exponential random variables with different rates λ_i has a hypoexponential distribution (Ross, 2014), eq. (1.9) can be written as follows (Stadler and Degnan,

2012):

$$P\left(\mathcal{G}_{\tau_i}, x|\mathcal{T}\right) = \int_v f_i(v_0, ..., v_{m_i})dv = \sum_{j=0}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i} (\lambda_{i,k} - \lambda_{i,j})}. \tag{1.10}$$

Therefore,

$$P\left(\mathcal{G}|\mathcal{T}\right) = \sum_{x\in\mathcal{Y}} H_{\ell_1}(x) \prod_{i=2}^{n-1} \sum_{j=0}^{m_i(x)} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i(x)} (\lambda_{i,k} - \lambda_{i,j})}, \tag{1.11}$$

we denoted m_i as $m_i(x)$ in eq. (1.11) to clarify that m_i depends on the ranked history x.

In general, we can uniquely specify an unranked or unrooted gene tree topology by the ranked gene tree topology. The probability of an unranked gene tree topology can be obtained by summing the probabilities of all ranked gene tree topologies that share that unranked topology. Similarly, the probability of an unrooted gene tree topology can be obtained by summing the probabilities of all unranked gene trees with the same unrooted topology.

Examples

Here we provide a few examples of calculation of gene tree probability for some ranked history x. Consider a species tree $\mathcal{T}$ and gene tree with matching ranked topology $((A,(B,(C,D)_4)_3)_2,(E,F)_5)$ (Figure 1.4C). We now calculate the probability of the ranked history $(1,2,2,2,2)$ in interval τ_2. Because four coalescences occur in interval τ_2, $m_2 = 4$ and $k_{2,j,z}$ is defined for $j = 0,1,2,3,4$ and $z = 1,2$. We have $k_{2,j,1} = (1,2,3,4,4)$ for $j = 0,1,...,4$ and $k_{2,j,2} = (1,1,1,1,2)$ for $j = 0,1,...,4$. Using $\lambda_{2,j} = \sum_{z=1}^{2} \binom{k_{2,j,z}}{2}$, we have $\lambda_{2,j} = (0,1,3,6,7)$ for $j = 0,1,...,4$. Thus, eq. (1.10) evaluates

to

$$P\left(\mathcal{G}_{\tau_2},(1,2,2,2,2)|\mathcal{T}\right)=\sum_{j=0}^{4}\frac{e^{-\lambda_{2,j}t_2}}{\prod\limits_{k=0,k\neq j}^{4}(\lambda_{2,k}-\lambda_{2,j})}=\frac{1}{126}-\frac{e^{-t_2}}{60}+\frac{e^{-3t_2}}{72}-\frac{e^{-6t_2}}{90}+\frac{e^{-7t_2}}{168},$$

where $t_2=s_1-s_2$ is the length of interval τ_2.

Similarly, we can compute the probabilities in intervals τ_3, τ_4, τ_5. No coalescence occurs in any of the intervals τ_3, τ_4, or τ_5. Therefore $m_i=0$ for $i=3,4,5$ and $\lambda_{3,0}=\binom{1}{2}+\binom{3}{2}+\binom{2}{2}=4$, $\lambda_{4,0}=\binom{1}{2}+\binom{1}{2}+\binom{2}{2}+\binom{2}{2}=2$, $\lambda_{5,0}=\binom{1}{2}+\binom{1}{2}+\binom{1}{2}+\binom{1}{2}+\binom{2}{2}=1$. The probabilities in these intervals are

$$\begin{aligned}
P\left(\mathcal{G}_{\tau_3},(1,2,2,2,2)|\mathcal{T}\right)&=e^{-\lambda_{3,0}t_3}=e^{-4t_3},\\
P\left(\mathcal{G}_{\tau_4},(1,2,2,2,2)|\mathcal{T}\right)&=e^{-\lambda_{4,0}t_4}=e^{-2t_4},\\
P\left(\mathcal{G}_{\tau_5},(1,2,2,2,2)|\mathcal{T}\right)&=e^{-\lambda_{5,0}t_5}=e^{-t_5}.
\end{aligned}$$

Given that the probability for the coalescence of $\ell_1=2$ lineages above the root appearing in the right order is $H_2=1$ (eq. (1.8)), the probability of the ranked history $(1,2,2,2,2)$ is equal to

$$\begin{aligned}
P\left(\mathcal{G},(1,2,2,2,2)|\mathcal{T}\right)=&H_2(x)\cdot\prod_{i=2}^{5}P\left(\mathcal{G}_{\tau_i},(1,2,2,2,2)|\mathcal{T}\right)\\
=&\left(\frac{1}{126}-\frac{e^{-t_2}}{60}+\frac{e^{-3t_2}}{72}-\frac{e^{-6t_2}}{90}+\frac{e^{-7t_2}}{168}\right)\cdot e^{-4t_3-2t_4-t_5},
\end{aligned}\tag{1.12}$$

where $t_i=s_{i-1}-s_i$.

Now consider a species tree $\mathcal{T}$ and gene tree $\mathcal{G}$ with nonmatching ranked topology $((A,(B,(C,D)_5)_4)_2,(E,F)_3)$ (Figure 1.4D). The values of $k_{i,j,z}$ in interval τ_2 are

$k_{2,j,1}=(1,2,2,3,4)$, $j=0,1,...,4$; $k_{2,j,2}=(1,1,2,2,2)$, $j=0,1,...,4$.

Thus, $\lambda_{2,j} = (0, 1, 2, 4, 7)$ for $j = 0, 1, ..., 4$, and the probability of the nonmatching ranked gene tree for the ranked history $(1, 2, 2, 2, 2)$ is

$$\begin{aligned} P\left(\mathcal{G}, (1,2,2,2,2)|\mathcal{T}\right) =& H_2(x) \cdot \prod_{i=2}^{5} P\left(\mathcal{G}_{\tau_i}, (1,2,2,2,2)|\mathcal{T}\right) \\ =& \left(\frac{1}{56} - \frac{e^{-t_2}}{18} + \frac{e^{-2t_2}}{20} - \frac{e^{-4t_2}}{72} + \frac{e^{-7t_2}}{630}\right) \cdot e^{-4t_3 - 2t_4 - t_5}, \end{aligned} \tag{1.13}$$

where $t_i = s_{i-1} - s_i$.

Chapter 2

Anomaly Zones

In 2006, Degnan and Rosenberg defined the concept of an *anomaly zone*: a subset of branch-length space for the species tree in which the most likely unranked gene tree has a topology differing from the species tree topology. A nonmatching gene tree topology that is more probable than the matching one was termed an *anomalous gene tree* (AGT) (Degnan and Rosenberg, 2006). In general, rooted labeled unranked or ranked gene tree topologies that are more probable than the labeled unranked or ranked gene tree topology matching the species tree are called *anomalous unranked gene trees* or *anomalous ranked gene trees* and are termed AGTs and $ARGT$s (Degnan *et al.*, 2012b), respectively. Similarly, unrooted gene trees that are more probable than the matching unrooted gene tree are termed *anomalous unrooted gene trees* ($AUGT$s) (Degnan, 2013). Species trees that have unrooted, unranked, or ranked anomalous gene trees are said to be in the unrooted, unranked, or ranked *anomaly zone* (AZ), respectively.

Following equations (1.12) and (1.13), the limiting probabilities for the matching and nonmatching ranked gene tree topologies for the ranked history $(1, 2, 2, 2, 2)$ when $t_2 \to \infty$ and $t_3, t_4, t_5 \to 0$ are $\frac{1}{126}$ and $\frac{1}{56}$, respectively. Thus, the ranked history

$(1, 2, 2, 2, 2)$ is more probable for the nonmatching ranked gene tree topology than for the matching ranked history when $t_2 \to \infty$ and $t_3, t_4, t_5 \to 0$. For sufficiently large t_2 and sufficiently small t_3, t_4, t_5, most of the probability of the ranked gene tree topology is concentrated on this ranked history, making the probabilities of the other ranked histories close to 0. Thus, the most probable ranked gene tree topology becomes discordant from the ranked species tree topology, forcing the species tree into the ranked anomaly zone.

An intuitive explanation for the existence of AGTs is that when rankings of coalescences are not taken into account, gene trees that are more symmetric can have more rankings than gene trees that are less symmetric (Degnan and Rosenberg, 2006; Rosenberg, 2013; Xu and Yang, 2016). As an extreme case, a gene tree with only one two-taxon clade, called a *caterpillar*, has only one possible ranking and can never be an AGT (Degnan and Rhodes, 2015).

This explanation leads to a similar question for ranked trees: does the most probable ranked gene tree match the species tree? In the case of four taxa, this turns out to be the case: although caterpillar species trees can have unranked anomalous gene trees, they cannot have ARGTs. However, for five or more taxa, ARGTs do exist (Degnan *et al.*, 2012a,b; Disanto and Rosenberg, 2014). The concept of anomalous gene trees has been further extended to consider AUGTs (Degnan, 2013). The concept of the anomaly zone can even be extended to phylogenetic networks (Zhu *et al.*, 2016). In particular, a gene tree is anomalous if it is more probable than any gene tree displayed by the network. Zhu *et al.* (2016) showed that three-taxon phylogenetic networks do not produce anomalies, but that symmetric phylogenetic networks with four leaves can produce anomalies.

Several properties of anomalous gene trees in different settings are known. In particular, every species tree topology with five or more taxa produces AGTs (Degnan and Rosenberg, 2006; Rosenberg, 2013). The analogous result for unrooted gene

trees is that every species tree topology with seven or more taxa produces AUGTs (Degnan, 2013). Rosenberg and Tao (2008) considered all sets of branch lengths that give rise to five-taxon AGTs. They found that the largest value possible for the smallest branch length in the species tree is greater in the five-taxon case (0.1934 coalescent time units) than in the previously studied case of four taxa (0.1568). This finding raises the question of whether species trees with more taxa are more likely to have AGTs. Studies for ARGTs (Degnan *et al.*, 2012a) showed that neither caterpillar nor pseudocaterpillar species trees have ARGTs, where a *pseudocaterpillar* can be obtained from a caterpillar $(\ldots(((A_1, A_2), A_3), A_4), \ldots A_n)$ by replacing $(((A_1, A_2), A_3), A_4)$ with $((A_1, A_2), (A_3, A_4))$ (Rosenberg, 2007). Strangely enough, although caterpillar gene trees cannot be anomalous unranked gene trees, they can be ARGTs. In addition, Disanto and Rosenberg (2014) showed that as the number of species $n \to \infty$, almost all ranked species trees give rise to ARGTs.

Evolutionary biologists have sometimes wondered how often anomalous gene trees arise in practice (Castillo-Ramírez and González, 2008; Zhaxybayeva *et al.*, 2009; Linkem *et al.*, 2016), because the existence of anomalous gene trees makes the method that chooses the most common gene tree as the estimate of the species tree statistically inconsistent in the anomaly zone. A recent empirical identification of the anomaly zone is for gibbons (Shi and Yang, 2018). In spite of the many analytic results known about the various types of anomalous gene trees, less is known about how often they arise in practice. This question is difficult to answer because it requires some knowledge of the empirical distribution of branch lengths in the species trees.

To study the probability that the species tree lies in an anomaly zone, we examine random species trees generated from a constant-rate birth-death process. The approach we use is to simulate the species tree while computing gene tree probabilities analytically for each simulated species tree. This simulation can help to understand

how often AGTs and ARGTs arise in practice, to the extent that birth-death processes are reasonable models for species trees and that we can understand typical birth-death process parameters. We additionally examine cross sections of anomaly zones to see how much overlap exists for different types of anomaly zones. This analysis shows that for larger trees, a species tree can simultaneously be in unranked and ranked anomaly zones.

We primarily focus on two types of gene trees: unranked and ranked gene trees, sometimes we will consider unrooted trees as well. We can view unrooted, unranked, and ranked gene trees as preserving increasing amounts of information about the underlying rooted trees with full branch length information.

We computed probabilities of ranked and unranked gene trees for species trees with five to eight taxa to find a subset of speciation interval length space in which a species tree has both AGTs and ARGTs. For plots comparing unrooted and unranked anomaly zones, see Degnan (2013).

2.1 Simulation design

We simulated species phylogenies under a constant-rate birth-death model. In this model, each species is equally likely to be the next to speciate. Each tree branch gives birth to a new branch at rate λ. Lineages can also go extinct at rate μ.

As explained in Section 1.4, because the length of a randomly selected interior branch in a Yule (rate λ) tree on n leaves is exponentially distributed with rate 2λ (Stadler and Steel, 2012), for $\lambda = 0.1$ and $\lambda = 1$, a species tree has a mean branch length of 5 and 0.5, respectively. We note that if all branch lengths were 0.5 coalescent units, then the species trees in the simulations would be outside of the unranked anomaly zone. A value of 0.5 coalescent units for an internal branch means

that two lineages have a probability of $1 - \exp(-0.5) \approx 39\%$ of coalescing within that branch, whereas for 5 coalescent units, the probability of coalescence exceeds 99%.

We let the speciation rate λ take the values of $0.1, 0.5$, and 1, and choose the extinction rate μ to depend on λ such that the turnover rate μ/λ is 0 or 0.5. Values of (n, λ, μ) were chosen to examine the effect of the species tree parameters on the existence of anomalous gene trees. For each combination (n, λ, μ), the distributions of unranked and ranked gene tree topologies were computed analytically for each simulated species tree. The probabilities of all possible unranked and ranked topologies were computed using *hybrid-coal* (Zhu and Degnan, 2017) and *PRANC*, respectively, conditional on a species tree generated under a constant-rate birth-death model with parameters (n, λ, μ). The presence of anomalous gene trees was then identified by comparing the analytical probabilities of the matching gene tree topology and the most probable nonmatching gene tree topology.

2.2 Five taxa

Figure 2.1A depicts a five-taxon species tree with interval lengths t_2, t_3, and t_4. The interval lengths determine how long or rapid the speciation was. The ranked topology shown is the only five-taxon species tree topology that possesses ARGTs. For fixed values of $t_4 = 0.05, 0.075, 0.1$, we computed the analytical probabilities of all 105 unranked and all 180 ranked gene tree topologies on a grid with $t_2 \in [0.01, 3]$ and $t_3 \in [0.01, 1]$. The anomaly zones in Figure 2.1B were identified by finding the set of values of t_2, t_3, and t_4 for which at least one nonmatching unranked or ranked gene tree topology has probability exceeding the probability of the corresponding matching gene tree topology. Figures 2.1B–D depict cross-sections of unranked and ranked anomaly zones for the five-taxon species tree in Figure 2.1A. For values of

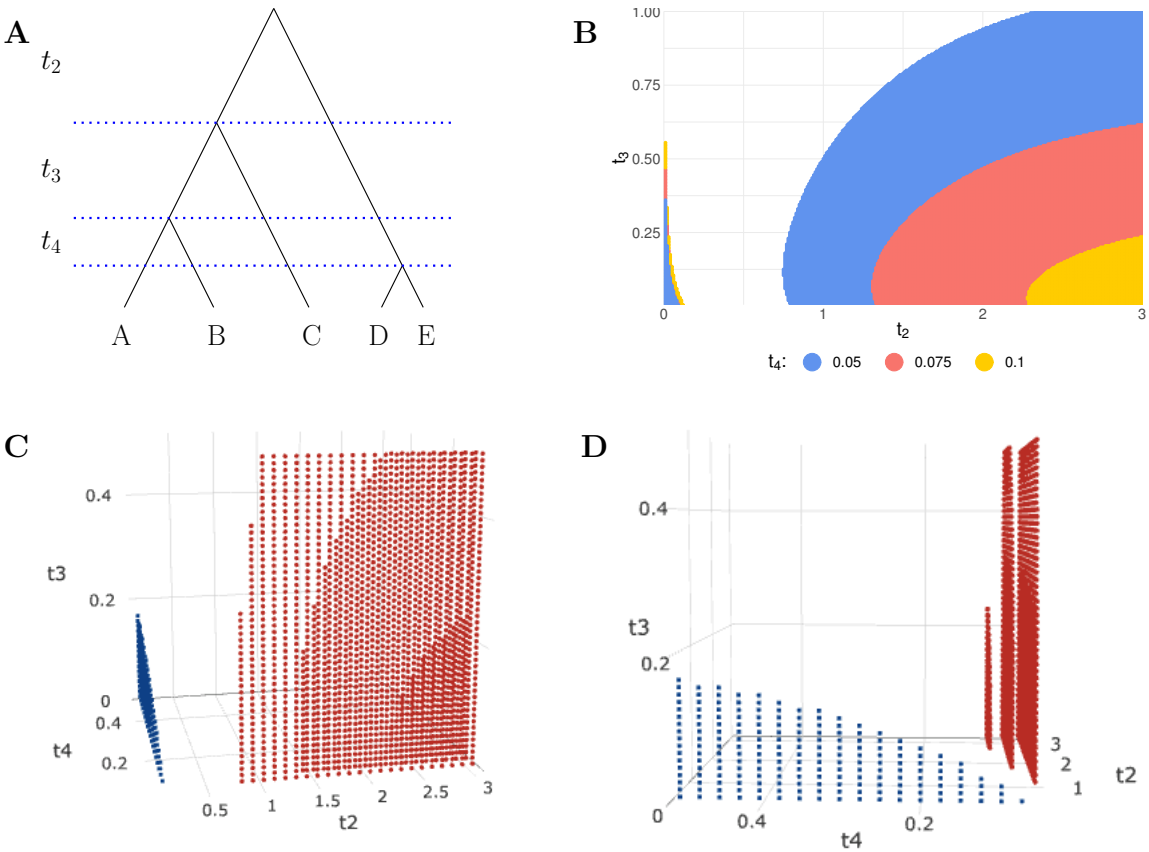


Figure 2.1: Five-taxon anomaly zones. (A) The only ranked five-taxon species tree topology that produces ARGTs. The same species tree, with a gene tree evolving inside, is shown in Figure 1.4A. (B) Slices of the unranked (on the left side) and ranked (on the right side) anomaly zones for the topology in (A). For fixed values of t_4, each shaded region represents pairs of speciation interval lengths (t_2, t_3) for which the most probable unranked (ranked) gene tree topology does not match the unranked (ranked) species tree topology. Each slice was generated by computing the probability distribution of gene tree topologies on a grid with $t_2 \in [0.01, 3]$ and $t_3 \in [0.01, 1]$, with increments of 0.01 for both variables. In the ranked case, the shaded region for a smaller t_4 contains the shaded region for a larger t_4. In the unranked case, the shaded region for a larger t_4 contains the shaded region for a smaller t_4. (C)–(D) Unranked (blue) and ranked (red) anomaly zones for different values of t_is. Each dot was generated by computing the probability distribution of gene tree topologies on a grid with $t_2 \in [0.01, 3]$ with increment of 0.1, $t_3 \in [0.01, 0.5]$ with increment of 0.01, and $t_4 \in [0.05, 0.5]$ with increment of 0.025.

t_2, t_3, and t_4 considered, we observe that the unranked and ranked anomaly zones do not overlap for five-taxon species trees. As t_4 becomes smaller, the ranked anomaly zone increases in size, whereas the size of the unranked anomaly zone decreases. Although for the values of t_i considered, we do not observe an overlap in unranked and ranked anomaly zones in the five-taxon case, these zones start to intersect for larger trees.

For the five-taxon case, it is not necessary to compute analytical probabilities of the entire gene tree distribution to study whether ranked and unranked anomaly zones overlap. As discussed in Degnan *et al.* (2012b), the only unranked topology that could produce anomalous ranked gene trees is the $\mathcal{T}_{RLL}$ tree $(((A,B)_3,C_2),(D,E)_4)$ (Figure 2.1A). There are three possible rankings of the unranked $\mathcal{T}_{RLL}$ tree topology. The 5-taxon species tree produces an anomalous ranked gene tree if for the species tree $\mathcal{T}_{RLL}$, the probability of any of the other two rankings is greater than the probability of the matching gene tree $\mathcal{G}_{RLL}$ (Degnan *et al.*, 2012b). As noted before, the probabilities of the ranked gene tree can be calculated by summing the probabilities of all possible ranked histories. Table 4 in Degnan *et al.* (2012b) presents the probabilities for each ranked history for the nonmatching gene trees $\mathcal{G}_{LRL} : (((A,B)_4,C_2),(D,E)_3)$ and $\mathcal{G}_{LLR} : (((A,B)_4,C_3),(D,E)_2)$. Because the ranked history set for $\mathcal{G}_{LLR}$ is a subset of the ranked histories of $\mathcal{G}_{LRL}$, it is sufficient to consider only the probability of $\mathcal{G}_{LRL}$ gene tree when checking for anomalousness. Therefore, we can write that the species tree $\mathcal{T}_{RLL}$ has ARGTs if

$$P[(((A,B)_4,C_2),(D,E)_3)|\mathcal{T}_{RLL}] > P[(((A,B)_3,C_2),(D,E)_4)|\mathcal{T}_{RLL}]. \tag{2.1}$$

It was shown in Rosenberg and Tao (2008) that only the unranked tree with two leaves on the one side and three on the other one can be AGTs for the species tree topology $(((A,B),C),(D,E))$. The probability of an unranked gene tree can be obtained by summing the probabilities of the ranked gene trees that share this unranked topology. Therefore, the probability of $((A,B),(C,(D,E)))$ gene tree will

be larger than the probability for the other two topologies $((B,C),(A,(D,E)))$ and $((A,C),(B,(D,E)))$. The species tree $\mathcal{T}_{RLL}$ has AGTs if

$$P[((A,B),(C,(D,E)))|\mathcal{T}_{RLL}] > P[(((A,B),C),(D,E))|\mathcal{T}_{RLL}], \tag{2.2}$$

where each of these gene trees has 3 possible rankings. By summing the probabilities of the possible ranking histories for each ranked topology and by denoting the interval lengths $t_i,\ i = 2,3,4$ by $x = e^{-t_2}, y = e^{-t_3}, z = e^{-t_4}$, respectively, for $0 < x, y, z < 1$, the species tree falls in the ranked anomaly zone if eq. (2.3) holds, and in the unranked anomaly zone if eq. (2.4) holds:

$$z\left(1 - \frac{2x}{3} - \frac{y(1+y/4)}{2} + \frac{2xy(x+y) + x^2y^2(x^2/8-1)}{9}\right) + \frac{2(x+y) - xy(1+x^2/6)}{3} - 1 > 0, \tag{2.3}$$

$$x + \frac{2y}{3} - xy\left(\frac{1}{2} + \frac{x^2}{9}\right) - 1 + \frac{xyz}{3}\left(2 - \frac{4y}{3} - \frac{13x}{6} + \frac{xy}{4}\left(\frac{13}{3} + x^2\right)\right) > 0. \tag{2.4}$$

These expressions can be used to determine whether a five-taxon species tree falls in the ranked or unranked anomaly zones instead of computing the gene tree distribution.

Using inequalities (2.3) and (2.4), Figures 2.2A–B depict cross-sections of unranked and ranked anomaly zones for the five-taxon species tree in Figure 2.1A. For the whole range of interval lengths $0 < x, y, z < 1$ (i.e., e^{-t_i}, $i = 2,3,4$) considered, we observe that the unranked and ranked anomaly zones do not overlap as well. In Figure 2.2 each dot was generated by computing the probability distribution of gene tree topologies on a grid of $(0,1)$ with increment of 0.02 for all variables. In general, the smaller increment of 0.005 was used to check whether the five-taxon anomaly zones could overlap. We did not find any of the values X, Y, Z for which the expressions (2.3) and (2.4) hold simultaneously.

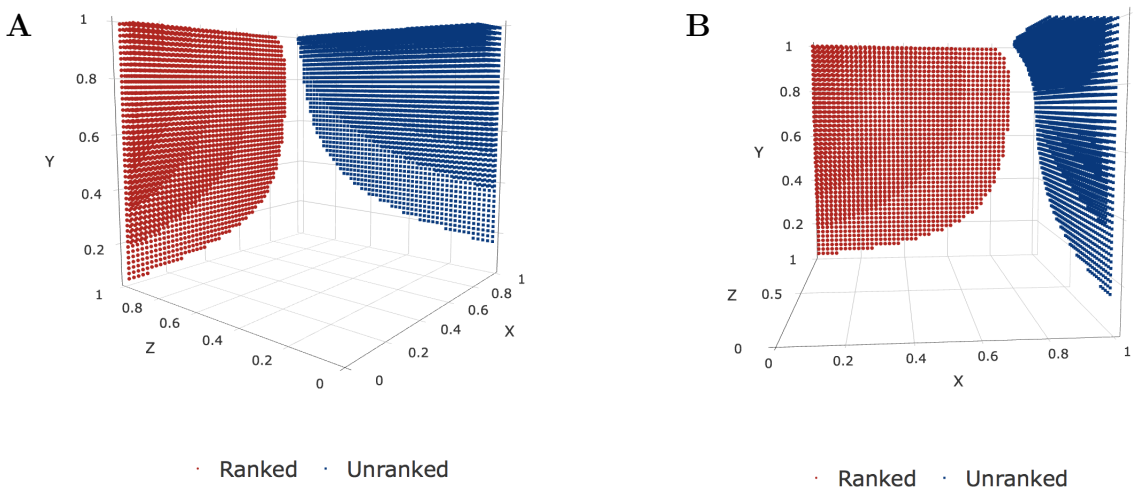


Figure 2.2: (A)–(B) Slices of the unranked (blue) and ranked (red) anomaly zones for the species tree depicted on the Figure 2.1A. Here X, Y, Z are e^{-t_2}, e^{-t_3}, and e^{-t_4}, respectively. Because the possible range for any interval length t_i is $(0, \infty)$, the possible range for e^{-t_i} is $(0, 1)$. Each dot was generated by computing the probability distribution of gene tree topologies on a grid of $(0, 1)$ with increment of 0.02 for all variables.

2.3 Six taxa

We next considered six-taxon trees. There exist six unlabeled tree shapes with six taxa. Excluding the caterpillar and pseudocaterpillar shapes, four of these, depicted in Figure 2.3, give rise to both AGTs and ARGTs. Figure 2.4 shows 2D cross-sections of unranked and ranked anomaly zones for the six-taxon species tree topologies in Figure 2.3. For ease of visualization, we consider only two different values, denoted by S and L, for the lengths of speciation intervals t_2, t_3, t_4, and t_5. For each combination of $S \in [0.005, 1]$ and $L \in [0.01, 2]$, we computed the distributions of unranked and ranked gene tree topologies, and the presence of AGTs and ARGTs was then identified by comparing the analytical probabilities of the matching gene tree topology and the most probable nonmatching gene tree topology.

We examined three interval lengths patterns: $LLSS$, $LSSL$, and $LSSS$ for the

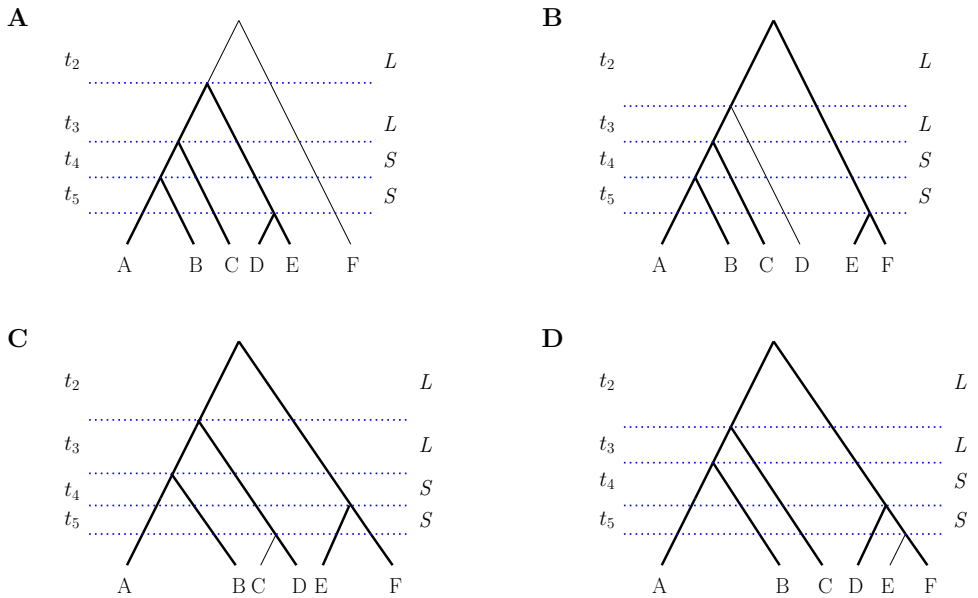


Figure 2.3: Representative labeled rankings of all six-taxon unlabeled species tree topologies, except the caterpillar and pseudocaterpillar. (A)–(D) Bold lines indicate a displayed five-taxon tree topology given in Figure 2.1A. We set some lengths of the speciation intervals to be equal to aid in visualization and computation. Two values L and S, measured in coalescent units, are used as interval lengths. All values of L are equal to each other and all values of S are equal to each other. The figures are not drawn to scale.

species tree topologies with rankings displayed in Figure 2.3. We chose these patterns because a longer branch near the root is more likely to result in ARGTs and two short consecutive intervals are likely to result in unranked anomalous gene trees based on Rosenberg (2013). For the values and three patterns considered, both ranked and unranked anomaly zones (AZs) do not overlap for the species tree topology in Figure 2.3C. The AZs corresponding to the species tree topologies with rankings displayed in Figure 2.3D have only a few values in common. The unranked anomaly zone is not continuous in all corresponding figures (i.e., 2.4J, 2.4K, and 2.4L), there is no unranked anomaly zone for very small $L < 0.1$ and for certain regions in S. For the less balanced species trees depicted in Figure 2.3A and Figure 2.3B, the behavior of the AZs changes drastically for different interval lengths patterns. As shown in

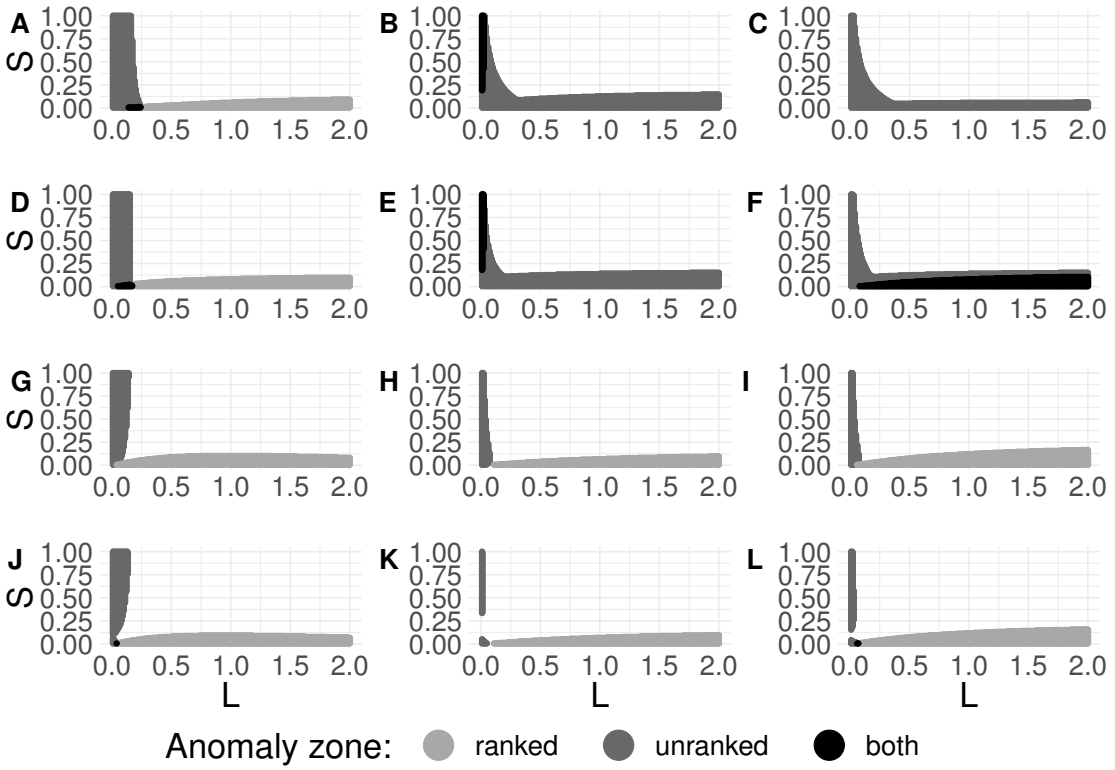


Figure 2.4: Two-dimensional cross-sections of unranked and ranked anomaly zones, each associated with a six-taxon species tree topology in the corresponding panel of Figure 2.3. (A)–(C) Species tree in Figure 2.3A. (D)–(F) Species tree in Figure 2.3B. (G)-(I) Species tree in Figure 2.3C. (J)–(L) Species tree in Figure 2.3D. For each species tree topology, 200 values of $L \in [0.01, 2]$ and 200 values of $S \in [0.005, 1]$ were used to identify the existence of anomalous gene trees. The interval lengths patterns are $LLSS$, $LSSL$, and $LSSS$ from left to right, respectively.

