Estimating parameters from Markov processes on trees

By

Brandon Legried

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Mathematics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2020

Date of final oral examination: August 5, 2020

The dissertation is approved by the following members of the Final Oral Committee:

David Anderson, Professor, Mathematics

Cécile Ané, Professor, Statistics and Botany

Sebastien Roch, Professor, Mathematics

Benedek Valko, Professor, Mathematics

Abstract

Reconstructing evolutionary relationships and traits between living and extinct taxa is a main goal within mathematical biology. The combinatorial and stochastic properties of the relevant models and algorithms have also been used to make progress on problems in statistical physics and networks as well. In this thesis, existing theories on both locus-based and sequence-based methods of reconstruction are extended to models featuring birth-death processes. The theory is extended in three separate ways.

In the first chapter, phylogenomic methods are considered. Phylogenomics is the estimation of species trees from multi-locus data, often by gene trees. Estimation is confounded by biological processes including incomplete lineage sorting (ILS), lateral gene transfer (LGT), and gene duplication and loss (GDL) where some gene trees differ from the species tree. We show that species trees are identifiable under GDL and some consensus-based methods are statistically consistent estimators under GDL. The results are further extended to DLCoal, a unified model for GDL and ILS.

In the second chapter, the sample complexity of gene trees is described when generated under GDL and DLCoal. That is, how many gene trees are needed to estimate the species tree with high probability? Here, the focus is on the linear birth-death process with constant rates over time. Through further probabilistic analysis of these models, the sample complexity bounds highlight the effect of these rates in both subcritical and supercritical regimes.

In the third chapter, we consider the problem of distance estimation under the TKF91 model of sequence evolution by indels (insertions and deletions) and substitutions on a

phylogeny. In this model, every site undergoes a constant linear birth-death process independently of all other sites. In the asymptotic regime where the expected total sequence length goes to infinity, we show that no consistent distance estimation is possible from sequence lengths alone. Formally, the distributions of pairs of sequence lengths at different distances cannot be distinguished with probability going to one.

Acknowledgments

First, I thank my advisor, Sebastien Roch, for his support, assistance, and passion. This work would not be possible without his guidance. He introduced me to many interesting problems that I will continue to pursue.

I also thank my committee members, Cécile Ané, David Anderson, and Benedek Valko. I appreciate their time, questions, and feedback in their consideration of this thesis as well as in-class instruction and discussion.

I thank my other co-authors, Erin K. Molloy, Tandy Warnow, Louis Fan, and Max Hill. Their interests, style, and feedback have marked my own research, writing, and presentation.

Finally, I thank my family and friends, especially John Moran, for their unwavering love, support, and trust in pursuing this endeavor.

Contents

Chapter 1

Introduction

Phylogeny estimation consists of the inference of an evolutionary tree from extant species data. It is a goal of both biologists and mathematicians to understand the relationships between taxa seen in the field today and possibly extinct taxa. A great body of mathematical and computational research has been performed and improved, consisting of theory, models, algorithms, and simulation experiments. See [Ste16], [War17], etc.

Phylogenetic trees are a graphical tool used to perform analysis. (However, more general graphs are used, see [HRS11].) The vertices indicate observed or unobserved species in the evolutionary history and the edges indicate evolutionary events between their incident vertices. Every species has observable traits that may be binary (e.g. 'horns' or 'no horns') or greater (e.g. DNA or amino acid sequences). Evolutionary events are comprised of mutations that influence these traits, and there are many ways to model them. Traditional sequence-based models include the Jukes-Cantor [JC69] and the generalized time-reversible model [Tav86].

In this work, we focus on the problem of tree reconstruction. The biological motivation is observing measured traits of contemporary taxa and using them to determine discrete relationships between the taxa and the evolutionary distance between them. Mathematically, we assume the evolutionary history of the species is depicted by a tree T and there is a statistical model of evolution along the tree. Then we try to deduce

the tree T from observations under the model.

Existing methods for estimating trees include probability- or likelihood-based methods [BSD+13], quartet-based methods [LKDA10], [MRB+14], [RSM19], concatenation [RS15], sequence-based methods [TKF91], [TKF92], [DR13], etc. In Chapters 2 and 3, we discuss quartet-based methods. In Chapter 4, we discuss sequence-based methods. For the reader, it is best to read Chapter 2 before reading Chapter 3, though Chapter 4 could be read first.

The models we use involve independent duplications and losses of data. For birth rate $\lambda \geq 0$ and death rate $\mu \geq 0$ over time, we make reference to the (constant-rate) linear birth-death process, see [Ken48], [AN72]. As an initial condition, suppose the randomly-sized population has a single ancestor and let $P_n(t)$ be the conditional probability that the population has n individuals at time t. Then the following system of differential equations is satisfied:

$$P_n'(t) = (n+1)\mu P_{n+1}(t) - n(\lambda+\mu)P_n(t) + (n-1)\lambda P_{n-1}(t), \quad n \geq 1$$
$$P_0'(t) = \mu P_1(t).$$

In Chapter 2, we address quartet-based summary methods in phylogenomics to reconstruct species trees where the root location is unknown, i.e. the tree is unrooted. Common model trees include the multi-species coalescent [RY03], uniform Poisson gene transfer [Gal07], and gene duplication and loss [ALS09], as well as models that unify these processes, see e.g. [RK12], [LGSC20]. We show that species trees in the gene duplication and loss (GDL) and the unified duplication, loss, and coalescence (DLCoal) models in [ALS09] and [RK12] are *identifiable* from phylogenomic data. Identifiability is the property that two different species trees must produce two different distributions

from which the data are generated. This property is desirable to understand a given algorithm's performance guarantees to estimating the tree. From the identifiability proof, we are also able to show two versions of ASTRAL [MRB+14], a well-known algorithm for generating unrooted species tree under the MSC model, are statistically consistent estimators of species tree under GDL and DLCoal in the sense that more independently sampled data eventually reconstructs the true species tree.

In Chapter 3, we use the results of the previous chapter to derive efficiency (i.e. *sample complexity*) estimates of the ASTRAL-based procedures. Following the approach in [SRM18], we give specific estimates of how much independently sampled data is enough to estimate the species tree with high probability. Determinants of this threshold amount of data include the birth and death rates of the duplication and loss processes, the depth of the tree, and the minimum edge length of the tree. In our estimates, we make clear the dependence on these parameters.

In Chapter 4, we change gears and discuss the traditional sequence-based reconstruction problem. Evolutionary traits are derived from DNA, so we consider the problem of reconstructing trees from independently evolving binary sequences. Our interest is in the INDEL (insertion-deletion) sequence evolution model of [TKF91] and [TKF92]. In it sites (i.e. nucleotides) independently evolve according to a Markovian substitution process in addition to Markovian insertions and deletions of new sites. A very complicated model with not many new results since its introduction, we examine an old claim in [Tha06] saying that tree reconstruction from *lengths* of sequences generated by the TKF91 model are sufficient to consistently reconstruct the phylogenetic tree T. We show that this is impossible, while introducing new methods to address the complications of INDEL processes.

1.1 Collaboration notes

The writing of Section 2.2 in Chapter 2 is based on the collaboration work in [LMWR19] with Erin Molloy and Tandy Warnow of the University of Illinois, Urbana-Champaign and Sebastien Roch. The work of Section 2.3 in Chapter 2 is a follow-up to the previous work in collaboration with Max Hill of the University of Wisconsin, Madison and Sebastien Roch, see [HLR20]. The writing of Chapter 3 is based on work in [HLR20] also done with Max Hill and Sebastien Roch. The writing of Chapter 4 is based on the collaboration work in [FLR20] with Wai-Tong Fan of Indiana University, Bloomington and Sebastien Roch.

Chapter 2

Phylogenomic methods

2.1 Introduction

In this chapter, we address phylogenomic techniques for reconstructing a phylogenetic tree. We focus on recovering the topology of the tree without regard to edge weights. Phylogenomics refers to the group of techniques that utilize the combinatorial and stochastic properties of entire genome. The *species tree* T_S depicts the evolution of the entire genome over time and thus the evolution of the species. For each gene g, its *gene tree* t_g depicts the evolution of the gene as a subset of the genome. One method that biologists use to reconstruct T_S from t_g for one or many genes g, as gene tree data is more easily obtained and characterized. Biologists possess the sequencing technology to sample hundreds or thousands of genes, and these sequences are used to generate the t_g. Then the t_g are processed in some way to estimate T_S. Practical overviews of this type of analysis are given in [Ste16], [War17], and [KK10].

However, the topologies of t_g can disagree with T_S and with one another. Gene tree heterogeneity is caused by various biological processes, including incomplete lineage sorting (ILS), lateral gene transfer (LGT), and gene duplication and loss (GDL), among others, see [Mad97]. To model these processes, we think of each gene tree as evolving "inside" the species tree according to some defined Markov process. Instances of species

and gene tree discordance caused by each of these processes are displayed in Figure 1.

Example 1 - Incomplete lineage sorting (ILS). The persistence of site mutations implies that evolutionary divergence might happen within a population of the same species. The multi-species coalescent (MSC) model introduced in [RY03] is a continuous-time backward-looking model to describe this process. At time 0, each leaf of the rooted species tree is assigned a single copy of the gene. As time increases, the Kingman coalescent process in [Kin82] is applied to each edge of the species tree as follows. Suppose there are n copies at the bottom of a given edge of length f (in coalescent time units). The first coalescent time T_1 is an exponential random variable with mean $\binom{n}{2}^{-1}$ in which a uniformly selected pair of copies coalesce. Once this coalescence occurs, there are $n-1$ remaining copies, where the next coalescent time T_2 is an exponential random variable with mean $\binom{n-1}{2}^{-1}$. Continue this process until time f. The remaining population of this edge is joined with the remaining population with the sibling edge and the Kingman coalescent process is then applied to their parent edge. In the top edge only, we condition on complete coalescence to yield the gene tree.

Much is known about estimating species trees in the presence of ILS. The most probable rooted tree for every four species coincides with the species tree [ADR11], implying the unrooted species tree topology is identifiable under the MSC from the gene tree distribution. This gives rise to quartet-based species tree reconstruction methods arising from combining gene trees, like BUCKy-pop [LKDA10] and ASTRAL [MRB$^+$14], [RSM19]. Identifiability from quartet identifiability implies these estimators are statistically conistent. Other approaches such as concatenation is not statistically consistent

and can be positively misleading (i.e. the estimate converges to the wrong tree with more samples), [RS15], [RNW18].

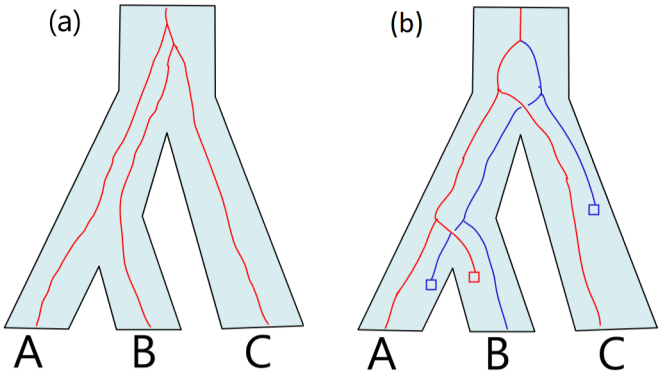

Figure 1: Gene trees generated from different models with the same species tree. (a) The left-most gene tree is generated by the multi-species coalescent model and produces the topology $(A,(B,C))$. The coalescence is deep in the sense that the copies of the gene associated to A and B failed to coalesce along the interior edge of the species tree. (b) The right-most gene tree is generated by a gene duplication and loss model and produces the topology $((A,C),B)$.

Example 2 - Gene duplication and loss (GDL). When analyzing a particular gene, some species may have multiple copies of that gene or have no copies of that gene. The linear birth-death model is a forward-looking model to describe this process. Following [ALS09], starting with a single copy at the top of the gene tree at time 0, the copy undergoes a linear birth-death process where copies are inserted at exponential rate λ and copies are deleted at exponential rate μ. Every copy that survives to a speciation event necessarily undergoes a bifurcation, resulting in a copy of that gene for each descendant edge. Lost copies are pruned from the gene tree, keeping with the intuition that we may only observe gene copies surviving to the present day. The resulting gene tree may be a *multi-copy gene tree* or *MUL-tree* where some species may

have multiple copies, and this data could be challenging to work with to predict the species tree T_S. However, even in gene tree realizations where one can get exactly one copy of each species in the MUL-tree, the gene tree can still fail to match the species tree.

Correspondingly little has been established about species tree estimation under GDL. Likelihood-based approaches such as PHYLDOG [BSD+13] have been developed, but they have not been proven to be statistically consistent. In this chapter, we adapt the methods utilizing ASTRAL in [MRB+14], [RSM19], and [DHN19] to estimators of the species tree from multi-copy gene tree distributions under GDL and prove that they are actually consistent as the number of sampled gene trees tends to infinity.

There is increasing realization in phylogenetics that forces like ILS, GDL, LGT, etc. should not be studied in isolation [RS15], [Deg18]. More complex models that combine these sources of gene-species tree discordance should be examined, see [MK09], [RK12], [LGSC20], etc. In Section 2.3, we consider a joint coalescent and branching process unifying ILS and GDL as introduced in [RK12]. The results on GDL on Section 2.2 will extend naturally to those in 2.3.

While gene trees can fail to match the species tree under any interpretation of them, the circumstances in Figure 1 are seemingly specific to where they seem less likely than realizations that produce topologies matching the species tree. For MSC [ADR11], when the species tree has three species with topology $((A, B), C)$, the gene tree is most likely to have the same topology, with the other two topologies $(B, (A, C))$ and $(A, (B, C))$ occurring with equal probability.

This is useful if we can independently sample many genes in the genome. Let k be

a positive integer. For each gene $g_i, i \in \{1, ..., k\}$, let

$$X_{g_i} = \mathbf{1}[T_{g_i} = T_S]$$

be the indicator for the event that the topology of t_{g_i} matches the topology of T_S. By the strong law of large numbers, we have

$$\frac{\sum_{i=1}^{k} X_{g_i}}{k} \to \mathbb{P}[T_{g_1} = T_S]$$

almost surely. If we define Y_{g_i} be the analogous indicator where we replace T_S with one of the non-matching topologies, then

$$\mathbb{E}X_{g_1} > \mathbb{E}Y_{g_1}$$

implies that as k goes to ∞, the sum of all X_{g_i} divided by k representing the proportion of the gene trees displaying the matching species tree is greater than that of Y_{g_i} with probability 1. So for MSC and the uniform Poisson model, a "majority-rule" consensus estimator for T_S from gene trees consistently reconstructs T_S. Additionally, this establishes *identifiability* of T_S from gene tree distributions t_g. The parameter T_S is said to be identifiable if differing T_S produce differing distributions of t_g. Establishing consistency of a method is important to understanding the method's performance guarantees.

Now, we recapitulate ASTRAL as applied to its native model, the MSC. The goal of ASTRAL is to consistently estimate an *unrooted* species tree for gene trees generated under MSC. A fundamental observation is that the topology of an unrooted species tree at least four species is determined by the displayed topologies of all quartets of species. Let $T = (V, E, L)$ be a tree with vertices V, edges E, and leaves L. In addition let S be the set of species bijectively labeling the leaves L. For each $S' \subset S$, let $T|_{S'} = (V', E', L')$ be the tree with leaf set L' corresponding to S' obtained by contracting edges (i.e.

deleting the edge and combining the incident vertices) until there are no vertices of degree 2. An example of this is given in Figure 2. Then we have the following theorem from [SS03].