Figures 2.4A and 2.4D, the unranked AZ occurs mostly for small values of L and larger values of S, while the ranked AZ occurs mostly for small values of S and large values of L. For the interval length pattern $LLSS$ this behavior is intuitive and similar to one observed in five-taxon case. In particular, the unranked AZ arises for the two consecutive short intervals on a path from tip to root, and the ranked AZ arises when the interval near the root of the tree is long. However, the anomaly zones behave differently for the pattern $LSSL$ for exactly the same ranked species tree topologies. The unranked AZ occurs more often than for the $LLSS$ patterns

as shown in Figures 2.4B and 2.4E. It is interesting that both AZs can happen for extremely small L and larger S. This might happen because the corresponding species tree topologies in Figures 2.3A and 2.3B are topologically closer to the five-taxon tree $\mathcal{T}_{RLL}$ that has both AGTs and ARGTs. For example, we can move the branch corresponding to the (D, E) clade closer to the root, therefore collapsing the F taxon and increasing the branch corresponding to the C taxon. Such "conversion" for the $LSSL$ pattern to the 5-taxon tree will make the interval closest to the root with length $L + S$ largest and other two intervals will have SL pattern for the next two intervals from past to present. This results in a ranked anomaly zone but it is interesting that an unranked AZ can occur for short L and long S. According to Figure 2.4C, the ranked anomaly zone does not occur for the species tree topology in Figure 2.3A when there are three consectutive intervals with equal lengths from tips to the root. However, the unranked AZ occurs in this case. For the species tree topology in Figure 2.3B, both AZs for the pattern $LSSS$ intersect for large L values and small values of S (Figure 2.4F).

Figures 2.5A–C display ranked and unranked anomaly zones for interval patterns $LLSS$, $LSSL$, and $LSSS$ from left to right, respectively, each associated with a six-taxon species tree topology in Figure 2.3C with ranking $(((A, B)_3, (C, D)_4)_2, (E, F)_5)$. It seems the ranking of the species tree does not greatly affect the anomaly zones for the $LLSS$ and $LSSL$ patterns. However, the ranked anomaly zone disappeared for this ranking and $LSSS$ pattern. Figures 2.5D–F display a scenario when only unranked anomaly zones occur for three interval patterns associated with a topology in Figure 2.3D with different ranking topology $(((A, B)_4, C)_2), ((E, F)_5, D)_3)$.

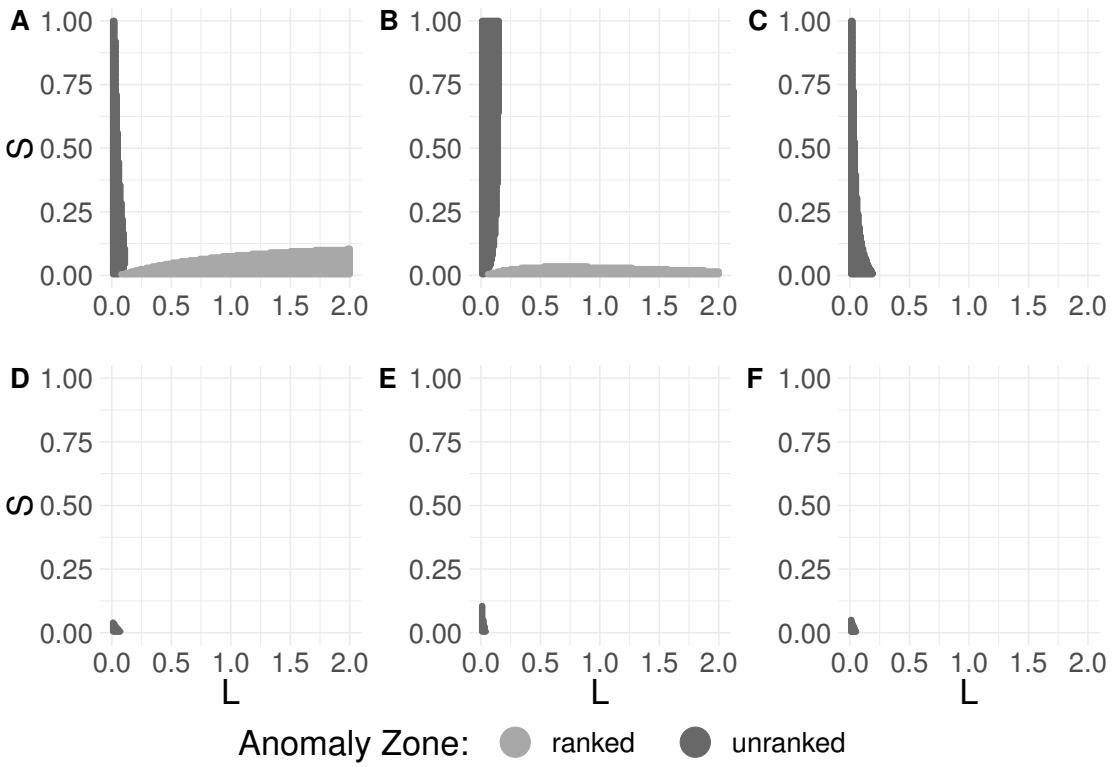


Figure 2.5: Two-dimensional cross-sections of unranked and ranked anomaly zones, each associated with a six-taxon species tree topology in the (C) or (D) panel of Figure 2.3. (A)–(C) Species tree in Figure 2.3C with different ranking topology $(((A,B)_3,(C,D)_4)_2,(E,F)_5)$. The interval patterns are $LLSS$, $LSSL$, and $LSSS$ from left to right, respectively. (D)–(F) Species tree in Figure 2.3D with different ranking topology $(((A,B)_4,C)_2),((E,F)_5,D)_3)$. The interval patterns are $LLSS$, $LSSL$, and $LSSS$ from left to right, respectively. For each species tree topology, 200 values of $L \in [0.01, 2]$ and 200 values of $S \in [0.005, 1]$ were used to identify the existence of anomalous gene trees.

Overall, the intersection of anomaly zones behaves differently for each topology and interval length pattern. The smallest overlap occurs for the more balanced species tree topologies. The anomaly zones are more sensitive for less balanced species tree topologies, in this case we observe large unranked anomaly zones and bigger intersections of two AZs than for the more balanced topologies.

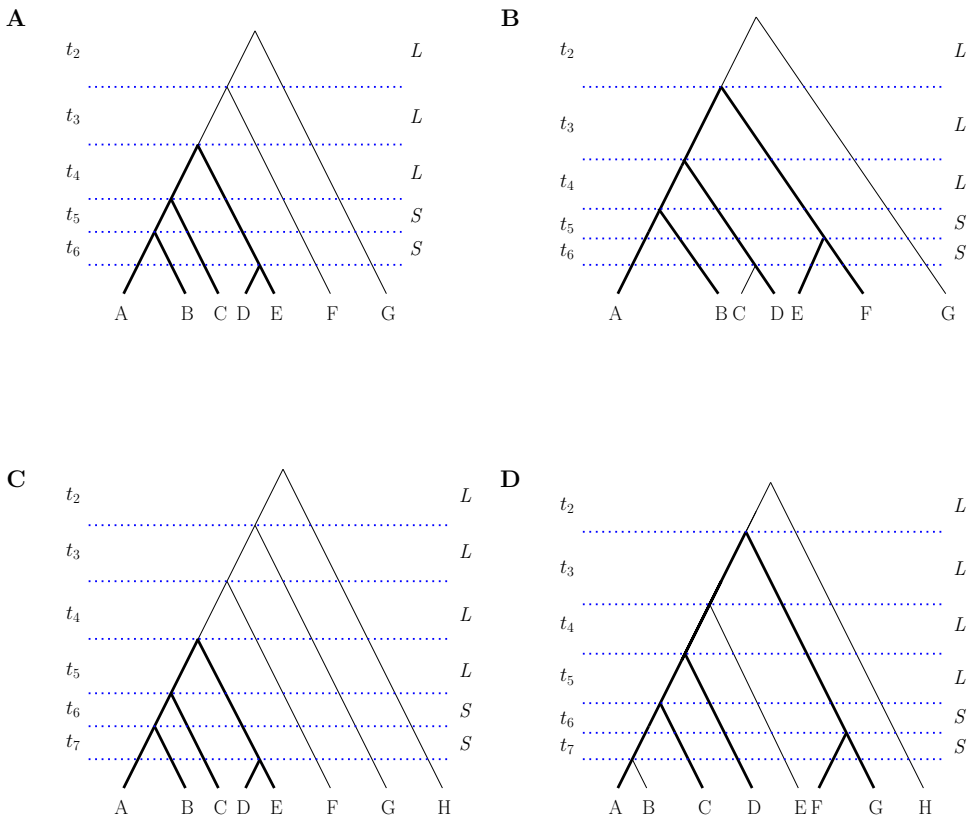


Figure 2.6: Representative labeled rankings of two seven-taxon and two eight-taxon species tree topologies that produce anomalous gene trees. (A)-(B) Seven taxa. (C)-(D) Eight taxa. Bold lines indicate a displayed five-taxon tree topology given in Figure 2.1A. Two values L and S, measured in coalescent units are used as interval lengths. We set some lengths of the speciation intervals to be equal to aid in visualization and computation. Two values L and S, measured in coalescent units, are used as interval lengths. All values of L are equal to each other and all values of S are equal to each other. The figures are not drawn to scale.

2.4 Seven and eight taxa

We next sought to examine scenarios with seven and eight taxa (Figure 2.6) to determine whether the interval-length cases giving rise to AGTs and ARGTs were similar to those seen in the case of six taxa.

The seven- and eight-taxon species trees were chosen so that they produce both

AGTs and ARGTs. To find such topologies, we used a "caterpillarization" technique of finding a short-short-long (SSL) pattern in three consecutive internal branches on a path from a tip to the root of the species tree, and setting all other branches to be long. In Degnan (2013), this technique was used to collapse taxa descended from long branches to be effectively a single taxon, making even a topologically balanced tree resemble a caterpillar when branch lengths are taken into account. More generally, the technique of setting some specific branches to be short and others to be long has been used frequently in identifying AGTs and ARGTs (Degnan and Rosenberg, 2006; Degnan *et al.*, 2009a, 2012b,a; Rosenberg, 2013).

Here we use "caterpillarization" to make seven- and eight-taxon trees resemble the five-taxon ranked tree $(((A, B)_3, C)_2, (D, E)_4)$, the only five-taxon ranked species tree that produces ARGTs. In particular, we consider cases in which a five-taxon species tree topology in Figure 2.1A is contained inside the larger trees. This five-taxon tree appears with bold font in larger tree topologies (Figures 2.3 and 2.6). Because the five-taxon tree in Figure 2.1A produces both AGTs and ARGTs, there exists a subset of branch lengths that makes larger trees also have AGTs and ARGTs simultaneously.

We observe a similar pattern in anomaly zones (Figure 2.7) for species tree topologies displayed in Figures 2.3A, 2.6A, and 2.6C. Each of these topologies was obtained from the five-taxon topology in Figure 2.1A by sequentially attaching an additional branch to the root.

Under the restriction that speciation intervals have one of two lengths, S and L, anomaly zones behave somewhat similarly in the cases of $n = 6, 7$, and 8. In particular, the species tree usually needs to have large values of L and small values of S to be in the ranked anomaly zone. However, the pattern is reversed for AGTs: to produce AGTs, L usually needs to be small whereas S may be relatively large.

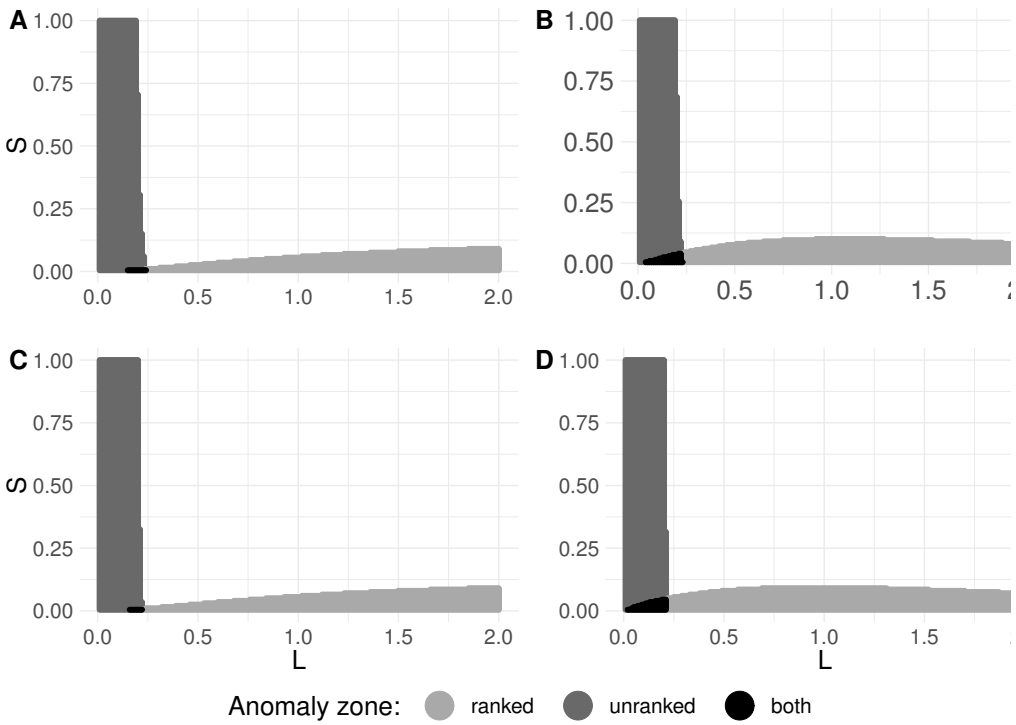


Figure 2.7: Two-dimensional cross-sections of unranked and ranked anomaly zones for associated seven- and eight-taxon species tree topologies in Figure 2.6. (A) Species tree in Figure 2.6A. (B) Species tree in Figure 2.6B. (C) Species tree in Figure 2.6C. (D) Species tree in Figure 2.6D. For each species tree topology, 200 values of $L \in [0.01, 2]$ and 200 values of $S \in [0.005, 1]$ were used to identify the existence of anomalous gene trees.

2.5 Simulation results

We performed simulations to explore the probability that random species trees have AGTs and ARGTs. In particular, we simulated 5000 species trees with $n = 5, 6, 7$, and 8-taxa under a constant-rate birth-death model using the *TreeSim* package in R (Stadler, 2011).

Figure 2.8 shows probabilities of the species tree being in the unranked and ranked anomaly zones in relation to the number of taxa n, speciation rate λ, and extinction rate μ. For both types of trees, the probability of a species tree being in an anomaly

zone increases with the number of taxa and with λ. For unranked trees, both results are intuitive: for increasing numbers of taxa, there are more possible ways to have consecutive short branches or intervals in a tree, a pattern typical of the unranked anomaly zone (Rosenberg, 2013). Increasing λ reduces the average branch length, making consecutive short branches more likely.

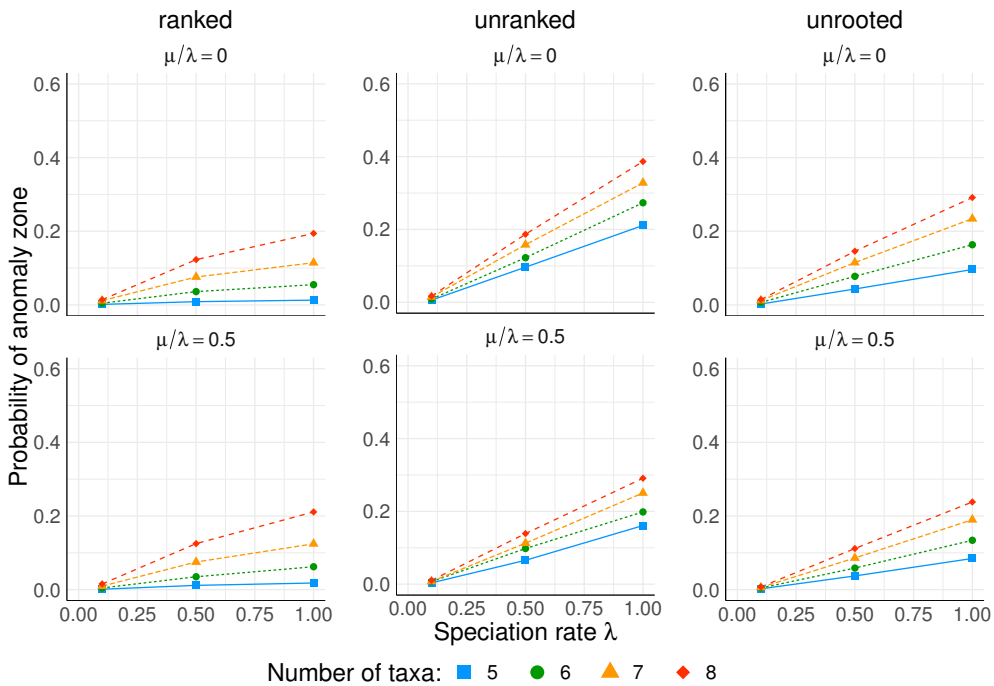


Figure 2.8: The impact of the speciation rate parameter λ and the turnover rate μ/λ on the existence of ranked, unranked, and unrooted anomaly zones. For each value of $n = 5, 6, 7$, and 8 taxa, 5000 species trees were simulated using a constant-rate birth-death process with rates $\lambda = 0.1, 0.5, 1$ and $\mu/\lambda = 0, 0.5$. For each combination of (n, λ, μ), the probability of the species tree being in the anomaly zone was computed from the 5000 trials.

We observed a different effect of the turnover rate μ/λ on the probability of producing unranked and ranked anomalous gene trees. The probability has a decreasing trend for the unranked anomaly zones and an increasing trend for the ranked anomaly zone as turnover rate increases. On average, branch lengths are longer as μ increases. In particular, a branch length near the root becomes longer, decreasing

the probabilities of AGTs but increasing the probabilities of ARGTs.

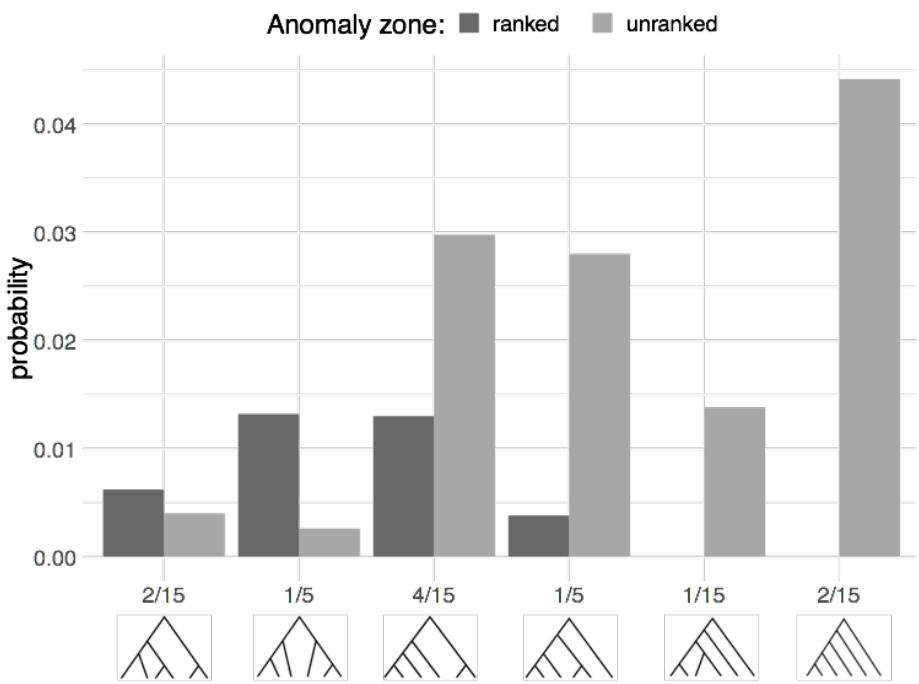


Figure 2.9: Conditional probabilities of ranked and unranked anomaly zones given species tree shape for all possible six-taxon unlabeled, unranked species tree topologies. The exact probabilities of tree shapes under the Yule birth process are displayed on the x-axis. The results are based on 5000 species trees simulated under the birth process with $n = 6$, $\lambda = 0.5$, and $\mu = 0$. Among the shapes with both AGTs and ARGTs, the third tree shape, with four taxa descended from one side of the root and two from the other, produces the largest combined frequency of AGTs and ARGTs. It is also the most probable shape under the birth process. Similar patterns occur for $\lambda = 0.1$ and $\lambda = 1$ (not shown).

We calculated the probabilities of ranked and unranked anomaly zones for specific five- and six-taxon tree topologies ($\lambda = 0.1, 0.5, 1$, $\mu = 0$, 5000 replicates) to investigate the frequency with which the different tree shapes give rise to AGTs and ARGTs. Under the Yule process, the probabilities of a caterpillar shape, pseudo-caterpillar shape, and the unranked version of the tree shape depicted in Figure 2.1A

for the five-taxon case are 1/3, 1/6, and 1/2, respectively. The conditional probabilities of a species tree being in the unranked anomaly zone given the shape are 7.42%, 0.87%, and 2.15% for the three shapes, respectively. Because neither caterpillar nor pseudocaterpillar species trees can produce ARGTs, the conditional probabilities of a species tree being in the ranked anomaly zone given the shape are 0%, 0%, and 0.77% for the three shapes, respectively.

Figure 2.9 shows conditional probabilities of ranked and unranked anomaly zones for all possible six-taxon topologies when $\lambda = 0.5$ and $\mu = 0$. Under the Yule process, the unranked tree shapes have probabilities 2/15, 1/5, 4/15, 1/5, 1/15, and 2/15 from left to right. AGTs arise more often for the caterpillar shape, whereas ARGTs arise more often for the second and third shapes (from left to right). The full probability of anomalous gene trees can be calculated using the law of total probability.

We also noticed that the probabilities of being in the unranked anomaly zone grow faster than those of the ranked anomaly zone as the speciation rate increases (Figure 2.10). For example, the probabilities that a species tree belongs to unranked and ranked anomaly zones are equal to 0.399 and 0.194, respectively, for $n = 8$, $\lambda = 1$, and $\mu = 0$. For an eight-taxon species tree, with $\lambda = 10$ and $\mu = 0$, these probabilities are equal to 0.909 and 0.267, respectively.

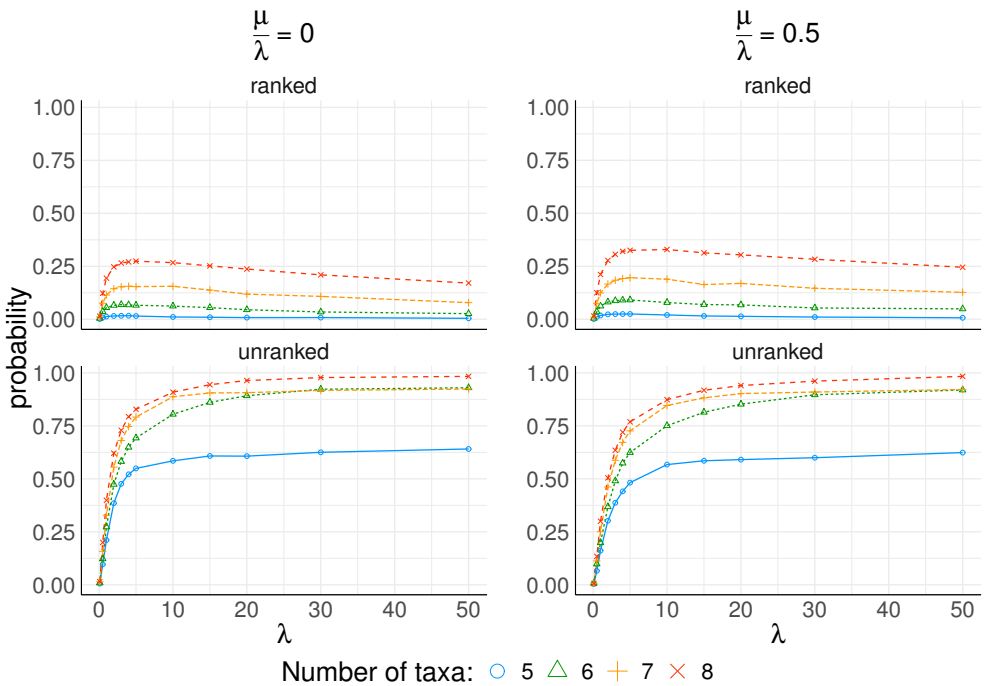


Figure 2.10: The impact of the speciation rate parameter $\lambda \in [0.1, 50]$ and the turnover rate $\mu/\lambda = 0$ and 0.5 on the existence of unranked and ranked anomaly zones. For each combination of (n, λ, μ), the probability of the species tree being in the anomaly zone was computed from 5000 species trees. Probabilities of the unranked anomaly zone appear to increase with λ, whereas probabilities of the ranked anomaly zone increase up to a certain value $\lambda \approx 5$, and then begin to decrease.

2.6 Discussion

The existence of anomalous gene trees poses challenges for inferring species trees from gene trees. We have studied AGTs and ARGTs for small trees, identifying cases in which a species tree possesses both types of anomalies (Figures 2.4, 2.7). We studied how the parameters of the species tree (n, λ, μ) simulated under a constant-rate birth-death process can affect the probability that a species tree is in the anomaly zone. We have shown that often, a species tree has lower probability to be in the ranked anomaly zone than in the unranked anomaly zone (Figures 2.8, 2.10).

We also ran our simulations with larger values of λ, observing that the probabilities of unranked anomaly zones grow faster than those of ranked anomaly zones as the speciation rate increases (Figure 2.10). The probability of a species tree being in the ranked anomaly zone for $n = 8$ reaches a peak near 27.4% and begins to decrease for approximately $\lambda > 5$. Probabilities of a species tree being in the unranked anomaly zone appear to increase with λ, but they are not approaching 1.

An intuitive reason that probabilities do not approach 1 for fixed n is that as λ increases, the probability increases that all coalescences occur more anciently than the root of the tree. This scenario does not always result in anomaly zones. For ranked trees, if the species tree is either a caterpillar or pseudocaterpillar, then there cannot be an ARGT, putting a limit on the probability that the species tree lies in the ranked anomaly zone when n is fixed. In the five-taxon case, ARGTs are more likely when interval τ_2, in which there are two populations (Figure 1.4A), is relatively large compared with other intervals. Increasing λ makes this condition less likely. For unranked species trees, if all coalescences occur above the root, then the species tree has AGTs if, and only if, the species tree does not have a maximally probable shape, where a maximally probable shape is one for which labeled topologies have the maximum number of possible rankings (Degnan and Rosenberg, 2006). For example, for five taxa, the tree $(((A, B), C), (D, E))$ has three rankings. Thus, if the species tree has this topology and all internal branches have length 0, then no other gene tree shape can be anomalous for it. In this case, as $\lambda \to \infty$, an unranked labeled gene tree topology approaches probability $r/180$, where r is the number of rankings for the gene tree.

For six taxa, the unlabeled tree shape whose labeled topologies have the maximum number of rankings has four taxa descended from one side of the root and two from the other side, as shown in Figure 2.3C, where the rooted subtrees on each side of the root themselves maximize the number of possible rankings. This scenario

results in an unlabeled tree with eight rankings and 45 labelings. Because there are 2700 ranked labeled topologies for $n = 6$ taxa, we therefore expect that as $\lambda \to \infty$, the probability of the species tree being in an unranked anomaly zone is at least $1 - (45 \cdot 8)/2700 = 13/15$. This value occurs because labeled unranked trees with this maximally probable shape are tied in probability for being the most probable when all coalescences occur more anciently than the root; as $\lambda \to \infty$, the probability approaches 13/15 that the species tree does not have the maximally probable shape, and therefore is in an unranked anomaly zone.

More generally, let T_n be an unlabeled species tree shape with the maximum number of rankings. For large λ, the probability of the species tree with n leaves being in an unranked anomaly zone has a lower bound of

$$1 - 2^{n-1-\sigma(T_n)} \prod_{i=1}^{n-1} [c_i(T_n) - 1]^{-1}, \tag{2.5}$$

where $\sigma(T_n)$ is the number of balanced internal vertices of T_n and $c_i(T_n)$ is the number of descendant leaves of interior vertex i, including the root as an interior vertex. The lower bound given in eq. (2.5) can be calculated as 1 minus the probability that the species tree under the Yule process has the shape that produces the largest number of rankings for a fixed labeling. For example, the lower bound for six-taxon species trees can be calculated as $1 - 2/15 = 13/15$. This lower bound in eq. (2.5) underestimates the probability of being in an anomaly zone for large λ because even labeled species trees with the maximally probable shape can have AGTs for some sets of branch lengths. It can be shown that this lower bound approaches 1 as $n \to \infty$ (see the next section for details).

In general, probabilities of both AGTs and ARGTs increase with the number of taxa. For example, going from five to eight taxa, the probability of an AGT approximately doubles, for both $\lambda = 0.5$ and $\lambda = 1$ at both levels of turnover. The probability of an ARGT increases by a factor of 10 to 15 going from five to eight

taxa at $\lambda = 0.5$ and $\lambda = 1$ at both levels of turnover (Figure 2.8).

An open question from Degnan *et al.* (2012b) was whether the most probable ARGT could have a different unranked topology from that of the species tree. In that study, examples of ARGTs had different rankings from the species tree but the same unranked topology. Here, in our simulation with different combinations of values (n, λ, μ), we have not found any cases where the most probable ranked gene tree and the species tree have different unranked topologies. However, we found a few cases where a gene tree within one step by nearest neighbor interchange — which has a different unranked topology from the species tree — has exactly the same ranked histories and probability as the ranked gene tree topology that matches the unranked species tree topology.

For example, for a species tree given in Figure 2.11, the two ranked gene trees have the same probabilities, because they have exactly the same values of $k_{i,j,z}$ and thus, the same values of $\lambda_{i,j}$ (see eq. (1.10) for details). The same result that at least one of the most probable ranked gene tree topologies must have the same unranked topology as the species tree was proved mathematically by Disanto *et al.* (2019). This result suggests that the "democratic vote" method used for ranked gene trees might be less misleading than in the unranked setting: if one takes the ranked gene tree (or gene trees, allowing for ties) that occurs most frequently in a large enough sample, then its unranked version is predicted to match the species tree, except possibly when another ranked gene tree is tied for being most probable.

We compute the probabilities of ranked and unranked gene tree topologies for all species trees with five to eight taxa to find a subset of speciation interval length space in which the species tree generates anomalous unranked and ranked gene trees. Studying the properties of anomalous gene trees, as well as examining connections between ranked and unranked anomaly zones, will help to find strategies for solving the problem posed during phylogenetic inference by the existence of anomalous gene

trees.

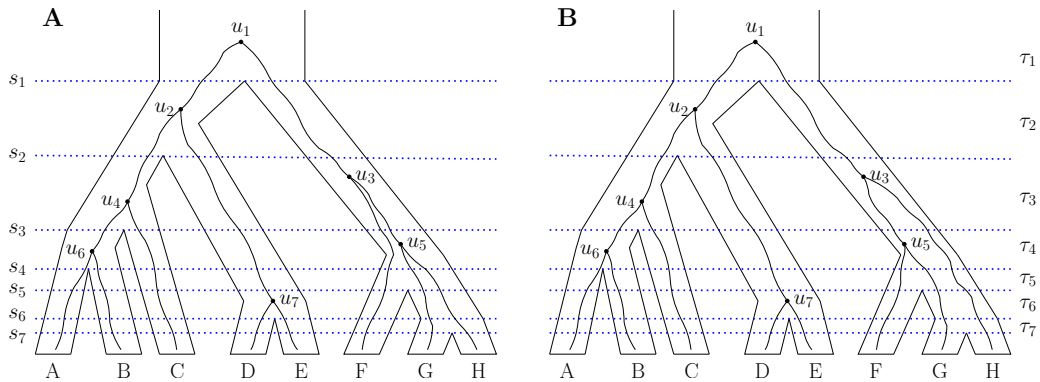


Figure 2.11: Gene trees evolving on an eight-taxon species tree. (A) Ranked gene tree $((((A,B)_6,C)_4,(D,E)_7)_2,((G,H)_5,F)_3)$ that shares the same unranked topology with that of the species tree. (B) Gene tree $((((A,B)_6,C)_4,(D,E)_7)_2,((F,G)_5,H)_3)$ that has a different unranked topology from the species tree. Note that the ranked gene tree $((((A,B)_6,C)_4,(D,E)_7)_2,((F,H)_5,G)_3)$ (not shown) has exactly the same probability as gene trees in (A) and (B) for the species tree depicted. For each $i = 1,2,...,7$, $s_i \geq 0$ denotes the time of the ith speciation, τ_i represents the interval between the $(i-1)$th and ith speciation events, t_i ($t_i = s_{i-1} - s_i$, $2 \leq i \leq 7$) represents the length of interval τ_i, and u_i represents the ith coalescence (node with rank i) in the gene tree. The species tree has ranked topology $((((A,B)_4,C)_3,(D,E)_6)_2,((G,H)_7,F)_5)$. For the species tree values $t_i = (0.29, 0.006, 0.041, 0.001, 0.022, 0.001)$, $i = 2,3,...,7$, the ranked gene trees in (A) and (B) are the most probable ranked gene trees, with probability 1.72404×10^{-5}.

2.7 The lower bound of a probability being in an unranked anomaly zone

The lower bound given in eq. (2.5) denotes the probability of the species tree with n leaves being in an unranked anomaly zone for large λ. In this section we prove that this lower bound approaches 1 as $n \to \infty$ and $\lambda \to \infty$.

Let T_n be a labeled species tree whose unlabeled shape maximizes the number

of rankings of its associated labeled topologies. For large λ, the probability of the species tree with n leaves being in an unranked anomaly zone has a lower bound of

$$1 - \frac{N_R \cdot R}{N_T}, \tag{2.6}$$

where N_R is the number of ways to label the unranked unlabeled tree with the maximum number of rankings, R is the number of rankings, and N_T is the number of ranked topologies for an n-taxon labeled tree.

A given unlabeled tree topology has $R = (n-1)!/\prod_{i=1}^{n-1}(c_i - 1)$ rankings, where c_i is the number of descendant leaves of interior vertex i, including the root as an interior vertex (Steel, 2016, p. 46). There are $N_R = n!2^{-\sigma}$ ways to label the tree with the maximum number of rankings, where σ is the number of balanced internal vertices (Steel, 2016). Because the number of ranked topologies for an n-taxon tree is $N_T = \prod_{i=2}^{n}\binom{i}{2} = n!(n-1)!/2^{n-1}$ (Brown, 1994; Steel, 2016), equation (2.6) leads to the following expression:

$$1 - \frac{n!2^{-\sigma(T_n)} \cdot (n-1)! \prod_{i=1}^{n-1}[c_i(T_n) - 1]^{-1}}{n!(n-1)!/2^{n-1}} = 1 - \frac{2^{n-1-\sigma(T_n)}}{(n-1)\prod_{i=2}^{n-1}[c_i(T_n) - 1]}, \tag{2.7}$$

equivalent to the expression (2.5).

An n-taxon labeled species tree T_n with the maximum number of rankings has $2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ taxa descended from one side of the root and $n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ from the other side (Harding, 1971, 1974; Hammersley and Grimmett, 1974) (Table 2.1). For an n-taxon tree, n must be between two powers of 2. Let $k \geq 0$ be an integer with $2^{k+1} < n \leq 2^{k+2}$. For a tree with the maximum number of rankings, one of the subtrees descended from T_n has at most 2^{k+1} leaves and has the number of leaves a power of 2. In particular, T_n with $2^{k+1} < n \leq 2^{k+2}$ leaves has $2^k \leq 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} \leq 2^{k+1}$ taxa descended from one side of the root and $2^k < n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} \leq 2^{k+1}$ from the other side (Table 2.1 and Figure 2.12). The tree rooted on each side of the

root of T_n itself maximizes the number of possible rankings for all labeled trees with the same number of leaves.

To prove that the lower bound approaches 1 as $n \to \infty$, we need to show that in eq. (2.7), $\prod_{i=2}^{n-1}[c_i(T_n) - 1]^{-1} \to 0$ and $2^{n-1-\sigma(T_n)}(n-1)^{-1} \leq 1$ as $n \to \infty$. We consider three cases: (1) $n = 2^{k+2}$, (2) n odd, and (3) n even and $n \neq 2^{k+2}$.

Consider a case with $n = 2^{k+2}$, $k = 0, 1, \ldots$. A completely balanced symmetric shape is the shape with the maximum number of rankings, with $\sigma(T_n) = n-1$. Thus, for $n = 2^{k+2}$, eq. (2.7) can be written as follows:

$$1 - \prod_{i=1}^{k+1}(2^{k-i+3} - 1)^{-2^{i-1}}. \tag{2.8}$$

The product in eq. (2.8) is the inverse product of the numbers of descendant leaves of all interior vertices, including the root as an interior vertex. That the lower bound for $n = 2^{k+2}$ approaches 1 as $k \to \infty$ is proven by Lemma 1.

Lemma 1: Let $c_i(T_n)$ be the number of descendant leaves of interior vertex i of a tree T_n, excluding the root. Then $\prod_{i=2}^{n-1}[c_i(T_n) - 1]^{-1} \to 0$ as $n \to \infty$.

Proof. Define c_i^* as

$$c_i^* = \begin{cases} 2, & \text{if } i \text{ is a cherry,} \\ 3, & \text{otherwise.} \end{cases}$$

The maximum number of cherries of an n-taxon tree is at most $n/2$. Hence,

$$\prod_{i=2}^{n-1}[c_i(T_n) - 1]^{-1} \leq \prod_{i=2}^{n-1}[c_i^*(T_n) - 1]^{-1} \leq 2^{-(n-2-n/2)} = 2^{-n/2+2},$$

where $n - 2 - n/2$ is the number of internal nodes excluding the root minus the maximum number of cherries. This quantity approaches 0 as $n \to \infty$, completing the proof. $\square$

Table 2.1: The n-taxon species trees with the maximum number of rankings for a labeled topology.

n	(ℓ, r)	n	(ℓ, r)	n	(ℓ, r)	n	(ℓ, r)
2	(1,1)	18	(10,8)	34	(18,16)	50	(32,18)
3	(2,1)	19	(11,8)	35	(19,16)	51	(32,19)
4	(2,2)	20	(12,8)	36	(20,16)	52	(32,20)
5	(3,2)	21	(13,8)	37	(21,16)	53	(32,21)
6	(4,2)	22	(14,8)	38	(22,16)	54	(32,22)
7	(4,3)	23	(15,8)	39	(23,16)	55	(32,23)
8	(4,4)	24	(16,8)	40	(24,16)	56	(32,24)
9	(5,4)	25	(16,9)	41	(25,16)	57	(32,25)
10	(6,4)	26	(16,10)	42	(26,16)	58	(32,26)
11	(7,4)	27	(16,11)	43	(27,16)	59	(32,27)
12	(8,4)	28	(16,12)	44	(28,16)	60	(32,28)
13	(8,5)	29	(16,13)	45	(29,16)	61	(32,29)
14	(8,6)	30	(16,14)	46	(30,16)	62	(32,30)
15	(8,7)	31	(16,15)	47	(31,16)	63	(32,31)
16	(8,8)	32	(16,16)	48	(32,16)	64	(32,32)
17	(9,8)	33	(17,16)	49	(32,17)	65	(33,32)

Note. — The tree with the maximum number of rankings splits into (left, right) subtrees with (ℓ, r) leaves. The n-taxon species tree with the maximum number of rankings T_n has $2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ taxa descended from one side of the root and $n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ from the other side.

For the other two cases, we use a series of lemmas. Table 2.2 depicts the results of Lemmas 2 and 3.

Lemma 2: Let $\sigma(T_n)$ be the number of balanced internal vertices in T_n, the tree with the maximal number of rankings. Then $\sigma(T_n) = n - k - 1$ when n is odd and $2^k < n < 2^{k+1}, k \geq 1$.

Proof. Let $C(k)$ be the statement that for odd n and $2^k < n < 2^{k+1}$, $\sigma(T_n) = n-k-1$. $C(k)$ is true for $k = 1$ as 3-taxon trees have one balanced internal vertex. Now we show that if $C(k)$ is true, then $C(k+1)$ is true for any $k \geq 1$.

We need to show that for odd n, $2^{k+1} < n < 2^{k+2}$, the number of balanced internal vertices is $\sigma(T_n) = n - (k+1) - 1 = n - k - 2$.

Among trees with $2^{k+1} < n < 2^{k+2}$ leaves, let T_n be a tree with the maximal number of rankings. Let $\ell(T_L)$ and $\ell(T_R)$ be the numbers of leaves in the trees rooted at the left and right immediate descendants of the root, respectively. Without loss of generality, let $\ell(T_L) = 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ and $\ell(T_R) = n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$.

T_L is a completely balanced symmetric tree, $\sigma(T_L) = 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - 1$. Because n is odd, T_R has an odd number of leaves with $2^k < n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} < 2^{k+1}$ for $2^{k+1} < n < 2^{k+2}$ (Figure 2.12).

Now, using an induction assumption that $C(k)$ is true, $\sigma(T_n) = \sigma(T_L) + \sigma(T_R) = 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - 1 + (n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - k - 1) = n - k - 2$. $\square$

Lemma 3: Let $\sigma(T_n)$ be the number of balanced internal vertices in T_n, the tree with the maximal number of rankings. Then $\sigma(T_n) \geq n - k - 1$ when n is even and $2^k < n \leq 2^{k+1}$, $k \geq 0$.

Proof. Let $C(k)$ be the statement that for even n and $2^k < n \leq 2^{k+1}$, $\sigma(T_n) \geq n - k - 1$. Obviously, $C(k)$ is true for $k = 0$ as 2-taxon trees have one balanced internal vertex ($\sigma(T_2) \geq 1$). Now we show that if $C(k)$ is true, then $C(k+1)$ is true for any $k \geq 0$.

We need to show that for even n, $2^{k+1} < n \leq 2^{k+2}$, the number of balanced internal vertices is $\sigma(T_n) \geq n - (k+1) - 1 = n - k - 2$.

Among trees with $2^{k+1} < n \leq 2^{k+2}$ leaves, let T_n be a tree with the maximal number of rankings. Let $\ell(T_L)$ and $\ell(T_R)$ be the numbers of leaves in the trees rooted at the left and right immediate descendants of the root, respectively. Without loss of generality, let $\ell(T_L) = 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$ and $\ell(T_R) = n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$.

T_L is a completely balanced symmetric tree, $\sigma(T_L) = 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - 1$. Because n is even, T_R has an even number of leaves with $2^k < n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} \leq 2^{k+1}$ for $2^{k+1} < n \leq 2^{k+2}$ (Figure 2.12).

Now, using an induction assumption that $C(k)$ is true, $\sigma(T_n) = \sigma(T_L) + \sigma(T_R) \geq 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - 1 + (n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} - k - 1) = n - k - 2$. □

Lemma 4: $2^{n-1-\sigma(T_n)}(n-1)^{-1} \leq 1$ as $n \to \infty$.

Proof. From Lemmas 2 and 3, it follows that $\sigma(T_n) \geq n - k - 1$ for $2^k < n \leq 2^{k+1}$ and $\log_2(n) - 1 \leq k < \log_2(n)$.

Consider two cases: $k = \log_2(n) - 1$ and $\log_2(n) - 1 < k < \log_2(n)$. If $k = \log_2(n) - 1$, then $\sigma(T_n) \geq n - \log_2(n)$ and

$$2^{n-1-\sigma(T_n)} \leq 2^{\log_2(n)-1} = 2^{\log_2(n)}/2 = n/2 \leq n - 1.$$

From $\log_2(n) - 1 < k < \log_2(n)$ and the fact that k is an integer, $k = \lfloor \log_2(n) \rfloor$ and $\sigma(T_n) \geq n - 1 - \lfloor \log_2(n) \rfloor$. Then, as $n \to \infty$

$$2^{n-1-\sigma(T_n)} \leq 2^{\lfloor \log_2(n) \rfloor} \leq 2^{\log_2(n-1)} = n - 1.$$

It follows that, as $n \to \infty$,

$$2^{n-1-\sigma(T_n)}(n-1)^{-1} \leq (n-1)/(n-1) = 1.$$

□

Theorem: The lower bound of the probability of the species tree with n leaves being in an unranked anomaly zone, as defined in eq. (2.7), approaches 1 as $n \to \infty$ and $\lambda \to \infty$.

Proof. The result immediately follows by Lemmas 1 and 4 in eq. (2.7). □

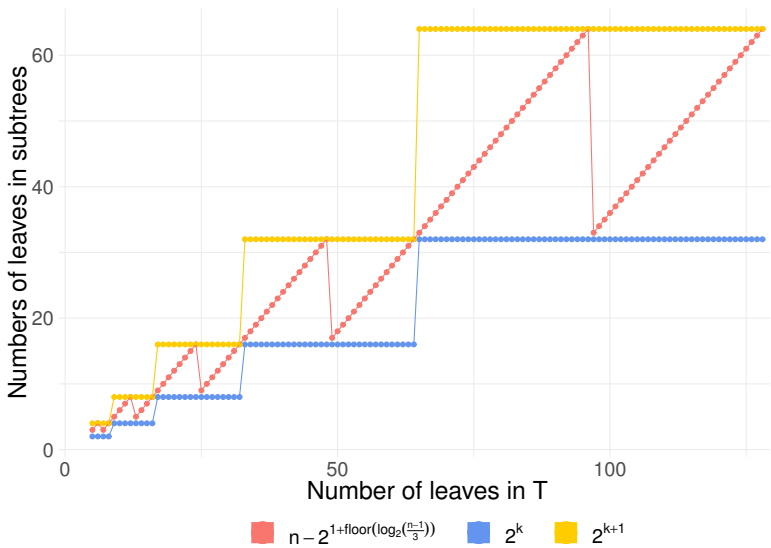


Figure 2.12: The values of $n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor}$, 2^k, and 2^{k+1} for a tree with $2^{k+1} < n \leq 2^{k+2}$ taxa. The tree with the maximum number of rankings has $2^k \leq 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} \leq 2^{k+1}$ taxa descended from one side of the root and $2^k < n - 2^{1+\lfloor \log_2[(n-1)/3] \rfloor} \leq 2^{k+1}$ from the other side.

Table 2.2: The number of balanced internal vertices $\sigma(T_n)$ in n-taxon species trees with the maximum number of rankings for a labeled topology.

n even			n odd		
n	$\sigma(T_n)$	$n-1-\sigma(T_n)$	n	$\sigma(T_n)$	$n-1-\sigma(T_n)$
2	1	0	3	1	1
4	3	0	5	2	2
6	4	1	7	4	2
8	7	0	9	5	3
10	7	2	11	7	3
12	10	1	13	9	3
14	11	2	15	11	3
16	15	0	17	12	4
18	14	3	19	14	4
20	17	2	21	16	4
22	18	3	23	18	4
24	22	1	25	20	4
26	22	3	27	22	4
28	25	2	29	24	4
30	26	3	31	26	4
32	31	0	33	27	5
34	29	4	35	29	5
36	32	3	37	31	5
38	33	4	39	33	5
40	37	2	41	35	5
42	37	4	43	37	5
44	40	3	45	39	5
46	41	4	47	41	5
48	46	1	49	43	5
50	45	4	51	45	5
52	48	3	53	47	5
54	49	4	55	49	5
56	53	2	57	51	5
58	53	4	59	53	5
60	56	3	61	55	5
62	57	4	63	57	5
64	63	0	65	58	6

Note. — For even n, $\sigma(T_n) \geq n-k-1$ (Lemma 3). For completely balanced and symmetric $n = 2^{k+2}$-taxon trees, $\sigma(T_n) = n-1$. For $n = 3 \cdot 2^{\lfloor \log_2(n)-1 \rfloor}$-taxon trees, $\sigma(T_n) = n-2$. For odd n, the number of balanced internal vertices is $\sigma(T_n) = n-1-\lfloor \log_2 n \rfloor$ (Lemma 2).

Chapter 3

Heuristic approaches for detecting anomaly zones

In this chapter, we introduce some heuristic approaches to infer whether species trees are in anomaly zones when it is intractable to compute the entire distribution of gene tree topologies. The number of possible tree topologies grows faster than exponentially with the number of species, make it computationally intensive to infer whether larger species trees (i.e., more than eight taxa) are in anomaly zones by calculating the gene tree distributions.

3.1 Theoretical probability of being in the unranked anomaly zone

Theoretically, the probability that a species tree is an anomaly zone can be obtained by integrating the distribution of species trees under the branching process, using the anomaly zone for limits. In the case of four-taxon unranked trees, the probability

under a pure birth model with rate λ is

$$\frac{1}{3}\int_0^\infty \int_0^{a(x)} 6\lambda^2 e^{-\lambda(2x+3y)}\, dy\, dx,$$

where the integrand is the joint density of the branch lengths (Stadler, 2011), the 1/3 term is the probability that the species tree has a caterpillar topology (i.e., only one two-taxon clade), and

$$a(x) = \log\left[\frac{2}{3} + \frac{3e^{2x} - 2}{18(e^{3x} - e^{2x})}\right] \tag{3.1}$$

is the boundary of the anomaly zone. Here x represents the more basal branch in the caterpillar species tree, and for $y < a(x)$, the species tree is in the anomaly zone (Degnan and Rosenberg, 2006). Even in the four-taxon case, however, the integral appears to not be analytically tractable, and requires either numerical or simulation methods to evaluate. For more species, the dimension of the integral would be $n-2$ since the probability requires integrating over all internal branches in the species tree, making the problem more difficult for larger trees. The boundary of the anomaly zone is also more complicated for larger trees (Rosenberg and Tao, 2008). Consequently, we have used the simulation approach.

3.2 Simulation design

Our simulation approach consisted of the following steps (1) generation of species trees under a constant rate birth-death model using *TreeSim* (Stadler, 2011), (2) computation of probabilities of gene trees for each species tree, (3) identification of the presence of AGTs by comparing the probability of the matching gene tree topology to that of the most probable nonmatching gene tree topology, (4) calculation of the proportion of species trees falling in anomaly zone. These are the same steps as in Chapter 2 except that in step (2), the probabilities of only a small subset of all possible gene trees are computed.

When generating species trees, we let the parameters take values in a biologically plausible range. In particular, the speciation rate λ takes the values of $0.1, 0.5, 1$, and the extinction rate μ depends on λ such that the turnover rate $\frac{\mu}{\lambda}$ is 0 or 0.5. This range of values of (λ, μ) was choosen to observe a moderate difference between gene and species trees topologies that allows examining the effect of the species tree parameters on the existence of AGTs.

Because all three types of AGTs can exist simultaneously for $n \geq 5$-taxon trees, we computed the distributions of gene tree topologies for the 5000 5-, 6-, 7-, and 8-taxon species trees. The results based on the $n \leq 8$–taxon trees in Figure 3.1 are the same as in Kim *et al.* (2019) because same species trees were used for these cases.

We proposed some heuristic methods for larger phylogenies and demonstrated their performance on 1000 simulated species trees with 9, 10, 11, and 12 taxa. Using *hybrid-coal* (Zhu and Degnan, 2017), probabilities of unranked gene tree topologies were computed, and from these, probabilities of unrooted topologies for $n \leq 8$ were found by summing probabilities of unranked topologies that share the same unrooted topology. We used *CalGTProb* command from *PhyloNet* (Yu *et al.*, 2012a) to compute probabilities of unrooted gene trees topologies for $n > 8$. The probabilities of ranked gene tree topologies were computed using *PRANC* (Kim *et al.*, 2019).

3.3 Proposed approaches

It is necessary to propose some heuristic approaches for nine and more taxa since computing probabilities of gene trees given a species tree for the entire distribution is computationally intensive. The number of tree topologies grows faster than exponentially with the number of species. For instance, there are 56,700, 1,587,600, and 57,153,600 ranked gene trees for 7, 8, and 9 species, respectively.

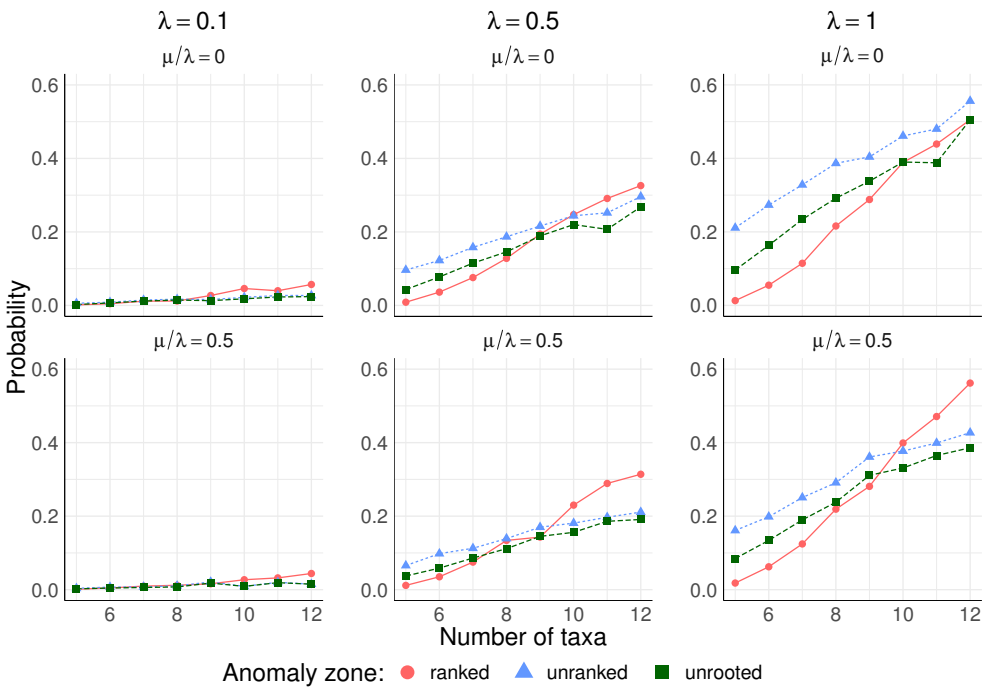


Figure 3.1: The impact of the number of taxa n on the existence of ranked, unranked, and unrooted anomaly zones given speciation λ and extinction μ rates. We simulated 5000 species trees for $n = 5, 6, 7, 8$ taxa and 1000 species trees for $n = 9, 10, 11, 12$ taxa using a constant rate birth-death process with rates $\lambda \in \{0.1, 0.5, 1\}$ and $\frac{\mu}{\lambda} \in \{0, 0.5\lambda\}$. For each combination of (n, λ, μ) the probabilities of the species tree being in each type of μ/λ anomaly zone were computed.

To see how topologically different ARGTs, AGTs, and AUGTs can be from species trees we consider the Robinson-Foulds (RF)(Robinson and Foulds, 1981) topological distance between the most probable gene tree topology and the species tree topology for different values of the speciation rate λ, extinction rate μ, and the number of taxa n (Table 3.1). The RF distance metric between two rooted trees is defined to be a number of distinct clades between two trees. The Robinson-Foulds distance can take only even integer values. The RF distance of zero between two rooted trees indicates that the trees have the same unranked labeled topology. The maximal RF distance between two rooted n-taxon trees is $2n-4$. For example, the RF distance between two 6-taxon trees can take a value of $0, 2, 4$, or 8. Larger values indicate more

topologically distant trees. The RF distances between pairs of trees in Figures 2.3A and B, Figures 2.3A and C, and Figures 2.3A and D are $4, 6$, and 4, respectively. For example, the symmetric difference between sets of clades between two trees in Figures 2.3A and D consists of four clades $\{(D, E), (((A, B), C), (D, E)), (E, F), ((D, E), F)\}$ that results in the RF distance of 4.

Table 3.1: Frequency distribution of the Robinson-Foulds distances between the most probable gene tree topology and the species tree topology. We simulated 5000 species trees for each combination of n, λ, and μ.

		$\frac{\mu}{\lambda} = 0$						$\frac{\mu}{\lambda} = 0.5$					
		unrooted			unranked			unrooted			unranked		
λ	n	0	2	4-8	0	2	4-10	0	2	4-8	0	2	4-8
0.1	5	99.76	0.24		99.42	0.58		99.78	0.22		99.64	0.36	
	6	99.32	0.68		99.12	0.86	0.02	99.56	0.44		99.30	0.70	
	7	98.80	1.08	0.12	98.54	1.32	0.14	99.40	0.58	0.02	99.24	0.72	0.04
	8	98.54	1.44	0.02	98.30	1.68	0.02	99.26	0.68	0.06	98.96	0.98	0.06
0.5	5	95.68	4.32		90.36	9.08	0.56	96.28	3.72		93.44	6.26	0.30
	6	92.24	7.36	0.40	87.76	11.14	1.10	94.14	5.60	0.26	90.92	8.40	0.68
	7	88.50	10.48	1.02	84.20	13.78	2.02	91.40	7.98	0.62	88.74	10.30	0.96
	8	85.42	13.12	1.46	81.32	16.30	2.38	88.84	10.24	0.92	86.10	12.50	1.40
1	5	90.38	9.62		78.90	18.54	2.56	91.54	8.46		83.90	14.42	1.68
	6	83.66	14.88	1.46	72.68	22.26	5.06	86.60	12.58	0.82	80.16	16.86	2.98
	7	76.60	19.78	3.62	67.20	25.46	7.34	81.02	16.50	2.48	74.94	20.44	4.62
	8	70.84	23.82	5.34	61.34	29.18	9.48	76.20	19.70	4.10	70.90	22.74	6.36

We observed that in the cases where a species tree produces AGTs or AUGTs, the majority of most probable gene tree topologies were not too distant from the species tree topology (RF-distance ≤ 2). In both unranked and unrooted cases for larger turnover μ/λ, the probabilities of anomalous trees are lower but there are more trees within RF ≤ 2 from a species tree topology. To reduce the computational complexity of determining anomaly zones, we consider only gene trees that are exactly one nearest neighbor interchange (NNI) away from the species tree. The NNI distance defined as the minimum number of nearest neighbor interchange moves that must be applied to one tree such that it becomes topologically equal to the other. One NNI move swaps two subtrees along a certain internal edge. The rightmost tree in Figure

3.2 is one NNI move away from the central tree but two NNI moves away from the leftmost tree. The central tree in Figure 3.2 is one NNI move away from the leftmost tree: the branch that connects the parent of species C and the root was fixed and two clades C and (D, E) were swapped.

Even if the most frequent gene tree topology is farther than one NNI step away from the species tree topology, it is likely that there is at least one other gene tree topology within one NNI from the species tree topology that has larger probability than the matching tree topology.

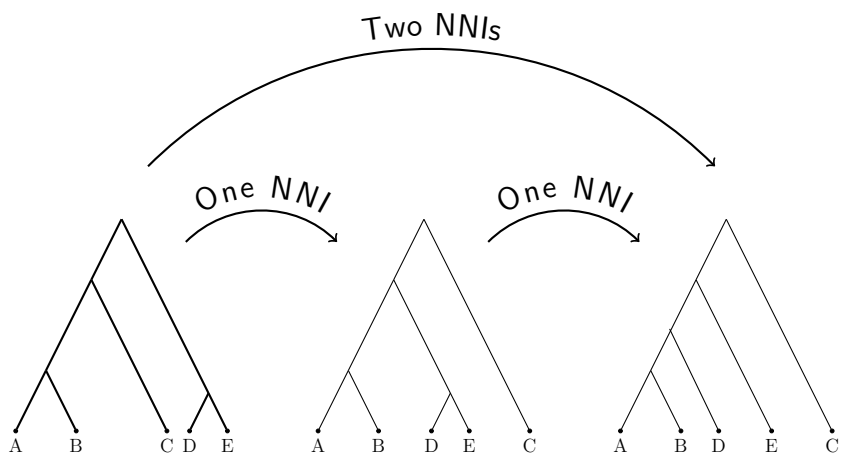


Figure 3.2: Central and rightmost six-taxon rooted trees within one and two NNI moves from the leftmost tree.

The above observations lead to a useful heuristic for unranked gene trees. Given a species tree, we compute probabilities of unranked gene trees that are exactly one NNI of the species trees. If one of these gene trees is anomalous, then the species tree is classified as being in an unranked anomaly zone. If none of the NNI unranked gene trees is in an unranked AZ, then the species tree is classified as not anomalous. Under this heuristic, false positives (judging a species tree to be in an AZ when it is not) cannot occur, but false negatives can occur when a specie tree has AGTs but all AGTs are more than one NNI move from the species tree. For n taxa, the heuristic

requires only computing $2n-4+1$ unranked gene tree probabilities (the plus 1 is for the matching tree) for each species tree. The heuristic for unrooted trees is the same as that for unranked trees except that $2n-6+1$ unrooted NNI trees (plus 1 for the matching tree) are used. We note that the heuristic will tend to underestimate the probability that the species tree is in an AZ.

Table 3.2 shows true positive rates of unrooted and unranked species trees that fall in their respective anomaly zones by computing probabilities of all gene tree topologies that are only one NNI step away from a species tree. Since we have found only a few cases in which all anomalous gene trees were more than two NNI steps away from the species tree topology, we propose that considering unranked or unrooted gene tree topologies within one NNI step from the species tree topology is a reasonable heuristic to infer the existence of AGTs or AUGTs.

Table 3.2: True positive rates of species trees that fall in the unrooted and unranked anomaly zones. We simulated 5000 species trees for each combination of n, λ, and μ.

		$\lambda(\mu=0)$			$\lambda(\frac{\mu}{\lambda}=0.5)$		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
unrooted	5	100.0	100.0	100.0	100.0	100.0	100.0
	6	100.0	100.0	99.88	100.0	99.66	100.0
	7	100.0	99.83	100.0	100.0	100.0	99.68
	8	100.0	100.0	99.73	100.0	99.82	99.75
unranked	5	100.0	100.0	99.91	100.0	100.0	100.0
	6	100.0	99.67	99.27	100.0	99.56	100.0
	7	100.0	99.75	99.76	100.0	100.0	99.76
	8	100.0	99.89	99.84	100.0	100.0	99.73

We use a different strategy in the search for ARGTs. Disanto *et al.* (2019) proved mathematically that at least one of the most probable ranked gene tree topologies must have the same unranked topology as the species tree (it is possible that several

conflicting trees are exactly tied for most probable). Based on this result, we propose to use only those ranked gene trees that have topologies that match the unranked species tree topology to check for anomalousness. This is an exact test with no false positives and no false negatives. This greatly reduces the number of tree probabilities to be computed when checking whether the species tree has an ARGT. For example, instead of computing 1,587,600 probabilities for the balanced 8-taxon tree, we need to compute only 80 probabilities, since there are 80 possible rankings for the balanced 8-taxon topology.

3.4 Limit of the anomaly zone

Another heuristic test was proposed to identify the unranked anomaly zone in larger trees by Linkem *et al.* (2016). They considered the limit of the anomaly zone $a(x)$ for the four-taxon caterpillar tree (Figure 3.3A) defined earlier (eq. (3.1)). Given that y and x are lengths of an internal branch and its immediate ancestor in the species tree, a four-taxon species tree falls in the unranked anomaly zone if it satisfies the condition $y < a(x)$. Linkem *et al.* (2016) proposed that this condition could be checked for any two consecutive branches in a tree within a larger species tree to conclude whether there is evidence of the unranked anomaly zone. In particular, their examination shows that several pairs of parent-child internodes in the skink phylogeny satisfy this condition, and that this species tree is therefore in the anomaly zone. This may explain a strong conflict between species trees inferred under the coalescent versus using the more traditional approach of concatenating multiple gene sequences and inferring a single tree.

We tested the Linkem *et al.* (2016) heuristic on small trees to estimate false positive and false negative rates. We then applied this approach on 5–8-taxon species trees since the exact error can be computed by analytically computing probabilities

Figure 3.3: (A) Using the Linkem *et al.* (2016) heuristic, the four-taxon species tree is said to be in unranked anomaly zone if two consecutive branches with lengths x and y in coalescent units, satisfy the anomaly zone condition $y < a(x)$. (B) Pairs of two internal consecutive branches (i.e., (b_1, b_2), (b_1, b_3), and (b_2, b_4)) in larger tree that can be checked for anomaly zone condition $y < a(x)$. If at least one pair satisfy condition, then the species tree is likely to be in an unranked anomaly zone.

for the entire gene tree distributions. We simulated 5000 species trees under a constant rate birth-death process for each combination of $\lambda = 0.1, 0.5, 1$ and $\mu/\lambda = 0, 0.5$. For each of these species trees, we calculated the probabilities of all possible unranked gene tree topologies and compared a species tree topology with the most probable gene tree topology to see whether a corresponding species tree fell in the unranked anomaly zone.

All pairs of consecutive internode branch lengths were used to check if at least one pair satisfied the anomaly zone limit condition $y < a(x)$ (Figure 3.3B). If there was evidence of the unranked anomaly zone based on this condition, the species tree was checked if it was in the unranked anomaly zone based on computing probabilities for the full gene tree distribution.

Table 3.3 depicts the percentages of species trees that were correctly identified (true positives) to be in the unranked anomaly zone. Table 3.3 also shows false positive percentages of trees that satisfy the anomaly zone condition $y < a(x)$ but are not in the anomaly zone.

Table 3.3: Percentages of species trees that were correctly identified to be in the anomaly zone by satisfying the anomaly zone limit condition $y < a(x)$, where y and x are branch lengths of an internal node and its parental node in the species tree. All consecutive pairs of internode branch lengths were used to check if at least one pair satisfying the anomaly zone limit condition $y < a(x)$. The table also depicts percentages of species trees that were incorrectly identified to be in the anomaly zone. There were 5000 species trees simulated for each combination of n, λ, and μ. The probabilities that species trees lie in unranked anomaly zones were computed based on the full gene tree distribution (true cases).

		$\lambda(\mu = 0)$			$\lambda(\frac{\mu}{\lambda} = 0.5)$		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
True positive	5	96.55	94.19	92.80	100.00	95.43	94.29
	6	97.73	95.26	94.51	100.00	93.83	94.46
	7	93.15	95.95	95.24	97.37	95.74	96.01
	8	92.94	95.61	95.45	98.08	93.96	97.32
False positive	5	0.00	2.16	7.06	0.06	1.52	4.76
	6	0.18	4.08	9.72	0.06	2.26	7.00
	7	0.12	4.60	11.76	0.02	3.44	9.28
	8	0.18	5.82	13.30	0.12	4.06	9.80

We observed that the false positive rate slowly increases with the number of taxa and speciation rate λ. Despite the relatively high false positive rate, the test is still useful for checking that a tree does not fall in an anomaly zone — if none of two consecutive branches on a path from the the root to a tip satisfy $y < a(x)$, then it is very unlikely that the species tree is in an unranked anomaly zone.