Theorem 1 (Quartet Theorem)**.** *For any two unrooted trees T, T' with species set S, let $\mathcal{Q}$ be the subsets of size at most 4 in S. Then T and T' have the same topology if and only if $T|_{S'}$ and $T'|_{S'}$ have the same topology for all $S' \in \mathcal{Q}$.*

If we are able to prove that unrooted quartets are identifiable from the gene tree distributions restricted to those quartets, then the quartet theorem implies that the entire unrooted species tree is identifiable from its unrooted gene tree distributions. By establishing identifiability, we can feel comfortable that a consistent estimator might be found.

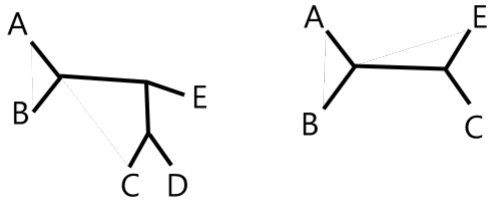

Figure 2: In this figure, we have $S = \{A, B, C, D, E\}$ and $S' = \{A, B, C, E\}$. The tree $T|_{S'}$ on the right is an example of a displayed quartet in S.

Now, we explain how to generate an unrooted gene tree under the MSC model when the inputted species tree is unrooted. When we say a tree is unrooted, we are leaving both the root location and the orientation of the tree as unknown. There are three steps to generate the tree. First, we acknowledge a not-known refinement of the tree by appending a single pendant edge (the *root edge*) to any already existing edge. This

produces a rooted tree for which the MSC process can generate a rooted tree. Second, generate a rooted tree using the MSC process for rooted trees. This produces a rooted gene tree, but it will not be useful to determine the root location of this gene tree. Finally, remove the root of the rooted gene tree to generate an unrooted gene tree. It is these unrooted gene trees that are given as input to ASTRAL.

To estimate a species tree on n species from an inputted set of unrooted gene trees, ASTRAL finds and outputs the species tree that agrees with the most quartet trees induced by the set of gene trees. By agreement, we mean a candidate species tree agrees with a quartet tree if that candidate displays the quartet. The authors define the following maximum quartet support species tree (MQSST) problem to find it:

> Given a set $\mathcal{T}$ of unrooted gene trees leaf-labeled with species S, find the tree T with leaf-labels S that maximizes $\sum_{q \in Q(T)} w(q, \mathcal{T})$, where $Q(T)$ is the set of quartets displayed by T and $w(q, \mathcal{T})$ is the number of trees in $\mathcal{T}$ that display q.

This is the exact mode of ASTRAL, which runs in $O(n^2x^2k)$ time where x is the number of bipartitions in T and k is the number of gene trees (see Theorem 1 of [MRB$^+$14]). As the number of bipartitions increases exponentially in n, the exact mode is not efficient. However, there is a heuristic version of ASTRAL that additionally takes $\mathcal{X}$, a set of bipartitions of S, as an input and restricts consideration of candidate trees T that display all the bipartitions in $\mathcal{X}$. Further, the *default* version of ASTRAL takes $\mathcal{X}$ to be the set of bipartitions displayed in all the input gene trees. In this case, we have $x = O(nk)$ so that the default mode of ASTRAL runs in $O(n^4k^3)$ time, which increases polynomially in n, also from Theorem 1 of [MRB$^+$14].

The other theorem from the original work on ASTRAL says the method is statistically consistent under MSC. We recall the proof of this theorem, returning to the stochastic properties of the MSC.

Theorem 2 (Consistency of ASTRAL - MSC)**.** *ASTRAL is a statistically consistent estimator of the species tree topology under the multi-species coalescent model, even when run in default mode.*

To prove this theorem, we need the following mathematical property about the MSC process applied to unrooted trees. The material here is not new (see [RY03], [ADR11], [Ste16]), but this proof serves as a gentle preview of proof methods of new results. Proving the analogous result to Lemma 1 for different models ensures that ASTRAL is a statistically consistent estimator for those models.

Lemma 1 (Consistency for unrooted quartets - MSC)**.** *Let T_S be an unrooted quartet with species A, B, C, D. For any rooting of this quartet, the generated unrooted gene tree t_g for any gene g has the same topology as T_S with probability greater than $1/3$, and the probabilities of the two alternative topologies are equal. Thus, the most probable gene tree matches T_S.*

Proof. Without loss of generality, assume the topology of T_S is $((A, B), (C, D))$ and the length of the interior edge is $f > 0$. Further, there are two basic cases of rooting T_S, one where the rooted version of T_S is *balanced* in that the root edge is appended to the interior edge and the other where the rooted version is *caterpillar* in that the root edge is appended to one of the pendant edges. The two possibilities are demonstrated in Figure 3.

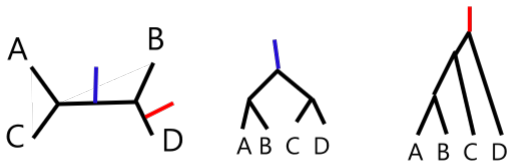

Figure 3: Starting with the unrooted quartet $((A, B), (C, D))$, the blue edge is an example of a rooting such that the result is a balanced tree (middle). The red edge is an example of a rooting such that the result is a caterpillar tree (right).

Balanced case. The root edge splits the interior edge into internal edges with lengths pf and qf where p, q are positive numbers summing to 1. Under the MSC, there are three disjoint events for which the unrooted gene tree quartet matches T_S. They are (i) the A and B copies coalesce along the interior branch of length pf; (ii) the A and B copies fail to coalesce there, but the C and D copies independently coalesce along the interior branch of length qf; and (iii) there is no coalescence beneath the interior root vertex, where the copies of A, B, C, D may be treated symmetrically and the topology t_g is chosen uniformly at random from the three possibilities. Recall that coalescent events along the branch of length pf occur as an exponential random variable with mean 1, so this implies

$$\mathbb{P}[T_g = T_S] = (1 - e^{-pf}) + e^{-pf}(1 - e^{-qf}) + \frac{1}{3}e^{-pf}e^{-qf} = 1 - \frac{2}{3}e^{-(p+q)f} = 1 - \frac{2}{3}e^{-f}.$$

The right-hand side is greater than $1/3$ for $f > 0$, implying the result. For the two alternative topologies, observe that the placement of the species C and D are exchangeable, so the probabilities of each are equal to each other. This probability is $\frac{1}{3}e^{-f}$.

Caterpillar case. Without loss of generality, suppose the root edge splits the pendant edge incident with the vertex labeled by D. Under the MSC, there are two disjoint events for which the unrooted gene quartet matches T_S. They are (i) the A and

B copies coalesce along their most recent common ancestral (MRCA) edge and (ii) the A and B copies fail to coalesce underneath the MRCA of A, B, C, so the three copies may be treated symmetrically. In (ii), the first coalescent event determines the topology of the unrooted quartet, and there are 3 choices of pairings. If no coalescent event occurs, then we have copies of A, B, C, D for which the unrooted topology is chosen uniformly at random. As before, coalescent events along the branch of length f occur as an exponential random variable with mean 1, so we have

$$\mathbb{P}[T_g = T_S] = 1 - e^{-f} + \frac{1}{3}e^{-f} = 1 - \frac{2}{3}e^{-f}.$$

The right-hand side is greater than 1/3 as before. For the two alternative topologies, the events that the root edge splits the pendant edge of C is exchangeable with splitting the pendant edge of D, so the probabilities of each are equal to each other. This probability is $\frac{1}{3}e^{-f}$. □

To prove this theorem, already proven in [ADR11], we need only the property proven in Lemma 1.

Proof. Let T_S be the true species tree and T denote a candidate tree. By proving the previous lemma, the strong law of large numbers implies that the majority-rule consensus method is statistically consistent for unrooted quartets. This is true for every quartet, i.e. each quartet topology induced by T_S has a higher probability than either of the alternative topologies. So for every quartet q and every candidate tree T, we have

$$w(q, T_S) \geq w(q, T)$$

with high probability. Thus, the *quartet score* $C(T) = \sum_{q \in \mathcal{Q}(T)} w(q, T)$ is maximized at $T = T_S$ with high probability. This maximum is unique following from the unique

maximum of the three topologies for each quartet in the lemma. So as the number of gene trees k tends to ∞, the probability of estimating the true species tree T_S tends to 1.

For the default mode, we observe that when the number of gene trees k is sufficiently large that at least one of them is identical to T_S with high probability. Thus the bipartitions in this particular gene tree are included in the default bipartition set $\mathcal{X}$, and T_S is inside the search space to be found by MQSST. This completes the proof of the theorem. □

For the rest of this chapter, we prove analogues to Lemma 1 for models involving gene duplication and loss. In the next section, we will discuss two ways to adapt multi-copy gene tree data to ASTRAL from the GDL model and show they are both statistically consistent estimators of T_S under this model. In the following section, we recall a unifying model of GDL and ILS and show ASTRAL is statistically consistent as well. In the last section, we discuss sample complexity and efficiency of ASTRAL under these new models.

2.2 Gene duplication and loss

In this section we focus on the gene duplication and loss (GDL) model of generating gene trees from the unrooted species tree T_S. Example 3 from the previous subsection specifies the progeny of copies to follow a linear birth-death process (see [Dur10], [Ken48]), but the results of this section also apply when the duplication-loss dynamics are subject to any Markovian birth-death process. In addition, the assumption that we start with a single copy at the top of the species tree is not needed. The reason for this is we will

cluster copies according to their descendants at an interior vertex of interest, and the Markovity of the birth-death process forgets the start of the process.

The resulting gene trees can have multiple copies or no copies at all, so we describe two ways to process these as single-copy trees. They are ASTRAL-one and ASTRAL-multi from [RSM19]. Suppose $\mathcal{T}$ is a collection of independently sampled multi-copy gene trees from the model.

- **ASTRAL-one:** For every multi-copy gene tree with at least one of each species, select a leaf copy of each species independently and uniformly at random and compute the displayed tree. This always results in a single-copy gene tree. If some species go extinct, then ignore this gene tree and move on to the next one.

- **ASTRAL-multi:** For every multi-copy gene tree, take all selections of a single copy from each species compute the displayed trees. These are all single-copy gene trees. The single-copy gene trees are not independent and a multi-copy gene tree with more selections gives a larger signal.

We define the quartet score for ASTRAL-multi, and the quartet score for ASTRAL-one will follow from it. Let S be the set of n species and R be the set of m individuals. The input are the multi-copy gene trees $\mathcal{T} = \{t_{g_i}\}_{i=1}^{k}$ where t_{g_i} is labeled by individuals in R_i. For any candidate species tree $\widetilde{T}$ labeled by S, an extended species tree $\widetilde{T}_{ext}$ labeled by R is built by adding to each leaf of $\widetilde{T}$ all individuals corresponding to that species as a polytomy. The quartet score of a candidate tree $\widetilde{T}$ with respect to $\mathcal{T}$ is then

$$Q_k\left(\widetilde{T}\right) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subset R_i} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_{g_i}^{\mathcal{J}}) \tag{2.1}$$

where $\mathbf{1}(T_1, T_2)$ is the indicator that T_1 and T_2 agree and $T_1^{\mathcal{J}}$ is the restriction of T_1 to individuals $\mathcal{J}$. The inner sum is analogous to $w(q, \mathcal{T})$ from the previous section in that the sum counts all possible selections of individuals from four different species.

The quartet score for ASTRAL-one is simpler in that we replace t_{g_i} with $\tilde{t}_{g_i}$ obtained by picking a uniform random gene copy of each species. Then $t_{g_i}^{\mathcal{J}}$ in (2.1) is replaced by $\tilde{t}_{g_i}^{\mathcal{J}}$.

We will show that these quartet scores are maximized at the true species tree. Before proceeding to consistency of these estimators, we will show that the species tree T_S is identifiable from its multi-copy gene trees.

2.2.1 Proof sketch

The key insight is that for each quartet of species $\mathcal{Q} = \{A, B, C, D\}$, the species quartet topology is identified by taking independently and uniformly sampling copies as in the ASTRAL-one procedure. The statement and proof of this fact is similar in spirit to Lemma 1, but the proof is more complicated.

Assume the species tree restricted to $\mathcal{Q}$ has topology $((A, B), (C, D))$. As in the proof of Lemma 1, we consider the two cases where the rooting of the quartet results in either a balanced or caterpillar quartet as in Figure 3. In the balanced tree case, we let R be the most recent common ancestor of all species in $\mathcal{Q}$. In the caterpillar tree case, we let R be the most recent common ancestor of A, B, C. Let a, b, c, d be the random gene copies in A, B, C, D, respectively. We make a few key observations about the balanced case, noting the caterpillar case is similar. (i) If a, b, c, d descend from different lineages in R, then by symmetry all three quartet topologies are equally

likely. (ii) If a and b descend from the same lineage in R (or c and d descend from the same lineage), the true species quartet topology occurs with probability 1. (iii) If a and b descend from different lineages and (for example) c descends from the same lineage as a, then the unrooted topology $ac|bd$ occurs with probability 1. The identifiability proof centers around accounting for each of these cases.

2.2.2 Identifiability

Now, we prove the theorem on identifiability.

Theorem 3 (Identifiability - GDL)**.** *Let T_S be a species tree with $n \geq 4$ leaves. Then the unrooted topology of T_S is identifiable from the distribution of multi-copy gene trees $\mathcal{T}$ under the GDL model over T_S.*

Proof. By Theorem 1, the unrooted topology is defined by its quartets. Let $\mathcal{Q} = \{A, B, C, D\}$ be four species in T and $T|_{\mathcal{Q}}$ be the restriction of T to $\mathcal{Q}$. Assume that the unrooted quartet topology is $AB|CD$. Let T_g be a multi-copy gene tree generated under the GDL model over T_S and $T_g|_{\mathcal{Q}}$ be the restriction of T_g to the gene copies a, b, c, d. Let $\mathbb{P}'$ be the probability measure conditioned on having at least one copy of each species in $\mathcal{Q}$, then independently pick a copy of each uniformly at random. Let q be the corresponding quartet topology under $T_g|_{\mathcal{Q}}$. We show that the most likely outcome is $q = ab|cd$. Now we split into the two rooting cases.

Balanced case. Let R be the most recent common ancestor of $\mathcal{Q}$ in T_S and I be the number of gene copies exiting R forward in time. Letting $\mathbb{P}'_I$ be conditional measure of $\mathbb{P}'$ further conditioned on I copies at R, the law of total probability implies

$$\mathbb{P}'[q = ab|cd] = \mathbb{E}'\left[\mathbb{P}'_I[q = ab|cd]\right].$$

The same expansions hold for $q = ac|bd$ and $q = ad|bc$. It suffices to prove

$$\mathbb{P}'_I[q = ab|cd] > \mathbb{P}'_I[q = ac|bd] \vee \mathbb{P}'_I[q = ad|bc] \tag{2.2}$$

almost surely for all $I \geq 1$. Here, we use the notation $z_1 \vee z_2 = \max\{z_1, z_2\}$.