We considered a similar test for the ranked anomaly zone. As discussed in Section 2.2, the expression (2.3) can be used to determine whether a five-taxon species tree falls in the ranked anomaly zones instead of computing the gene tree distribution. We use this expression to determine candidates for being in a ranked anomaly zone by checking every three consecutive speciation intervals in larger rooted species trees. Unfortunately, this method has high false positive and low true positive rates (data

not shown), and cannot be used as a quick test whether a species tree falls into the ranked anomaly zone.

3.5 Results for 9-12 taxa

Figures 3.1 and 3.4 show probabilities of the species tree being in the unranked, unrooted, and ranked anomaly zones for different combinations of the number of taxa n, speciation rate λ, and extinction rate μ. For all types of trees, the probability of being in an anomaly zone increases with the number of taxa and with λ in this range. Increasing λ makes consecutive short branches more likely to appear in the species tree, which explains the increasing trend in probabilities of the unranked and unrooted anomaly zones.

We also observed the opposite effect of the turnover rate μ/λ on the probability of producing unranked and unrooted versus ranked anomalous gene trees. On average, branches closer to the root are longer than other branches in a tree as the turnover rate increases (Table 3.4). This leads to a longer speciation interval near the root and explains the decreasing trend in probability in the unranked and unrooted anomaly zones and the increasing trend in the ranked anomaly zone as turnover rate increases since longer branches near the root can produce ARGTs when other branches are short.

To reduce the computational complexity, we use the heuristic described in Section 3.3 of considering unranked and unrooted gene tree topologies within one NNI step from the species tree topology to infer the existence of AGTs and AUGTs for larger trees ($n > 8$). For ranked gene tree topologies, we consider only those that share the same unranked topology with the species tree.

We used Venn diagrams to visualize results obtained from analyzing larger trees.

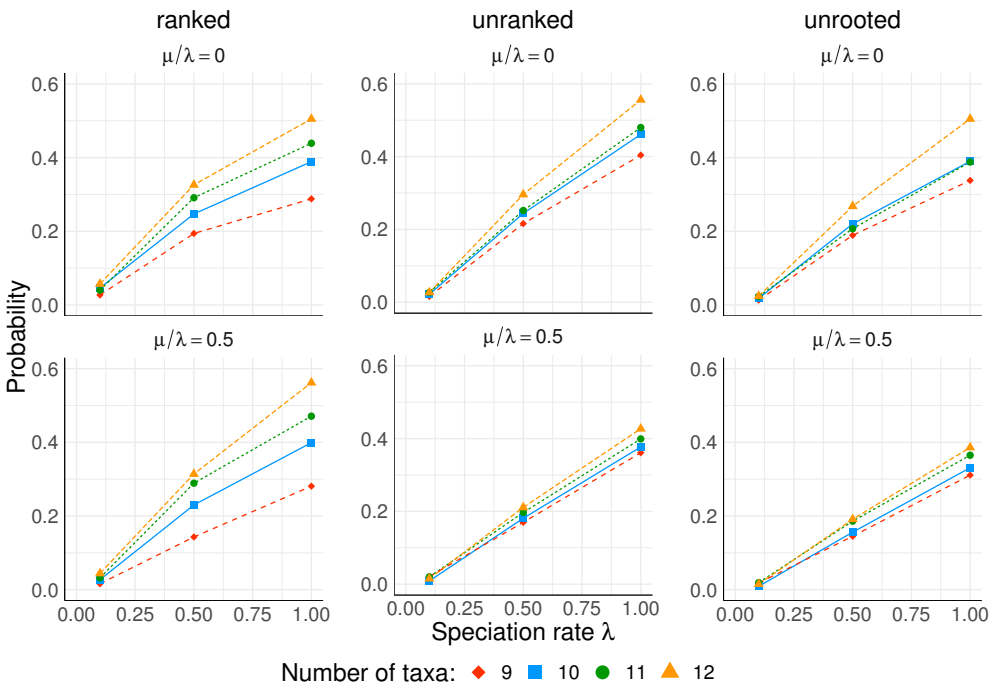


Figure 3.4: The impact of the speciation rate parameter λ and turnover rate μ/λ on the existence of ranked, unranked, and unrooted anomaly zones. We simulated 1000 species trees for $n = 9, 10, 11, 12$ taxa using a constant rate birth-death process with rates $\lambda \in \{0.1, 0.5, 1\}$ and $\frac{\mu}{\lambda} \in \{0, 0.5\lambda\}$. For each combination of (n, λ, μ) the probabilities of the species tree being in each type of μ/λ anomaly zone were computed.

Figures 3.5 and 3.6 depict the relationships between unrooted, unranked, and ranked anomaly zones (AZ_{UGT}, AZ_{GT} and AZ_{RGT}). Each slice represents the number of species trees in the anomaly zone. We observe that for low speciation rates, AZ_{UGT} is often a subset of AZ_{GT}, and there are not many species trees in any of the three anomaly zones. However, as λ increases, species trees start to produce anomalous gene trees more often in each type of anomaly zone, and the proportion of trees in the intersection of two or more anomaly zones also increases (Figures 3.5B, 3.5D). As shown in Figure 3.6, turnover rate does not make a substantial difference to the relationships between different types of anomaly zones.

Table 3.4: Average length and average proportion of the intervals in the tree. We generated 10000 8-taxon trees under the constant rate birth-death process with speciation rate $\lambda = 1$ and extinction rates $\mu = 0$, 0.5. The proportion for each tree was calculated by dividing the interval length by the sum of the interval lengths in the tree. The speciation intervals are represented by t_2 to t_8 from past to present, respectively.

	$\mu = 0$		$\frac{\mu}{\lambda} = 0.5$	
t	length	proportion	length	proportion
t_2	0.50	0.26	0.84	0.30
t_3	0.33	0.19	0.51	0.20
t_4	0.25	0.15	0.35	0.15
t_5	0.20	0.12	0.26	0.12
t_6	0.17	0.10	0.20	0.09
t_7	0.14	0.09	0.16	0.07
t_8	0.12	0.08	0.13	0.06

Overall, using the proposed heuristic approach for $n > 8$, species trees produce more AGTs than AUGTs and ARGTs (Figures 3.1, 3.4, 3.5). Species trees more often fall in AZ_{RGT} than in AZ_{GT} when the speciation rate λ is small, but as λ increases they start to produce more AGTs than ARGTs. The unrooted and unranked anomaly zones have more trees in common, and the ranked anomaly zone is more separated from them.

We computed probabilities of anomaly zones for extreme values of $\lambda \in [1, 30]$ and noticed that the probabilities of unranked and unrooted anomaly zones grow much faster than that of the ranked anomaly zone as the speciation rate increases (Figure 3.7). In particular, the probability that a species tree belongs to unrooted, unranked, and ranked anomaly zones is equal to 0.338, 0.404, and 0.285, respectively, for $n = 9$, $\lambda = 1$, $\mu = 0$. For 9-taxon species tree with $\lambda = 10$ and $\mu = 0$, the corresponding probabilities are equal to 0.796, 0.914, and 0.417, respectively. For extreme speciation rates, probabilities of unranked and unrooted anomaly zones

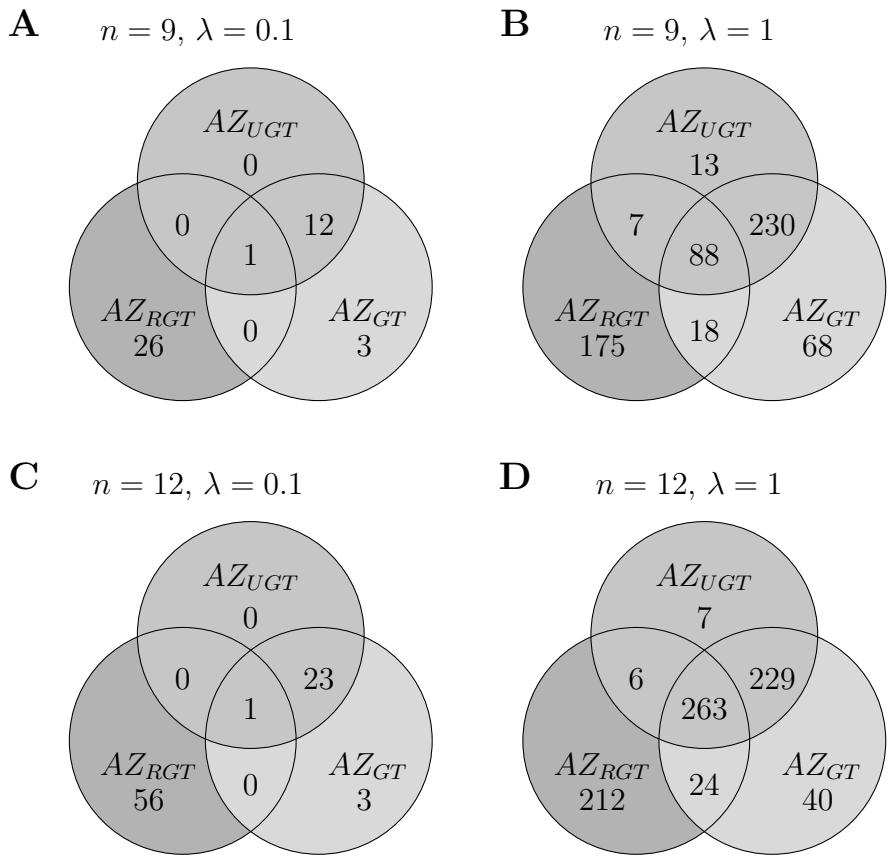


Figure 3.5: Relationships between unrooted, unranked, and ranked anomaly zones. We considered 1000 species trees with birth parameters $\lambda = 0.1, 1$ for 9 and 12 taxon species trees. In each Venn diagram, each slice represents a number of anomalous trees found. Note that figures are not drawn to scale.

grow much faster than that of the ranked anomaly zone with the speciation rate (Figure 3.7).

It is hard to detect what kind of species tree shapes tend to fall more often in the certain types of anomaly zones. We calculate the Colless index (Colless, 1982) for 9-taxon species trees, simulated under the constant rate birth process with $\lambda = 1$, in the ranked and unranked anomaly zones (Figure 3.8). More balanced trees tend to fall in the ranked anomaly zone and more imbalanced into the unranked anomaly zone. The average Colless statistic is 9.92 for the ranked anomaly zone and 12.39 for

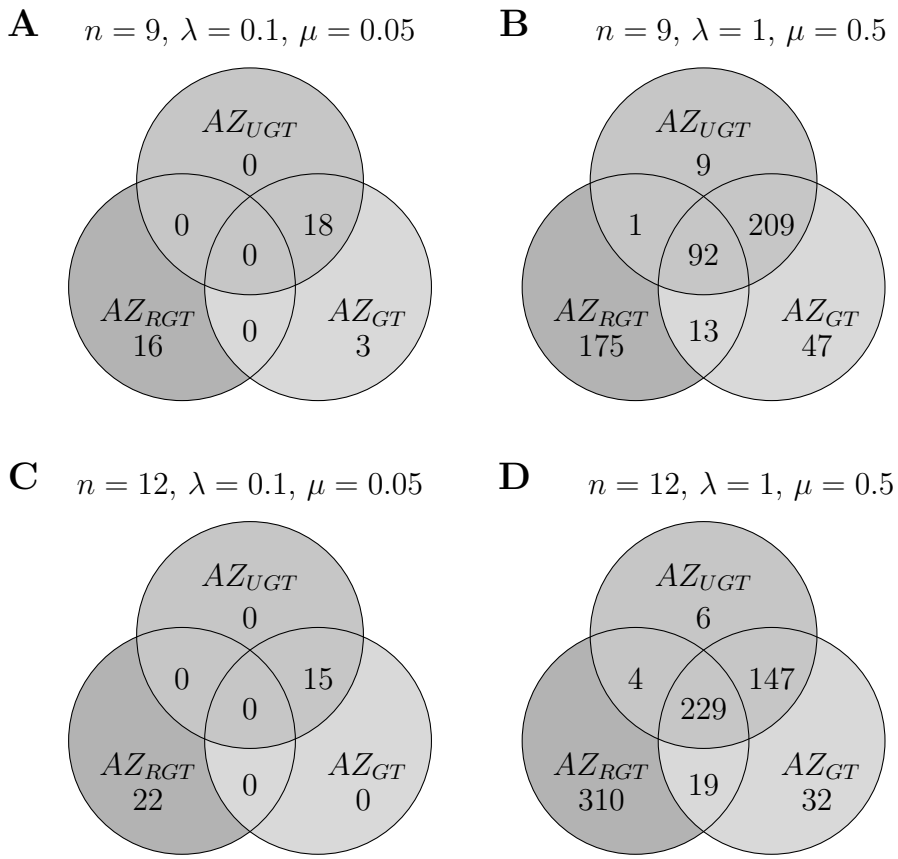


Figure 3.6: Relationships between unrooted, unranked, and ranked anomaly zones. We considered 1000 species trees with birth parameters $\lambda = 0.1, 1$ and turnover rates $\frac{\mu}{\lambda} = 0.5$ for 9 and 12 taxon species trees. In each Venn diagram, each slice represents a number of anomalous trees found. Note that figures are not drawn to scale.

the unranked anomaly zone. For the 12-taxon species trees (Figure 3.9), the average Colless statistics are 16.96 and 20.01 for the ranked and unranked anomaly zone, respectively. The bars corresponding to the ranked AZ are higher than the bars for the unranked AZ for the smaller Colless values, whereas the bars corresponding to the unranked AZ are higher than the bars for the ranked AZ for the larger Colless values.

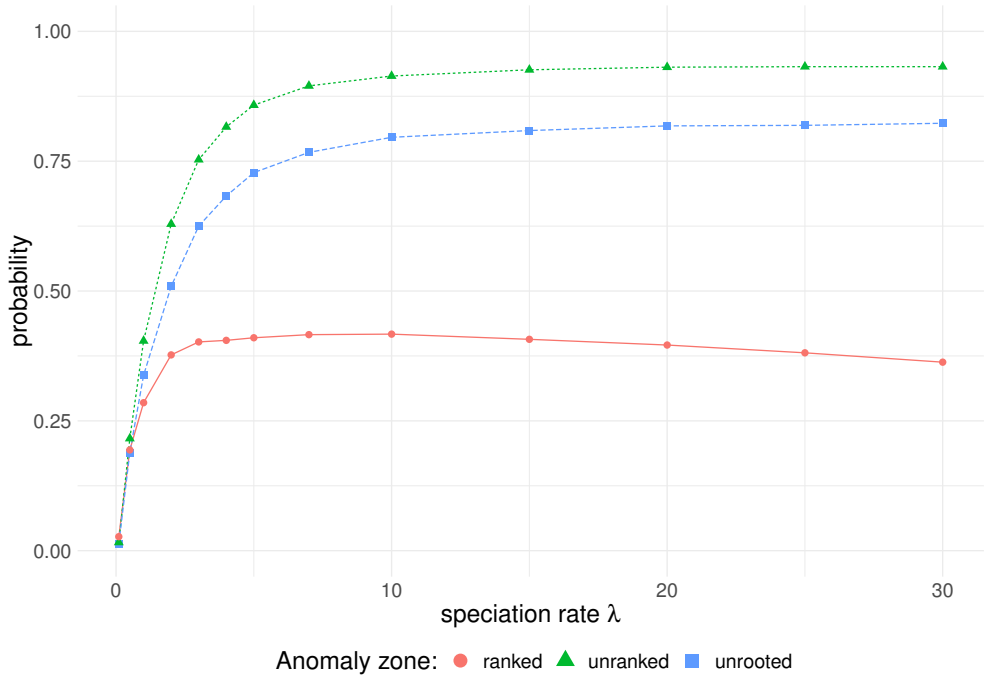


Figure 3.7: The impact of the speciation rate parameter $\lambda \in \{0.1, 30\}$ on the existence of unranked, unrooted and ranked anomaly zones. The probabilities of the 9-taxon species trees being in the anomaly zone were computed. Probabilities of unranked anomaly zone appear to increase with λ, whereas probabilities of ranked anomaly zone increase up to a certain values, and then begin to decrease for approximately $\lambda \geq 10$.

3.6 Discussion

Although the theoretical possibility of anomalous gene trees has been known over a decade, it has not been clear how often this phenomenon occurs in practice. This section addresses this question, albeit indirectly, by estimating how often AGTs, ARGTs, and AUGTs occur under the widely-used birth-death models of speciation.

The probability of the species tree having an anomalous gene tree increases with the speciation rate and the number of species sampled. Our simulation was based on unconstrained tree heights for the species tree, so that sampling more species meant

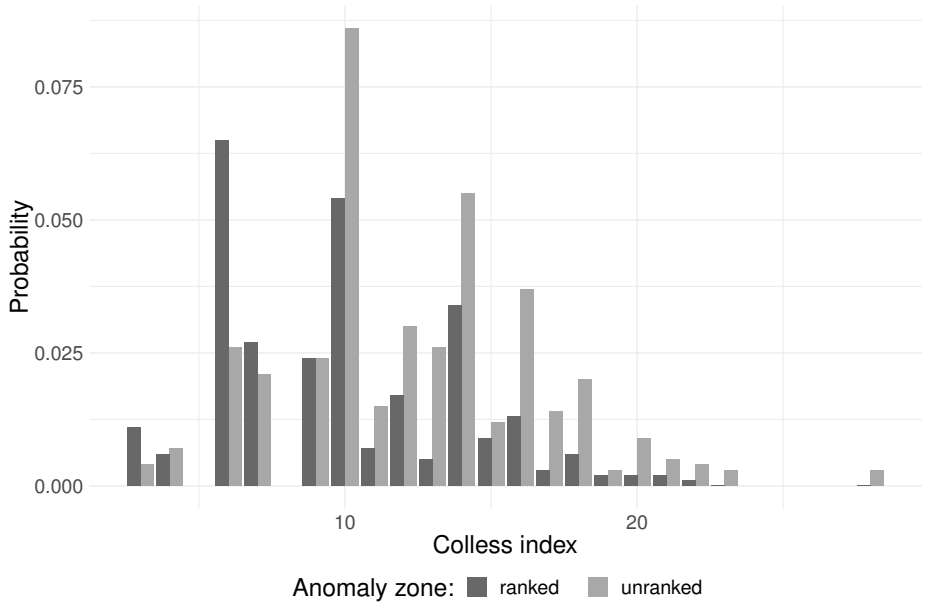


Figure 3.8: The probability of species trees being in the unranked and ranked anomaly zones. 1000 9-taxon species trees were simulated under the birth process with $\lambda = 1$. The Colless statistic was computed for all trees in the anomaly zone. The larger the value, the more imbalanced the tree is.

that the total expected height of the tree also increased. Even under this design, probabilities of being in anomaly zones increased with more taxa. In practice, the question of how often anomaly zones arise is still difficult to answer, but if something is known about the speciation rate, particularly in coalescent units, then this study can help to answer that question. We did not study the effect of sampling more densely within a clade. When more taxa are sampled within a clade, then the total height of the species tree is kept constant but more branches are added, making the branches shorter and more likely to produce at least AGTs and AUGTs. We leave such effects of taxon sampling to future work.

Knowing how often and what kind of anomalous gene trees can generate species trees can help to design valid simulation studies. Our proposed heuristic approaches

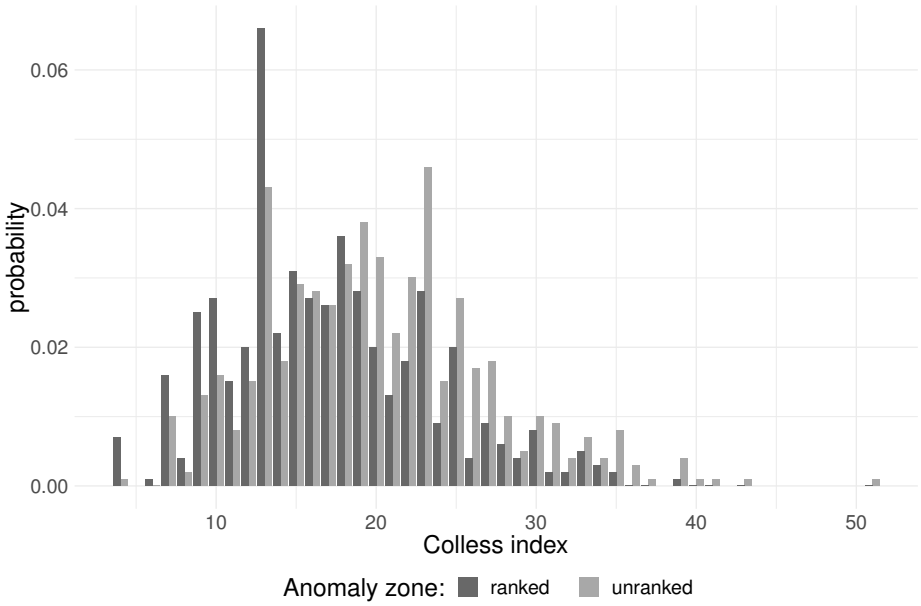


Figure 3.9: The probability of species trees being in the unranked and ranked anomaly zones. 1000 12-taxon species trees were simulated under the birth process with $\lambda = 1$. The Colless statistic was computed for all trees in the anomaly zone. The larger the value, the more imbalanced the tree is.

are generally useful for anomaly zone calculation with high true positives and no false positives. Our simulation study revealed that the most probable tree often is not topologically far from the species tree. When the most probable tree is far from the species tree (in terms of RF distance), there are usually other AGTs or AUGTs that are closer to the species tree, even if they are not the highest probability gene trees. Therefore, using the nearest neighbor interchange branch rearrangement technique, we found that considering only unrooted and unranked gene trees within one NNI move from the species tree topology is a good heuristic to infer the existence of anomalous unrooted and unranked gene trees, respectively. This heuristic underestimates the probability that the species tree is in an anomaly zone due to there being few false negatives but no false positives. This tends to reinforce our conclusion that

there is a high probability that a species tree can have AGTs or AUGTs for moderately high levels of speciation. Less balanced tree shapes also tend to have higher probabilities of AGTs and AUGTs (Figures 3.8 and 3.9). If standard birth-death models underestimate levels of imbalance in real phylogenies (Mooers and Heard, 1997; Bortolussi *et al.*, 2005; Stadler *et al.*, 2016), then we can expect probabilities of AGTs and AUGTs to be even higher than these simulations suggest.

Chapter 4

Maximum likelihood species tree estimation

As discussed in Chapter 1.2, the multispecies coalescent model (MSC) is often used to model discordance between species trees and gene trees representing the evolutionary relationships among a set of species and genes, respectively. The MSC is widely used to infer species trees directly from sequence data (Huelsenbeck and Ronquist, 2001; Heled and Drummond, 2009; Bryant *et al.*, 2012; Chifman and Kubatko, 2014; Yang, 2015). It is extensively used to make inferences from unrooted gene tree topologies (Larget *et al.*, 2010; Liu and Yu, 2011; Mirarab *et al.*, 2014), rooted unranked gene tree topologies (Liu *et al.*, 2009b, 2010; Wu, 2012; Pei and Wu, 2017), or from gene trees with branch lengths (Liu *et al.*, 2009b; Kubatko *et al.*, 2009; Mossel *et al.*, 2011).

4.1 Description

Let $\mathcal{T}$ be an n-taxon rooted species tree with branch lengths. Assuming that we have observed a collection of N independent ranked gene trees $\mathcal{G}_i, i = 1, 2, \ldots, N$,

the maximum likelihood (ML) species tree is

$$\mathcal{T}_{ML} = \underset{\mathcal{T}}{\operatorname{argmax}}\, P[\mathcal{G}_1, ..., \mathcal{G}_N|\mathcal{T}] = \underset{\mathcal{T}}{\operatorname{argmax}} \prod_{i=1}^{N} P[\mathcal{G}_i|\mathcal{T}] \tag{4.1}$$

The probability of the ranked gene tree $P(\mathcal{G}|\mathcal{T})$ is described elsewhere (Degnan *et al.*, 2012b; Stadler and Degnan, 2012; Kim *et al.*, 2019).

Initially we developed *PRANC* to compute the analytical probabilities of the ranked gene trees given a species tree. However, *PRANC* has other useful options to work with phylogenetic trees. One of the main options is to find a species tree $\mathcal{T}$ with branch lengths in coalescent units that maximizes the likelihood given by eq. (4.1). To estimate the ML species tree, *PRANC* uses the following steps:

1. Process the initial species tree. If the tree has the branch lengths specified in coalescent units, treat the tree as a ranked tree. Find a set of speciation interval lengths that maximizes the likelihood. If the branch lengths are not specified in the tree, generate all possible rankings. Randomly select a subset of ranked trees (by default, $2n$). Define the speciation interval length between the $(i-1)$th and ith speciation events as $t_i = s_{i-1} - s_i$, where s_i is the time of the interior node of rank i. For each of these trees, initialize each interval length t_i to 1.0 and find a set of speciation interval lengths that maximizes the likelihood. Pick the tree $\mathcal{T}$ with the highest likelihood. Note that several initial trees can be given.

2. Obtain all trees that are one nearest neighbor interchange (NNI) away from $\mathcal{T}$, where a move is determined by swapping two subtrees that are separated by an internal edge. NNI will be discussed in detail in Section 4.3. For each of these unranked trees, generate all possible ranked trees. Randomly select a subset of ranked trees (by default, $2n$). Find the speciation interval lengths that maximizes the likelihood of the ranked gene trees and pick the one with

the highest likelihood. If this tree has a larger likelihood, then set $\mathcal{T}$ to this tree.

3. Repeat step 2 until convergence or until all trees within k (by default, $k = 5$) NNI steps are explored.

4. Calculate the branch lengths of the inferred tree. Note that we set the time to the most recent internal node to 0.1 coalescent units.

We optimize the interval lengths using Brent's method (Brent, 1973) that will be discussed in Section 4.4.1. We compute the initial likelihood for the tree obtained in step 1. Then, we change each length one at a time, fixing the other lengths. We allow the length to be in the interval $[0.001, 6]$. We randomly pick interval orders for optimization. After m rounds of such optimizations (by default, m is set to the number of taxa), the optimal tree is reported. As a better alternative, we propose to use limited memory algorithm for bound constrained optimization (Byrd *et al.*, 1995) for the interval lengths optimization that will be discussed in Section 4.4.2. To further reduce the computational time, *PRANC* can compute likelihoods of a randomly chosen small subset of ranked trees (by default, $n/2$). If at least one of the obtained likelihoods is better than the threshold, *PRANC* will compute the likelihoods of a larger subset of ranked trees (by default, up to $2n$).

4.2 Finding a starting tree

It is important to find a good starting tree topology in exploring tree space. *PRANC*'s ability to locate the ML species tree quickly highly depends on the choice of the starting tree. Because it is better to have an initial tree that is topologically close to the true tree, we propose two methods for picking a starting tree. The first method is a hill-climbing algorithm that will penalize long branches of the species

tree that have multiple lineages persisting through multiple species divergence events (discussed in the next subsection). Although this method can provide a list of good candidate trees for a species tree, we proved that it is statistically inconsistent as shown in Subsection 4.2.2. Another method is based on the greedy consensus method (Bryant, 2003) which is also statistically inconsistent (Degnan *et al.*, 2009a). In this consensus method, clades are first ordered according to the number of times they appear in the sample of gene trees, then the consensus tree is formed combining clades which occurred in at least a certain percentage of the resampled trees. The greedy consensus method considers the clades in order of the frequency with which they have appeared, adding to the consensus tree any which are compatible with it until the tree is fully resolved. We propose a modification to the greedy consensus method. Our method outputs a tree with a maximal score, where the score is determined by summing the frequencies of compatible clades. In such a way, the program will include all most supported clades in the estimated tree and not only those compatible with the most frequent clade.

4.2.1 Minimizing the ancient coalescence criterion

For a ranked gene tree $\mathcal{G}_i$ given a species tree $\mathcal{T}$, the minimizing ancient coalescence (MAC) method is based on minimizing the sum of the minimum number of extra lineages withing each time interval τ_i (Stadler and Degnan, 2012)

$$m(\mathcal{G}_i|\mathcal{T}) = \sum_{i=1}^{n-1}(g_i - (i+1)). \tag{4.2}$$

To compute g_i, let $lca(u_j)$ be the least common ancestor node on the species tree for a node u_j on the gene tree. Here $lca(u_j)$ is the node with the largest rank on the species tree which is ancestral to all species with the same labels as descended from u_j in the gene tree.

Minimizing 4.2 as a criterion for the species tree will tend to penalize long edges

of the species tree and thus will allow less ILS. The expected value of the MAC criterion for the species tree $\mathcal{T}$ can be defined as

$$E_m(\mathcal{T}) = \sum_{i=1}^{\mathcal{G}_{num}} m(\mathcal{G}_i|\mathcal{T})P(\mathcal{G}_i|\mathcal{T}). \tag{4.3}$$

A species tree candidate $\mathcal{T}_{cand}$ with the smallest expected value is chosen as an estimate of the species tree $\mathcal{T}$ as the number of gene trees $\mathcal{G}_{num} \to \infty$. If at least one species tree candidate produces a smaller expected value than the true tree $\mathcal{T}$, then the MAC criterion gives an incorrect estimate of the true tree as the number of gene trees increases; that is, the MAC criterion is not statistically consistent.

The idea of the MAC criterion is based on the minimizing deep coalescence (MDC) criterion (Maddison, 1997) which minimizes the number of total extra lineages in all branches of the species tree. However, the MAC criterion is designed for ranked gene trees, whereas the MDC criterion disregards the rankings of the gene trees. Than and Rosenberg (2011a) showed how to compute the MDC cost in detail and proved the inconsistency of the MDC criterion. Figure 4.1 shows the computation of both the MDC and MAC costs.

4.2.2 Statistical inconsistency of MAC

Trees with three leaves

The MAC criterion for the ranked gene trees has a similar idea as the MDC criterion for the unranked gene trees. Because a ranked three-taxon tree is exactly the same as its unranked version, the result for three taxon trees that holds for MDC also holds for MAC. It was shown by Than and Rosenberg (2011a) that MDC is consistent for the trees with three leaves. Thus, MAC is consistent for three taxon trees as well.

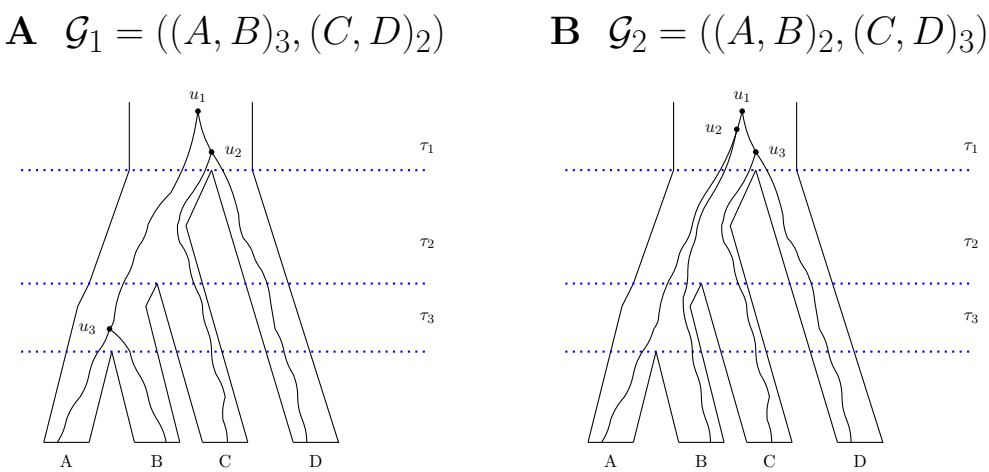


Figure 4.1: Computing the minimizing deep coalescence (MDC) cost and the minimizing ancient coalescence (MAC) cost for the four taxon gene trees evolving inside the species tree $(((A,B)_3,C)_2,D)$. (A) Both the MDC cost $\alpha(\mathcal{G}_1|\mathcal{T}) = 1$ and the MAC cost $m(\mathcal{G}_1|\mathcal{T}) = 1$ represent the total number of extra lineages in all the branches of $\mathcal{T}$. The lineage corresponding to the C taxon failed to coalesce earlier in the interval τ_2 leading to the discordance of the gene tree and underlying species tree and the extra lineage in the interval τ_1. (B) The MDC cost $\alpha(\mathcal{G}_2|\mathcal{T}) = 1$ and the MAC cost $m(\mathcal{G}_2|\mathcal{T}) = 3$. These two costs are different because the MDC criterion does not distinguish between the rankings of the same gene tree topology, whereas the MAC criterion is designed for ranked gene trees. The lineage corresponding to the B taxon failed to coalesce earlier in the interval τ_3 and then also in the interval τ_2, and thus producing two extra lineages. The lineage corresponding to the C taxon also failed to coalesce earlier in the interval τ_2 and thus producing one extra lineage.

Trees with four leaves

There are 18 rooted ranked trees on four leaves. To study the consistency of the MAC criterion, it is sufficient to consider only one labeling for each unlabeled species tree topology. For example trees $(((A,B),C),D)$ and $(((B,C),A),D)$ are equivalent up to permuting labels. We distinguish species trees $((A,B)_2,(C,D)_3)$ and $((A,B)_3,(C,D)_2)$ because they have different rankings. We assume that the species tree is either $\mathcal{T}_7 = (((A,B)_3,C_2),D)$, $\mathcal{T}_1 = ((A,B)_3,(C,D)_2))$, or $\mathcal{T}_2 = ((A,B)_2,(C,D)_3)$. We check the consistency of the MAC criterion on trees with

four leaves by an exhaustive approach, that is, by directly computing the expected ancient coalescence cost for every species tree candidate and comparing it with the corresponding cost of the true species tree. The probabilities of ranked gene tree topologies for the species trees $\mathcal{T}_1$ and $\mathcal{T}_7$ are given in Tables 1 and 2 of Degnan *et al.* (2012b), respectively. The ancient coalescence score of each gene tree $\mathcal{G}_i$ conditioning on a species tree candidate $\mathcal{T}_i$ is shown in Table 4.1.

Table 4.1: Ancient coalescence scores of each gene tree $\mathcal{G}_i$ conditioning on a species tree candidate $\mathcal{T}_i$.

candidate $\mathcal{T}_i$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$	$\mathcal{G}_7$	$\mathcal{G}_8$	$\mathcal{G}_9$	$\mathcal{G}_{10}$	$\mathcal{G}_{11}$	$\mathcal{G}_{12}$	$\mathcal{G}_{13}$	$\mathcal{G}_{14}$	$\mathcal{G}_{15}$	$\mathcal{G}_{16}$	$\mathcal{G}_{17}$	$\mathcal{G}_{18}$
$\mathcal{T}_1 = ((AB)_3(CD)_2)$	0	1	3	3	3	3	1	1	3	3	3	3	3	3	3	3	2	2
$\mathcal{T}_2 = ((AB)_2(CD)_3)$	1	0	3	3	3	3	2	2	3	3	3	3	3	3	3	3	1	1
$\mathcal{T}_3 = ((AC)_3(BD)_2)$	3	3	0	1	3	3	3	3	1	1	3	3	3	3	2	2	3	3
$\mathcal{T}_4 = ((AC)_2(BD)_3)$	3	3	1	0	3	3	3	3	2	2	3	3	3	3	1	1	3	3
$\mathcal{T}_5 = ((AD)_3(BC)_2)$	3	3	3	3	0	1	3	3	3	3	1	1	2	2	3	3	3	3
$\mathcal{T}_6 = ((AD)_2(BC)_3)$	3	3	3	3	1	0	3	3	3	3	2	2	1	1	3	3	3	3
$\mathcal{T}_7 = (((AB)_3C)_2D)$	1	3	2	3	3	2	0	1	1	2	3	3	1	2	3	3	3	3
$\mathcal{T}_8 = (((AB)_3D)_2C)$	1	3	3	2	2	3	1	0	3	3	1	2	3	3	1	2	3	3
$\mathcal{T}_9 = (((AC)_3B)_2D)$	2	3	1	3	3	2	1	2	0	1	3	3	1	2	3	3	3	3
$\mathcal{T}_{10} = (((AC)_3D)_2B)$	3	2	1	3	2	3	3	3	1	0	2	1	3	3	3	3	1	2
$\mathcal{T}_{11} = (((AD)_3B)_2C)$	2	3	3	2	1	3	2	1	3	3	0	1	3	3	1	2	3	3
$\mathcal{T}_{12} = (((AD)_3C)_2B)$	3	2	2	3	1	3	3	3	2	1	1	0	3	3	3	3	1	2
$\mathcal{T}_{13} = (((BC)_3A)_2D)$	2	3	2	3	3	1	1	2	1	2	3	3	0	1	3	3	3	3
$\mathcal{T}_{14} = (((BC)_3D)_2A)$	3	2	3	2	3	1	3	3	3	3	3	3	1	0	2	1	2	1
$\mathcal{T}_{15} = (((BD)_3A)_2C)$	2	3	3	1	2	3	2	1	3	3	1	2	3	3	0	1	3	3
$\mathcal{T}_{16} = (((BD)_3C)_2A)$	3	2	3	1	3	2	3	3	3	3	3	3	2	1	1	0	2	1
$\mathcal{T}_{17} = (((CD)_3A)_2B)$	3	1	2	3	2	3	3	3	2	1	2	1	3	3	3	3	0	1
$\mathcal{T}_{18} = (((CD)_3B)_2A)$	3	1	3	2	3	2	3	3	3	3	3	3	2	1	2	1	1	0

Caterpillar species tree

Table 2 of Degnan *et al.* (2012b) contains the probabilities of 18 ranked gene trees given that the true species tree is $\mathcal{T}_7 = (((A, B)_3, C_2), D)$. It can be observed that $P(\mathcal{G}_1|\mathcal{T}_7) = P(\mathcal{G}_8|\mathcal{T}_7)$, $P(\mathcal{G}_2|\mathcal{T}_7) = P(\mathcal{G}_4|\mathcal{T}_7) = P(\mathcal{G}_5|\mathcal{T}_7) = P(\mathcal{G}_{11}|\mathcal{T}_7) = P(\mathcal{G}_{12}|\mathcal{T}_7) = P(\mathcal{G}_{15}|\mathcal{T}_7) = P(\mathcal{G}_{16}|\mathcal{T}_7) = P(\mathcal{G}_{17}|\mathcal{T}_7) = P(\mathcal{G}_{18}|\mathcal{T}_7)$, $P(\mathcal{G}_3|\mathcal{T}_7) = P(\mathcal{G}_6|\mathcal{T}_7) = P(\mathcal{G}_{10}|\mathcal{T}_7) = P(\mathcal{G}_{14}|\mathcal{T}_7)$, and $P(\mathcal{G}_9|\mathcal{T}_7) = P(\mathcal{G}_{13}|\mathcal{T}_7)$. Plugging the probabilities and ancient coales-

cence costs given in Table 4.1 for the species tree $\mathcal{T}_7$, we have

$$E_m(\mathcal{T}_7) = \sum_{i=1}^{18} m(\mathcal{G}_i|\mathcal{T}_7)P(\mathcal{G}_i|\mathcal{T}_7) =$$
$$=2P(\mathcal{G}_1|\mathcal{T}_7) + 9P(\mathcal{G}_2|\mathcal{T}_7) + 4P(\mathcal{G}_3|\mathcal{T}_7) + P(\mathcal{G}_7|\mathcal{T}_7) + 2P(\mathcal{G}_9|\mathcal{T}_7). \tag{4.4}$$

Let $\mathcal{T}_{cand}$ be a species tree candidate different from $\mathcal{T}$. We need to prove that $E_m(\mathcal{T}_7) < E_m(\mathcal{T}_{cand})$ for all possible other 17 candidates. Table 4.2 shows the differences of the expected values $E_m(\mathcal{T}_7) - E_m(\mathcal{T}_{cand})$. It is straightforward to check that all differences in this table are strictly less than 0 for all positive real values of the interval lengths t_2 and t_3. Therefore, $E_m(\mathcal{T}_7) < E_m(\mathcal{T}_{cand})$ for all other candidates $\mathcal{T}_{cand} \neq \mathcal{T}_7$. The MAC criterion is statistically consistent in this case.

Symmetric species trees

Table 1 of Degnan *et al.* (2012b) contains the probabilities of 18 ranked gene trees given that the true species tree is $\mathcal{T}_1 = ((A,B)_3,(C,D)_2))$. The corresponding probabilities for $\mathcal{T}_2 = ((A,B)_2,(C,D)_3)$ are similar to those in Table 1, except that the probabilities in the first two rows are swapped with each other, and the probabilities in the seventh and eighth rows are swapped with the probabilities in the last two rows. Tables 4.3 and 4.4 show the differences of the expected values $E_m(\mathcal{T}_1) - E_m(\mathcal{T}_{cand})$ and $E_m(\mathcal{T}_2) - E_m(\mathcal{T}_{cand})$, respectively. It is straightforward to check that differences of the expected values between these true trees and other symmetric candidate trees (the first six rows) in these two tables are strictly less than 0 for all positive real values of t_2 and t_3. However, for some possible values of t_2 and t_3, the differences of the expected values between these true trees and caterpillar candidate trees (the last twelve rows) can be greater than 0. For example, in the Table 4.3, the difference of the expected values

$$E_m[((AB)_3(CD)_2)] - E_m[(((BD)_3A)_2C)] = e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$$

is positive for $t_2 = 0.01, t_3 = 0.01$ and negative for $t_2 = 0.01, t_3 = 0.5$. Therefore, the MAC criterion is not statistically consistent in these two cases.

For species trees with three leaves or for asymmetric species trees with four leaves, we have shown that the MAC criterion is statistically consistent. However, it is not statistically consistent for symmetric species trees with four leaves. The result is somewhat interesting, since Than and Rosenberg (2011a) proved that the MDC criterion is statistically consistent for species trees with three leaves or for symmetric species trees with four leaves; and is not statistically consistent for asymmetric four-leaf species trees.

Table 4.2: The differences of the expected values $E_m(\mathcal{T}_7) - E_m(\mathcal{T}_{cand})$, where $\mathcal{T}_7 = (((AB)_3C)_2D)$ has a caterpillar shape.

species tree candidate $\mathcal{T}_{cand}$	$E_m(\mathcal{T}_7) - E_m(\mathcal{T}_{cand})$
$((AB)_3(CD)_2)$	$e^{-t_2} - \frac{2 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{6} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 1$
$((AB)_2(CD)_3)$	$e^{-t_2} + \frac{e^{-t_2} \cdot e^{-t_3}}{6} + \frac{11 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 2$
$((AC)_3(BD)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{4 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$((AC)_2(BD)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{11 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$((AD)_3(BC)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{11 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$((AD)_2(BC)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{4 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$(((AB)_3C)_2D)$	0
$(((AB)_3D)_2C)$	$e^{-t_2} - \frac{2 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{6} + \frac{e^{-3 \cdot t_2} \cdot e^{-t_3}}{2} - 1$
$(((AC)_3B)_2D)$	$e^{-t_3} - 1$
$(((AC)_3D)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{4 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{e^{-3 \cdot t_2} \cdot e^{-t_3}}{2} - 3$
$(((AD)_3B)_2C)$	$e^{-t_2} + \frac{e^{-t_2} \cdot e^{-t_3}}{6} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 2$
$(((AD)_3C)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$(((BC)_3A)_2D)$	$e^{-t_3} - 1$
$(((BC)_3D)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{4 \cdot e^{-t_3}}{3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{e^{-3 \cdot t_2} \cdot e^{-t_3}}{2} - 3$
$(((BD)_3A)_2C)$	$e^{-t_2} + \frac{e^{-t_2} \cdot e^{-t_3}}{6} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 2$
$(((BD)_3C)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$(((CD)_3A)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$(((CD)_3B)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + \frac{e^{-t_2} \cdot e^{-t_3}}{2} + \frac{5 \cdot e^{-3 \cdot t_2} \cdot e^{-t_3}}{6} - 3$

Simulation

We simulate 100 $n = 5, 6, 7, 8$-taxon species trees under the constant rate birth-death model with speciation rate $\lambda = 1$ and extinction rate $\mu = 0$. For each species tree,

Table 4.3: The differences of the expected values $E_m(\mathcal{T}_1) - E_m(\mathcal{T}_{cand})$, where $\mathcal{T}_1 = ((AB)_3(CD)_2)$ has a symmetric shape.

species tree candidate $\mathcal{T}_{cand}$	$E_m(\mathcal{T}_1) - E_m(\mathcal{T}_{cand})$
$((AB)_3(CD)_2)$	0
$((AB)_2(CD)_3)$	$e^{-t_3} - 1$
$((AC)_3(BD)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AC)_2(BD)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AD)_3(BC)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AD)_2(BC)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$(((AB)_3C)_2D)$	$e^{-t_2} - \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 1$
$(((AB)_3D)_2C)$	$e^{-t_2} - \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 1$
$(((AC)_3B)_2D)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((AC)_3D)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((AD)_3B)_2C)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((AD)_3C)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((BC)_3A)_2D)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((BC)_3D)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((BD)_3A)_2C)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((BD)_3C)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((CD)_3A)_2B)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{3 \cdot e^{-t_3}}{2} + e^{-t_2} \cdot e^{-t_3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$(((CD)_3B)_2A)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{3 \cdot e^{-t_3}}{2} + e^{-t_2} \cdot e^{-t_3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 3$

we generate $100, 500$, and 1000 gene trees using *hybrid-lambda* (Zhu *et al.*, 2015). To estimate the MAC species tree, we used the hill-climbing approach described in the following steps:

1. In the sample of n_g gene trees, consider each gene tree as a species tree candidate and compute the MAC score for each of these gene trees. In particular, set $\mathcal{T}_j$ to G_j and then compute the MAC score for the candidate $\mathcal{T}_j$ as $\sum_i^{n_g} m(G_i|\mathcal{T}_j)$ for each j.

2. Select $n_c = 10$ candidate species trees with the lowest MAC score and remove the branch lengths. Record the lowest value of the MAC score x_{mac}.

3. Obtain all trees that are one nearest neighbor interchange (NNI) away from

Table 4.4: The differences of the expected values $E_m(\mathcal{T}_2) - E_m(\mathcal{T}_{cand})$, where $\mathcal{T}_2 = ((AB)_2(CD)_3)$ has a symmetric shape.

species tree candidate $\mathcal{T}_{cand}$	$E_m(\mathcal{T}_2) - E_m(\mathcal{T}_{cand})$
$((AB)_3(CD)_2)$	$e^{-t_3} - 1$
$((AB)_2(CD)_3)$	0
$((AC)_3(BD)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AC)_2(BD)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AD)_3(BC)_2)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$((AD)_2(BC)_3)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{7 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{6} - 3$
$(((AB)_3C)_2D)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{3 \cdot e^{-t_3}}{2} + e^{-t_2} \cdot e^{-t_3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$(((AB)_3D)_2C)$	$\frac{2 \cdot e^{-t_2}}{3} + \frac{3 \cdot e^{-t_3}}{2} + e^{-t_2} \cdot e^{-t_3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 3$
$(((AC)_3B)_2D)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((AC)_3D)_2B)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((AD)_3B)_2C)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((AD)_3C)_2B)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((BC)_3A)_2D)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((BC)_3D)_2A)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((BD)_3A)_2C)$	$\frac{2 \cdot e^{-t_2}}{3} + e^{-t_3} + e^{-t_2} \cdot e^{-t_3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 3$
$(((BD)_3C)_2A)$	$e^{-t_2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{5 \cdot e^{-2 \cdot t_2} \cdot e^{-t_3}}{9} - 2$
$(((CD)_3A)_2B)$	$e^{-t_2} - \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 1$
$(((CD)_3B)_2A)$	$e^{-t_2} - \frac{e^{-t_3}}{2} + \frac{2 \cdot e^{-t_2} \cdot e^{-t_3}}{3} + \frac{e^{-2 \cdot t_2} \cdot e^{-t_3}}{18} - 1$

each of these n_c trees. For each of these unranked trees, generate all possible ranked trees.

4. Compute MAC scores for obtained ranked trees and extract $n_c = 10$ trees with the lowest MAC score.

5. If the lowest MAC score obtained in step 4 is smaller than the lowest MAC score in step 2, update x_{mac} and repeat steps 2-4 until x_{mac} keeps improving or until all trees within k (by default, $k = 5$) NNI steps are explored.

6. Save $n_c = 10$ candidate trees with the lowest MAC scores in the file.

7. Calculate the Robinson-Foulds distance between n_c candidates and the original species tree from which gene trees were simulated. Record two RF values:

one corresponds to the candidate with the lowest MAC score (ties resolved randomly) and another RF value corresponds to the minimum of RF scores among all $n_c = 10$ final candidates.

It is clear that larger trees have more topologies and thus, it is harder to obtain topologically close inferred tree to the actual species tree as the number of taxa increases. The normalized Robinson-Foulds distance accounts for this scenario since it is computed by dividing by the maximum $2n - 4$. For example, if for 100 5-taxon species trees there are 70 inferred trees with $RF = 0$, 20 trees with $RF = 2$, 8 trees with $RF = 4$, and 2 trees with $RF = 6$, then the normalized RF distance is $(\frac{2}{6} \cdot 20 + \frac{4}{6} \cdot 8 + \frac{6}{6} \cdot 10)/100 = 0.14$. If for 100 8-taxon species trees there are only 56 inferred trees with $RF = 0$, and $20, 14, 8, 0, 0, 2$ trees with $RF = 2, 4, 6, 8, 10, 12$, respectively, then the normalized RF distance is $(\frac{2}{12} \cdot 20 + \frac{4}{12} \cdot 14 + \frac{6}{12} \cdot 8 + \frac{12}{12} \cdot 2)/100 = 0.14$. The two normalized RF distances in the 5- and 8-taxon cases are the same but the distribution of the RF distances are quite different.

Figure 4.2 shows the normalized Robinson-Foulds distances between the inferred trees and true species trees. We consider two cases with one and five NNI moves, respectively. In the case with five NNI moves, we also compute the RF score among $n_c = 10$ species tree candidates with lowest MAC scores. It is better to propose several candidates because one of them can be topologically close to the true tree. This makes sense especially if we use the list of candididates as starting trees for ML methods. Moreover, the scores for different candidates might be not that different from each other to strongly prefer one tree over the other.

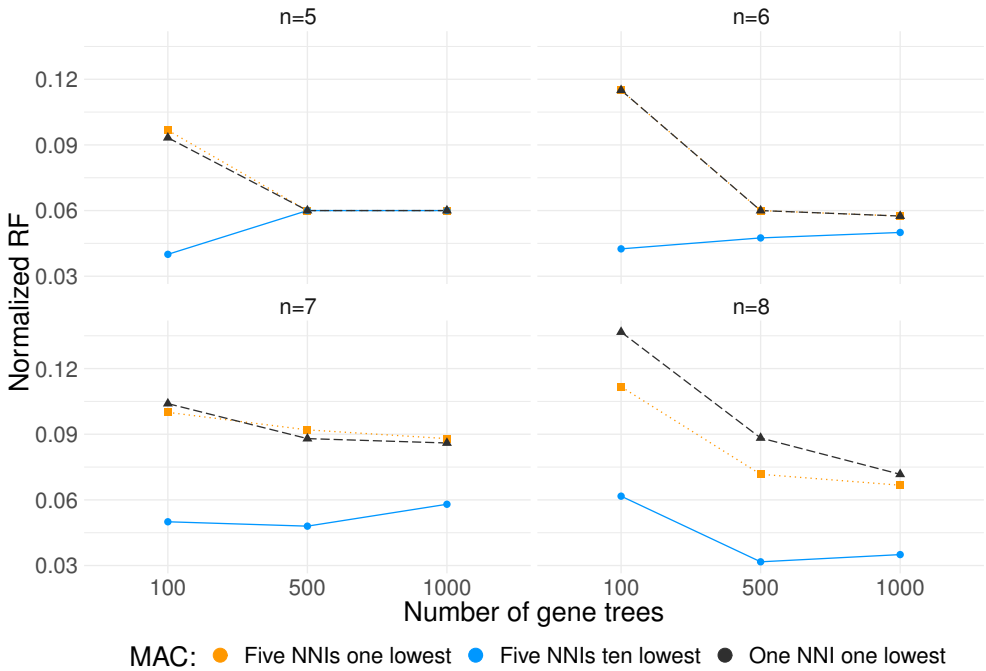


Figure 4.2: The normalized Robinson-Foulds distances between the inferred trees by MAC criterion and true species trees. The pure birth process with $\lambda = 1$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. The number of NNI moves does not play significant role in improving tree estimate. It is better to propose the list of n_c candidates with the lowest MAC scores instead of just one best tree in terms of MAC score.

4.2.3 Greedy consensus tree and Maximum clade frequency consensus tree

One common way to estimate a phylogeny from a collection of gene trees is to reconstruct the species tree from the most supported clades in gene trees. Both majority rule and greedy consensus methods are often used to obtain a summary tree from a sample of trees. It was shown that both are statistically inconsistent (Degnan *et al.*, 2009a). As noted before, clades are first ordered according to the number of times they appear in the sample of gene trees, then the consensus tree is formed combining clades which occurred in at least a certain percentage of the resampled

trees. For example, the majority rule consensus tree consists only of those clades that are present in at least 50% of the trees. In constructing a greedy consensus tree the most frequent clade is picked and then the other clades are sequentially added to the resulting tree that are compatible with the rest of the clades until the tree is fully resolved (Bryant, 2003). Two clades C_1 and C_2 are compatible clades if $C_1 \subset C_2$, or $C_2 \subset C_1$, or they are disjoint.

Let's consider the following 7 trees in the Newick format:

```
(((((F,B),C),D),A),E);
(((((C,B),F),D),A),E);
(((((A,C),D),F),B),E);
(((((D,B),C),A),F),E);
(((((D,E),C),A),B),F);
(((((D,E),C),A),B),F);
(((((D,E),C),A),B),F);
```

the corresponding clades are ordered below according to their frequency (support) in the sample:

```
A|B|C|D|E|F|: 7
A|B|C|D|F|: 4
A|B|C|D|E|: 3
A|C|D|E|: 3
C|D|E|: 3
D|E|: 3
B|C|D|F|: 2
B|C|F|: 2
A|B|C|D|: 1
A|C|D|F|: 1
```

```
B|C|D|: 1
A|C|D|: 1
B|D|: 1
B|F|: 1
A|C|: 1
B|C|: 1
```

Here, for example, both clades (B, C) and (B, F) occur once in the sample of 7 trees. The clade combining these two, i.e. $B|C|F$, occurs twice. The greedy consensus tree starts from the most supported clade $A|B|C|D|F|$ after the clade that includes all leaves (i.e. $A|B|C|D|F|E$). Then, the next two compatible clades with $A|B|C|D|F|$ in the ordered list are $B|C|D|F|$ and $B|C|F|$. Then, the greedy consensus will randomly decide which clade $B|F|$ or $B|C|$ to include in the tree since both of them are compatible with the selected clades and both have same frequency. For example, let's pick $B|C|$. To get a fully resolved tree, the taxon E will be added as an outgroup. Therefore, the resulting greedy consensus tree is $(((((C, B), F), D), A), E)$;.

We propose a modification to the greedy consensus method: *maximum clade frequency consensus* (MCFC) method. MCFC outputs a tree with a maximal score, where the score is constructed from the frequencies of compatible clades. The method will include all most supported clades in the estimated tree and not only those compatible with the most frequent clade. For example, in this the sample of 7 trees, the greedy consensus tree $(((((C, B), F), D), A), E)$; has the score of $7 + 4 + 2 + 2 + 1 = 16$. The MCFC will output $(((((D, E), C), A), B), F)$; tree with the score of $7 + 3 + 3 + 3 + 3 = 19$. However, it will also output the usual greedy consensus tree $(((((C, B), F), D), A), E)$;. Ties are resolved randomly.

To find the tree with the maximal score, the algorithm constructs the compatibility matrix of clades (Table 4.5) and recursively searches for the sequence of compatible clades with the highest score. The current version of the MCFC is not that fast

as the usual greedy consensus method. For example, consider a simulation process where we have 50 species trees and for each of these trees we generate a sample of 100 gene trees and find the consensus tree. It approximately takes one minute to estimate 50 greedy consensus trees and ten minutes to estimate 50 MCFC trees.

Table 4.5: The matrix of compatibility of the clades

clades	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$	$\mathcal{C}_7$	$\mathcal{C}_8$	$\mathcal{C}_9$	$\mathcal{C}_{10}$	$\mathcal{C}_{11}$	$\mathcal{C}_{12}$	$\mathcal{C}_{13}$	$\mathcal{C}_{14}$	$\mathcal{C}_{15}$	$\mathcal{C}_{16}$
$C_1 = A\|B\|C\|D\|E\|F\|$	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$C_2 = A\|B\|C\|D\|F\|$	1	**1**	0	0	0	0	1	1	1	1	1	1	1	1	1	1
$C_3 = A\|B\|C\|D\|E\|$	1	0	**1**	1	1	1	0	0	1	0	1	1	1	0	1	1
$C_4 = A\|C\|D\|E\|$	1	0	1	**1**	1	1	0	0	0	0	0	1	0	1	1	0
$C_5 = C\|D\|E\|$	1	0	1	1	**1**	1	0	0	0	0	0	0	0	1	0	0
$C_6 = D\|E\|$	1	0	1	1	1	**1**	0	1	0	0	0	0	0	1	1	1
$C_7 = B\|C\|D\|F\|$	1	1	0	0	0	0	**1**	1	0	0	1	0	1	1	0	1
$C_8 = B\|C\|F\|$	1	1	0	0	0	1	1	**1**	0	0	0	0	0	1	0	1
$C_9 = A\|B\|C\|D\|$	1	1	1	0	0	0	0	0	**1**	0	1	1	1	0	1	1
$C_{10} = A\|C\|D\|F\|$	1	1	0	0	0	0	0	0	0	**1**	0	1	0	0	1	0
$C_{11} = B\|C\|D\|$	1	1	1	0	0	0	1	0	1	0	**1**	0	1	0	0	1
$C_{12} = A\|C\|D\|$	1	1	1	1	0	0	0	0	1	1	0	**1**	0	0	1	0
$C_{13} = B\|D\|$	1	1	1	0	0	0	1	0	1	0	1	0	**1**	0	1	0
$C_{14} = B\|F\|$	1	1	0	1	1	1	1	1	0	0	0	0	0	**1**	1	0
$C_{15} = A\|C\|$	1	1	1	1	0	1	0	0	1	1	0	1	1	1	**1**	1
$C_{16} = B\|C\|$	1	1	1	0	0	1	1	1	1	0	1	0	0	0	1	**1**

By default, *PRANC* uses MCFC tree as a starting tree. If the score for the usual greedy consensus tree is not maximal, *PRANC* will use both greedy and MCFC trees as starting trees. Because *PRANC* can output the MCFC tree along with usual greedy consensus tree implemented in *PHYLIP*, we expect that *PRANC* will outperform the *PHYLIP* in the moderate-size simulation as shown in Figure 4.3.

4.3 Topological tree space

Using an exhaustive tree search, we can find the species tree topology which will maximize the likelihood of the sample of gene trees. However, searching in whole

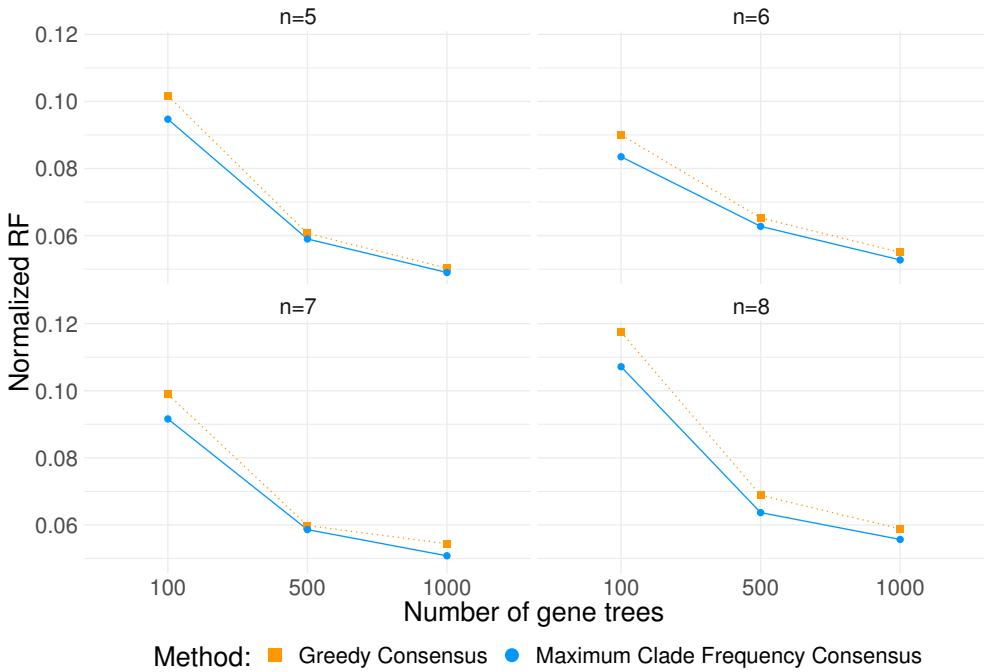


Figure 4.3: The normalized Robinson-Foulds distances between the estimated trees and true species trees. The greedy consensus and maximum clade frequency consensus (MCFC) methods were used to estimate the species tree. The pure birth process with $\lambda = 1$ was used to generate 1000 species trees for each number of species n, where $n = 5, 6, 7$, and 8.

tree space is computationally expensive. Thus, the feasible approach is to consider several trees that are topologically close to the starting species tree and compute the likelihood for these trees. If the better tree topology is found, the algorithm now uses this updated tree and considers several trees that are topologically close to it. This process is repeated until a better tree cannot be found or until the maximum number of iterations is reached.

Several methods exist that allow searching tree space through local pertubations to the starting tree. We will use a branch swapping rearrangment technique called *Nearest neighbor interchange* (NNI) (Felsenstein, 2004). For the n-taxon tree, one NNI move generates $2n - 4$ neighbor trees by swapping a subtree on one side of the

branch with a subtree on the other side. Figure 4.4 displays six rooted 5-taxon trees generated from the central tree with one step NNI.

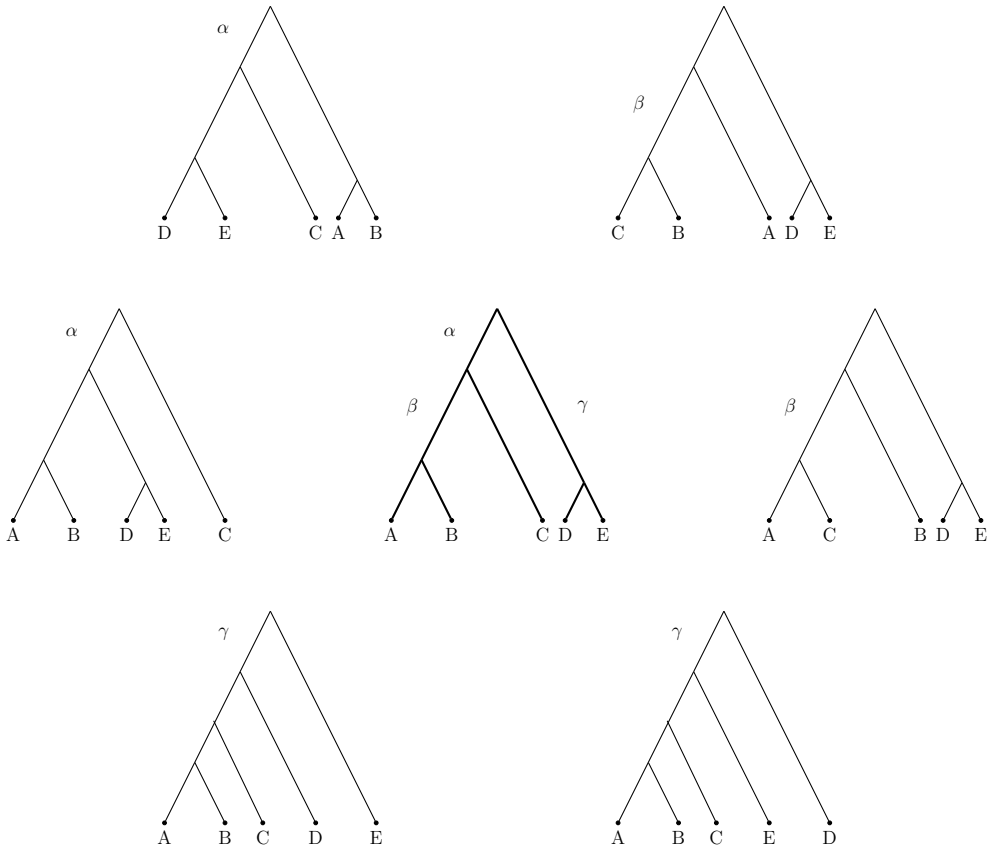


Figure 4.4: Six rooted trees within one NNI move from the central tree with branches α, β, and γ. To generate each tree within one NNI move, a certain branch in the central tree was fixed and then the corresponding subtrees were swapped.

One downside in exploring the tree space using NNI technique is that we may see the same tree more than once. For example, if one of the trees in Figure 4.4 is better than the central tree, we will select that tree and will generate 6 one step NNI trees from it. Of course, one of the trees will have the same topology as the central tree. Another downside of such heuristic methods, not only NNI, is that such methods are not guaranteed to find the optimal tree.

By default, *PRANC* explores species tree topologies within $k = 5$ NNI moves

from the initial starting tree.

4.4 Branch optimization

It is hard to maximize the likelihood (minimize the negative log likelihood) given in eq. (4.1) using derivatives as would be required using the standard optimization algorithms for finding the optimum like Newton-Raphson. Therefore, to optimize speciation inteval lengths in the tree we use Brent's algorithm (Brent, 1973) that does not use derivatives. However, Brent's method is not suitable well for the function minimization with respect to multiple variables because it can minimize the function of the one variable. It is faster to determine the direction of the gradient when all set of variables is used. We use a negative log-likelihood as a funtion to be minimized $f(x)$ and we assume that each speciation interval length $t_i \in [a, b] = [0.001, 6], i = 2, 3, \ldots, n-1$. We use a constant-rate birth-death model to generate the species trees. As shown in Table 3.4 under the pure birth process the average interval lengths are $\frac{1}{2\lambda}, \frac{1}{3\lambda}, \frac{1}{4\lambda}, \ldots$ coalescent units going forward in time from the root to the tips. For example, the average lengths for the six intervals for the 8-taxon tree generated under the Yule model with $\lambda = 0.1$ are vary between $\frac{1}{2 \cdot 0.1} = 5$ for the first interval near the root to $\frac{1}{7 \cdot 0.1} = 1.43$ for the last interval above the most recent cherry.