Let $i_x \in \{1, ..., I\}$ be the ancestral lineage of $x \in \{a, b, c, d\}$ in R. Then

$$\begin{aligned}
\mathbb{P}'_I[q = ab|cd] &= \mathbb{P}'_I[i_a = i_b \cup i_c = i_d] + \mathbb{P}'_I[q = ab|cd \text{ and } \{i_a, i_b, i_c, i_d\} \text{ are distinct}] \\
&= \mathbb{P}'_I[i_a = i_b] + \mathbb{P}'_I[i_c = i_d] - \mathbb{P}'_I[i_a = i_b, i_c = i_d] \\
&\quad + \mathbb{P}'_I[q = ab|cd \text{ and } \{i_a, i_b, i_c, i_d\} \text{ are distinct}].
\end{aligned} \tag{2.3}$$

Noting that $\{i_a = i_c \neq i_d = i_b\} = \{i_b \neq i_a = i_c \neq i_d\} \cap \{i_a \neq i_b = i_d \neq i_c\}$, we similarly have

$$\begin{aligned}
\mathbb{P}'_I[q = ac|bd] &= \mathbb{P}'_I[i_b \neq i_a = i_c \neq i_d] + \mathbb{P}'_I[i_a \neq i_b = i_d \neq i_c] \\
&\quad - \mathbb{P}'_I[i_a = i_c \neq i_d = i_b] + \mathbb{P}'_I[q = ac|bd \text{ and } \{i_a, i_b, i_c, i_d\} \text{ are distinct}] \\
&\leq \mathbb{P}'_I[i_b \neq i_a = i_c \neq i_d] + \mathbb{P}'_I[i_a \neq i_b = i_d \neq i_c] \\
&\quad + \mathbb{P}'_I[q = ac|bd \text{ and } \{i_a, i_b, i_c, i_d\} \text{ are distinct}].
\end{aligned} \tag{2.4}$$

By symmetry of the GDL process above R under $\mathbb{P}'_I$, the last terms on the right-hand side of (2.3) and (2.4) are equal. The first two terms on the right-hand side of (2.4) are equal by independence and exchangeability of the pairs (i_a, i_b) and (i_c, i_d) under $\mathbb{P}'_I$. Putting it together, this implies

$$\begin{aligned}
\mathbb{P}'_I[q = ab|cd] - \mathbb{P}'_I[q = ac|bd] &\geq \mathbb{P}'_I[i_a = i_b] + \mathbb{P}'_I[i_c = i_d] \\
&\quad - \mathbb{P}'_I[i_a = i_b, i_c = i_d] - 2\mathbb{P}'_I[i_b \neq i_a = i_c \neq i_d].
\end{aligned}$$

The events $i_a = i_b$ and $i_c = i_d$ are independent conditioned on I. Also,

$$\begin{aligned}\mathbb{P}'_I[i_b \neq i_a = i_c \neq i_d] &= \mathbb{P}'_I[i_a = i_c | i_a \neq i_b, i_c \neq i_d]\mathbb{P}'_I[i_a \neq i_b]\mathbb{P}'_I[i_c \neq i_d] \\ &= \frac{1}{I}\mathbb{P}'_I[i_a \neq i_b]\mathbb{P}'_I[i_c \neq i_d],\end{aligned}$$

where we use the fact that given the choice of i_a, the choice of i_c is uniform and independent of i_a. Letting $x = \mathbb{P}'_I[i_a = i_b]$ and $y = \mathbb{P}'[i_c = i_d]$, this implies

$$\mathbb{P}'_I[q = ab|cd] - \mathbb{P}'_I[q = ac|bd] \geq x + y - xy - \frac{2}{I}(1-x)(1-y) =: h(x,y).$$

For fixed y, the function $h(x,y)$ is linear in x with a non-negative slope and $h(1,y) = 1$, so $h(\cdot, y)$ achieves its minimum at the smallest value allowed for x. The same holds for y. In the next lemma, we show that i_a and i_b are positively correlated to where $x \geq 1/I$. Here, we use the notation $z_1 \wedge z_2 = \min\{z_1, z_2\}$

Lemma 2. *Almost surely, we have $x \wedge y \geq 1/I$.*

Proof. For each $j \in \{1, ..., I\}$, let N_j be the number of gene copies descending from the jth lineage in R that survive to the most recent common ancestor R' of A and B. Conditioning on $\{N_j\}_j$, the choice of a and b is independent, with i_a and i_b being picked proportionally to the corresponding N_j. This is because the gene copies in R' are equally likely to have given rise to a (and similarly for b). The law of total probability implies

$$\mathbb{P}'_I[i_a = i_b] = \mathbb{E}'_I\left[\mathbb{P}'_I[i_a = i_b | (N_j)_j]\right] = \mathbb{E}'_I\left[\frac{\sum_{j=1}^I N_j^2}{\left(\sum_{j=1}^I N_j^2\right)}\right].$$

The quadratic mean is greater than the arithmetic mean, so the fraction inside the expectation on the right-hand side is at least $1/I$. The procedure is the same for y. This completes the proof of the lemma. □

Since h is minimized at $x = y = 1/I$, we have

$$h(x,y) \geq h\left(\frac{1}{I}, \frac{1}{I}\right) = \frac{2}{I} - \frac{1}{I^2} - \frac{2(I-1)^2}{I^3} = \frac{3I-2}{I^3},$$

where the right-hand side is positive for all $I \geq 1$. This finishes the proof of (2.2) in the balanced case.

Remark. We actually have the stronger quantitative bound

$$h(x,y) \geq \frac{1}{4}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right), \tag{2.5}$$

proven as follows. Across (x,y) such that $x \wedge y \geq 1/2$, we have h is bounded below by its value at $x = y = 1/2$, and

$$h(x,y) \geq h\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{4} + \frac{1}{2I} \geq \frac{1}{4}\left(x - \frac{1}{I}\right) \vee \left(y - \frac{1}{I}\right).$$

Across (x,y) such that $y < 1/2$, we have

$$\begin{aligned} h(x,y) &= \left(x - \frac{1}{I}\right)(1-y) + \left(y - \frac{1}{I}\right)(1-x) + \frac{1}{I}x(1-y) + \frac{1}{I}y(1-y) \\ &> \frac{1}{2}\left(x - \frac{1}{I}\right) \end{aligned}$$

by replacing the $1 - y$ in the first term by $1/2$ and observing the other terms are non-negative. Then across (x,y) such that $x \wedge y < 1/2$, we have

$$h(x,y) \geq \frac{1}{2}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right),$$

implying the bound. This result could be used to strengthen identifiability and consistency results by quantifying sample complexity.

Caterpillar case. In this case, assume the topology of the rooted quartet is $T^{\mathcal{Q}} = (((A,B),C),D)$ and let R be the most recent common ancestor of A, B, C in $T^{\mathcal{Q}}$, and

let I be the number of gene copies exiting R. As before, it suffices to prove (2.2) almost surely. Let $i_x \in \{1, ..., I\}$ be the ancestral lineage of $x \in \{a, b, c\}$ in R. Then

$$\mathbb{P}'_I[q = ab|cd] = \mathbb{P}'_I[i_a = i_b] + \mathbb{P}'_I[q = ab|cd \text{ and } i_a, i_b, i_c \text{ are distinct}]$$

and

$$\mathbb{P}'_I[q = ac|bd] = \mathbb{P}'_I[i_b \neq i_a = i_c] + \mathbb{P}'_I[q = ac|bd \text{ and } i_a, i_b, i_c \text{ are distinct}].$$

By symmetry again, the last terms on the right-hand side are the same. Define $x = \mathbb{P}'_I[i_a = i_b]$ and proceed similarly as in the balanced case. This implies

$$\begin{aligned}\mathbb{P}'_I[q = ab|cd] - \mathbb{P}'_I[q = ac|bd] &= \mathbb{P}'_I[i_a = i_b] - \mathbb{P}'_I[i_b \neq i_a = i_c] \\ &= x - (1 - x)\mathbb{P}'_I[i_a = i_c | i_a \neq i_b] = x - \frac{1}{I}(1 - x) =: g(x).\end{aligned}$$

The function g attains its minimum value at the minimum value of x, which by Lemma 2 is $1/I$. Then

$$\mathbb{P}'_I[q = ab|cd] - \mathbb{P}'_I[q = ac|bd] \geq g\left(\frac{1}{I}\right) = \frac{1}{I} - \frac{1}{I} + \frac{1}{I^2} = \frac{1}{I^2}.$$

The right-hand side is positive for all $I \geq 1$, completing the proof in the caterpillar case. Combining this result with the quartet theorem, this completes the proof. □

Follow-up to previous remark. The more quantitative bound

$$g(x) \geq x - \frac{1}{I}$$

holds for the caterpillar case, which has similar behavior to the balanced case.

In this proof of identifiability, we have proved an analogue to Lemma 1 which is essentially proving statistical consistency of the ASTRAL-one pipeline. For each gene tree T_{g_i}, pick in each species a random gene copy and run ASTRAL on the set of modified gene trees.

Theorem 4 (Consistency of ASTRAL-one under GDL)**.** *ASTRAL-one is statistically consistent under the GDL model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL-one converges to T_S almost surely, when run in exact mode or in its default constrained version.*

Proof. We start with the exact version. The input is the collection of multi-copy gene trees $\mathcal{T} = \{t_{g_i}\}_{i=1}^{k}$ where t_{g_i} are labeled by individuals (i.e. gene copies) R_i. Recall the quartet score is

$$Q_k\left(\widetilde{T}\right) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subset R_i} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, \tilde{t}_{g_i}^{\mathcal{J}}).$$

Under the GDL model, by independence of the gene trees, $Q_k\left(\widetilde{T}\right)/k$ converges almost surely to its expectation simultaneously for all unrooted species tree topologies over S.

For any species $A \in S$ and any gene tree $\tilde{t}_{g_i}$, let A_i be the gene copy in A on $\tilde{t}_{g_i}$ if it exists and let $\mathcal{E}_i^A$ be the event that it exists. Then for the quartet of species $\mathcal{Q} = \{A, B, C, D\}$, let $\mathcal{Q}_i$ and $\mathcal{E}_i^{\mathcal{Q}}$ be the associated quartet of gene copies and event that some gene copies of all the species exist. Then

$$\begin{aligned}\mathbb{E}\left[\frac{Q_k\left(\widetilde{T}\right)}{k}\right] = &\sum_{\mathcal{Q}=\{A,B,C,D\}} \mathbb{E}\left[\mathbf{1}(\widetilde{T}_{ext}^{\mathcal{Q}_1}, \tilde{t}_{g_1}^{\mathcal{Q}_1})\middle|\mathcal{E}_1^{\mathcal{Q}}\right]\mathbb{P}[\mathcal{E}_1^{\mathcal{Q}}] \\ &+ \sum_{\mathcal{Q}=\{A,B,C,D\}} \mathbb{E}\left[\mathbf{1}(\widetilde{T}_{ext}^{\mathcal{Q}_1}, \tilde{t}_{g_1}^{\mathcal{Q}_1})\middle|\left(\mathcal{E}_1^{\mathcal{Q}}\right)^c\right]\mathbb{P}[\left(\mathcal{E}_1^{\mathcal{Q}}\right)^c]. \qquad (2.6)\end{aligned}$$

If $\mathcal{E}_1^{\mathcal{Q}}$ fails to occur, then there is no complete quartet to sample and the indicator equals 0. So there is no contribution to the sum from the quartet $\mathcal{Q}$ in the first gene tree. So the second series sums to 0.

Now, let a, b, c, d be random gene copies on the first gene tree sample t_{g_1} in species A, B, C, D. Then if q is the topology of t_1 restricted to a, b, c, d (out of the three), then

the expectation in the first sum coincides with $\mathbb{P}'[q = \widetilde{T}^{\mathcal{Q}}]$, i.e.

$$\mathbb{E}\left[\mathbf{1}(\widetilde{T}^{\mathcal{Q}_1}_{ext}, \tilde{t}^{\mathcal{Q}_1}_{g_1}) \middle| \mathcal{E}^{\mathcal{Q}}_1\right] = \mathbb{P}'[q = \widetilde{T}^{\mathcal{Q}}].$$

From (2.2) from the identifiability proof, this expectation has a unique maximum at the true species tree, i.e. $\widetilde{T} = T_S$. Maximizing this quantity together with (2.6) implies that almost surely $Q_k\left(\widetilde{T}\right)$ is maximized by T_S as k goes to ∞. This completes the proof for the exact version.

The default version is statistically consistent by the same extension as for MSC. As the number of gene trees sampled tends to infinity, the true species tree will appear as one of the input gene trees almost surely. So ASTRAL-one returns the true species tree topology almost surely as the number of sampled gene trees increases. □

2.2.3 ASTRAL-multi

Next, we prove that ASTRAL-multi is statistically consistent. The proof is similar to the identifiability proof.

Theorem 5 (Consistency of ASTRAL-multi under GDL)**.** *ASTRAL-multi, where copies of a gene in a species are treated as multiple alleles within the species, is statistically consistent under the GDL model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL-multi converges to T_S almost surely, when run in exact mode or in its default constrained version.*

Proof. We start with the exact mode. As before, the input are the gene trees $\mathcal{T} = \{t_{g_i}\}_{i=1}^{k}$ with t_{g_i} labeled by individuals from R_i. Recall the quartet score of a candidate tree $\widetilde{T}$

with respect to $\mathcal{T}$ is

$$Q_k\left(\widetilde{T}\right) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subset R_i} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_{g_i}^{\mathcal{J}}).$$

For any quartet of gene copies $\mathcal{J} = \{a, b, c, d\}$, define $m(\mathcal{J})$ to be the corresponding set of species. A result of [RSM19] is that if $\mathcal{J}$ has cardinality less than 4, meaning at least two of the copies are of the same species, then those quartets are *trivial* in the sense that they contribute mass 0 to the quartet score. So to maximize $Q_k\left(\widetilde{T}\right)$, it suffices to maximize

$$\widetilde{Q}_k\left(\widetilde{T}\right) = \sum_{i=1}^{k} \sum_{\substack{\mathcal{J}=\{a,b,c,d\}\subset R_i \\ |m(\mathcal{J})|=4}} \mathbf{1}\left(\widetilde{T}_{ext}^{\mathcal{J}}, t_i^{\mathcal{J}}\right). \tag{2.7}$$

Again, by independence of the gene trees, the sample average $\widetilde{Q}_k\left(\widetilde{T}\right)/k$ converges almost surely to its expectation simultaneously for all candidate trees over S. The expectation is

$$\begin{aligned}\mathbb{E}\left[\frac{\widetilde{Q}_k\left(\widetilde{T}\right)}{k}\right] &= \mathbb{E}\left[\sum_{\substack{\mathcal{J}=\{a,b,c,d\}\subset R_i \\ |m(\mathcal{J})|=4}} \mathbf{1}\left(\widetilde{T}_{ext}^{\mathcal{J}}, t_i^{\mathcal{J}}\right)\right] \\ &= \sum_{\mathcal{Q}=\{A,B,C,D\}} \mathbb{E}\left[\sum_{\mathcal{J}:R_1:m(\mathcal{J})=\mathcal{Q}} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_1^{\mathcal{J}})\right], \end{aligned} \tag{2.8}$$

where we used the fact that the sampled gene trees are independent and identically distributed (even though the quartets within a gene tree are not). Now we introduce some notation. Let $\mathcal{N}_{AB|CD}^{\mathcal{Q}}$ (respectively $\mathcal{N}_{AC|BD}^{\mathcal{Q}}, \mathcal{N}_{AD|BC}^{\mathcal{Q}}$) be the number of choices consisting of one gene copy in t_{g_1} from each species in $\mathcal{Q}$ whose corresponding restriction $t_{g_1}^{\mathcal{Q}}$ agrees with $AB|CD$ (respectively, $AC|BD, AD|BC$). Then the quantity inside the expectation on the right-hand side of (2.8) equals $\mathbb{E}\left[\mathcal{N}_{\widetilde{T}^{\mathcal{Q}}}^{\mathcal{Q}}\right]$.