4.4.1 Brent's method

Brent's method combines the bisection, secant, and inverse quadratic interpolation methods. Brent's method seeks a local minimum of a function $f(x)$ in an interval $[a, b]$. The method finds an approximation x to the point at which f attains its minimum, and returns value of the function at x. The algorithm never evaluates f at two points closer than $abstol = \epsilon \cdot |x| + tol$, where tol is not smaller than

the machine precision and ϵ is usually not smaller than the square root of machine precision. We use $\epsilon = 1e-06$ and $tol = 1e-06$. The choice of $[a, b]$ has the biggest effect on *PRANC*'s execution time using Brent's method. By making both ϵ and tol bigger we may further reduce the execution time because the negative log-likelihood will be evaluated at smaller numbers of values.

At the beginning, *PRANC* takes the unranked ultrametric tree topology and initializes all interval lengths to 1.0. *PRANC* then randomly picks interval orders for optimization. After m rounds of such optimizations (by default, m is set to the number of taxa), the final tree is reported. Because only the interval lengths have an impact on the likelihood and not individual branch lengths, we can optimize the interval lengths first. We then select the cherry with the largest rank in the tree and set the length from the tips to their parent to 0.1 coalescent units. After this, all individual branch lengths of the optimal tree can be calculated.

4.4.2 Limited memory algorithm for bound constrained optimization

A limited memory algorithm for bound constrained optimization (L-BFGS) introduced in Byrd *et al.* (1995) is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) using a limited amount of computer memory. L-BFGS is based on the gradient projection method and uses a limited memory BFGS matrix to approximate the Hessian of the objective function. The method is a popular algorithm for parameter estimation in machine learning. The algorithm does not require the second derivative of the objective function. For each parameter the lower and upper bounds must be given. In our case, the same range $[0.001, 6]$ was given for the each interval. We can compute the function derivative numerically as h approaches zero

(h is set to $1e-12$) as shown in eq. (4.5):

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}. \tag{4.5}$$

We used the default stopping criteria defined in Byrd *et al.* (1995). Given that f is a function that calculates the negative log likelihood, we can minimize it using eq. (4.5) as the derivative. It should be noted that L-BFGS is better suited for the negative log likelihood minimization than the Newton-Raphson-like algorithms and Brent's method, because L-BFGS can minimize across multiple variables at the same time and the boundaries for the allowed values that parameters can take can be defined in L-BFGS. In this way we ensure that the negative log likelihood will not be computed at non-positive values.

4.5 Simulation

We simulate 100 $5-8$-taxon species trees under the Yule model with speciation rate $\lambda = 0.5$ and 1. We use *TreeSim* (Stadler, 2011) to generate species trees and *hybrid-lambda* (Zhu *et al.*, 2015) to simulate $100, 500, 1000$ gene trees for each species tree. We mainly compared *PRANC*'s performance with *ASTRAL* (Mirarab *et al.*, 2014) and *STELLS2* (Pei and Wu, 2017). *ASTRAL*, *STELLS2*, and *PRANC* estimate species trees from a sample of unrooted, unranked, and ranked gene tree topologies, respectively. *PRANC* computes maximum likelihood by using exact probabilities of ranked gene trees. *STELLS2* computes maximum likelihood by using approximate probabilities of unranked gene trees. There is an earlier version of *STELLS2*, i.e. *STELLS*, that uses exact probabilities of unranked gene trees. However, the author of the both programs claims in Pei and Wu (2017) that *STELLS2* performs comparably with *STELLS* in terms of inference but runs much faster. *ASTRAL* estimates the species tree by minimizing the quartet distance between unrooted gene tree topologies

and the species tree. It is very fast and has low estimation error. Because *ASTRAL* takes unrooted trees as an input, we first added an outgroup to the rooted species trees which is 10 coalescent units away from the other species. We then used *hybrid-lambda* to simulate gene trees. We ran *ASTRAL* on a set of unrooted trees. Then the estimated unrooted species trees inferred by *ASTRAL* were rooted on the outgroup, and the outgroup was dropped to get rooted trees without the outgroup.

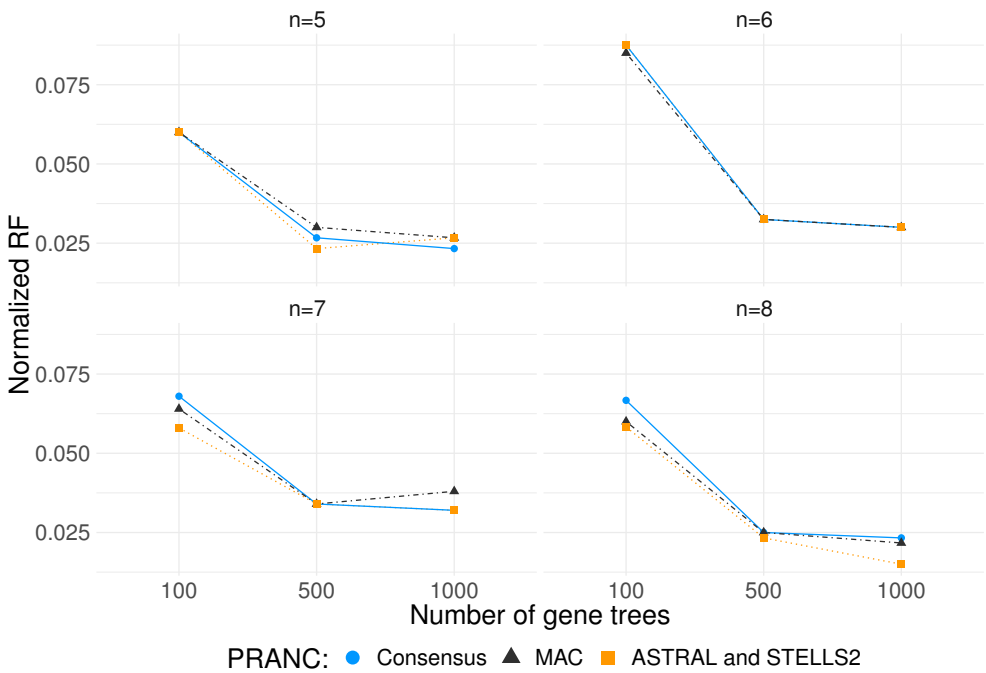


Figure 4.5: The normalized Robinson-Foulds distances between the inferred trees by *PRANC* and true species trees. The pure birth process with $\lambda = 1$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. Greedy consensus, tree with the lowest MAC score, and trees estimated by *ASTRAL* and *STELLS2* were used as starting trees for *PRANC*. Brent's branch optimization was used.

Figure 4.5 shows the normalized Robinson-Foulds distances between the inferred trees by *PRANC* and true species trees. The greedy consensus tree, the tree with the lowest MAC score, and trees estimated by *ASTRAL* and *STELLS2* were used as starting trees for *PRANC*. The interval lengths were optimized using Brent's method.

The pure birth process with $\lambda = 1$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. We can conclude that the greedy consensus tree is a reasonable choice for the starting tree.

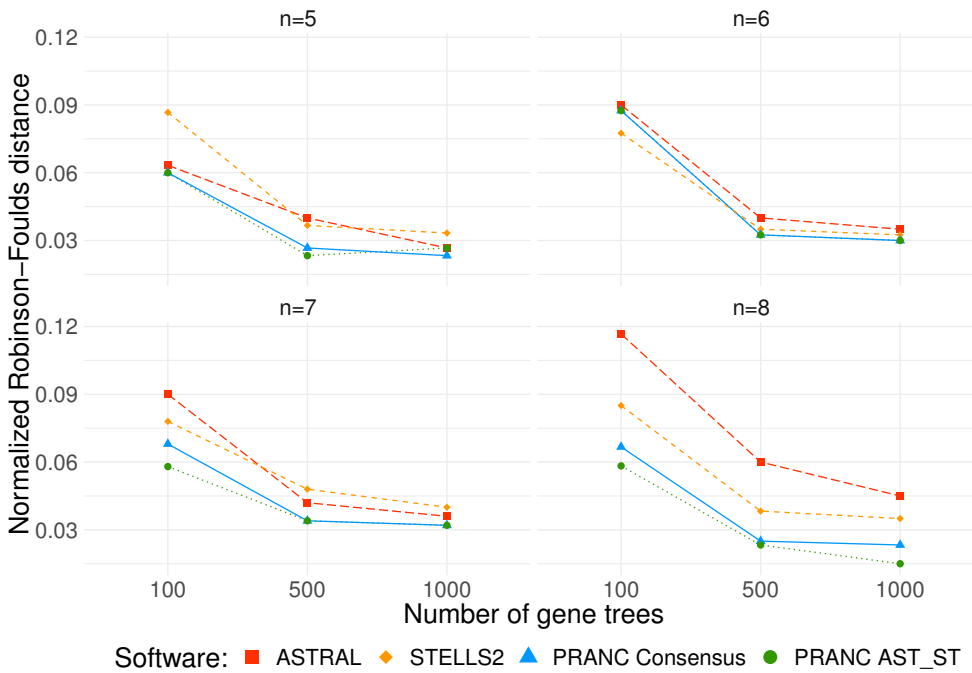


Figure 4.6: The normalized Robinson-Foulds distances between the inferred trees by *ASTRAL*, *STELLS2*, and *PRANC* and true species trees. The pure birth process with $\lambda = 1$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. Greedy consensus and trees estimated by *ASTRAL* and *STELLS2* were used as starting trees for *PRANC*. Brent's method was used to optimize branch lengths with $m = n$ rounds. We considered up to $2n$ rankings of starting tree and up to $2n$ rankings of each unranked species tree candidate. Because *ASTRAL* takes unrooted trees as an input, we first added an outgroup to the rooted species trees which is 10 coalescent units away from the other species. We then used *hybrid-lambda* to simulate gene trees. We ran *ASTRAL* on a set of unrooted trees. Then the estimated unrooted species trees inferred by *ASTRAL* were rooted on the outgroup, and the outgroup was dropped to get rooted trees without the outgroup.

Figures 4.6 and 4.7 show the normalized Robinson-Foulds distances between the inferred trees by *ASTRAL*, *STELLS2*, *PRANC* and true species trees. Both greedy consensus and maximum clade frequency consensus trees were used as starting trees

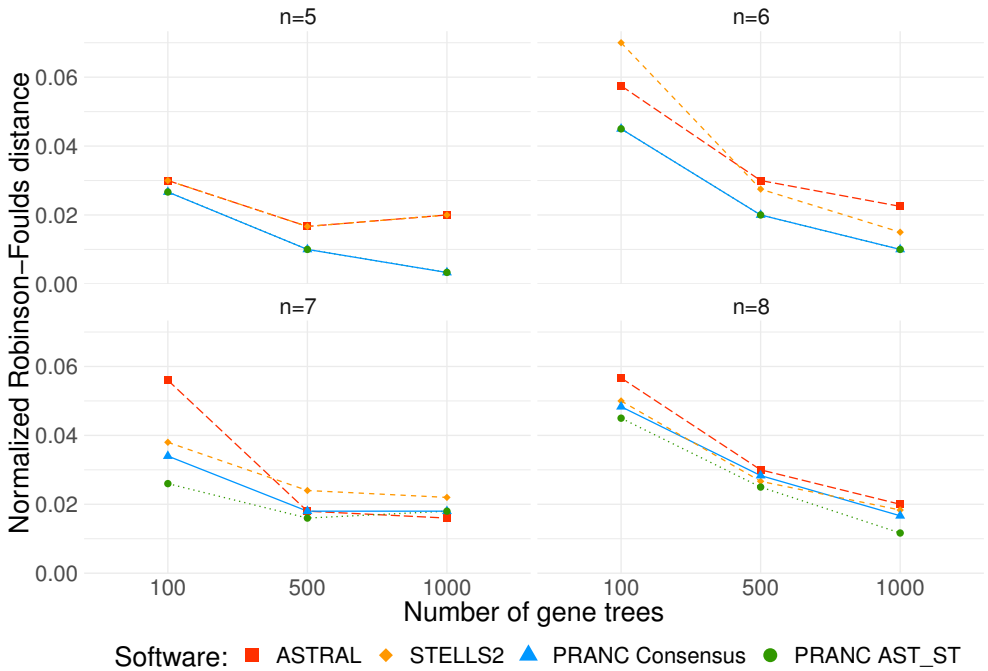


Figure 4.7: The normalized Robinson-Foulds distances between the inferred trees by *ASTRAL*, *STELLS2*, and *PRANC* and true species trees. The pure birth process with $\lambda = 0.5$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. Greedy consensus and maximum clade frequency consensus were used as starting trees for *PRANC*. L-BFGS method was used to optimize branch lengths. We considered all rankings of starting tree and up to $2n$ rankings of each unranked species tree candidate as described in Section 2 in the main text. Because *ASTRAL* takes unrooted trees as an input, we first added an outgroup to the rooted species trees which is 10 coalescent units away from the other species. We then used *hybrid-lambda* to simulate gene trees. We ran *ASTRAL* on a set of unrooted trees. Then the estimated unrooted species trees inferred by *ASTRAL* were rooted on the outgroup, and the outgroup was dropped to get rooted trees without the outgroup.

for *PRANC*. The interval lengths were optimized either by Brent's method or by L-BFGS method. In general, *PRANC* runs faster when using L-BFGS method than using Brent's method for the branch lengths estimation. The pure birth processes with $\lambda = 0.5$ and $\lambda = 1$ were used to generate 100 species trees for each number of species n, where $n = 5, 6, 7$, and 8. In most cases, *PRANC* outperforms both *ASTRAL* and *STELLS2*.

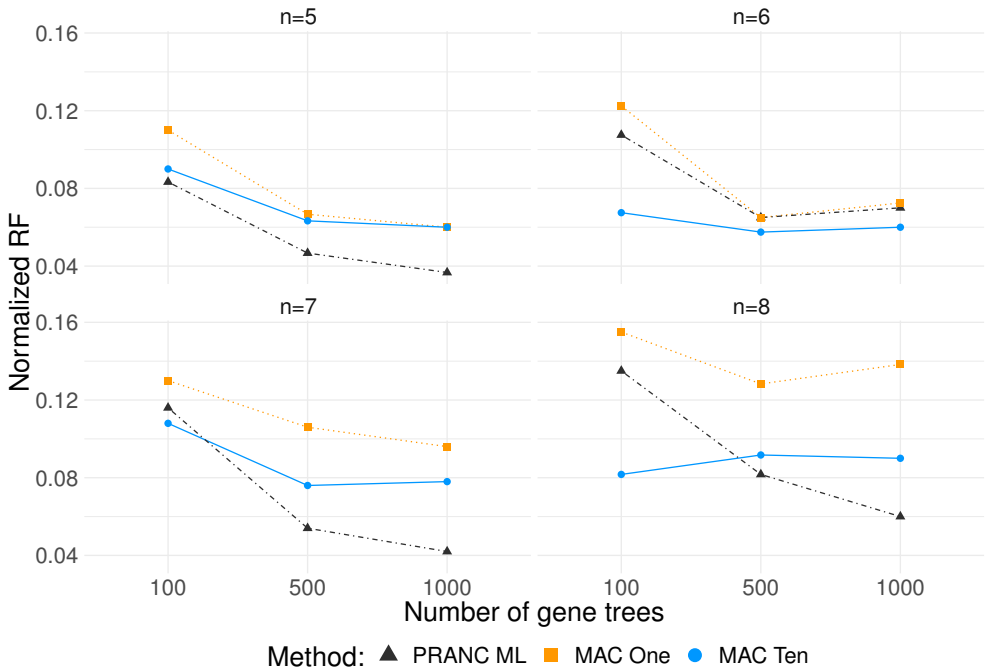


Figure 4.8: The normalized Robinson-Foulds distances between the inferred trees by MAC and *PRANC* and true species trees. The pure birth process with $\lambda = 1$ was used to generate 100 species trees for each number of species n, where $n = 5, 6, 7,$ and 8. *BEAST* was used to estimate gene trees from sequences, and then both MAC and *PRANC* were used to estimate the species tree. Brent's branch optimization was used.

Previously, we used simulated gene trees from the species tree. We decided to test the performance of *PRANC* in cases where gene trees might be estimated first from sequences. We simulated 100 species trees under the Yule process with $\lambda = 1$ for $n = 5, 6, 7,$ and 8 taxa. For each species tree we generated $100, 500,$ and 1000 gene trees, respectively. We used seqGen (Rambaut and Grass, 1997) to simulate sequences for each gene tree under the Generalised time-reversible (GTR) substitution model (Tavaré, 1986). GTR is the Markov model of DNA sequence evolution that assumes

different rates of substitution for each pair of nucleotides. GTR also assumes different frequencies of occurrence of nucleotides. *BEAST* (Suchard *et al.*, 2018) was used to estimate gene trees from the obtained sequences. Figure 4.8 shows the normalized Robinson-Foulds distances between the inferred trees and true species trees obtained by MAC and *PRANC*. The result is worse than the result depicted in Figure 4.5. This happens because we do not estimate species trees from true gene trees here, rather the species trees are estimated from the estimated gene trees. We do not account for the uncertainty in the gene trees here.

4.5.1 Rank dissimilarity measure between ranked topologies

We studied *PRANC*'s accuracy by computing the Robinson-Foulds distance of estimated and true species trees. The Robinson-Foulds (RF) metric is a way to measure the distance between unrooted or unranked trees. It cannot compare ranked trees. An interesting question to ask is how often *PRANC* can infer the correct ranked species tree topology?

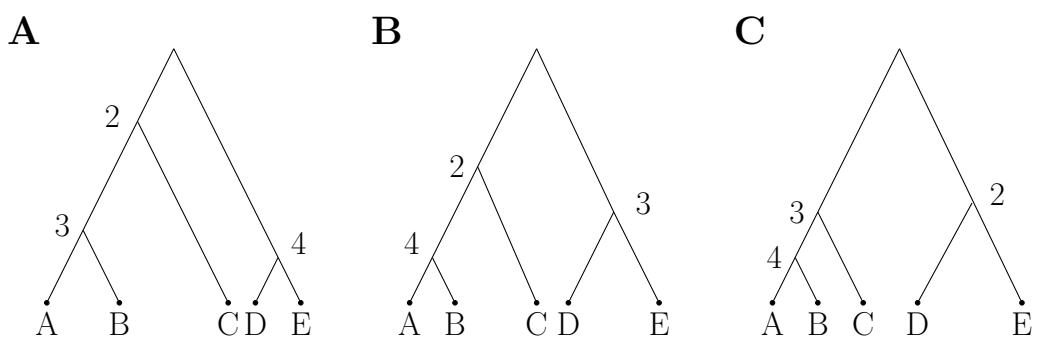


Figure 4.9: All possible rankings for the tree with unranked topology $(((A, B), C), (D, E))$. The rank dissimilarity measures between pairs of trees in (A)-(B), (A)-(C), (B)-(C) are 2, 6, and 2, respectively.

Given two n-taxon ranked trees with the same unranked topology, we define the rank dissimilarity measure d_R between trees as the square root of the sum of squares

Table 4.6: Squared rank dissimilarity measures of 100 estimated species trees that have the same unranked topologies as corresponding true species trees.

n	$\mathcal{G}_{num}$	0	2	4	6	8	10	12
5	100	94	6					
	1000	98	2					
6	100	65	27	4	3	0	1	
	1000	85	14	0	1			
7	100	65	27	4	3	0	1	
	1000	75	23	1	1			
8	100	41	36	9	10	3	0	1
	1000	59	27	3	7	4		

of the ranks of nodes that have the same set of descendant species (eq. (4.6)). The rank measure of zero indicates that two trees have same ranked topology.

$$d_R = \sqrt{\sum_{i=1}^{n-2}(U_i - V_i)^2}, \tag{4.6}$$

where U_i and V_i are ranks of the nodes that have the same set of descendant species. Note that the rank measure is defined for the trees with same unranked topology. In cases, when trees have different unranked topologies, it is sufficient to only use RF distance to compare unranked topologies.

For trees shown in Figure 4.9(A) and 4.9(B), the squared rank measure is

$$d_R^2 = (U_{AB}-V_{AB})^2+(U_{ABC}-V_{ABC})^2+(U_{DE}-V_{DE})^2 = (3-4)^2+(2-2)^2+(4-3)^2 = 2,$$

the value of 2 is indicating that the ranks of two nodes are swapped in the corresponding trees. For trees shown in Figure 4.9(A) and 4.9(C), the squared rank dissimilarity measure is

$$d_R^2 = (U_{AB}-V_{AB})^2+(U_{ABC}-V_{ABC})^2+(U_{DE}-V_{DE})^2 = (3-4)^2+(2-3)^2+(4-2)^2 = 6.$$

In this particular case, all ranks of corresponding nodes are different. Table 4.6 shows squared rank dissimilarity measures of 100 estimated species trees that

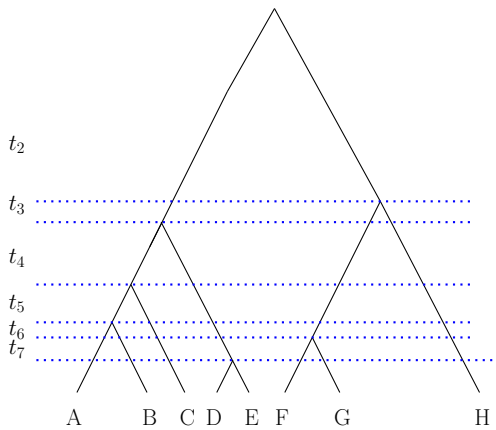
have the same unranked topologies as true species trees. *PRANC* can recover the correct ranked species tree topology in most cases. It is harder to estimate the correct ranked topology for larger number of species because the number of possible rankings increases with the number of species. The results from Table 4.6 indicate that the likelihood of the gene trees computed given the species tree with the correct unranked but incorrect ranked topology may still be bigger than the likelihood obtained under any incorrect unranked topology.

4.5.2 Accuracy of the interval lengths and internal branch lengths estimation

As shown in the previous sections, *PRANC* is able to recover the correct unranked topology of the species tree accurately. *PRANC* also does a good job in recovering the correct ranked topology. Another question that may be asked is how well *PRANC* estimates the interval or internal branch lengths corresponding to the evolutionary times between speciation events.

Figure 4.10 shows estimated coalescent interval lengths of a particular 8-taxon species tree generated under the Yule process with $\lambda = 0.5$. *PRANC*, *STELLS2*, and *ASTRAL* were used to estimate interval lengths t_i, $i = 2, ..., 7$. The interval lengths must be inferred from the ultrametric trees. Because both *STELLS2* and *ASTRAL* report non-ultrametric trees, we make them ultrametric by extending all the external branches of the inferred trees to match the external branch with the greatest total time. For example, we can extend external branches in the tree in Figure 1.1(D) to obtain an ultrametric tree displayed in 1.1(A).

PRANC primarly estimates interval lengths and then calculates internal branch lengths from them. For convenience, the time of the most recent clade is set to 0.1 but could be set to any other value because it does not affect the probabilities. The



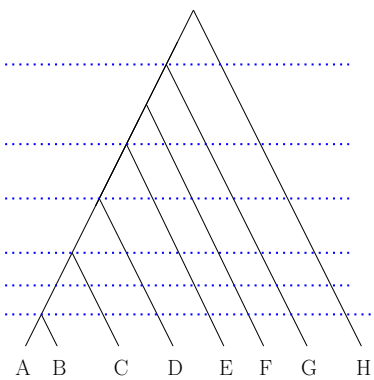
	True	PRANC	STELLS2	ASTRAL
t_2	1.5695	1.6936	1.5622	1.2799
t_3	0.3132	0.0010	0.1482	0.1953
t_4	0.2759	0.4478	0.8400	0.5723
t_5	0.2508	0.2312	0.3691	0.2362
t_6	0.5929	0.6370	0.2179	0.0319
t_7	0.0176	0.0016	0.0674	0.2836
MSE		0.0221	0.2140	0.2238

Figure 4.10: Estimated coalescent interval lengths $t_i, i = 2, \ldots, 7$ of the 8-taxon species tree. *PRANC*, *STELLS2*, and *ASTRAL* were used to estimate the species tree from 1000 gene trees. The best estimated length for each interval is shown in blue. *PRANC* gives the lowest mean squared error (MSE). Note that the correct ranked topology of the species tree was given to *PRANC*.

times of the nodes are noninformative. For example, in Figure 4.10 the estimated interval length between the parent of the clade (D, E) with rank 7 and parent of the clade (F, G) with rank 6 is 0.001625. If we set the time of the most recent clade (D, E) to 0.1 then the time of (F, G) will be calculated as $0.1 + 0.001625 = 0.101625$, and the external branch lengths of both species F and G will be set to 0.101625 coalescent units. However, the lengths of external branches do not affect the lengths of the internal branches that are of main interest in phylogenetic inference. In this example, the internal branch length for the clade (F, G) is calculated as the sum of interval lengths t_3, t_4, t_5, and t_6 which *PRANC* aims to estimate either using L-BFGS or Brent's methods.

Figure 4.11 shows estimated branch lengths between true tree and estimated trees by *PRANC*, *STELLS2*, and *ASTRAL*, respectively. Note that for caterpillar trees the internal branch lengths and the lengths of the corresponding intervals are the same. Figures 4.10 and 4.12 show estimated interval lengths and internal branch lengths between true tree and estimated trees by three programs. Note that *PRANC*

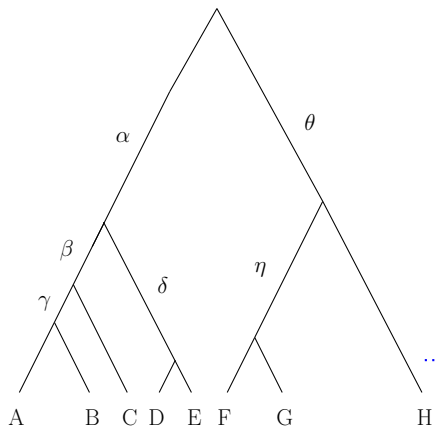
estimated the branch lengths for the correct ranked topology. *PRANC* gives lowest mean squared error of the interval/branch lengths of this single tree. Figure 4.13 shows estimated branch lengths in the case when *PRANC* estimated the incorrect ranked topology. In particular, ranks of the three nodes in the estimated tree are different from that in the true species tree. Still *PRANC* gives the lowest mean squared error.



True length	PRANC	STELLS2	ASTRAL
0.3361	0.2930	0.5491	0.3325
0.8951	0.8721	1.0313	0.9165
0.2817	0.3002	0.5077	0.2874
0.2124	0.2073	0.3735	0.2055
0.1520	0.1300	0.3190	0.1432
0.0441	0.0939	0.2649	0.0482
MSE	0.00011	0.03627	0.00095

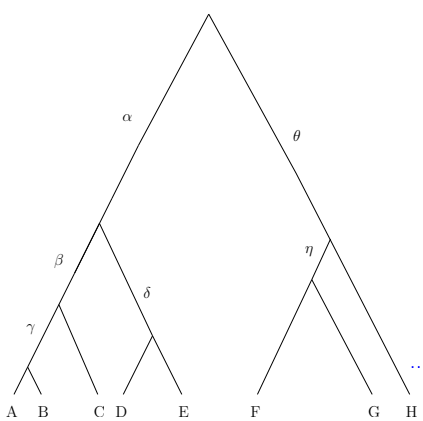
Figure 4.11: Estimated internal branch lengths of the 8-taxon caterpillar-shape species tree. *PRANC*, *STELLS2*, and *ASTRAL* were used to estimate the species tree from 1000 gene trees. On average, *PRANC* gives better branch lengths estimates which is reflected in the lowest mean squared error.

On average, will *PRANC* estimate lengths more accurately than *ASTRAL* and *STELLS2*? To answer this question, we consider 100 estimated species trees by *PRANC*, *STELLS2*, and *ASTRAL*. Note that all 100 inferred trees have same unranked topologies as their corresponding species trees. In particular, we generated 100 five- and eight-taxon trees under the Yule model with the birth rate $\lambda = 0.5$ and $\lambda = 1$. In each case, 100 or 1000 gene trees were generated from each species tree and have been used to estimate the species tree internal branch lengths. For each inferred tree, we calculated the mean squared error between true and estimated internal branch lengths. We then report average, standard deviation, minimum, and maximum of the mean squared errors averaged over the 100 species trees considered.



True length	PRANC	STELLS2	ASTRAL
$\alpha = 1.5695$	1.6946	1.5622	1.2799
$\beta = 0.5891$	0.4478	0.9882	0.7676
$\gamma = 0.2508$	0.2312	0.5870	0.2681
$\delta = 1.4328$	1.3176	1.6425	1.3193
$\eta = 1.1372$	1.3170	1.2090	0.8085
$\theta = 1.8827$	1.6936	1.7104	1.4753
MSE	0.0196	0.0585	0.0672

Figure 4.12: Estimated internal branch lengths α, β, γ, etc. of the 8-taxon species tree. *PRANC*, *STELLS2*, and *ASTRAL* were used to estimate the species tree from 1000 gene trees. The best estimated length for the each internal branch is shown in blue. Although *PRANC* gives the better estimates for the majority of intervals in Figure 4.10, it does not give the most closest estimates for the majority of the branch lengths. However, on average, *PRANC* gives better branch lengths estimates which reflected in the lowest mean squared error (MSE).



True length	PRANC	STELLS2	ASTRAL
$\alpha = 0.2785$	0.2579	0.4000	0.2293
$\beta = 0.3561$	0.3569	0.4859	0.3975
$\gamma = 0.2218$	0.2238	0.5223	0.2298
$\delta = 0.4168$	0.3806	0.5887	0.4349
$\eta = 0.0632$	0.1064	0.3251	0.0512
$\theta = 0.5636$	0.5330	0.6319	0.6133
MSE	0.00076	0.03745	0.00119

Figure 4.13: Estimated internal branch lengths α, β, γ, etc. of the 8-taxon species tree. *PRANC*, *STELLS2*, and *ASTRAL* were used to estimate the species tree from 1000 gene trees. Note that *PRANC* estimated the incorrect ranked topology. In particular, ranks of the three nodes $((A, B), C)$, (D, E), and (F, G) in the estimated tree are different from that in the true species tree. Still *PRANC* gives the lowest mean squared error.

Table 4.7: Average of 100 mean squared errors of the internal branch lengths between estimated and true trees. On average, among three software considered, *STELLS2* gives worst estimates, both *ASTRAL* and *PRANC* give good estimates for the branch lengths. Among three programs, *PRANC* gives best estimates.

λ	G_{num}	n	statistic	*PRANC*	*STELLS2*	*ASTRAL*
$\lambda = 0.5$	100	5	$\bar{x}$	0.08007	2.66780	0.12940
			s	0.21642	18.944	0.41323
			$[min(x), max(x)]$	[0.00017,1.59254]	[0.01874,187.785]	[0.00057,3.55976]
		8	$\bar{x}$	0.02968	0.37330	0.06491
			s	0.06807	1.38995	0.14601
			$[min(x), max(x)]$	[0.00060,0.62408]	[0.01352,10.7410]	[0.00063,1.32526]
	1000	5	$\bar{x}$	0.00702	0.07543	0.02759
			s	0.01716	0.05037	0.13813
			$[min(x), max(x)]$	[0.00008,0.10929]	[0.02791,0.37597]	[0.00003,1.30013]
		8	$\bar{x}$	0.00267	0.10178	0.00586
			s	0.00495	0.35839	0.01346
			$[min(x), max(x)]$	[0.00004,0.02916]	[0.02561,3.63738]	[0.00017,0.11605]
$\lambda = 1$	100	5	$\bar{x}$	0.01552	0.06637	0.01944
			s	0.02238	0.0353	0.02821
			$[min(x), max(x)]$	[0.00017,0.13902]	[0.00924,0.18836]	[0.00013,0.18083]
		8	$\bar{x}$	0.00684	0.05539	0.01431
			s	0.00698	0.02953	0.01362
			$[min(x), max(x)]$	[0.00069,0.04875]	[0.01555, 0.17789]	[0.00134, 0.08833]
	1000	5	$\bar{x}$	0.00151	0.05390	0.00237
			s	0.00347	0.01936	0.00414
			$[min(x), max(x)]$	[0.00002,0.02936]	[0.02338,0.12207]	[0.00002,0.02291]
		8	$\bar{x}$	0.00077	0.04169	0.00154
			s	0.00109	0.01684	0.00222
			$[min(x), max(x)]$	[0.00002,0.00794]	[0.01578,0.10378]	[0.00016,0.01788]

The results are shown in Table 4.7.

On average, *PRANC* estimates branch lengths more accurately than *ASTRAL* and *STELLS2*. As expected, using 1000 gene trees instead of 100 trees results in more accurate estimates for all three programs. It is interesting that, on average, *ASTRAL* does a better job in lengths estimation than *STELLS2* which preserves more information about rooted trees.

Table 4.8 shows the average differences between corresponding interval lengths of the inferred trees and the species trees when 1000 gene trees were used for estimation. We observed that *PRANC* does not have difficulty in estimating certain interval lengths when a sufficient number of gene trees is given. However, we note that *PRANC* provides slightly better estimates for the interval lengths closest to the root

Table 4.8: Average difference between corresponding interval lengths of the inferred tree and species tree using 1000 gene trees.

n	λ	t_2	t_3	t_4	t_5	t_6	t_7
5	0.5	0.0501	0.0341	0.0366			
	1	0.0213	0.0210	0.0259			
8	0.5	0.0425	0.0285	0.0237	0.0215	0.0230	0.0272
	1	0.0168	0.0138	0.0137	0.0151	0.0172	0.0202

when $\lambda = 1$ than when $\lambda = 0.5$.

4.5.3 Discussion

One of the main issues during the phylogenetic inference is the noise in input data that may come from the mutation and incomplete lineage sorting processes (Maddison, 1997). It was observed by Huang *et al.* (2010) that using more information contained in gene trees (e.g., topology and branch lengths as opposed to just topology) may not improve accuracy. So far, most coalescent-based methods were using either gene tree topologies alone or both topologies and coalescent times. The approach of using ranked gene trees may be viewed as an intermediate between these other two approaches. The hope is to develop methods based on the ranked gene trees that can handle the noise in gene trees.

We introduce *PRANC* software that has different options to work with the ranked gene trees. *PRANC* can analytically compute the probabilities of ranked gene trees given a species tree in Newick format, following eq. (1.7). The program has an option to compute the probability of an unranked gene tree by summing the probabilities of all ranked gene trees that share the corresponding unranked topology. We improved the numerical results by adding the probabilities of the ranked histories in ascending order, enabling the smallest-magnitude values to accumulate before interacting with

larger-magnitude values.

PRANC can be used for inference as well. It performs a heuristic search from the initial trees to find a maximum likelihood species tree. There is a trade-off between *PRANC*'s estimation accuracy and its computational time. The speed of the program mainly depends on the choice of initial tree and the number of rankings considered for each unranked species tree candidate. We tested *PRANC*'s performance under different settings. In general, it is sufficient to consider $2n$ rankings for each unranked n-taxon species tree candidate. More rankings can be considered to improve accuracy at the expense of speed. Both greedy consensus and introduced maximum clade frequency consensus can be used as starting trees (both implemented in *PRANC*). By default, *PRANC* searches for a better tree within 5 NNI moves from the starting tree. L-BFGS is used to optimize branch lengths. On average, under these settings, *PRANC* can estimate an 8-taxon species tree from 100 gene trees in less than 5 minutes. It takes seconds, a few minutes, ten-fifteen minutes, and an hour to estimate a $5, 6, 7$- and 8-taxon species tree from 1000 gene trees, respectively.

Chapter 5

Empirical Studies

We apply the methods we propose to two real biological datasets. Anomalous gene trees are sometimes serve as a potential explanation for conflicting relationships in phylogenomic studies. An empirical anomaly zone has been reported in Scincidae skinks study (Linkem *et al.*, 2016) and in an empirical study of Palaeognathae flightless birds (Cloutier *et al.*, 2019). We examined a 16-taxon Scincidae skink phylogeny using heuristic methods proposed in Section 3.3. We evaluated the performance of maximum likelihood inference based on ranked gene trees (*PRANC*) by application to a real biological dataset of gibbons (Carbone *et al.*, 2014; Veeramah *et al.*, 2015). We compare the computational and accuracy performance of *PRANC* with that of *BPP* (Yang, 2015), *STELLS2* (Pei and Wu, 2017), and *ASTRAL* (Mirarab *et al.*, 2014). *STELLS2* and *ASTRAL* infer a species tree from unranked and unrooted gene tree topologies, respectively, whereas *BPP* infers a phylogenetic tree directly from DNA sequences.

5.1 The anomaly zone of skinks

Linkem *et al.* (2016) used the coalescence based *MP-EST* (Liu *et al.*, 2010) method to estimate a species tree from 429 estimated gene trees. They then used the inferred species tree topology from *MP-EST* and estimated median branch lengths by summarizing the posterior distribution from *BPP* (Yang, 2015). The estimated 16-taxon skink phylogeny by Linkem *et al.* (2016) (see Table 3 in Linkem *et al.* (2016)) is shown in Figure 5.1 and illustrated below in the Newick format:

```
(((((((Mesoscincus_manguae:2.704236,Scincus_scincus:2.704236):0.26527,
(Eurylepis_taeniolatus:2.945401,Chalcides_ocellatus:2.945401):0.024105):
0.10778,(Plestiodon_fasciatus:2.948328,Ophiomorus_raithmai:2.948328):
0.128958):0.14189,Brachymeles_bonitae:3.219176):0.79165,
(((Emoia_caeruleocauda:2.97096,Lygosoma_brevicaudis:2.97096):0.309054,
Mabuya_unimarginata:3.280014):0.701966,(((Sphenomorphus_variegatus:1,
Lobulia_elegans:1):0.49652,Tytthoscincus_parvus:1.49652):2.02657,
Sphenomorphus_tridigitus:3.52309):0.45889):0.028846):0.005778,
Typhlosaurus_sp:4.016604):0.781678,Xantusia_vigilis:4.798282);
```

We computed unranked probabilities of the $2n - 4 = 28$ gene trees that are one NNI step away from the species tree. We found that there are four anomalous gene trees, and the probability of the unranked matching tree, $1.3987e-06$, is lower than the probability of the most probable unraked nonmatching tree, $2.66115e-06$, which confirms that this species tree is in the unranked anomaly zone.

We also found that the probability of the unrooted matching tree $8.972783e-06$ is lower than the probability of the highest probability unrooted nonmatching tree $1.719202e-05$, which confirms that this species tree is also in the unrooted anomaly zone. Among $2n-6 = 26$ one step NNI 16-taxon unrooted trees, four are anomalous.

However, computing probabilities of the $73,920$ 16-taxon ranked gene trees that

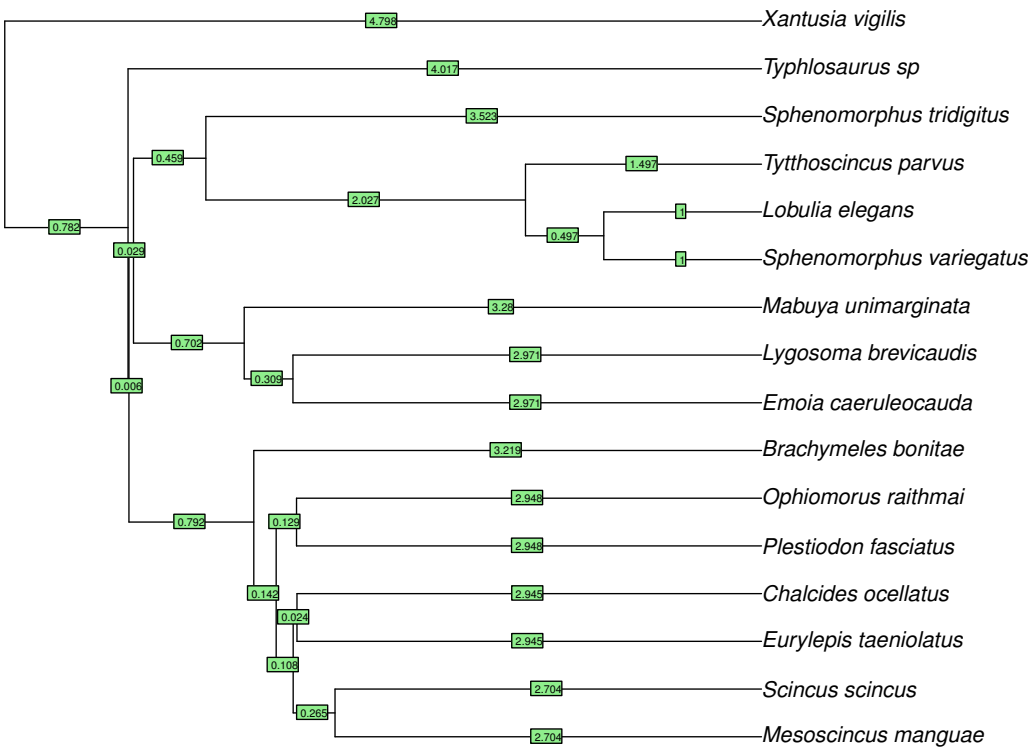


Figure 5.1: The estimated 16-taxon skink phylogeny by *MP-EST*. The tree topology and estimated internal branch lengths (in coalescent units) is based on the results from Linkem *et al.* (2016). We extended external branches to obtain an ultrametric tree. The external branch lengths are noninformative.

share the same unranked topology as species tree did not indicate that the species tree is in the ranked anomaly zone. The most probable ranked gene tree the has same ranked topology as the species tree and this probability is $8.166337e-09$.

This result agrees with the simulation results in Section 3.5 that unranked and unrooted anomaly zones are more closely related with each other than either one of them with ranked anomaly zone.

5.2 DNA sequences for five gibbons

We explore the performance of *PRANC* in inferring correct species tree for a genome-scale dataset consisting of 5 species of gibbons: Hoolock leuconedys (B, HLE), Hylobates moloch (Hm, HMO), Hylobates pileatus (Hp, HPL), Nomascus leucogenys (N, NLE), and Symphalangus syndactylus (S, SSY). This dataset was generated by Carbone *et al.* (2014) and Veeramah *et al.* (2015) and consists of $12,143$ noncoding loci, each of $1,000$ bp in length. Except for the human outgroup (O, hg19) representing a human genome, multiple lineages were sampled for each species: 2 for HMO and for HPL, and 4 for HLE, NLE, and SSY. The sample locus is shown below:

```
17 1000
HPL_Domino.A    GGTCCATGCAACACTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HPL_Domino.B    GGTCCATGCAACACTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HLE_Drew.A      GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HLE_Drew.B      GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
SSY_Monty.A     GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HLE_Maung.A     GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
SSY_Monty.B     GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HLE_Maung.B     GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
SSY_Karenina.A  GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
SSY_Karenina.B  GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HMO_Madena.B    GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
NLE_Asteriks.B  GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
hg19            GGTAAATGCAACATTGAGTGTGAAATTGATAAATAGTCCTCAA..
HMO_Madena.A    GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
NLE_Asteriks.A  GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
NLE_Vok.B       GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
NLE_Vok.A       GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
```

We selected one lineage per species for the further analysis: HPL_Domino.A, HLE_Drew.A, SSY_Karenina.A, HMO_Madena.A, NLE_Vok.A, and outgroup hg19. The sequences of the selected subset of five species plus the outgroup are shown below:

```
6 1000
HPL_Domino.A    GGTCCATGCAACACTGAGTGTGAAATTGATAAGTAGTCCTCAA..
HLE_Drew.A      GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
SSY_Karenina.A  GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
hg19            GGTAAATGCAACATTGAGTGTGAAATTGATAAATAGTCCTCAA..
HMO_Madena.A    GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
NLE_Vok.A       GGTCCATGCAACATTGAGTGTGAAATTGATAAGTAGTCCTCAA..
```

We used *ASTRAL* (Mirarab *et al.*, 2014), *STELLS2* (Pei and Wu, 2017), and *PRANC* to estimate species trees from a sample of unrooted, unranked, and ranked gene tree topologies, respectively. The Bayesian method *BPP* (Yang, 2015) was used to estimate species trees directly from such noncoding loci consisting of the alignments of six species.

We used *IQ-TREE* (Nguyen *et al.*, 2014) software to estimate 6-taxon unrooted gene trees from DNA sequences under the general time reversible substitution model (*GTR*) with unequal rates and unequal base frequency (Tavaré, 1986). These unrooted trees were used as an input for *ASTRAL* (Mirarab *et al.*, 2014). Then the estimated unrooted species tree by *ASTRAL* was rooted on the outgroup, and the outgroup was dropped to get a rooted 5-taxon species tree estimate.

We rooted the obtained 6-taxon unrooted trees from *IQ-TREE* on the human outgroup hg19, and then droped the outgroup to get 5-taxon rooted gene trees and make them ultrametric.

We cannot assume that the molecular clock assumption is reasonable for any

rooted tree so we need test for the validity of this assumption. Under the null hypothesis, the branch lengths of the rooted tree are constrained such that all of the tips can be drawn at a single time plane. Under the alternative hypothesis, each branch is allowed to vary independently. The likelihood ratio test statistic is $-2\log L = 2(\log L_0 - \log L_1) \sim \chi_{n-2}$, where n is the number of tips, and L_0 and L_1 are the likelihoods under the null and alternative hypotheses, respectively (Felsenstein, 2004). To conduct the likelihood ratio test, we run DNA maximum likelihood programs *Dnaml* and *Dnamlk* (Felsenstein, 2013). For each locus, we obtain the unrooted tree topology from *IQ-TREE*. The outgroup is then dropped. The sequences with the outgroup dropped are used to infer the tree again using both *Dnaml* (not assuming a clock) and *Dnamlk* (assuming a clock). The tree searches using Dnaml and Dnamlk use the tree topology found from *IQ-TREE* as a constraint, so that branch lengths are optimized to maximize the likelihood with and without a clock. This allows testing each gene tree for the molecular clock assumption using the likelihood ratio test. If the p-value is greater than 0.05, we keep the tree with the estimated topology by *IQ-TREE* and estimated branch lengths by *Dnamlk* for further analysis. Using the likelihood ratio test, we obtained $10,706$ ultrametric trees out of $12,143$ loci. We use *STELLS2* and *PRANC* to estimate a species tree from these rooted gene trees. The greedy consensus tree was used as the starting tree for *PRANC*.

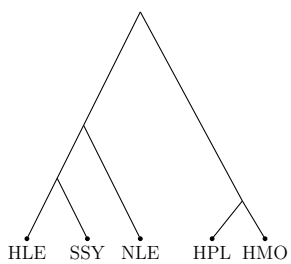


Figure 5.2: The species tree topology estimated from the gibbon data from all non-coding loci using *BPP* by Shi and Yang (2018).

Since it is a real dataset we do not know a true species tree. However, Shi and Yang (2018) estimated the species tree from all loci using *BPP* displayed in Figure 5.2, therefore we refer to this tree as the correct tree and compare all our estimates to it. Note that we run *BPP* on slightly different data since we sampled one lineage per species, thus reducing the original dataset from 16 lineages to 5 plus human outgroup.

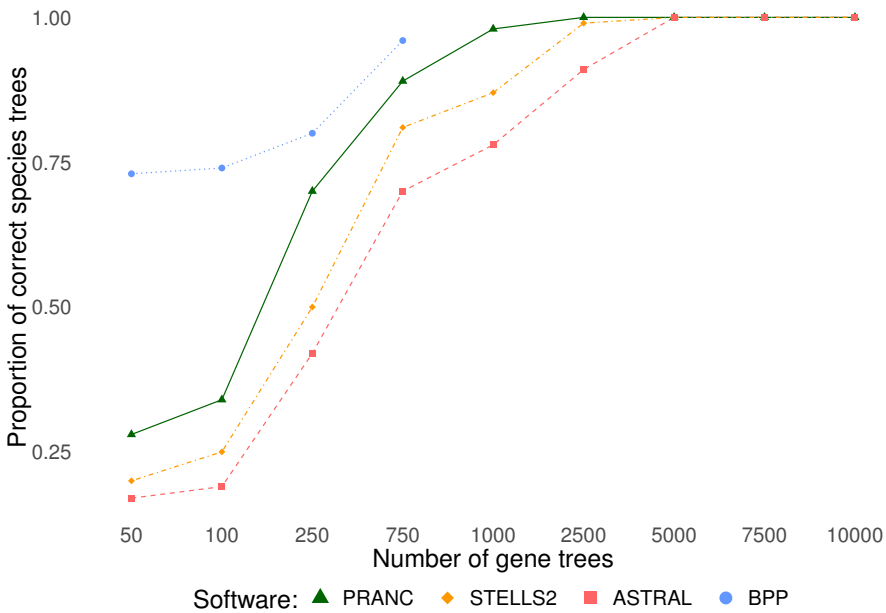


Figure 5.3: The proportion of correct species trees in a gibbon dataset obtained by four different methods plotted against the number of gene trees. The gene trees were considered as ranked for *PRANC*, unranked for *STELLS2*, and unrooted for *ASTRAL*. The greedy consensus tree was used as a starting tree for *PRANC*. To get an estimated rooted species tree from *ASTRAL*, we added an outgroup to the 5-taxon unranked gene trees. Then the estimated unrooted species tree by *ASTRAL* was rooted on the outgroup, and the outgroup was dropped to get a rooted 5-taxon tree. DNA sequences and no trees were used for *BPP*. The results for *BPP* were computed using up to 750 gene trees.

Figure 5.3 shows the proportion of correct species trees in a gibbon dataset obtained by four methods plotted against the number of gene trees. To select a subset of k gene trees out of $10,706$ trees, we first randomly select k indicies $\in [1, 10,706]$.

We then select sequences, unrooted trees, and rooted trees corresponding to these indicies. This is done to run four different programs on the same subsets of gene trees (sequences corresponding to these trees for *BPP*).

We observed that all methods converge as the number of genes increases to the correct species tree topology. The results shown in figure 5.3 are intuitive. As expected, for small number of genes, the Bayesian *BPP* that calculates the posterior probabilities of different species trees from DNA sequences is able to recover more correct species trees than other three programs. Unrooted, unranked, and ranked trees preserve increasing amounts of information about the rooted trees with specified branch lengths, respectively.

In terms of approximate computational time, *BPP* can estimate a 5-taxon species tree from 100 loci in about 1 hour and 500 loci in about $3-4$ hours whereas it takes seconds to estimate a 5-taxon species tree from 100 or 500 gene trees with *ASTRAL*, *STELLS2*, and *PRANC*.

Table 5.1 shows *PRANC*, *STELLS2*, and *ASTRAL* estimates of the internal branch lengths and the 95% confidence intervals of bootstrap estimates. We calculated 1000 bootstrap estimates using a sample of $10,706$ gene trees. All *PRANC* estimates do not agree with *STELLS2* estimates. *ASTRAL* gives similar estimates for the two branch lengths to *PRANC* estimates and one estimate that is close to the *STELLS2* estimate. Because true lengths are unknown, we can not judge which estimates are closer to the true values.

Table 5.1: Estimates of internal branch lengths in coalescent units of the species tree topology in Figure 5.2. The 95% confidence intervals of internal branch lengths are calculated by the bootstrap method. *PRANC*, *STELLS2*, and *ASTRAL* were used to get 1000 bootstrap estimates. Due to high computational time *BPP* was not used. Top row in this table represents clades for which internal branch lengths were estimated.

method	statistic	(HLE,SSY)	((HLE,SSY),NLE)	(HPL,HMO)
PRANC	$\hat{\theta}$	0.0382	0.0493	1.3930
	$(\hat{\theta}_{(0.025)}, \hat{\theta}_{(0.975)})$	(0.0271,0.0496)	(0.0401,0.0576)	(1.3617,1.4321)
STELLS2	$\hat{\theta}$	0.2176	0.2251	1.7873
	$(\hat{\theta}_{(0.025)}, \hat{\theta}_{(0.975)})$	(0.2048,0.2260)	(0.2167,0.2340)	(1.7446,1.8320)
ASTRAL	$\hat{\theta}$	0.0377	0.0378	1.7954
	$(\hat{\theta}_{(0.025)}, \hat{\theta}_{(0.975)})$	(0.0252,0.0503)	(0.0260,0.0490)	(1.7498,1.8426)

Chapter 6

Conclusions and Future work

A probability distribution of gene trees defines a unique species tree under the multi-species coalescent model. Many statistically consistent methods have been developed to estimate a species tree from a collection of gene trees (Kubatko *et al.*, 2009; Liu *et al.*, 2009b, 2010; Liu and Yu, 2011; Wu, 2012; Mirarab *et al.*, 2014; Pei and Wu, 2017). Surprisingly, for certain species tree topologies and branch lengths, the most likely gene tree can be different from the species tree. Such a tree is called an anomalous gene tree (Degnan and Rosenberg, 2006). The existence of anomalous gene trees gave insight for finding when other methods of species tree estimation could be misleading in regions of branch length space resembling, but not identical to, the anomaly zone (Kubatko and Degnan, 2007; Degnan *et al.*, 2009b; Wang and Degnan, 2011; Than and Rosenberg, 2011b). Although likelihood-based methods, including Bayesian methods (Liu and Pearl, 2007; Heled and Drummond, 2010; Flouri *et al.*, 2018), are not misled by anomalous gene trees, the recognition of the possibility of such trees, and especially that they do not exist for three taxa on rooted trees or four taxa on unrooted trees, motivated the development of numerous two-staged methods using rooted triples or quartets (e.g., Ewing *et al.*, 2008; DeGiorgio and Degnan, 2010; Liu *et al.*, 2010; Larget *et al.*, 2010; Mirarab *et al.*, 2014). The concept of the

anomaly zone has also been useful for designing simulation studies to test species tree inference methods in challenging regions of parameter space (Kubatko and Degnan, 2007; Liu and Edwards, 2009; Liu *et al.*, 2009c; DeGiorgio and Degnan, 2010; Shekhar *et al.*, 2018). Although the theoretical possibility of anomalous gene trees has motivated many methods, the extent that they arise in practice is less clear. We address this question by estimating how often anomalous gene trees occur under the widely-used birth-death models of speciation. We consider three types of anomaly zones, each corresponding to different types of gene trees: unrooted, unranked, and ranked gene trees. The study of various types of anomaly zones can lead to the discovery of the cases when such zones do not overlap with each other. Because the number of possible tree topologies grows faster than exponentially with the number of species, it is necessary to propose reasonable heuristic approaches to infer whether larger species trees (i.e., more than eight taxa) are in anomaly zones.

The multispecies coalescent model has emerged as a powerful framework that allows computing the likelihood of gene trees for a given species tree. Most of the methods discussed above infer species trees from unrooted or unranked gene tree topologies. To preserve some information about branch lengths in the gene trees, one can use ranked gene trees. Instead of using branch lengths directly, a ranked gene tree depicts the temporal order of the nodes of the gene tree together with topological relationships among gene lineages. Thus, methods based on ranked gene trees should provide more robust and informative estimates than methods based on the topological relationships alone. Such methods might also be robust to errors in estimating gene tree branch lengths. Moreover, the probabilities of unrooted and unranked gene trees can be directly obtained from the probabilities of ranked gene trees. Degnan *et al.* (2012b) and Stadler and Degnan (2012) derived the probability distribution of ranked gene tree topologies and proposed several ways to calculate the probabilities of ranked gene trees. Using these probabilities the likelihood of gene trees for a given species tree can be determined. We introduce a software *PRANC*

that estimates a maximum likelihood species tree from a sample of ranked gene tree topologies. The algorithm searches in tree topology and branch length space for a species tree with the highest likelihood.

6.1 Conclusions

In this section we summarize the main contributions of this dissertation, which are fully discussed in the previous chapters.

6.1.1 Anomaly zones

We studied how the parameters of a species tree simulated under the birth-death models of speciation can affect the probability that the species tree has different types of anomalous gene trees. We found that with more than five species, it is possible for species trees to be in the intersection of the unrooted, unranked, and ranked anomaly zones. We found the cases when such zones do not overlap with each other for certain species tree topologies and/or branch lengths, meaning that methods based on one type of gene tree might provide more robust estimates than methods that use other types of gene trees. We observed that the probability of being in all types of anomaly zones increases with more taxa and with higher speciation rates. For unranked trees, both results are intuitive: for increasing numbers of taxa, there are more possible ways to have consecutive short branches or intervals in a tree, a pattern typical of the unranked anomaly zone. Increasing λ reduces the average branch length, making consecutive short branches more likely. We observed a different effect of the turnover rate μ/λ on the probability of producing unranked and unrooted versus ranked anomalous gene trees. The probability has a decreasing trend for the unranked anomaly zones and an increasing trend for the ranked anomaly

zone as turnover rate increases. On average, branch lengths become longer as μ increases. In particular, a branch length near the root becomes longer, decreasing the probabilities of anomalous unranked and unrooted gene trees but increasing the probabilities of anomalous ranked gene trees. We found that the probabilities of unranked and unrooted anomaly zones are higher and grow much faster than those of ranked anomaly zones as the speciation rate increases. We also found that more balanced trees tend to fall in the ranked anomaly zone and more imbalanced into the unranked anomaly zone. We derived a lower bound of the probability of the species tree having unranked anomalous gene trees for large speciation rates and proved that it approaches 1 as both speciation rate and number of species approach to infinity.

6.1.2 Heuristic approaches for detecting anomaly zones

We introduced heuristic approaches to infer whether species trees have unranked or unrooted anomalous gene trees when it is impractical to compute the entire distribution of gene tree topologies. We found that the most frequent tree often is not topologically far from the species tree. Thus, we only need to search in a nearby space of the true tree to find whether it has anomalous gene trees. This heuristic underestimates the probability that the species tree is in an anomaly zone due to there being few false negatives but no false positives. We found that for biologically plausible speciation rates, $0 < \lambda < 1$, unranked and unrooted anomaly zones intersect with each other more than with the ranked anomaly zone. Because at least one of the most probable ranked gene tree topologies must have the same unranked topology as the species tree, we propose to use only those ranked gene trees that have unranked topologies that match the unranked species tree topology to check for anomalousness. We also investigated the validity of the method of comparing pairs of parent-child internodes in a phylogenetic tree for anomalousness by using the limit of the unranked anomaly zone in larger trees. We observed that the false

positive rate slowly increases with the number of taxa and speciation rate λ. Despite the relatively high false positive rate, the test is still useful for checking that a tree does not fall in an anomaly zone. We demonstrated the discussed heuristic methods on the biological dataset of skinks.

6.1.3 Maximum likelihood species tree estimation

We developed an algorithm to get a maximum likelihood estimate of a species tree from ranked gene trees. We introduce a software suite *PRANC* that takes a set of ranked gene tree topologies and searches for the maximum likelihood species tree. *PRANC* is the first method we are aware of that uses ranked gene trees to infer species trees. To estimate the maximum likelihood species tree, *PRANC* picks the tree with the highest likelihood from the set of initial trees and generates all trees that are one nearest neighbor interchange move away from the initial tree. Then, for each of these trees, the program generates all possible rankings and finds the branch lengths that maximize the likelihood of the gene trees. The process is repeated until the stopping criterion is met. Various numerical techniques were used to improve the accuracy, robustness, and efficiency of the algorithm. We tested *PRANC*'s performance under different settings. There is a trade-off between the estimation accuracy and the computational time. The speed of the program highly depends on the choice of the initial tree, the number of nearest neighbor interchange moves considered, the number of rankings for each unranked species tree candidate, and on the allowable range of branch lengths. Mainly, *PRANC*'s computational speed depends on the number of rankings considered. For balanced topologies, far more rankings exist than for less balanced topologies. Computing the likelihood of gene trees for every possible ranked topology is not efficient for $n > 7$-taxon trees. We observed that in most cases, the values of likelihoods for different rankings of the same unranked species tree are close to each other. Therefore, we proposed to compute likelihoods

of a randomly chosen small subset of rankings. If at least one of the obtained likelihoods is larger than the threshold, *PRANC* will compute the likelihoods for a larger subset of rankings. On average, under default settings, in simulation studies, it takes seconds, a few minutes, ten-fifteen minutes, and an hour to estimate a $5, 6, 7$- and 8-taxon species tree from 1000 gene trees, respectively. It takes less than 5 minutes to estimate an 8-taxon species tree from 100 gene trees.

Because it is better to have an initial tree that is topologically close to the true tree, we proposed two methods for picking a starting tree. The first method is a hill-climbing algorithm that penalizes long branches of the species tree which has multiple lineages persisting through multiple species divergence events. Although this method can provide a list of good candidate trees for a species tree, we proved that it is statistically inconsistent. Another method is based on the greedy consensus method. The usual greedy consensus method considers the clades in order of the frequency with which they have appeared, adding to the consensus tree any which are compatible with it until the tree is fully resolved. We proposed a modification to the greedy consensus method. Our method, called maximum clade frequency consensus, outputs a tree with a maximal score, where the score is determined by summing the frequencies of compatible clades. In such a way, the method will include all most supported clades in the estimated tree and not only those compatible with the most frequent clade.

We showed in simulation and the biological study of gibbons that *PRANC* had better accuracy than competing methods.

6.2 Future work

The main shortcoming of likelihood-based methods is their computational complexity, which rises exponentially with the number of species in a tree. In particular,

since *PRANC* evaluates the probability of each ranked gene tree, it runs slower than *STELLS2* and *ASTRAL*. To speed up the computation, multiple jobs were run on the supercomputer at the UNM Center for Advanced Research Computing. It is possible to utilize parallelization in *PRANC* to scale it for larger species trees.

Because the maximum likelihood calculation of a species tree formed a major computational bottleneck in the inference, introducing pseudo-likelihood methods for ranked gene trees for species tree inference might significantly reduce the computational time while providing good estimation accuracy. One of the main future directions for *PRANC* is to estimate phylogenetic networks rather than trees. A phylogenetic network is the usual species tree with horizontal edges that captures the inheritance of genetic material through gene flow (Meng and Kubatko, 2009; Yu *et al.*, 2012b; Solís-Lemus and Ané, 2016; Wen *et al.*, 2018) (Figure 6.2). Since the concept of the anomaly zone can even be extended to phylogenetic networks (Zhu *et al.*, 2016), it is interesting to study the regions of branch lengths and inheritance probabilities that may produce anomalous ranked gene trees.

6.2.1 Parallelization

Certain functions of *PRANC* can be parallelized. For example, when searching in tree space, each n-taxon species candidate can produce $k = 2n - 4$ unranked trees that are one NNI move away from the original tree. Then each of the obtained k trees produces $2n - 4$ unranked trees, and so on. This process can be parallelized. For each unranked tree all possible rankings are generated and then the branch lengths of each of these ranked trees are optimized. It is costly to optimize branch lengths for every possible ranking, especially for larger trees. For example, for an 8-taxon tree the number of rankings can range from 1 to 80 whereas for a 12-taxon tree the number of rankings can range from 1 to $211,200$. We can distribute probabilities computations for a subset of species tree rankings across multiple processors. Finally,

since probabilities of all gene trees in a sample are computed for a given species tree, we can assign each thread or processor to work on a certain subset of gene trees.

6.2.2 A maximum pseudo-likelihood approach

The fact that any four-taxon species tree can produce anomalous unranked gene trees but cannot produce anomalous ranked gene trees give us an opportunity to develop methods based on ranked gene trees that are not affected by ranked anomaly zones. An n-taxon species tree with internal branch lengths $b_i, i = 1, \ldots, n-2$ has $\binom{n}{4}$ rooted ranked quartets. For example, a five-taxon species tree in Figure 6.1A contains $\binom{5}{4} = 5$ ranked quartets: $(((A, B)_3, C)_2, D)$, $(((A, B)_3, C)_2, E)$, $((A, B)_2, (D, E)_3)$, $((A, C)_2, (D, E)_3)$, and $((B, C)_2, (D, E)_3)$. For the last three quartets, we require that (D, E) should have a highest rank since it has a highest rank in the original five-taxon tree. Each quartet has two internal branch lengths that can be calculated from the internal branch lengths of original species tree. For example, the quartet in Figure 6.1D is one of the possible quartets for the species tree in Figure 6.1A that has internal branch lengths $q_1 = b_1 + b_2$ and $q_2 = b_3$ that can be calculated from the corresponding lengths in the species tree.

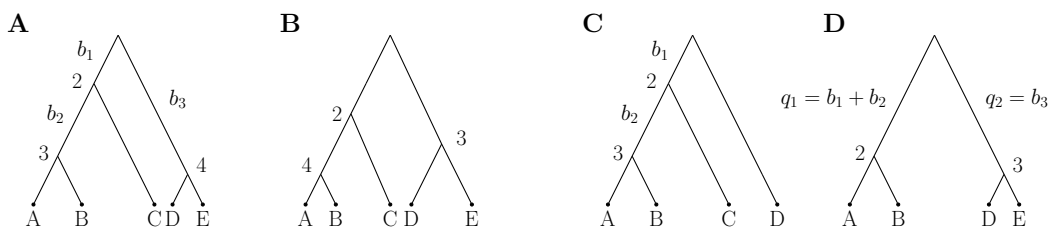


Figure 6.1: Five-taxon trees with different rankings and two representative quartets. (A)-(B) Ranked five-taxon trees with internal branch lengths b_1, b_2, and b_3. (C) Caterpillar four-taxon tree. The tree is a ranked quartet for both trees in (A) and (B). (D) Balanced four-taxon tree. The tree is a ranked quartet only for the tree in (A).