In a similar spirit to (2.2), we show the right-hand side of (2.8) is maximized at the true species tree $T^{\mathcal{Q}}$. Without loss of generality, assuming the true quartet is $AB|CD$, we want to show

$$\mathbb{E}[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] > \mathbb{E}[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \vee \mathbb{E}[\mathcal{N}^{\mathcal{Q}}_{AD|BC}]. \tag{2.9}$$

Then the law of large numbers implies that almost surely we have $\widetilde{Q}_k\left(\widetilde{T}\right)/k$ is maximized at $\widetilde{T} = T_S$ as k goes to ∞. Now we establish (2.9) when $\mathcal{Q} = \{A, B, C, D\}$ and the unrooted quartet topology in T_S is $AB|CD$. As with the previous section, there are two cases for the rooting of the quartet of T_S, balanced or caterpillar.

Balanced case. Let R be the most recent common ancestor of $\mathcal{Q}$ in T_S and let I be the number of gene copies exiting R. For $j \in \{1, \ldots, I\}$, let $\mathcal{A}_j$ be the number of gene copies in A descending from j in R, and similarly define $\mathcal{B}_j$, $\mathcal{C}_j$ and $\mathcal{D}_j$.

By the law of iterated expectation, $\mathbb{E}[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] = \mathbb{E}[\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}]]$. We show that

$$\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] > \mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \vee \mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AD|BC}] \tag{2.10}$$

almost surely for any $I \geq 1$. By symmetry, we may define $X^{=} := \mathbb{E}_I[\mathcal{A}_j\mathcal{B}_j] = \mathbb{E}_I[\mathcal{A}_1\mathcal{B}_1]$, $Y^{=} := \mathbb{E}_I[\mathcal{C}_j\mathcal{D}_j] = \mathbb{E}_I[\mathcal{C}_1\mathcal{D}_1]$. These are the number of choices of copies of a and b such that $i_a = i_b = j$ and choices of copies of c and d such that $i_c = i_d = j$, for each $j \in \{1, ..., I\}$. Similarly, we may define $X^{\neq} \equiv \mathbb{E}_I[\mathcal{A}_j\mathcal{B}_k] = \mathbb{E}_I[\mathcal{A}_1]\mathbb{E}_I[\mathcal{B}_1]$ as well as $Y^{\neq} \equiv \mathbb{E}_I[\mathcal{C}_j\mathcal{D}_k] = \mathbb{E}_I[\mathcal{C}_1]\mathbb{E}_I[\mathcal{D}_1]$. These are the number of choices of copies of a and b such that $i_a = j$ and $i_b = k$ and choices of copies of c and d such that $i_c = j$ and $i_d = k$ for differing $j, k \in \{1, ..., I\}$. Then define X to be the expected number of pairs consisting of a single gene copy from A and B, which equals $X = IX^{=} + I(I-1)X^{\neq}$. There is a similar expression for Y, defined to be the expected number of pairs consisting of a single gene copy from C and D.

Counting the number of possibilities for events expressed in (2.3) and (2.4), we have

$$\begin{aligned}\mathbb{E}'[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] &= (IX^{=})Y + X(IY^{=}) - (IX^{=})(IY^{=}) \\ &\quad + \frac{1}{3}I(I-1)(I-2)(I-3)\mathbf{1}(I \geq 4)X^{\neq}Y^{\neq}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}'[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] &= 2I(I-1)X^{\neq}(I-1)Y^{\neq} - I(I-1)X^{\neq}Y^{\neq} + X(IY^{=}) - (IX^{=})(IY^{=}) \\ &\quad + \frac{1}{3}I(I-1)(I-2)(I-3)\mathbf{1}(I \geq 4)X^{\neq}Y^{\neq}.\end{aligned}$$

Similarly to the previous section, subtracting yields

$$\begin{aligned}&\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] - \mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \\ &\quad \geq (IX^{=})Y + X(IY^{=}) - (IX^{=})(IY^{=}) - I(I-1)X^{\neq}[2(I-1)Y^{\neq}] \\ &\quad = XY\left[x + y - xy - 2(1-x)(1-y)\frac{1}{I}\right],\end{aligned}$$

where here we define $x = \frac{IX^{=}}{X}, y = \frac{IY^{=}}{Y}$. It suffices to show that almost surely, $x, y \geq 1/I$. That is implied by the following positive correlation result.

Lemma 3. *Almost surely,* $X^{=} \geq X^{\neq}$.

This implies

$$x = \frac{IX^{=}}{IX^{=} + I(I-1)X^{\neq}} \geq \frac{IX^{=}}{IX^{=} + I(I-1)X^{=}} = \frac{1}{I}.$$

Proof. For $j \in \{1, \ldots, I\}$, let N_j be the number of gene copies at the divergence of the most recent common ancestor of A and B that are descending from j in R. Then, for $j \in \{1, \ldots, I\}$, since $\mathcal{A}_j$ and $\mathcal{B}_j$ are conditionally independent given $(N_j)_j$ under $\mathbb{E}_I$, it follows that

$$X^{=} = \mathbb{E}_I[\mathbb{E}_I[\mathcal{A}_j\mathcal{B}_j \,|\, (N_j)_j]] = \mathbb{E}_I[(N_j\alpha)(N_j\beta)] = \alpha\beta\mathbb{E}_I[N_j^2],$$

where α (respectively β) is the expected number of gene copies in A (respectively B) descending from a single gene copy in the most recent common ancestor of A and B under $\mathbb{E}_I$. Similarly, for $j \neq k \in \{1, \ldots, I\}$,

$$X^{\neq} = \mathbb{E}_I[\mathbb{E}_I[\mathcal{A}_j\mathcal{B}_k \,|\, (N_j)_j]] = \mathbb{E}_I[(N_j\alpha)(N_k\beta)] = \alpha\beta\mathbb{E}_I[N_jN_k] \leq \alpha\beta\mathbb{E}_I[N_j^2],$$

by Cauchy-Schwarz and $\mathbb{E}_I[N_j^2] = \mathbb{E}_I[N_k^2]$ This completes the proof of the lemma.

□

Following the end steps of proof of the identifiability theorem, showing $x, y \geq 1/I$ is enough to establish (2.9). This completes the proof in the balanced case.

Caterpillar case. Assume without loss of generality that $T^{\mathcal{Q}} = (((A, B), C), D)$ and let R be the most recent common ancestor of A, B, C (but not D) in T. We want to establish (2.10) in this case. For $i = 1, 2, 3$, let $\mathcal{N}_{AB|CD}^{\mathcal{Q},\{i\}}$ (respectively $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{i\}}$) be the number of choices consisting of one gene copy from each species in $\mathcal{Q}$ whose corresponding restriction on $t^{\mathcal{Q}}$ agrees with $AB|CD$ (respectively $AC|BD$) and where, in addition, copies of A, B, C descend from i distinct lineages in R. We make five observations:

- Contributions to $\mathcal{N}_{AB|CD}^{\mathcal{Q},\{2\}}$ necessarily come from copies in A and B descending from the same lineage in R, together with a copy in C descending from a distinct lineage and any copy in D. Similarly for $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{2\}}$.

- Moreover $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{1\}} = 0$ almost surely, as in that case the corresponding copies from A and B have a common ancestor below R.

- Arguing similarly to the identifiability theorem, by symmetry we have the equality $\mathbb{E}_I[\mathcal{N}_{AB|CD}^{\mathcal{Q},\{3\}}] = \mathbb{E}_I[\mathcal{N}_{AC|BD}^{\mathcal{Q},\{3\}}]$.

- For $j \in \{1, \ldots, I\}$, let $\mathcal{A}_j$ be the number of gene copies in A descending from j in R, and similarly define $\mathcal{B}_j$, $\mathcal{C}_j$. Let $\mathcal{D}$ be the number of gene copies in D. Then, under the conditional probability $\mathbb{P}_I$, $\mathcal{D}$ is independent of $(\mathcal{A}_j, \mathcal{B}_j, \mathcal{C}_j)_{j=1}^I$. Moreover, under $\mathbb{P}_I$, $(\mathcal{C}_j)_{j=1}^I$ is independent of $(\mathcal{A}_j, \mathcal{B}_j)_{j=1}^I$.

- Similarly to the balanced case, by symmetry we define $X^= \equiv \mathbb{E}_I[\mathcal{A}_j\mathcal{B}_j] = \mathbb{E}_I[\mathcal{A}_1\mathcal{B}_1]$, $X^{\neq} \equiv \mathbb{E}_I[\mathcal{A}_j\mathcal{B}_k] = \mathbb{E}_I[\mathcal{A}_1]\mathbb{E}_I[\mathcal{B}_1]$ for all $j, k \leq I$ with $j \neq k$. Define also $X = IX^= + I(I-1)X^{\neq}$, $Y \equiv \mathbb{E}_I[\mathcal{C}_1]$ and $Z \equiv \mathbb{E}_I[\mathcal{D}]$.

Then

$$\begin{aligned}\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] &= \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{1\}}_{AB|CD}] + \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{2\}}_{AB|CD}] + \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{3\}}_{AB|CD}] \\ &\quad IX^=YZ + I(I-1)X^=YZ + \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{1\}}_{AB|CD}]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] &= \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{2\}}_{AC|BD}] + \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{3\}}_{AC|BD}] \\ &= I(I-1)X^{\neq}YZ + \mathbb{E}_I[\mathcal{N}^{\mathcal{Q},\{3\}}_{AC|BD}].\end{aligned}$$

Subtracting gives

$$\begin{aligned}&\mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] - \mathbb{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \\ &= IX^=YZ + I(I-1)X^=YZ - I(I-1)X^{\neq}YZ\end{aligned}$$

where we see the right-hand side is positive by Lemma 3. This completes the proof in the caterpillar case and the proof of the theorem. □

2.3 Unified gene duplication, loss, and coalescence model

In this section, we introduce the unified coalescence, duplication, and loss model in [RK12], called DLCoal. The purpose of the model is to describe multi-copy gene trees that are subject to both ILS from the first section and GDL from the last section. Again, our input is a collection $\mathcal{T} = \{t_i\}_{i=1}^{k}$ of k unrooted gene trees from the unrooted species tree T_S. A key stipulation of DLCoal is that duplications result in daughter lineages in which coalescence is conditioned to occur with probability 1. The biological reasoning behind this assumption is that a mutation or polymorphism that applies to one copy of the gene is highly negatively correlated with other mutations to its sibling copy. Gene trees are generated in two steps under DLCoal:

- *Step 1. Birth-death process of gene duplication and loss with daughter edges.* In this section, trees generated by this process are called *locus* trees. Locus trees are generated from T_S and then pruned by the same forward-in-time process in GDL. Every edge may be labeled as a mother segment or daughter segment.

- *Step 2. Coalescent process on a locus tree.* In this section, trees generated in this step are called *gene* trees. Gene trees are generated by a backward-in-time coalescent process within the locus tree. Copies at the bottom of a directed mother segment in the locus tree undergo the Kingman coalescent process for a time equal to the length of the directed edge in coalescent time units. Copies at the bottom of a directed daughter segment in the locus tree necessarily coalesce before reaching the top of the edge. Note that losses in Step 1 can result in locus tree edges with

partial mother/daughter segments. Continuing along ancestral edges, this process yields a gene tree whose leaves are in bijective correspondence with the leaves of the locus tree. As usual, the multi-copy gene trees $\{t_i\}_{i=1}^k$ are assumed independent and identically distributed. An realization of this process is given in Figure 4.

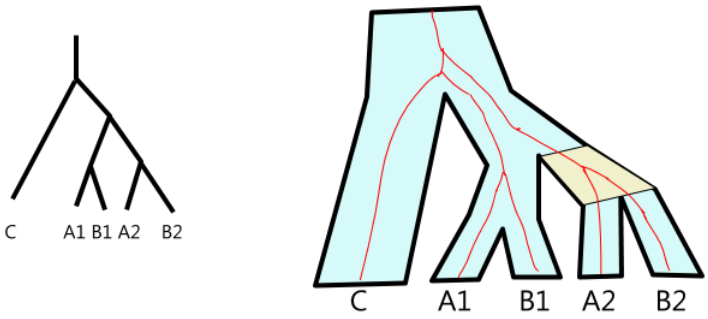

Figure 4: The left display is a generated locus tree generated from the rooted species topology $(C,(A,B))$ in which there are two copies of species A and two copies of species B. Suppose the most recent common ancestral vertex of these four copies is a duplication, and there are no further duplications or losses. In the right display, the daughter segment is given in yellow. While coalescence did occur in the ancestral edge of $A1$ and $B1$, it was not ensured by the conditioning. Full coalescence did not occur in the ancestral edge of $\{A1, B1, A2, B2\}$. Of the four selections of A and B for ASTRAL-one to make, all four selections in the *locus* tree display the true species tree, while only two display the true species tree in the *gene* tree.

As with the GDL model, we can process the multi-copy gene trees in this section using ASTRAL-one and ASTRAL-multi. We prove identifiability of T_S from gene tree distributions first. Some ideas used in the proofs are similar, though new ideas are used to accommodate Step 2.

Theorem 6 (Identifiability - DLCoal)**.** *Let T_S be a species tree with $n \geq 4$ leaves. Then the unrooted topology of T_S is identifiable from the distribution of multi-copy gene trees $\mathcal{T}$ under the DLCoal model over T_S.*

Then consistency of ASTRAL-one and ASTRAL-multi follows.

Theorem 7 (Consistency of ASTRAL-one under DLCoal)**.** *ASTRAL-one is statistically consistent under the DLCoal model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL-one converges to T_S almost surely, when run in exact mode or in its default constrained version.*

Theorem 8 (Consistency of ASTRAL-multi under DLCoal)**.** *ASTRAL-multi, where copies of a gene in a species are treated as multiple alleles within the species, is statistically consistent under the DLCoal model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL-multi converges to T_S almost surely, when run in exact mode or in its default constrained version.*

2.3.1 DLCoal Identifiability - balanced case

The high-level idea is the same as in the previous section, but we will use a slightly different conditioning that adapts better to ASTRAL-multi. Define $\mathcal{X} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ to be the count of the number of copies in species $\mathcal{Q} = \{A, B, C, D\}$ and let $\mathbb{P}'$ be the probability measure subject to this conditioning. While we only reconstruct an unrooted species tree, both locus trees and gene trees are generated from the species tree. On a restriction to four species, as with GDL there are two cases of root location to consider: balanced with rooted topology $((A, B), (C, D))$ or caterpillar with rooted topology $(((A, B), C), D)$. To simplify notation, define the following events for gene copies a, b, c, d from A, B, C, D:

$$Q_1 = \{q = ab|cd\},\ Q_2 = \{q = ac|bd\},\ Q_3 = \{q = ad|bc\}.$$

Proof. We start with the balanced case.

Balanced case. Let R be the most recent common ancestor of $\mathcal{Q}$ in the species quartet $T^{\mathcal{Q}}$ and I be the number of locus copies exiting R (forward in time). Let $\mathbb{P}''$ be the probability measure indicating conditioning on I as well as $\mathcal{X}$. For any selection of copies (a, b, c, d) from each species in the quartet, let $i_x \in \{1, ..., I\}$ be the ancestral lineage of $x \in \{a, b, c, d\}$ in R. By the law of total probability, we have

$$\mathbb{P}[Q_1] = \mathbb{E}[\mathbb{P}''[Q_1]]. \tag{2.11}$$

Hence, in order to show identifiability of the species quartet, it is sufficient to show that

$$\mathbb{P}[Q_1] > \mathbb{P}[Q_2] \vee \mathbb{P}[Q_3]$$

when the copies of (A, B, C, D) are chosen uniformly at random. Let $\vec{1}$ denote the vector of all ones and $<$ refer to the lexicographic (dictionary) ordering. If $\mathcal{X} < \vec{1}$ (that is, at least one of (A, B, C, D) fails to have a copy to select), then ASTRAL-one selects Q_1, Q_2, Q_3 each with probability 0. So we consider the case $\mathcal{X} \geq \vec{1}$. To prove the above inequality, it is sufficient to prove the following proposition.