There are 18 possible ranked topologies of any four-taxon tree. The probabilities

of each 18 ranked gene trees given a species tree can be calculated under coalescent model. The probability distributions of ranked gene trees given a caterpillar or balanced species tree are given in Tables 1 and 2 in Degnan *et al.* (2012b). Let RQ_i denote the rooted ranked quartet in the species tree with the internal branch lengths q_1 and q_2. Let $x_{1i}, x_{2i}, ..., x_{18i}$ be the counts of 18 ranked quartets occuring in ranked gene trees, then the counts $x_{1i}, x_{2i}, ..., x_{18i}$ have a multinomial distribution

$$p(x_{1i}, x_{2i}, ..., x_{18i}|RQ_i) = \frac{\sum_{j=1}^{18} x_{ji}}{x_{1i}!x_{2i}! \cdot ... \cdot x_{18i}!} p_{1i}^{x_{1i}} p_{2i}^{x_{2i}} \cdot ... \cdot p_{18i}^{x_{18i}}, \tag{6.1}$$

where $p_{ji}, j = 1, 2, \ldots, 18$ are probabilities of the ranked rooted quartets in the gene tree. For example, given $RQ_i = (((A, B)_3, C)_2, D)$ in the species tree, the probability p of ranked quartet $((A, B)_2, (C, D)_3)$ in some gene tree is $e^{-t_3}e^{-3t_2}/18$ using Table 2 in Degnan *et al.* (2012b) where t_2 and t_3 are the interval lengths which can be converted into the branch lengths. Then the pseudo-likelihood of a species tree $\mathcal{T}$ given a sample of N ranked gene trees $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N$ can be written as

$$L(\mathcal{T}|\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N) = \prod_{i=1}^{\binom{n}{4}} \Big[\frac{\sum_{j=1}^{18} x_{ji}}{x_{1i}!x_{2i}! \cdot ... \cdot x_{18i}!} p_{1i}^{x_{1i}} p_{2i}^{x_{2i}} \cdot ... \cdot p_{18i}^{x_{18i}} \Big]. \tag{6.2}$$

It is a pseudo-likelihood rather than likelihood since we approximate the joint probability distribution of ranked quartets in the species tree by the product of corresponding marginal probabilities ignoring the covariance structure among the quartets. Maximizing the pseudo-likelihood given in eq. (6.2) with respect to the species tree topology and branch lengths will allow estimating larger species trees faster than using the traditional full likelihood approach.

6.2.3 Phylogenetic networks

A phylogenetic tree, the classical representation of evolution of a set of organisms, is unable to represent some biological processes such as hybridization, recombination,

or horizontal gene transfer. A rooted phylogenetic network can be repsented as a rooted directed graph (Figure 6.2). It is the usual species tree with horizontal edges that captures the inheritance of genetic material through gene flow (Meng and Kubatko, 2009; Yu *et al.*, 2012b; Solís-Lemus and Ané, 2016; Wen *et al.*, 2018). A phylogenetic network might be more suitable to model evolutionary history of certain sets of organisms such as bacteria who transfer their DNA regions across species boundaries.

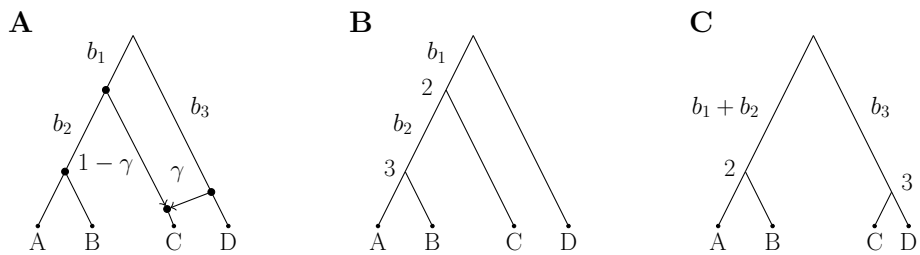


Figure 6.2: An example of rooted phylogenetic network with one hybridization event and resulting species trees. (A) Four-taxon rooted phylogenetic network with one hybridization event. (B) The underlying species tree obtained by assuming that gene flow affected less than 50% of genes ($\gamma < 0.5$). (C) The underlying species tree obtained by assuming that gene flow affected more than 50% of genes ($\gamma > 0.5$).

Under the hybridization event and the rooted network in Figure 6.2, each gene from taxon C has probability γ to inherit the genetic material of D taxon's parent and probability $1-\gamma$ to inherit the genetic material of the (A, B) taxon's parent and be related more to (A, B) than to D. The probability of a gene tree given a species network is computed by considering the coalescence process inside each branch of the species network. For example, the probability of the gene tree $((A, C)_2, (B, D)_3)$ under the network displayed in Figure 6.2 is

$$(1-\gamma)e^{-b_2}\left(\frac{3}{2}e^{-b_1} - \frac{3}{2}e^{-3b_1} + e^{-3b_1}\right)\frac{1}{9} + \gamma\frac{1}{18}e^{-(b_1+b_2+b_3)}.$$

Being able to reconstruct a phylogenetic network characterized by the topology, branch lengths, and inheritance probabilities from a sample of ranked gene trees will allow uncovering more complex relationships between species.

Appendix A

A PRANC software

We originally developed *PRANC* in *C++* to calculate the probability distribution of ranked gene trees given a species tree. We later extended it to infer maximum likelihood trees. However, *PRANC* has other useful options to work with phylogenetic trees presented in Table A1. All input files should be in the Newick format. All trees are treated as rooted binary trees. A species tree should be ultrametric (leaves of the tree are all equidistant from the root). The taxon names of gene trees should match the taxon names of the corresponding species tree.

PRANC can compute the analytical probabilities of ranked gene trees given a species tree in Newick format. The program has an option to compute the probability of an unranked gene tree by summing the probabilities of all ranked gene trees that share the corresponding unranked topology. It has an option to output symbolic probabilities followed by ranked histories. User can run PRANC as shown below:

```
pranc -rprob <species-tree-file> <ranked-gene-tree-file>
pranc -uprob <species-tree-file> <unranked-gene-tree-file>
pranc -sym <species-tree-file> <ranked-gene-tree-file>
```

PRANC also can output the "democratic vote" ranked or unranked tree topology.

Table A1: List of the main options available in *PRANC*

option	description
-rprob	calculates probabilities of ranked gene tree topologies
-uprop	calculates probabilities of unranked gene tree topologies
-sym	outputs symbolic probabilities of ranked gene tree topologies
-like_nonni	calculates ML interval lengths of a given species tree topology
-like_nni_brent	estimates ML species tree given a starting tree. Brent's method is used for branch lengths optimization
-like_nni	estimates ML species tree given a starting tree. L-BFGS method is used for branch lengths optimization
-rtopo	outputs ranked tree topologies and frequencies of the topologies
-utopo	outputs unranked tree topologies and frequencies of the topologies
-write	outputs tree with ranks instead of branch lengths
-rank_trees	outputs all ranked topologies that share given unranked topology
-mac	outputs species tree MAC score
-cons	outputs greedy consensus tree without branch lengths
-mcfc	outputs MCFC tree without branch lengths
-rank_dist	calculates rank dissimilarity measure between two trees
-coal_ints	calculates interval lengths
-br_len	calculates internal branch lengths

Using the following code, the program outputs two files: one with ranked or unranked topologies for each tree, and another with unique topologies and their frequencies,

```
pranc -rtopo <input-file>
pranc -utopo <input-file>
```

To output the maximum clade frequency consensus (MCFC) tree and MAC score of a species tree, one can run

```
pranc -mcfc <ranked-gene-tree-file>
pranc -mac <species-tree-file> <ranked-gene-tree-file>
```

To estimate a species tree from a sample of ranked gene trees, the user can run the command below (the starting tree should be provided). The program processes the initial species trees and picks the one with the highest likelihood $\mathcal{T}$. Then it searches

a space of unranked trees to find trees that are one nearest neighbor interchange (NNI) away from $\mathcal{T}$. After that, *PRANC* searches for the speciation interval lengths that maximizes the likelihood of the ranked gene trees using L-BFGS (Byrd et al., 1995) method. The process is repeated until convergence (-diff) or until all trees within k NNI steps are explored (-nni). At the end, *PRANC* calculates the branch lengths of the inferred tree.

```
pranc -like_nni <starting-species-tree-file> -rgt <ranked-gene-tree-file>
pranc -like_nni <starting-species-tree-file> -rgt <ranked-gene-tree-file>
-nni 5 -diff 0.1 -startsubset 3 -initsubset 2 -maxsubset 4  -lb 0.001
-ub 6 -tol 1e-10 -tiplen 0.1
Default settings:
1) the maximum number of NNI moves considered from the starting tree:
-nni 5
2) the difference between old and current negative log-likelihoods:
-diff 0.1
3) the number of rankings of the starting tree:
-startsubset All possible rankings
4) the number of initial rankings considered of each
unranked species tree candidate, if at least one of the negative
log-likelihoods corresponding to -initsubset rankings is smaller
than the threshold, the program will consider more rankings set
by -maxsubset option:
-initsubset Number of taxa
5) the number of maximum rankings considered of each
unranked species tree candidate:
-maxsubset 2*Number of taxa
6) allowed lower bound of the speciation interval lengths:
-lb 0.001
7) allowed upper bound of the speciation interval lengths:
```

```
-ub 6
8) tolerance of the L-BFGS method:
-tol 1e-10
9) the time of the most recent clade is set to 0.1:
-tiplen 0.1
```

The user may fix a topology of the species tree and only estimate the branch lengths. The program calculates maximum likelihood branch lengths using L-BFGS (Byrd et al., 1995) optimization technique for a given species tree topology. *PRANC* changes all lengths at the same time. It allows the lengths to be in the interval $[0.001, 6]$ (-lb, -ub) coalescent units. Note that *PRANC* first computes maximum likelihood speciation intervals, and then translates them to the branch lengths by setting the time of the most recent internal node to 0.1 (-tiplen) coalescent unit. The tolerance is controlled by the -tol option.

```
pranc -like_nonni <species-tree-file> -rgt <ranked-gene-tree-file>
pranc -like_nonni <species-tree-file> -rgt <ranked-gene-tree-file>
-lb 0.001 -ub 6 -tol 1e-10 -tiplen 0.1
Default settings:
-lb 0.001
-ub 6
-tol 1e-10
-tiplen 0.1
```

For detailed examples with file examples and shown output visit `https://github.com/anastasiiakim/PRANC`.

Appendix B

Here we provide a few representative scripts that were used for simulations at various stages of research. We selected less trivial scripts to show. More scripts will be uploaded to `https://github.com/anastasiiakim/PRANC`. We ran most of the simulations on the supercomputer at the UNM Center for Advanced Research Computing. A job *pbs* file may look like this:

```
## University of New Mexico Center for Advanced Research Computing
#PBS -l nodes=1:ppn=8
#PBS -l walltime=48:00:00
#PBS -N some_job_name
#PBS -S /bin/bash
#PBS -n ${taxa}_${lambda}_${mu}_${iter}
# Load modules for R, python, other software installed in this module
module load r-3.4.1-gcc-4.8.5-python2-gzeg24m
cd wheeler-scratch  # $PBS_O_WORKDIR output dir for this machine

/users/anakim/unranked/some_script $taxa $lambda $mu $iter
```

We can run this script from the *wheeler-scratch* directory:

```
qsub -v taxa=8,lambda=1,mu=0,iter=500 some_sample.pbs
```

B Bash Scripts

B.1 Anomaly zone detection

The bash script below represents how anomalous gene trees can be detected:

```
#!/bin/bash
OUTPUT=/dev/null

# Inputs for this run:

NumTaxa=$1
BDlambda=$2
BDmu=$3
NumIter=$4

echo "Starting run on $(date), Job number: $PBS_JOBID on host $(hostname)"
echo "NumTaxa=$NumTaxa"
echo "BDlambda=$BDlambda"
echo "BDmu=$BDmu"
echo "NumIter=$NumIter"

EXE_DIR="/users/anakim/bin"
PRANC_DIR="users/anakim/module/BIN"

RSCRIPTS_DIR="/users/anakim/rankedtrees"
ST_DIR="/users/anakim/nni/SThightaxa"
SCRATCHDIR=${NumTaxa}_${BDlambda}_${BDmu}_${NumIter}_$(echo $PBS_JOBID | cut -d"." -f1)
mkdir ./$SCRATCHDIR
```

```
# copy all R_scripts to the scratch folder to write output files in that place
cp $RSCRIPTS_DIR/R_sort.r $SCRATCHDIR
cp $RSCRIPTS_DIR/R_rf_dist.r $SCRATCHDIR
cp $RSCRIPTS_DIR/gtuniqtrees.txt $SCRATCHDIR
cp $RSCRIPTS_DIR/gtuniqtops.txt $SCRATCHDIR
cp $ST_DIR/ST_${NumTaxa}_${BDlambda}_${BDmu}.txt $SCRATCHDIR

cd $SCRATCHDIR

filename="ST_${NumTaxa}_${BDlambda}_${BDmu}.txt"

zero=0;
while read line;
do
    echo "$line" > ST.txt
    rm -f anomalous_trees.txt

    #gtuniqtops.txt is optional file that represents topologies of gene trees
    $PRANC_DIR/pranc -rprob ST.txt gtuniqtrees.txt gtuniqtops.txt

# R_sort.r R script will print a number that indicates
# whether we have anomalous ranked gene trees
varname=`Rscript R_sort.r`
echo "$varname"
if [[ $varname -eq $zero ]]; then
cp ST.txt intree
#run PHYLIP treedist function, PHYLIP is interactive program
$EXE_DIR/treedist << EOF >> $OUTPUT
D
R
```

```
1
2
L
S
Y
EOF
    rm -f intree
    rm -f intree2
    # treedist produces output file: outfile
    # the RF distance can be extracted using R
    Rscript R_rf_dist.r >> $OUTPUT
    rm -f outfile
fi
done < $filename

# Collect results, do not remove $SCRATCHDIR if unsure that the results are saved
RESULTDIR=${NumTaxa}_${BDlambda}_${BDmu}_${NumIter}-some_results
mkdir ../$RESULTDIR
cp some_file_1.txt some_file_2.txt some_file_3.txt ../$RESULTDIR

cd ../

rm -r $SCRATCHDIR

exit 0
```

B.2 Selecting trees with the lowest MAC score

The script corresponding to the described simulation process in Section 4.2.

```
#!/bin/bash
OUTPUT=/dev/null

# Inputs for this run:
NumTaxa=$1
BDlambda=$2
BDmu=$3
NumGTs=$4
Lb=$5
Ub=$6

EXE_DIR="/users/anakim/bin"
RSCRIPTS_DIR="/users/anakim/likelihood"
ST_DIR="/users/anakim/likelihood/SpeciesTrees"
BIN_DIR="/users/anakim/module/BIN"
SCRATCHDIR=mac_nni-${NumTaxa}_${BDlambda}_${BDmu}_${NumGTs}_${Lb}_${Ub}
mkdir ./$SCRATCHDIR

cp $RSCRIPTS_DIR/R_check_scores.r $SCRATCHDIR
cp $RSCRIPTS_DIR/R_check_scores_again.r $SCRATCHDIR
cp $ST_DIR/ST_${NumTaxa}_${BDlambda}_${BDmu}.txt $SCRATCHDIR

cd $SCRATCHDIR
sed -n -e ${Lb},${Ub}p ST_${NumTaxa}_${BDlambda}_${BDmu}.txt > ST_part.txt
filename="ST_part.txt"
```

```
while read line;
do
    echo "$line" > ST.txt
    $EXE_DIR/hybrid-Lambda -spcu ST.txt -num ${NumGTs} -seed 2802 -o sim_gts.txt
    sed "s/_1//g" sim_gts.txt_coal_unit > sim_gts.txt

    rm -f count_min_ac.txt
    gt_file="sim_gts.txt"
    while read line;
    do
        echo "$line" > ST_candidate.txt
        var_count=`$BIN_DIR/pranc -mac ST_candidate.txt sim_gts.txt`
        echo "$var_count" >> count_min_ac.txt
    done < $gt_file

    rm -f candidates.txt
    Rscript R_check_scores.r

    COUNTER=0
        while [ ! -e temp_rf.txt ] && [ $COUNTER -lt 5 ];
        do

    let COUNTER=COUNTER+1
    rm -f ranked_cands.txt
    file_st_cands_equal_cost="candidates.txt"
    while read line;
    do
        echo "$line" > cand.txt
        $BIN_DIR/pranc -rank_trees cand.txt
        # uncomment if want to select only 5 random rankings
```

```
        #shuf -n 5 res.txt >> ranked_cands.txt
        cat res.txt >> ranked_cands.txt
        rm -f res.txt
    done < $file_st_cands_equal_cost

    rm -f fin_count.txt
    gt_file="ranked_cands.txt"
        while read line;
        do
                echo "$line" > ST_candidate.txt
                var_count=`$BIN_DIR/pranc -mac ST_candidate.txt sim_gts.txt`
                echo "$var_count" >> fin_count.txt
        done < $gt_file


    Rscript R_check_scores_again.r
    echo $COUNTER
    done

cat temp_rf.txt >> rf_lowestscore.txt
rm -f temp_rf.txt
rm -f scores.txt
rm -f fin_trees.txt

done < $filename

exit 0
```

B.3 Maximum likelihood estimation

The script corresponding to the described simulation process in Section 4.1.

```
#!/bin/bash
OUTPUT=/dev/null

# Inputs for this run:
NumTaxa=$1
BDlambda=$2
BDmu=$3
NumIter=$4
Lb=$5
Ub=$6

EXE_DIR="/users/anakim/bin"
RSCRIPTS_DIR="/users/anakim/likelihood"
ST_DIR="/users/anakim/likelihood/SpeciesTrees"
BIN_DIR="/users/anakim/module/BIN"
SCRATCHDIR=maxlike_nni-${NumTaxa}_${BDlambda}_${BDmu}_${NumIter}_${Lb}_${Ub}
mkdir ./$SCRATCHDIR

cp $RSCRIPTS_DIR/R_remove_branches_pick_btw_cons.r $SCRATCHDIR
cp $RSCRIPTS_DIR/RF_calcnorm.r $SCRATCHDIR
cp $ST_DIR/ST_${NumTaxa}_${BDlambda}_${BDmu}.txt $SCRATCHDIR

cd $SCRATCHDIR
sed -n -e ${Lb},${Ub}p ST_${NumTaxa}_${BDlambda}_${BDmu}.txt > ST_part.txt

filename="ST_part.txt"
```

```
NOW=$(date +"%T") # measure execution time
echo "Job started at  $NOW"

while read line;
do
    echo "$line" > ST.txt
    # simulate gene trees from a species tree
    $EXE_DIR/hybrid-Lambda -spcu ST.txt -num ${NumIter} -seed 2802
    sed "s/_1//g" OUT_coal_unit > sim_gts.txt
cp sim_gts.txt intree

# get a greedy consensus tree from PHYLIP
$EXE_DIR/consense << EOF >> $OUTPUT
R
Y
EOF

# get MCFC tree
$BIN_DIR/pranc -cons sim_gts.txt
cat outGreedyCons.txt >> pranc_greedy_cons.txt
Rscript $RSCRIPTS_DIR/R_remove_branches_pick_btw_cons.r
rm -f intree
rm -f outfile
rm -f outtree

# estimate ML tree
$BIN_DIR/pranc -like_nni unranked_tree.txt -rgt sim_gts.txt

rm -f unranked_tree.txt
```

```
done < $filename

# report rank diss. measure between trees
$BIN_DIR/pranc -rank_dist ST_part.txt outWithNniMLTopo.txt
Rscript RF_calcnorm.r

NOW=$(date +"%T")

echo "Job ended at  $NOW"
exit 0
```

B.4 Gibbons dataset

The control file for BPP and bash script corresponding to the results of chapter 5. The control file A01.bpp.ctl:

```
         seed =  -1

      seqfile = bpp_seqs_sample.txt
     Imapfile = gibbons.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0 * fixed species tree
        speciestree = 1   * speciestree pSlider ExpandRatio ShrinkRatio

  speciesmodelprior = 1  * 0: uniform LH; 1:uniform rooted trees;
                         * 2: uniformSLH; 3: uniformSRooted

 species&tree = 5  B  S  N  P M
                   1  1 1  1   1
                (((B, S),N), (P, M));
       diploid =   1  1  1  1  1

      usedata = 1  * 0: no data (prior); 1:seq like

    cleandata = 1    * remove sites with ambiguity data (1:yes, 0:no)?

   thetaprior = 3 0.002    # invgamma(a, b) for theta
     tauprior = 3 0.004    # invgamma(a, b) for root tau &
                           # Dirichlet(a) for other tau's
```

```
*     heredity = 1 4 4
*    locusrate = 1 5

      finetune =  1: 5 0.001 0.001  0.001 0.3 0.33 1.0  # finetune for GBtj, GBspr,
                                                  # theta, tau, mix, locusrate, seqerr

         print = 1 0 0 0   * MCMC samples, locusrate, herediyscalars, Genetrees
        burnin = 2000
      sampfreq = 2
       nsample = 20000
```

The script:

```
#!/bin/bash
OUTPUT=/dev/null

# Inputs for this run:
NumIter=$1
NumCount=$2

EXE_DIR="/users/anakim/bin"
GIBBONS_DIR="/users/anakim/gibbons"
BIN_DIR="/users/anakim/module/BIN"
STELLS_DIR="/users/anakim/STELLS2"
ASTRAL_DIR="/users/anakim/ASTRAL/Astral"
BPP_DIR="/users/anakim/bpp/src"
SCRATCHDIR=gibbons_pranc_stell_astral_bpp
mkdir ./$SCRATCHDIR

cp $GIBBONS_DIR/R_pick_btw_two_cons.r $SCRATCHDIR
```

```
cp $GIBBONS_DIR/R_subsample_bpp.r $SCRATCHDIR
cp $GIBBONS_DIR/R_astral.r $SCRATCHDIR
cp $GIBBONS_DIR/RF_pranc_stells_astral_bpp.r $SCRATCHDIR
cp $GIBBONS_DIR/bpp_sequences.txt $SCRATCHDIR
cp $GIBBONS_DIR/A01.bpp.ctl $SCRATCHDIR
cp $GIBBONS_DIR/gibbons.Imap.txt $SCRATCHDIR
cp $GIBBONS_DIR/gibbons_gene_trees.txt $SCRATCHDIR
cp $GIBBONS_DIR/gibbons_topo.txt $SCRATCHDIR

cd $SCRATCHDIR

COUNTER=0
while [ $COUNTER -lt ${NumCount} ];
do

let COUNTER=COUNTER+1

Rscript R_subsample_bpp.r ${NumIter}
$BPP_DIR/bpp --cfile A01.bpp.ctl > o.txt
tail -n 1 o.txt > o1.txt
sed 's/;.*/;/' o1.txt > o.txt
sed 's/#//g' o.txt > o1.txt
sed 's/ //g' o1.txt >> bpp_inferred_trees.txt

cp gene_trees_sample.txt intree
$EXE_DIR/consense << EOF >> $OUTPUT
R
Y
EOF
```

```
$BIN_DIR/ppranc -cons gene_trees_sample.txt

Rscript R_pick_btw_two_cons.r
rm -f intree
rm -f outfile
rm -f outtree

$BIN_DIR/ppranc -like_nni unranked_tree.txt -rgt gene_trees_sample.txt
rm -f unranked_tree.txt

Rscript R_astral.r
java -jar $ASTRAL_DIR/astral.5.6.3.jar -i gibbons_unrooted_trees.txt -o out.tre
cat out.tre >> astral_res_top.txt
rm -f gibbons_unrooted_trees.txt

$STELLS_DIR/stells-v2-1-0-linux64 -g gene_trees_sample.txt  > output_temp.txt
tail -2 output_temp.txt > temp.txt
head -n 1 temp.txt > output.txt
sed -n -e 's/^.*tree: //p' output.txt >> stells_inferred_trees.txt

rm -f gene_trees_sample.txt

done

Rscript RF_pranc_stells_astral_bpp.r

exit 0
```

References

Bortolussi, N., Durand, E., Blum, M., and François, O. (2005). apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, 22, 363–364.

Brent, R. (1973). *Algorithms for minimization without derivatives.* Prentice-Hall, Englewood Clifts, New Jersey.

Brown, J. K. M. (1994). Probabilities of evolutionary trees. *Syst. Biol.*, 43, 78–91.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science*, 61, 163–184.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, 29(8), 1917–1932.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5), 1190–1208.

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., Meyer, T. J., Herrero, J., Roos, C., Aken, B., *et al.* (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, 513, 195.

Castillo-Ramírez, S. and González, V. (2008). Factors affecting the concordance between orthologous gene trees and species tree in bacteria. *BMC Evol. Biol.*, 8(1), 300.

Chifman, J. and Kubatko, L. (2014). Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23), 3317–3324.

Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., and Edwards, S. V. (2019). Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. *Syst. Biol.*, 68, 937–955.

Colless, D. H. (1982). Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool.*, 31, 100–104.

Darwin, C. (1872). *The origin of species by means of natural selection.* J. Murray.

DeGiorgio, M. and Degnan, J. H. (2010). Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol. Biol. Evol.*, 27, 552–569.

DeGiorgio, M. and Degnan, J. H. (2014). Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst. biol.*, 63(1), 66–82.

Degnan, J. H. (2013). Anomalous unrooted gene trees. *Syst. Biol.*, 62, 574–590.

Degnan, J. H. and Rhodes, J. A. (2015). There are no caterpillars in a wicked forest. *Theor. Popul. Biol.*, 105, 17–23.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet.*, 2, 762–768.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59, 24–37.

Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009a). Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.*, 58(1), 35–54.

Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009b). Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.*, 58, 35–54.

Degnan, J. H., Rosenberg, N. A., and Stadler, T. (2012a). A characterization of the set of species trees that produce anomalous ranked gene trees. *IEEE/ACM Trans. Comput. BiolBioinform.*, 9(6), 1558–1568.

Degnan, J. H., Rosenberg, N. A., and Stadler, T. (2012b). The probability distribution of ranked gene trees on a species tree. *Math. Biosci.*, 235, 45–55.

Disanto, F. and Rosenberg, N. A. (2014). On the number of ranked species trees producing anomalous ranked gene trees. *IEEE/ACM Trans. Comput. BiolBioinform.*, 11, 1229–1238.

Disanto, F., Miglionico, P., and Narduzzi, G. (2019). On the unranked topology of maximally probable ranked gene tree topologies. *J. Math. Biol.*, 79, 1205–1225.

Ewing, G. B., Ebersberger, I., Schmidt, H. A., and Von Haeseler, A. (2008). Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.*, 8, 118.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6), 368–376.

Felsenstein, J. (1983). Statistical inference of phylogenies. *Journal of the Royal Statistical Society: Series A (General)*, 146(3), 246–262.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Felsenstein, J. (2013). PHYLIP (Phylogeny Inference Package) version 3.695. *Distributed by the author*.

Fisher, R. A. (1930). *The genetical theory of natural selection*. The Clarendon Press, Oxford.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with bpp using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35, 2585–2593.

Ford, D., Matsen, F. A., and Stadler, T. (2009). A method for investigating relative timing information on phylogenetic trees. *Syst. biol.*, 58(2), 167–183.

Hamilton, M. (2011). *Population genetics*. John Wiley & Sons.

Hammersley, J. M. and Grimmett, G. R. (1974). *Maximal solutions of the generalized subadditive inequality.* Stochastic geometry (E. F. Harding and D. G. Kendall eds.). John Willey and Sons, London.

Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Probab.*, 3, 44–77.

Harding, E. F. (1974). *The probabilities of the shapes of randomly bifurcating trees.* Stochastic geometry (E. F. Harding and D. G. Kendall eds.). John Willey and Sons, London.

Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27(3), 570–580.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27, 570–580.

Huang, H., He, Q., Kubatko, L. S., and Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.*, 59(5), 573–583.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, 23(2), 183–201.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8), 754–755.

Jiang, X., Edwards, S. V., and Liu, L. (2019). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *bioRxiv*, page 860809.

Kendall, D. G. *et al.* (1948). On the generalized" birth-and-death" process. *The annals of mathematical statistics*, 19(1), 1–15.

Kim, A., Rosenberg, N. A., and Degnan, J. H. (2019). Probabilities of unranked and ranked anomaly zones under birth-death models. *Mol. Biol. Evol.*, 37(5), 1480–1494.

Kingman, J. F. (1982a). On the genealogy of large populations. *Journal of applied probability*, 19(A), 27–43.

Kingman, J. F. C. (1982b). The coalescent. *Stochastic processes and their applications*, 13(3), 235–248.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Syst. biol.*, 56(1), 17–24.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7), 971–973.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). Bucky: gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26(22), 2910–2911.

Leaché, A. D. and Rannala, B. (2011). The accuracy of species tree estimation under simulation: a comparison of methods. *Syst. biol.*, 60(2), 126–137.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., Lpez-Bautista, J. M., Zuccarello, G. C., and Clerck, O. D. (2014). Dna-based species delimitation in algae. *European Journal of Phycology*, 49(2), 179–196.

Linkem, C. W., Minin, V. N., and Leache, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). *Syst. Biol.*, 65, 465–477.

Liu, L. (2008). Best: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21), 2542–2543.

Liu, L. and Edwards, S. V. (2009). Phylogenetic inference in the anomaly zone. *Syst. Biol.*, 58, 452–460.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, 56, 504–514.

Liu, L. and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Syst. Biol.*, 60(5), 661–667.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009a). Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, 53(1), 320–328.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009b). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, 58(5), 468–477.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009c). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, 58, 468–477.

Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, 10, 302.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. biol.*, 46(3), 523–536.

McCormack, J. E., Huang, H., and Knowles, L. L. (2009). Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. biol.*, 58(5), 501–508.

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.*, 75(1), 35–45.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17), i541–i548.

Mooers, A. O. and Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *The quarterly review of Biology*, 72, 31–54.

Mossel, E., Roch, S., and Sly, A. (2011). On the inference of large phylogenies with long branches: How long is too long? *Bulletin of mathematical biology*, 73(7), 1627–1644.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, 32, 268–274.

Pei, J. and Wu, Y. (2017). STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. *Bioinformatics*, 33, 1789–1797.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36, 541–562.

Popovic, L. (2004). Asymptotic genealogy of a critical branching process. *Ann. Appl. Probab.*, 14(4), 2120–2148.

Rambaut, A. and Grass, N. C. (1997). Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3), 235–238.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4), 1645–1656.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.*, 53, 131–147.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.*, 100, 56–62.

Rosenberg, N. A. (2006). The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. *Ann. Comb.*, 10, 129–146.

Rosenberg, N. A. (2007). Counting coalescent histories. *J. Comput. Biol.*, 14, 360–377.

Rosenberg, N. A. (2013). Discordance of species trees with their most likely gene trees: A unifying principle. *Mol. Biol. Evol.*, 30, 2709–2713.

Rosenberg, N. A. and Tao, R. (2008). Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst. Biol.*, 57, 131–140.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Schafer, S. F. and Krekorian, C. (1983). Agonistic behavior of the galápagos tortoise, geochelone elephantopus, with emphasis on its relationship to saddle-backed shell shape. *Herpetologica*, pages 448–456.

Shekhar, S., Roch, S., and Mirarab, S. (2018). Species tree estimation using astral: how many genes are enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 15, 1738–1747.

Shi, C.-M. and Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35, 159–179.

Solís-Lemus, C. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*, 12(3).

Stadler, T. (2011). Simulating trees on a fixed number of extant species. *Syst. Biol.*, 60, 676–684.

Stadler, T. and Degnan, J. H. (2012). A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Algorithm. Mol. Biol.*, 7, 338–355.

Stadler, T. and Steel, M. (2012). Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *J. Theor. Biol.*, 297, 33–40.

Stadler, T., Degnan, J. H., and Rosenberg, N. A. (2016). Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? *Syst. Biol.*, 65, 628–639.

Steel, M. (2016). *Phylogeny: discrete and random processes in evolution*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(1), vey016.

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2), 437–460.

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. (2006). Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3), 1511–1520.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26(2), 119–164.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2), 57–86.

Than, C. V. and Rosenberg, N. A. (2011a). Consistency properties of species tree inference by minimizing deep coalescences. *Journal of Computational Biology*, 18, 1–15.

Than, C. V. and Rosenberg, N. A. (2011b). Consistency properties of species tree inference by minimizing deep coalescences. *J. Comput. Biol.*, 18, 1–15.

Veeramah, K. R., Woerner, A. E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., Carbone, L., Wall, J. D., and Hammer, M. F. (2015). Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*, 200, 295–308.

Volkov, I., Banavar, J. R., Hubbell, S. P., and Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424(6952), 1035–1037.

Wakeley, J. (2009). Coalescent theory. *Roberts & Company*.

Wang, Y. and Degnan, J. H. (2011). Performance of matrix representation with parsimony for inferring species from gene trees. *Stat. Appl. Genet. Mol.*, 10, 21.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature microbiology*, 1(9), 1–8.

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring phylogenetic networks using phylonet. *Syst. Biol.*, 67, 735–740.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2), 97.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution: International Journal of Organic Evolution*, 66, 763–775.

Xu, B. and Yang, Z. (2016). Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204, 1353–1368.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z. (2015). The bpp program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5), 854–865.

Yu, Y., Degnan, J. H., and Nakhleh, L. (2012a). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.*, 8, e1002660.

Yu, Y., Degnan, J. H., and Nakhleh, L. (2012b). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS genetics*, 8(4).

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46), 16448–16453.

Zhaxybayeva, O., Doolittle, W. F., Papke, R. T., and Gogarten, J. P. (2009). Intertwined evolutionary histories of marine synechococcus and prochlorococcus marinus. *Genome Biol. Evol.*, 1, 325–339.

Zhu, J., Yu, Y., and Nakhleh, L. (2016). In the light of deep coalescence: Revisiting trees within networks. *BMC Bioinformatics*, 17, 415.

Zhu, S. and Degnan, J. H. (2017). Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.*, 66, 283–298.

Zhu, S., Degnan, J. H., Goldstien, S. J., and Eldon, B. (2015). Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC bioinformatics*, 16(1), 292.