Proposition 1 (Quartet identifiability under DLCoal - balanced case)**.** *Let* $x = \mathbb{P}''[i_a = i_b]$ *and* $y = \mathbb{P}''[i_c = i_d]$*. On the events* $\mathcal{X} \geq \vec{1}$ *and* $I \geq 1$*, we have almost surely*

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{12}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right).$$

The next lemma establishes $x \wedge y \geq 1/I$, similarly to 3. The difference in the statements comes down to conditioning on specific $\mathcal{X}$ versus conditioning on the event $\{\mathcal{X} \geq \vec{1}\}$.

Lemma 4. *Let* $x = \mathbb{P}''[i_a = i_b]$ *and* $y = \mathbb{P}''[i_c = i_d]$. *For each* $\mathcal{X} \geq \vec{1}$ *and* $I \geq 1$, *we have almost surely*

$$x \wedge y \geq \frac{1}{I}.$$

Proof. Conditioning on a particular $\mathcal{X} \geq \vec{1}$ and on $\{N_j\}_j$, the choice of a and b in the locus tree is still independent. So i_a and i_b are picked proportionally to the N_j's by symmetry. Then

$$x = \mathbb{P}''[i_a = i_b] = \mathbb{E}'' \left[\frac{\sum_{j=1}^{I} N_j^2}{\left(\sum_{j=1}^{I} N_j\right)^2} \right] \geq \frac{1}{I}.$$

The same holds for y, completing the proof of the lemma. □

Ancestral locus configurations

Conditioned on $\mathcal{X}$ and I, we will characterize the occurrence of Q_1, Q_2, Q_3 based on how $i_x, x \in \{a, b, c, d\}$ are selected at the root R. Then, in a worst-case scenario, we will analyze coalescent events of the coalescent process above R. For an arbitrary quartet (a, b, c, d), we relate the likelihood of Q_1, Q_2, Q_3 under each of the following events:

$$\begin{aligned}
E &= a - b - c - d \\
F_{ab} &= ab - c - d \qquad F_{ac} = ac - b - d \qquad F_{ad} = ad - b - c \\
F_{bc} &= bc - a - d \qquad F_{bd} = bd - a - c \qquad F_{cd} = cd - a - b \\
G_{ab} &= ab - cd \qquad G_{ac} = ac - bd \qquad G_{ad} = ad - bc \\
H_{abc} &= abc - d \qquad H_{abd} = abd - c \qquad H_{acd} = acd - b \qquad H_{bcd} = bcd - a \\
K &= abcd,
\end{aligned} \tag{2.12}$$

where $-$ indicates separate lineages at R for the chosen copies from A, B, C, D. For example, the event E indicates that the i_a, i_b, i_c, i_d are distinct. These events are disjoint and mutually exhaustive. Letting $\mathcal{E}$ run across all the above events, the law of total probability implies

$$\mathbb{P}''[Q_i] = \sum_{\mathcal{E}} \mathbb{P}''[Q_i|\mathcal{E}]\mathbb{P}''[\mathcal{E}]. \tag{2.13}$$

Reduction to coalescence above R

For the rooted locus quartet implied by the four copies a, b, c, d, let $\mathcal{NC}$ be the event that no coalescent event occurs beneath R between the four corresponding lineages. The following lemma shows that conditioning on $\mathcal{NC}$ reduces the probability of Q_1 while increasing that of Q_2.

Lemma 5. *For any I and any $\mathcal{X} \geq \vec{1}$ and any event $\mathcal{E} \in \{E, F_{ab}, \ldots, G_{ab}, \ldots, H_{abc}, \ldots\}$,*

$$\mathbb{P}''[Q_1|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] \quad \textit{and} \quad \mathbb{P}''[Q_i|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}], \quad i \in \{2, 3\},$$

almost surely.

Proof. For Q_1, the law of total probability implies

$$\mathbb{P}''[Q_1|\mathcal{E}] = \mathbb{P}''[\mathcal{NC}^c|\mathcal{E}] + \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}]\,\mathbb{P}''[\mathcal{NC}|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}],$$

where we used that Q_1 is guaranteed under $\mathcal{NC}^c$. Similarly

$$\mathbb{P}''[Q_i|\mathcal{E}] = \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}]\,\mathbb{P}''[\mathcal{NC}|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}],$$

for $i \in \{2, 3\}$. □

The event K will play a special role in the proof and we treat it separately. For the other terms, combining (2.13) and Lemma 5, we have

$$\begin{aligned}\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] &= (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K] \\ &\quad + \sum_{\mathcal{E}\neq K} \left(\mathbb{P}''[Q_1|\mathcal{E}] - \mathbb{P}''[Q_2|\mathcal{E}]\right) \mathbb{P}''[\mathcal{E}] \\ &\geq (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K] \\ &\quad + \sum_{\mathcal{E}\neq K} \left(\mathbb{P}''[Q_1|\mathcal{E}\cap\mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{E}\cap\mathcal{NC}]\right) \mathbb{P}''[\mathcal{E}].\end{aligned}$$

To prove Proposition 1, we derive an explicit bound on this last sum.

Under $\mathbb{P}''$, the events E, H are symmetric in the sense that switching the roles of a and c or the roles of a and d does not change the conditional probability of Q_1 and Q_2. Hence

$$\mathbb{P}''[Q_1|\mathcal{E}\cap\mathcal{NC}] = \mathbb{P}''[Q_2|\mathcal{E}\cap\mathcal{NC}], \quad \forall \mathcal{E} \in \{E, H\}$$

and using this above we get

$$\begin{aligned}\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] \geq (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K] \\ + \sum_{j\in\{ab,ac,ad\}} \left(\mathbb{P}''[Q_1|G_j\cap\mathcal{NC}] - \mathbb{P}''[Q_2|G_j\cap\mathcal{NC}]\right) \mathbb{P}''[G_j] \\ + \sum_{j\in\{ab,\ldots,cd\}} \left(\mathbb{P}''[Q_1|F_j\cap\mathcal{NC}] - \mathbb{P}''[Q_2|F_j\cap\mathcal{NC}]\right) \mathbb{P}''[F_j]. \qquad (2.14)\end{aligned}$$

The F and G events

We now consider the events $\{F_{ab}, F_{cd}, F_{ac}, F_{bd}\}$ and $\{G_{ab}, G_{ac}\}$.

Event probabilities In the next lemma, we compute the probabilities of a given locus tree quartet satisfying the events in $\{F_{ab}, \ldots, F_{cd}, G_{ab}, G_{ac}\}$.

Lemma 6. *Let $x = \mathbb{P}''[i_a = i_b]$ and $y = \mathbb{P}''[i_c = i_d]$. For $I \geq 2$ and any $\mathcal{X} \geq \vec{1}$, the following hold almost surely:*

$$\begin{aligned}
\mathbb{P}''[F_{ab}] &= \frac{I-2}{I}x(1-y) \\
\mathbb{P}''[F_{cd}] &= \frac{I-2}{I}(1-x)y \\
\mathbb{P}''[F_{ac}] = \mathbb{P}''[F_{bd}] &= \frac{I-2}{I(I-1)}(1-x)(1-y) \\
\mathbb{P}''[G_{ab}] &= \frac{I-1}{I}xy \\
\mathbb{P}''[G_{ac}] &= \frac{1}{I(I-1)}(1-x)(1-y)
\end{aligned}$$

Proof. The calculations for F_{ab} and F_{cd} are similar, except that we condition on different events. Indeed, note that

$$\begin{aligned}
\mathbb{P}''[F_{ab}] &= \mathbb{P}''[F_{ab}|i_a = i_b, i_c \neq i_d]\mathbb{P}''[i_a = i_b]\mathbb{P}''[i_c \neq i_d] \\
\mathbb{P}''[F_{cd}] &= \mathbb{P}''[F_{cd}[i_a \neq i_b, i_c = i_d]\mathbb{P}''[i_a \neq i_b]\mathbb{P}''[i_c = i_d].
\end{aligned}$$

The conditional probability of F_{ab} is then obtained by considering that given the placement of the pair (i_c, i_d) among the I ancestral lineages, the shared lineage $i_a = i_b$ has $I-2$ choices where they do not intersect $\{i_c, i_d\}$. The result in the statement follows. Similarly, for F_ι with $\iota \in \{ac, bd\}$, we have

$$\mathbb{P}''[F_\iota] = \mathbb{P}''[F_\iota|i_a \neq i_b, i_c \neq i_d]\mathbb{P}''[i_a \neq i_b]\mathbb{P}''[i_c \neq i_d].$$

In this case, out of $I(I-1)$ choices for i_a and i_b, the choice of i_c is determined and there are $I-2$ remaining choices for i_d, implying the result.

We use the same principle for G_{ab} and G_{ac}. Keeping this in mind, we have

$$\begin{aligned}
\mathbb{P}''[G_{ab}] &= \mathbb{P}''[G_{ab}|i_a = i_b, i_c = i_d]\mathbb{P}''[i_a = i_b]\mathbb{P}'_I[i_c = i_d] \\
\mathbb{P}''[G_{ac}] &= \mathbb{P}''[G_{ac}|i_a \neq i_b, i_c \neq i_d]\mathbb{P}''[i_a \neq i_b]\mathbb{P}''[i_c \neq i_d],
\end{aligned}$$

and we proceed as before to get the result. □

Using the previous lemma, we collect further bounds on the probabilities of events at the root of the locus tree.

Lemma 7. *Letting again* $x = \mathbb{P}''[i_a = i_b]$ *and* $y = \mathbb{P}''[i_c = i_d]$*, the following statements hold.*

(a) If $I = 2$ *then*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq \left(x - \frac{1}{2}\right) \wedge \left(y - \frac{1}{2}\right)$$

(b) If $I \geq 3$ *and* $x \wedge y \geq 1/2$*, then*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq 1/8.$$

(c) If $I = 2$*,*

$$\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] = 0.$$

(d) If $I \geq 3$ *and* $x \wedge y \leq 1/2$*, then*

$$\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \geq \frac{1}{4}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right).$$

Proof. By Lemma 6, $\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] = \frac{1}{2}(x + y - 1)$ which implies part (a).

To prove (b), observe that by Lemma 6 again,

$$\begin{aligned}
\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] &= \frac{I-1}{I}xy - \frac{1}{I(I-1)}(1-x)(1-y) \\
&\geq \frac{1}{2}\left(\frac{I-1}{I}y - \frac{1}{I(I-1)}(1-y)\right) \\
&\geq \frac{1}{4}\left(\frac{I-1}{I} - \frac{1}{I(I-1)}\right) \\
&= \frac{1}{4}\left(\frac{I-2}{I-1}\right)
\end{aligned}$$

where the inequalities are justified by the assumption $x \wedge y \geq 1/2$. Since $I \geq 3$, it follows that $\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq 1/8$.

To prove (c) and (d), observe that by Lemma 6,

$$\begin{aligned}
&\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \\
&= \frac{I-2}{I}\left(x(1-y) + y(1-x)\right) - 2\frac{I-2}{I(I-1)}(1-x)(1-y) \\
&= \frac{I-2}{I(I-1)}\left((1-y)\left((I-1)x - (1-x)\right) + (1-x)\left((I-1)y - (1-y)\right)\right) \\
&= \frac{I-2}{I(I-1)}\left((1-y)(Ix-1) + (1-x)(Iy-1)\right).
\end{aligned}$$

Clearly if $I = 2$, the right-hand side is zero, which proves (c). Furthermore, since $x, y \geq 1/I$ by Lemma 4, it follows that both $(1-y)(Ix-1) \geq 0$ and $(1-x)(Iy-1) \geq 0$, and therefore

$$\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \geq \frac{I-2}{I(I-1)}(1-u)(Iv-1)$$

for $(u, v) \in \{(x, y), (y, x)\}$. Taking $u = \min(x, y)$ and $v = \max(x, y)$ gives

$$\begin{aligned}
\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] &\geq \frac{I-2}{I(I-1)}(1 - \min(x,y))(I\max(x,y) - 1) \\
&\geq \frac{I-2}{I-1}(1 - \min(x,y))\left(\max(x,y) - \frac{1}{I}\right) \\
&\geq \frac{1}{2}(1 - \min(x,y))\left(\max(x,y) - \frac{1}{I}\right) \\
&\geq \frac{1}{4}\left(\max(x,y) - \frac{1}{I}\right)
\end{aligned}$$

which implies (d). □

Conditional probabilities of quartet topologies In the following lemma, we give expressions for $\mathbb{P}''[Q_i | \mathcal{E} \cap \mathcal{NC}]$ across the events $\{F_{ab}, F_{cd}, F_{ac}, F_{bd}\}$ and $\{G_{ab}, G_{ac}\}$.

Lemma 8. *(a) For any I and any $\mathcal{X} \geq \vec{1}$, we have*

$$\mathbb{P}''[Q_1|F_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_1|F_{cd} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|F_{ac} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|F_{bd} \cap \mathcal{NC}] := \phi''_+$$

and

$$\mathbb{P}''[Q_2|F_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|F_{cd} \cap \mathcal{NC}] = \mathbb{P}''[Q_1|F_{ac} \cap \mathcal{NC}] = \mathbb{P}''[Q_1|F_{bd} \cap \mathcal{NC}] := \phi''_-.$$

(b) For any I and any $\mathcal{X} \geq \vec{1}$, we have

$$\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] = 1$$

and

$$\mathbb{P}''[Q_2|G_{ac} \cap \mathcal{NC}] = 1.$$

Proof. (a) The quantities ϕ''_+ and ϕ''_- are indeed well-defined as above by symmetry. (b) By switching the roles of b and c, we observe that $\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|G_{ac} \cap \mathcal{NC}]$. To see why $\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] = 1$, we again examine the topology above the root with leaves ab and cd. At least one of these leaves descends from a daughter edge, which implies Q_1 is constructed with probability 1. This completes the proof of the lemma. □

The following lemma establishes that, conditioned on F_{ab} and $\mathcal{NC}$, the difference in probability between Q_1 and Q_2 is at least $1/3$.

Lemma 9. *For $I \geq 1$ and any $\mathcal{X} \geq \vec{1}$, we have*

$$\phi''_+ - \phi''_- \geq \frac{1}{3}.$$

Proof. By definition of ϕ''_+ and ϕ''_-, it suffices to show $\mathbb{P}''[Q_1|F_{ab} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|F_{ab} \cap \mathcal{NC}] \geq 1/3$. Conditioned on $F_{ab} \cap \mathcal{NC}$, no coalescent event between the chosen lineages occurs

beneath R. So, we examine the topology of the locus tree above the root with leaf set being the three leaves implied by F_{ab}. Using the law of total probability, we condition further across the three possible rooted locus topologies on the three leaves ab, c, and d. Let τ_i be the rooted topology in which character i is the outgroup of the triple. Using Newick tree format, for example we have $\tau_{ab} = ((c, d), ab)$. Then

$$\mathbb{P}''[Q_1|F_{ab} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|F_{ab} \cap \mathcal{NC}] = \frac{1}{3} \sum_i \left(\mathbb{P}''[Q_1|\tau_i, F_{ab} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|\tau_i, F_{ab} \cap \mathcal{NC}] \right),$$

where we used the fact that $\mathbb{P}''[\tau_i|F_{ab} \cap \mathcal{NC}] = 1/3$ for each i. Now we compute the summands. If $i = ab$, then either ab descends from a daughter lineage or the pair (c, d) descends from a daughter lineage, meaning we observe Q_1 with probability 1 and Q_2 with probability 0. In the other two cases, let $p > 0$ be the probability that a and b coalesce along the pendant edge for ab. If they do not coalesce along the pendant edge, then the lineages from a and b live in the same population as that of c. Then there is probability $1/3$ that the first coalescing pair among a, b, c is a, b. So the probability of observing Q_1 is $p + \frac{1}{3}(1 - p)$. There is probability $1/3$ that the first coalescing pair among a, b, c is a, c, so the probability of observing Q_2 is $\frac{1}{3}(1 - p)$. Then

$$\phi''_+ - \phi''_- = \frac{1}{3} \left(1 - 0 + 2 \left(p + \frac{1}{3}(1 - p) - \frac{1}{3}(1 - p) \right) \right) = \frac{1}{3}(1 + 2p) \geq \frac{1}{3}.$$

□

Proof of Proposition 1

With that we can prove Proposition 1.

Proof of Proposition 1. In the $I = 1$ case, $\mathbb{P}''[K] = 1$ so

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] = \mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K] > 0,$$

where we used that, under $K \cap \mathcal{NC}$, the quartets Q_1 and Q_2 occur with equal probability under $\mathbb{P}''$. Since $x - 1/I = y - 1/I = 0$, the claim follows.

For $I \geq 2$, (2.14) and Lemma 8 implies that

$$\begin{aligned}
\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] &\geq (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K] \\
&\quad + \mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] + \left(\phi''_+ - \phi''_-\right)\left(\mathbb{P}''[F_{ab}] + \mathbb{P}''[F_{cd}]\right) \\
&\quad - \left(\phi''_+ - \phi''_-\right)\left(\mathbb{P}''[F_{ac}] + \mathbb{P}''[F_{bd}]\right) \\
&> \mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \\
&\quad + \left(\phi''_+ - \phi''_-\right)\left(\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}]\right),
\end{aligned}$$

where again we used that, under $K \cap \mathcal{NC}$, the quartets Q_1 and Q_2 occur with equal probability. If $I = 2$, then by Lemma 9 and Lemma 7 parts (a) and (c), this leads to

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right).$$

If $I \geq 3$, then by Lemma 7 parts (b) and (d),

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \begin{cases} 1/8 & \text{if } x \wedge y \geq 1/2 \\ \frac{1}{12}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right) & \text{if } x \wedge y \leq 1/2 \end{cases}$$

It follows that $\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{12}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right)$, finishing the proof of the main claim in the balanced case. □

□

2.3.2 Caterpillar case

We now consider the caterpillar case. Without loss of generality, assume the species tree restricted to $\mathcal{Q}$ has topology $(((A, B), C), D)$. Let R be the most recent common

ancestor of A, B, C and let I be the number of locus copies exiting R (forward in time). Let $\mathbb{P}''$ be the probability measure indicating conditioning on I and $\mathcal{X}$. Let $i_x \in \{1, ..., I\}$ be the ancestral lineage of $x \in \{a, b, c\}$ in R. As with the balanced case, if $\mathcal{X} < \vec{1}$, then ASTRAL-one selects Q_1, Q_2, Q_3 each with probability 0. To prove

$$\mathbb{P}[Q_1] > \mathbb{P}[Q_2] \vee \mathbb{P}[Q_3],$$

it is sufficient to prove Proposition 2 below for $\mathcal{X} \geq \vec{1}$.

Proposition 2 (Quartet identifiability: Caterpillar case)**.** *Let* $x = \mathbb{P}''[i_a = i_b]$. *On the events* $I \geq 1$ *and* $\mathcal{X} \geq \vec{1}$*, we have almost surely*

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{3}\left(x - \frac{1}{I}\right).$$

Similarly to the balanced case, in order to prove this proposition we consider the following events:

$$E = a - b - c$$

$$G_{ab} = ab - c \qquad G_{ac} = ac - b \qquad G_{bc} = bc - a$$

$$K = abc,$$

where $-$ indicates separation of lineages in R of the chosen copies of A, B, C. Letting $\mathcal{E}$ run across all events, the law of total probability implies

$$\mathbb{P}''[Q_i] = \sum_{\mathcal{E}} \mathbb{P}''[Q_i|\mathcal{E}]\mathbb{P}''[\mathcal{E}]. \tag{2.15}$$

Let $\mathcal{NC}$ be the event that no coalescent event occurs beneath R between the three lineage corresponding to a, b, c.

Analogues to Lemmas 4, 5, 6, 7, and 8 hold with similar proofs.

Lemma 10. *Let $x = \mathbb{P}''[i_a = i_b]$. On the events $\mathcal{X} \geq \vec{1}$ and $I \geq 1$, we have almost surely*

$$x \geq \frac{1}{I}.$$

Lemma 11. *Let the species tree be a rooted caterpillar on four leaves A, B, C, D. For all $I \geq 1$ and $\mathcal{X} \geq \vec{1}$, and any event $\mathcal{E} \in \{E, G_{ab}, G_{ac}, G_{bc}\}$,*

$$\mathbb{P}''[Q_1|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] \quad \textit{and} \quad \mathbb{P}''[Q_i|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}], \quad i \in \{2, 3\}.$$

Lemma 12. *For $I \geq 2$ and any $\mathcal{X} \geq \vec{1}$, let $x = \mathbb{P}''[i_a = i_b]$. Then the following hold:*

$$\mathbb{P}''[G_{ab}] = \frac{I-1}{I}x$$
$$\mathbb{P}''[G_{ac}] = \mathbb{P}''[G_{bc}] = \frac{1}{I}(1-x).$$

Lemma 13. *On $I \geq 1$, almost surely*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] = x - \frac{1}{I}.$$

Lemma 14. *For any I and any $\mathcal{X} \geq \vec{1}$, we have*

$$\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|G_{ac} \cap \mathcal{NC}] := \psi''_+$$

and

$$\mathbb{P}''[Q_2|G_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_1|G_{ac} \cap \mathcal{NC}] := \psi''_-.$$

The G events

The following lemma bounds the conditional probability difference for the G events.

Lemma 15. *On the events $I \geq 1$ and $\mathcal{X} \geq \vec{1}$, we have almost surely*

$$\psi''_+ - \psi''_- \geq \frac{1}{3}.$$

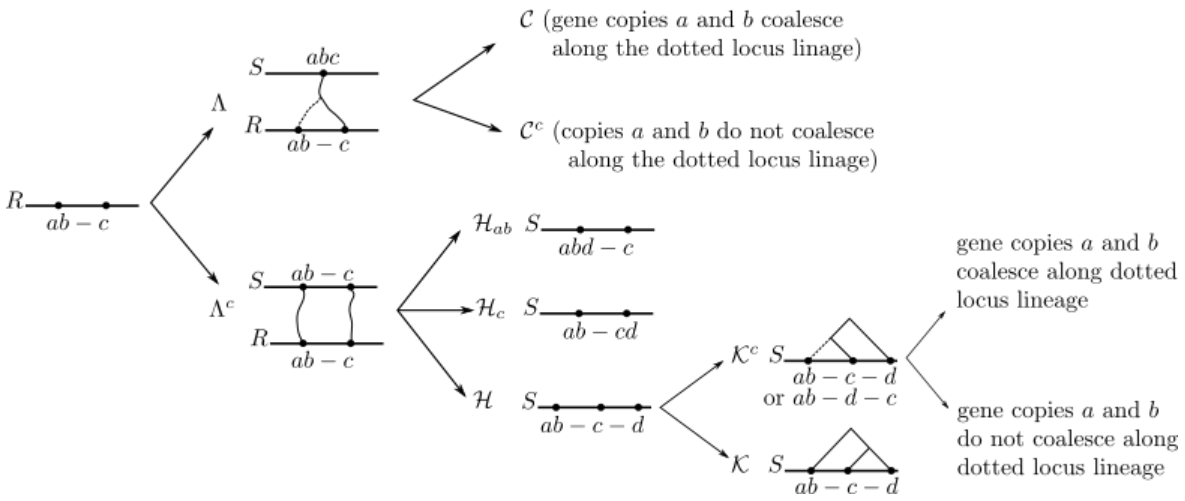

Figure 5: Flowchart for case analysis in Lemma 15.

Proof. By definition of ψ''_+ and ψ''_-, it suffices to show that $\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|G_{ab} \cap \mathcal{NC}] \geq 1/3$. The proof of this inequality involves decomposing $G_{ab} \cap \mathcal{NC}$ into a number of subcases, depicted in Figure 5, and computing the probabilities of Q_1 and Q_2 in each subcase. Let S be the most recent common ancestor of $\mathcal{Q}$ in the species tree. Let Λ be the event that the ab and c individuals in R descend from a common ancestor in S and let $q = \mathbb{P}''[\Lambda|G_{ab} \cap \mathcal{NC}]$. There are two cases:

1. *(Condition on Λ)* Let $\mathcal{C}$ be the event that gene copies a and b coalesce above R and below the MRCA of loci i_{ab} and i_c. Let $q' = \mathbb{P}''[\mathcal{C}|\Lambda, G_{ab}, \mathcal{NC}]$. We claim that $q' \geq 1/2$. To see this, observe that conditional on $G_{ab} \cap \mathcal{NC}$ the loci i_{ab} and i_c share the same ancestral locus at S only if there occurred a duplication event between S and R which is ancestral to both of them. Therefore with probability at least $1/2$, the gene copies a, b coalesce along their shared pendant edge in the rooted topology between R and S, proving the claim. Furthermore, it is obvious that

$$\mathbb{P}''[Q_1|\mathcal{C}, \Lambda, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{C}, \Lambda, G_{ab}, \mathcal{NC}] = 1. \tag{2.16}$$

On the other hand, conditional on $\mathcal{C}^c$, the copies of a, b and c enter the same population and are then symmetric, and hence

$$\mathbb{P}''[Q_1|\mathcal{C}^c, \Lambda, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{C}^c, \Lambda, G_{ab}, \mathcal{NC}] = 0. \tag{2.17}$$

2. *(Condition on Λ^c)* Let $\mathcal{H}_j$ be the event that copies d and j share the same ancestor in the locus tree at S, and define $\mathcal{H} = (\mathcal{H}_{ab} \cup \mathcal{H}_c)^c$ and $r = \mathbb{P}''[\mathcal{H}|\Lambda^c, G_{ab}, \mathcal{NC}]$. Then by symmetry, $\mathbb{P}''[\mathcal{H}_{ab}|\Lambda^c, G_{ab}, \mathcal{NC}] = \mathbb{P}''[\mathcal{H}_c|\Lambda^c, G_{ab}, \mathcal{NC}] = \frac{1-r}{2}$. By a further symmetry argument similar to that made in Case 1 we have

$$\mathbb{P}''[Q_1|\mathcal{H}_{ab}, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{H}_{ab}, \Lambda^c, G_{ab}, \mathcal{NC}] \geq 0, \tag{2.18}$$

where the inequality accounts for the possibility that the lineages from a and b coalesce between R and S. Let τ be the topology of the locus tree restricted to the copies ab, c, d and restricted to the portion above S (and suppressing nodes of degree 2). Conditioned on $\mathcal{H}_c$, we have $\tau = (ab, cd)$, so it must be the case that either ab descends from a daughter lineage or cd descends from a daughter lineage, meaning we observe Q_1 with probability 1 and Q_2 with probability 0. Therefore

$$\mathbb{P}''[Q_1|\mathcal{H}_c, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{H}_c, \Lambda^c, G_{ab}, \mathcal{NC}] = 1. \tag{2.19}$$

It remains to consider the case $\mathcal{H} = (\mathcal{H}_{ab} \cup \mathcal{H}_c)^c$. Let $\mathcal{K}$ be the event that $\tau = (ab, (c, d))$. By symmetry, the three possible topologies are equally likely, so $\mathbb{P}''[\mathcal{K}|\mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = 1/3$. Conditioned on $\mathcal{K}$, either ab descends from a daughter lineage or the pair (c, d) descends from a daughter lineage, and therefore

$$\mathbb{P}''[Q_1|\mathcal{K}, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{K}, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = 1. \tag{2.20}$$

Conditioned on $\mathcal{K}^c$, let p denote the probability that gene copies a and b coalesce along the pendant edge for ab in the rooted triple above S. Then

$$\mathbb{P}''[Q_1|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}]p + \frac{1}{3}(1-p)$$

and

$$\mathbb{P}''[Q_2|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = \frac{1}{3}(1-p),$$

and hence

$$\mathbb{P}''[Q_1|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] \geq p, \tag{2.21}$$

where again the inequality accounts for the possibility that the lineages from a and b coalesce between R and S.

Finally, applying the law of total probability and using equations (2.16)-(2.21) gives

$$\begin{aligned}\mathbb{P}''[Q_1|G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|G_{ab}, \mathcal{NC}] &\geq q'q + \left(\frac{1-r}{2} + \frac{1}{3}r + \frac{2}{3}pr\right)(1-q) \\ &\geq \frac{1}{2}q + \frac{1}{3}(1-q) = \frac{1}{3},\end{aligned}$$

where the inequality follows from $q' \geq 1/2$ and $r \geq 0$. □

Proofs of Proposition 2 and Theorems 6 and 7

With that, the proof of Proposition 2 is similar to that of Proposition 1.

In both the balanced and caterpillar cases, observe that $\mathbb{P}''[Q_1] - \mathbb{P}''[Q_3] = \mathbb{P}''[Q_1] - \mathbb{P}''[Q_2]$ by switching the roles of c and d. By Propositions 1 and 2, all species quartet topologies are identifiable and hence we have verified Theorem 6.

Theorem 7 then follows, along similar lines as with the analogous theorem for GDL, from the law of large numbers.

2.3.3 ASTRAL-multi

In this section, we give a proof of Theorem 5.

Proof of Theorem 5. Let $\mathcal{N}_{AB|CD}$ (respectively $\mathcal{N}_{AC|BD}, \mathcal{N}_{AD|BC}$) be the number of choices consisting of one gene copy in the gene tree from each species in $\mathcal{Q} = \{A, B, C, D\}$ whose corresponding restriction in t_1 agrees with $AB|CD$ (respectively $AC|BD$, $AD|BC$). Similarly to the proof for ASTRAL-one, it suffices to show that

$$\mathbb{E}[\mathcal{N}_{AB|CD}] > \mathbb{E}[\mathcal{N}_{AC|BD}] \vee \mathbb{E}[\mathcal{N}_{AD|BC}]. \tag{2.22}$$

Letting again $\mathcal{X} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$, by taking expectation with respect to I in Propositions 1 and 2, we have on the event $\mathcal{X} \geq \vec{1}$ that

$$\mathbb{P}[q = AB|CD \,|\, \mathcal{X}] > \mathbb{P}[q = AC|BD \,|\, \mathcal{X}] \vee \mathbb{P}[q = AD|BC \,|\, \mathcal{X}], \tag{2.23}$$

where q is the topology of a uniformly chosen quartet among A, B, C, D. Let $\mathcal{M} = \mathcal{ABCD}$ be the number of quartet choices and let $q_i, i = 1, \ldots, \mathcal{M}$ be the corresponding topologies ordered arbitrarily. Because q is a uniform choice, we have

$$\mathbb{P}[q = AB|CD \,|\, \mathcal{X}] = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \mathbb{P}[q_i = AB|CD \,|\, \mathcal{X}], \tag{2.24}$$

and similarly for the other topologies. Since

$$\mathcal{N}_{AB|CD} = \sum_{i=1}^{\mathcal{M}} \mathbf{1}\{q_i = AB|CD\},$$

and similarly for the other topologies, taking expectations and using (2.23) and (2.24) gives (2.22) as claimed. □

Chapter 3

Sample complexity

3.1 Introduction

In this chapter, we follow up on the phylogenomic methods chapter by discussing the sample complexity of ASTRAL-one and ASTRAL-multi. In the native application of ASTRAL to gene trees generated under the multi-species coalescent (MSC) model [MRB$^+$14], recall that the time units required to estimate the species tree is $O(nm|\mathcal{X}|^n)$, where n is the number of species, m is the number of gene trees sampled, and $|\mathcal{X}|$ is the number of bipartitions checked to solve the maximum quartet support species tree (MQSST) problem. As reviewed in the introduction, ASTRAL is a statistically consistent estimator of the species tree given gene trees generated under MSC under either the exact or default constrained version.

The *sample complexity* of a method refers to the number gene trees m required to estimate the target parameter with high probability. Results of this type are given for the multi-species coalescent (MSC) model in [SRM18], but no such result has been derived under either the GDL or DLCoal models. The novel contribution of this chapter is the sufficient number of multi-copy gene trees to estimate the species tree T_S with high probability. The proof highlights the somewhat counterintutive role played by the duplication and loss rates in the branching process underlying the GDL model.

From the proof of Lemma 1, shorter branches in coalescent-time units increase the probability of gene tree discordance and thus decrease the probability of estimating the species tree. Quartet scores follow a multinomial distribution in the number of quartets, so the frequencies of topologies concentrate strongly around their mean frequencies. In [SRM18], the authors show that the sufficient and necessary numbers of gene trees for the exact version of ASTRAL are each $O(f^{-2}\log(n))$. Here, f is taken to be the minimal branch length throughout the species tree, in coalescent time units. As with the previous section, we assume the given gene trees are *true*. Their theorems are given as follows.

Theorem 9 (Sufficient number of gene trees - MSC)**.** *Let T_S be an unrooted species tree on $n \geq 4$ leaves with minimum branch length $f < \log\left(\sqrt{2}\right)$. Then for any $\epsilon > 0$, the exact version of ASTRAL returns the true species tree with probability at least $1-\epsilon$ if the number of input-error free gene trees satisfies*

$$m > \frac{9}{2}\log\left(4\binom{n}{4}\epsilon^{-1}\right)\frac{1}{(1-e^{-f})^2}.$$

Note in the statement of this theorem that we suppose $f < \log\left(\sqrt{2}\right)$. Since we are interested in what occurs as $f \searrow 0$, it does not hurt to make this stipulation.

For f close to 0, we use the fact that $(1-e^{-f})/f \nearrow 1$, so that when $(1-e^{-f})/f \geq \sqrt{0.9}$. Observe that

$$\begin{aligned}\frac{9}{2}(1-e^{-f})^{-2}\log\left(4\frac{n(n-1)(n-2)(n-3)}{24}\epsilon^{-1}\right)\\ \leq \frac{9}{2}\frac{1}{0.9}f^{-2}\log\left(\frac{n^4}{6\epsilon}\right) = 20\log\left(\frac{n}{6\epsilon}\right)\frac{1}{f^2},\end{aligned}$$

so a further sufficient condition on number of gene trees is

$$m > 20\log\left(\frac{n}{6\epsilon}\right)\frac{1}{f^2}.$$

Intuitively, by taking the minimum branch length to 0, we are increasing the probability that ASTRAL makes a single error in that at least one of the quartets in the n species has the wrong quartet topology. By sampling enough gene trees to cover the worst branch in the species tree, we will have sampled enough gene trees to cover all quartets in the species tree.

Theorem 10 (Necessary number of gene trees - MSC)**.** *(i) Let T_S be a species tree with $n = 4$ leaves and f be the length of the single internal branch. Then for any $\epsilon \in [0, 0.5)$, there exists $f_0 > 0$ and a constant d_ϵ such that for all $m \leq d_\epsilon / f^2$ with $f \leq f_0$, the probability that the exact version of ASTRAL outputs a wrong quartet is at least ϵ.*

(ii) For any n and $\epsilon \in [0, 0.5)$, there exist a species tree T_S with n leaves and shortest branch length f such that when ASTRAL is performed with $m \leq d'_\epsilon / f^2$ gene trees for some constant d'_ϵ, the probability that the exact version of ASTRAL reconstructs a wrong tree is at least ϵ.

These two results show that at least $O(f^{-2} \log(n))$ are necessary to generate the species tree. Thus the asymptotics on the required number of gene trees follow a function of the order $O(f^{-2} \log(n))$ as $f \to 0$.

Here, we recall the basic steps to the proof of Theorem 9 for MSC-only from [SRM18].

Proof of Theorem 9. In this proof, we give an upper bound on the probability of ASTRAL making an error in terms of m, n, f. Let $k = \binom{n}{4}$ be the number of quartets in the species tree. A sufficient condition for ASTRAL returning the correct species tree is that for all quartets $i \in [k] := \{1, ..., k\}$, the true quartet topology is observed most frequently in the m gene trees. Letting n_{i1} be the number of gene trees displaying the

true quartet and n_{i2}, n_{i3} be the numbers displaying the alternative quartets, the sufficient condition is $n_{i1} > n_{i2} \vee n_{i3}$. Conversely, for ASTRAL to produce an error, at least one quartet must have an alternative topology in more of the gene trees. We compute an upper bound on this probability using Hoeffding's inequality and a union bound.

Under MSC, the probability that the quartet q_i in gene tree indexed by $i \in [k]$ has the true quartet topology q_{i1} is $p_i = 1 - \frac{2}{3}e^{-d_i}$ where d_i is the length of the interior edge by Lemma 1. The other two topologies have matching probility $r_i = (1 - p_i)/2$. We consider the difference

$$\delta_i = p_i - \frac{1}{3}.$$

Let $\delta = \min_i \delta_i$ and f denote the minimum branch length in the species tree. We are interested in measuring with what probability $n_{ij}/m, j \in \{1, 2, 3\}$ are close to their respective probabilities. Define the events

$$A_i = \{|\frac{n_{i1}}{m} - p_i| < w\}$$
$$B_i = \{|\frac{n_{i2}}{m} - r_i| < w\}$$
$$C_i = \{|\frac{n_{i3}}{m} - r_i| < w\}$$

for any $w > 0$. Let $D_i = A_i \cap (B_i \cup C_i)$, which is the event that n_{i1}/m and one of $n_{ij}/m, j \in \{2, 3\}$ are close to their respective probabilities. The following lemma shows that D_i for all $i \in [k]$ is sufficient for ASTRAL to return the true species tree with probability 1 when w is less than $\delta/2$.

Lemma 16. *For any $w < \delta/2$, ASTRAL returns the true species tree if D_i occurs for all $i \in [k]$.*

Proof. ASTRAL returns the true species tree if $n_{i1} > n_{i2} \vee n_{i3}$ for all $i \in [k]$. Suppose D_i holds, and without loss of generality that it is because B_i holds. In the worst case for ASTRAL to return the true species tree, this means

$$\frac{n_{i1}}{m} = p_i - w \ \frac{n_{i2}}{m} = r_i - w,$$

where we observe $n_{i1} > n_{i2}$. It remains to show $n_{i1} > n_{i3}$. We have

$$\frac{n_{i3}}{m} = 1 - \frac{n_{i1} + n_{i2}}{m} = r_i + 2w.$$

Now, if $w < \delta_i/2$ for all i, then

$$r_i + 2w = \frac{1 - p_i}{2} + 2w < p_i - w$$

because $p_i > r_i$ from Lemma 1. So we have $n_{i1} > n_{i3}$, as needed. □

Using the previous lemma, we derive a sufficient number of gene trees.

Lemma 17. *Suppose the input to ASTRAL is m true gene trees generated under MSC. Then for every $\epsilon > 0$, ASTRAL reconstructs the true species tree with probability at least $1 - \epsilon$ if*

$$m > 2 \log \left(4 \binom{n}{4} \epsilon^{-1} \right) \frac{1}{\delta^2}. \tag{3.1}$$

Proof. Let E be the event that ASTRAL does not return the true species tree. Then Lemma 16 gives the union bound

$$\mathbb{P}[E] \leq \mathbb{P}\left[\cup_{i=1}^{k} D_i^c\right].$$

Since $D_i^c \subset A_i^c \cup B_i^c$, monotonicity and subadditivity of measure imply

$$\mathbb{P}[E] \leq \mathbb{P}\left[\cup_{i=1}^{k} A_i^c \cup B_i^c\right] \leq \sum_{i=1}^{k} \left(\mathbb{P}[A_i^c] + \mathbb{P}[B_i^c]\right). \tag{3.2}$$

To bound the probabilities of A_i^c and B_i^c we use Hoeffding's inequality, see [Lug04], [Ver18]:

Theorem 11 (Hoeffding's inequality)**.** *Let $X_i, i \in [m]$ be a sequence of independent and identically distributed Bernoulli random variables with success probability p and $S_m = \sum_{i=1}^m X_i$ be the sum. Then for $\epsilon > 0$ smaller than p, we have*

$$\mathbb{P}\left[\frac{S_m}{m} - p < -\epsilon\right] \leq e^{-2m\epsilon^2}$$

and

$$\mathbb{P}\left[\frac{S_m}{m} - p > \epsilon\right] \leq e^{-2m\epsilon^2}.$$

To finish proving Lemma 17, we have $n_{i1} = \sum_{g=1}^m X_{ig}$ where we define X_{ig} to be independent Bernoulli random with parameter p_i, with a similar definition for n_{i2}. Putting the two above versions of Hoeffding's inequality together, w sufficiently small implies

$$\mathbb{P}[A_i^c] = \mathbb{P}\left[|\frac{n_{i1}}{m} - p_i| < w\right] \leq 2e^{-2mw^2}.$$

Using a similar bound for $\mathbb{P}[B_i^c]$, plugging this into the right-hand side of (3.2) implies

$$\mathbb{P}[E] \leq \binom{n}{4} 4e^{-2mw^2}. \tag{3.3}$$

In order for the right-hand side to be smaller than ϵ, solving implies

$$m \geq \frac{1}{2w^2} \log\left(4\binom{n}{4}\epsilon^{-1}\right),$$

and the lemma follows from $w < \delta/2$. This completes the proof of the lemma. □

With the bound in Lemma 16, we find that a sufficient condition on the number of gene trees is indeed

$$m > 20 \log\left(\frac{n}{6\epsilon}\right) \frac{1}{f^2},$$

completing the proof of Theorem 9.

□

The novel contribution to this chapter is to give a sufficient number of gene trees for the exact version of ASTRAL-one when they are generated under the GDL and DLCoal models. The GDL model is a special case of DLCoal when zero ILS occurs, so our results focus on DLCoal. The methods in this work are new compared to those in other sample complexity problems due to the properties of the branching process governing GDL. In particular, while identifiability and consistency results of the previous chapter rely only on Markovity of the branching process, our results address specifically the constant linear birth-death process with birth rate $\lambda \geq 0$ and death rate $\mu \geq 0$, see [Ken48], [AN72], [Ste16]. The upper bound we give will be $O(f^{-2}\log(n))$, but the relationship of this upper bound with λ, μ, and the tree depth Δ is emphasized in addition to f and n.

Finally, this work does not attempt to address the default constrained version of ASTRAL-one. The default constrained version requires at least as many gene trees because as we sample them, we must wait for every bipartition in the species tree to appear among the gene tree samples. So this necessary bound for the exact version of ASTRAL is also a necessary bound for the default constrained version. However, [SRM18] only gives a sufficient number of gene trees of the form $O(f^{-(n-3)})$, using the method of bipartition covers, see [UWR16]. This data requirement is much higher than $O(f^{-2}\log(n))$, speaking to the greater challenge of deriving the upper bound for the

default constrained version of ASTRAL.

3.2 Sufficient number of gene trees under DLCoal

As before, we let f be the minimum branch length, n be the number of species in the unrooted species tree T_S, and m be the number of multi-copy gene trees. Let $\lambda \geq 0$ and $\mu \geq 0$ be the birth and death rates in the linear birth-death process and Δ be the depth of the species tree. The *depth* is the maximum length of the paths going from the root to the leaves of the tree. We prove the following theorem.

Theorem 12 (Sufficient number of gene trees of ASTRAL-one - DLCoal)**.** *Let T_S be a model unrooted species tree whose minimum branch length f is finite and assume gene trees are generated under the DLCoal model. Then, for any $\epsilon > 0$, there are universal positive constants C, C' such that the exact version of ASTRAL-one returns the true species tree with probability $1 - \epsilon$ if the number of input error-free gene trees satisfies*

$$m \geq C' \frac{1}{f^2} \frac{e^{C|\mu-\lambda|\Delta}}{\left(1 - \frac{\lambda}{\mu} \wedge \frac{\mu}{\lambda}\right)^C} \log\left(\frac{n}{\epsilon}\right).$$

In the proof, we rely heavily on the quantitative bounds in Propositions 1 and 2. The sample complexity result uses a union bound over all quartets and builds on the analysis of the previous chapter. In particular, we refine the analysis of the event K in the identifiability proof.

3.2.1 The branching process

We highlight the role of a number of parameters in the sample complexity: the shortest branch length in the species tree, f; the depth of the species tree, Δ; and the duplication

and loss rates, λ and μ. These parameters enter the analysis through three quantities of significance:

- *Coalescence of a pair of lineages on an edge:* In the standard coalescent in the proof of Lemma 1, the probability that a pair of lineages has coalesced by time f is

$$\gamma := 1 - e^{-f}.$$

- *Survival probability of a quartet:* For a quartet $\mathcal{Q} = \{A, B, C, D\}$, let $\mathcal{X}_{\mathcal{Q}} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be the number of gene copies in the corresponding species. The smallest probability over all quartets that a gene family contains a copy in each species is denoted by

$$\sigma := \min_{\mathcal{Q}} \mathbb{P}\left[\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\right].$$

- *Expected number of lineages at a vertex:* For any vertex R in the species tree, let I_R be the number of copies at R in a single gene family. The largest expectation of I_R over all vertices will be denoted by

$$\alpha := \max_{R} \mathbb{E}[I_R].$$

These last two quantities can be bounded using the following formulas derived in [Ken48] and expanded on in [Ste16, Section 9.2] for the relevant results in the phylogenetic context. Let $p_n(t)$ be the probability that a single copy at time 0 evolves into n copies at time t. Then

$$p_n(t) = \begin{cases} \frac{\mu}{\lambda} q(t) & n = 0 \\ e^{-(\lambda-\mu)t}(1 - p_0(t))^2 q(t)^{n-1} & n > 0 \end{cases} \tag{3.4}$$

where

$$q(t) = \begin{cases} \lambda \frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}} & \lambda = \mu \\ \\ \frac{\lambda t}{1+\lambda t} & \lambda \neq \mu. \end{cases} \tag{3.5}$$

In practical applications, it is unrealistic to see $\mu = \lambda$, so the critical case is left to the reader. The following lemma determines the bounds of σ using these quantities.

Lemma 18. *The following hold:*

- *When $\mu > \lambda$,*

$$\sigma \geq \left(\frac{1}{e^{(\mu-\lambda)\Delta}} \left(1 - \frac{\lambda}{\mu} \right) \right)^4.$$

- *When $\lambda > \mu$,*

$$\sigma \geq \left(1 - \frac{\mu}{\lambda} \right)^4.$$

Proof. Let $\mathcal{Q} = \{A, B, C, D\}$ be a quartet and let as before $\mathcal{X}_{\mathcal{Q}} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be the number of gene copies in the corresponding species. Assume the species tree topology on $\mathcal{Q}$ is balanced, as the argument in the caterpillar case being similar. The probability that $\mathcal{A} \geq 1$ is given by

$$\mathbb{P}[\mathcal{A} \geq 1] = 1 - \frac{\mu}{\lambda} q(\Delta).$$

It can be checked that, whether $\mu > \lambda$ or $\mu < \lambda$, the function $q(t)$ is increasing in t. Conditioned on $\{\mathcal{A} \geq 1\}$, there is at least one copy in each species tree vertex along the path between the root and A. Letting $\mathcal{I}$ be the number of copies at the root, the Markov property implies

$$\mathbb{P}[\mathcal{D} \geq 1 \,|\, \mathcal{A} \geq 1] = \frac{\mathbb{P}[\mathcal{D} \geq 1 \,|\, \mathcal{I} \geq 1]}{\mathbb{P}[\mathcal{I} \geq 1 \,|\, \mathcal{A} \geq 1]} = \mathbb{P}[\mathcal{D} \geq 1 \,|\, \mathcal{I} \geq 1] \geq 1 - \frac{\mu}{\lambda} q(\Delta).$$

Repeating this argument for B and C gives

$$\begin{aligned}\mathbb{P}\left[\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\right] &= \mathbb{P}[\mathcal{A} \geq \vec{1}]\mathbb{P}[\mathcal{D} \geq \vec{1}|\mathcal{A} \geq \vec{1}]\mathbb{P}[\mathcal{B} \geq \vec{1}|\mathcal{A}, \mathcal{D} \geq 1]\mathbb{P}[\mathcal{C} \geq \vec{1}|\mathcal{A}, \mathcal{B}, \mathcal{D} \geq \vec{1}] \\ &\geq \left(1 - \frac{\mu}{\lambda}q(\Delta)\right)^4.\end{aligned}$$

From the next lemma, we are able to bound the right-hand side. When $\lambda > \mu$, $q(t) \to 1$ as $t \to +\infty$. So $1 - \frac{\mu}{\lambda}q(\Delta) \geq 1 - \frac{\mu}{\lambda}$. On the other hand, when $\mu > \lambda$,

$$\begin{aligned}1 - \frac{\mu}{\lambda}q(\Delta) &= \lambda\frac{1 - e^{-(\lambda-\mu)\Delta}}{\lambda - \mu e^{-(\lambda-\mu)\Delta}} \\ &= \frac{\mu - \lambda}{\mu e^{(\mu-\lambda)\Delta} - \lambda} \\ &\geq \frac{1}{e^{(\mu-\lambda)\Delta}}\left(1 - \frac{\lambda}{\mu}\right).\end{aligned}$$

□

Lemma 19. *We have*

$$\alpha \leq 1 \vee e^{(\lambda-\mu)\Delta}.$$

Proof. From our assumption that there is a single lineage at the root of the species tree, we can compute the time elapsed between the root vertex and any of its descendants. If the time elapsed between the root and a vertex U is d, then the expected number of lineages at U is $e^{(\lambda-\mu)d}$. The lemma then follows from the fact that $d \leq \Delta$ by considering separately the cases $\mu > \lambda$ and $\lambda > \mu$, as

$$e^{(\lambda-\mu)d} \leq \begin{cases} 1 & \lambda < \mu \\ e^{(\lambda-\mu)\Delta} & \lambda > \mu. \end{cases}$$

□

3.2.2 The effective number of samples

A concern for computing sample complexity is distinguishing between samples and effective samples. For a given quartet $\mathcal{Q}$, if a sampled gene tree fails to have copies of one of the species, then the true species quartet cannot be reconstructed by ASTRAL-one and this sample must be thrown out. We quantify the required number of effective samples here. For a quartet $\mathcal{Q}$, let $\mathcal{K}_{\mathcal{Q}}$ be the set of gene trees such that each species in $\mathcal{Q}$ has at least one gene copy. For $k \geq 1$, let

$$\mathcal{S}_k = \{|\mathcal{K}_{\mathcal{Q}}| \geq k : \forall \mathcal{Q}\}.$$

Lemma 20. *For any $k \geq 1$ and $\epsilon > 0$, we have that $\mathbb{P}[\mathcal{S}_k] \geq 1 - \epsilon$ provided*

$$m \geq \left\{\frac{2k}{\sigma}\right\} \vee \left\{\frac{8}{\sigma^2}\log\frac{n}{\epsilon}\right\}.$$

Proof. For any $\mathcal{Q}$, by the definition of $\mathcal{K}_{\mathcal{Q}}$, if m is the number of loci then

$$\mathbb{E}|\mathcal{K}_{\mathcal{Q}}| = \mathbb{E}\sum_{\ell=1}^{m} \mathbf{1}(\mathcal{X}_{\mathcal{Q}} \geq \vec{1}) \geq m\sigma.$$

Noting that σ concerns a minimal probability across quartets in a given gene tree, σ is independent of k. So we take k to be large enough that $\frac{1}{2}m\sigma \geq k$. By Hoeffding's inequality,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{S}_k^c\right] &\leq \sum_{\mathcal{Q}} \mathbb{P}\left[|\mathcal{K}_{\mathcal{Q}}| < k\right] \\
&\leq \sum_{\mathcal{Q}} \mathbb{P}\left[|\mathcal{K}_{\mathcal{Q}}| < \frac{1}{2}m\sigma\right] \\
&\leq \sum_{\mathcal{Q}} \mathbb{P}\left[\mathbb{E}|\mathcal{K}_{\mathcal{Q}}| - |\mathcal{K}_{\mathcal{Q}}| > \frac{1}{2}m\sigma\right] \\
&\leq n^4 \exp\left(-2\frac{(m\sigma/2)^2}{m}\right) \\
&\leq \epsilon,
\end{aligned}$$

if

$$m \geq \frac{2}{\sigma^2} \log \frac{n^4}{\epsilon},$$

and since $\epsilon < 1$ this inequality holds whenever

$$m \geq \frac{8}{\sigma^2} \log \frac{n}{\epsilon}.$$

This finishes the proof of the lemma. □

The K event

Using the notation the previous chapter, fix a quartet of species $\mathcal{Q} = \{A, B, C, D\}$, let $\mathbb{P}'$ denote the conditional probability given the events $\{\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\}$, and define $\delta' = \mathbb{P}'[Q_1] - 1/3$. Since $\mathbb{P}'[Q_2] = \mathbb{P}'[Q_3]$, we have

$$\delta' = \mathbb{P}'[Q_1] - \frac{\mathbb{P}'[Q_1] + \mathbb{P}'[Q_2] + \mathbb{P}'[Q_3]}{3} = \frac{2}{3}\left(\mathbb{P}'[Q_1] - \mathbb{P}'[Q_2]\right) \tag{3.6}$$

We seek to bound the right-hand side.

Assume first that the species tree restricted to $\mathcal{Q}$ is balanced. Letting $\mathbb{P}'_i$ indicate $\mathbb{P}'$ conditioned on $\{I = i\}$, by the proof of Proposition 1 we have

$$\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] \geq (\mathbb{P}'_i[Q_1|K] - \mathbb{P}'_i[Q_2|K])\mathbb{P}'_i[K]$$

where we took expectations over $\mathcal{X}_{\mathcal{Q}}$. Because on the event $\mathcal{NC}$, Q_1 and Q_2 are equally likely by symmetry, we are left with

$$\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] \geq \mathbb{P}'_i[K \cap \mathcal{NC}^c]. \tag{3.7}$$

To bound the right-hand side, define the event $\mathcal{C}_{ab}$ that the lineages picked from A and B uniformly at random coalesce beneath R. Note that $\mathcal{C}_{ab}$ implies $\{i_a = i_b\}$. The event $K \cap \mathcal{NC}^c$ is implied by $\mathcal{C}_{ab} \cap \{i_c = i_d = i_a\}$, which are conditionally independent conditioned on R. Then by Lemma 4, we have

$$\mathbb{P}'_i[K \cap \mathcal{NC}^c] \geq \mathbb{P}'_i[i_c = i_d]\frac{1}{i} \geq \frac{1}{i^2}.$$

Thus we have

$$\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] \geq \mathbb{P}'_i[\mathcal{C}_{ab}]\frac{1}{i^2}.$$

By a similar argument in the caterpillar case, we have

$$\begin{aligned}\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] &\geq \mathbb{P}'_i[K \cap \mathcal{NC}^c] \\ &\geq \mathbb{P}'_i[\mathcal{C}_{ab}]\,\frac{1}{i} \\ &\geq \mathbb{P}'_i[\mathcal{C}_{ab}]\,\frac{1}{i^2}. \end{aligned} \tag{3.8}$$

It remains to bound $\mathbb{P}'_i[\mathcal{C}_{ab}]$.

Lemma 21. *We have*

$$\mathbb{P}'_i[\mathcal{C}_{ab}] \geq \left\{\gamma \wedge \frac{1}{8}\right\}\frac{1}{i}.$$

Proof. Similarly to the proof of Lemma 2, for copy ℓ at R, let N_ℓ be the number of its descendant copies at R', the most recent common ancestor of A and B, and let $J = \sum_{\ell=1}^{i} N_\ell$. There are two cases for N_ℓ:

1. In the case $N_\ell = 1$ and $\{i_a = i_b = \ell\}$, let $Y_\ell = 1$ if the lineages from a and b coalesce before R, and 0 otherwise. Under the standard coalescent, the probability

of that coalescent event is at least γ. Here, we are working under the bounded coalescent in which we condition on full coalescence by the top of the edge. In the case that a daughter edge is ancestral to the lineages from a and b, the additional conditioning on complete coalescence only increases the probability that a and b coalesce before R. So γ remains a lower bound.

2. In the case $N_\ell \geq 2$ and $\{i_a = i_b = \ell\}$, let $Z_\ell = 1$ if the lineages from a and b coalesce below R, and 0 otherwise. Since $N_\ell \geq 2$, there is at least one duplication below ℓ and below R'. By symmetry, there is probability at least $1/2$ that the first duplication produces a daughter edge with at least half of the descendants of ℓ below it at R'. Under $\{i_a = i_b = \ell\}$, there is then a probability at least $1/4$ that a and b descend from copies at R' below that daughter edge. So overall there is probability at least $1/8$ that $Z_\ell = 1$ in that case.

Putting these two cases together , we get

$$\begin{aligned}
\mathbb{P}'_i[\mathcal{C}_{ab}] &= \mathbb{E}'_i\left[\mathbb{E}'_i\left[\sum_{\ell:N_\ell=1}\left(\frac{N_\ell}{J}\right)^2 Y_\ell + \sum_{\ell:N_\ell>1}\left(\frac{N_\ell}{J}\right)^2 Z_\ell \,\middle|\, (N_\ell)_{\ell=1}^i\right]\right] \\
&= \mathbb{E}'_i\left[\sum_{\ell:N_\ell=1}\left(\frac{N_\ell}{J}\right)^2 \mathbb{E}'_i\left[Y_\ell \,|\, N_\ell\right] + \sum_{\ell:N_\ell>1}\left(\frac{N_\ell}{J}\right)^2 \mathbb{E}'_i\left[Z_\ell \,|\, N_\ell\right]\right] \\
&\geq \left\{\gamma \wedge \frac{1}{8}\right\} \mathbb{E}'_i\left[\sum_{\ell:N_\ell=1}\left(\frac{N_\ell}{J}\right)^2 + \sum_{\ell:N_\ell>1}\left(\frac{N_\ell}{J}\right)^2\right] \\
&\geq \left\{\gamma \wedge \frac{1}{8}\right\} \frac{1}{i},
\end{aligned}$$

as in Lemma 4, proving the claim. □

In the next lemma, we finally give the needed lower bound for δ'.

Lemma 22. *We have*

$$\delta' \geq \frac{2}{3}\left\{\gamma \wedge \frac{1}{8}\right\}\frac{\sigma^3}{\alpha^3}.$$

Proof. By (3.6), (3.7), (3.8), and Lemma 21,

$$\begin{aligned}\delta' &= \frac{2}{3}\left(\mathbb{P}'[Q_1] - \mathbb{P}'[Q_2]\right)\\ &\geq \frac{2}{3}\left\{\gamma \wedge \frac{1}{8}\right\}\mathbb{E}'\left[\frac{1}{I^3}\right]\\ &\geq \frac{2}{3}\left\{\gamma \wedge \frac{1}{8}\right\}\frac{1}{\mathbb{E}'\left[I\right]^3},\end{aligned} \tag{3.9}$$

where the last line follows from Jensen's inequality. Moreover

$$\begin{aligned}\alpha &\geq \mathbb{E}[I]\\ &= \mathbb{E}\left[I\middle|\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\right]\ \mathbb{P}\left[\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\right] + \mathbb{E}\left[I\middle|\mathcal{X}_{\mathcal{Q}} < \vec{1}\right]\ \mathbb{P}\left[\mathcal{X}_{\mathcal{Q}} < \vec{1}\right]\\ &\geq \mathbb{E}'[I]\,\sigma.\end{aligned}$$

Plugging back into (3.9) finishes the proof of the lemma. □

3.2.3 Finishing analysis

Now we prove the theorem.

Proof of Theorem 12. The following, adapted from Lemmas 16 and 17 in [SRM18], gives a bound on the number of gene tree samples k required to reconstruct the correct species tree with probability $1-\epsilon$ in terms of δ'

$$k > 8\log\left(\frac{n}{\epsilon}\right)\frac{1}{(\delta')^2}.$$

By Lemmas 20 and 22, it suffices to have

$$m \geq \left\{ \frac{16}{\sigma} \log\left(\frac{n}{\epsilon}\right) \frac{1}{(\frac{2}{3}\left\{\gamma \wedge \frac{1}{8}\right\}\frac{\sigma^3}{\alpha^3})^2} \right\} \vee \left\{ \frac{8}{\sigma^2} \log\left(\frac{n}{\epsilon}\right) \right\}$$
$$\geq \frac{16}{\sigma} \log\left(\frac{n}{\epsilon}\right) \frac{1}{(\frac{2}{3}\left\{\gamma \wedge \frac{1}{8}\right\}\frac{\sigma^3}{\alpha^3})^2} \geq \frac{2304\alpha^6}{\sigma^7\gamma^2} \log\left(\frac{n}{\epsilon}\right).$$

The claim follows then from Lemmas 18 and 19. □

Chapter 4

Sequence-based methods with INDEL

4.1 Introduction

In this chapter, we address the more traditional tree reconstruction approach using sequences. To connect this result to the previous two chapters, we can imagine this tree reconstruction approach being used to generate the gene trees t_g for many genes g.

The evolutionary tree is thought to be inferred from biological sequences in each extant species. These sequences might consist of DNA, amino acids, etc. These could be used to reconstruct gene trees from the previous chapter by modeling the sequence evolution of particular genes or to reconstruct the tree from the entire sequence directly. In this chapter, reconstruction includes estimation of the rooted tree topology together with its edge lengths. The length of an edge reflects the distance between the two sequences at the ends of the edge.

Assuming a fixed rooted tree T, we assume sequences evolve forward-in-time from the root ancestral sequence to the leaf sequences. Substitution-only models stipulate that the sequences are fixed length $L \in \mathbb{Z}_+$ with sites labeled from the alphabet Σ and the sites evolve independently according to some Markovian process on Σ. For example, the

Jukes-Cantor model [JC69] on a sequence of length L has an alphabet $\Sigma = \{A, C, G, T\}$ and a transition rate matrix

$$Q = \nu \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix},$$

where ν is the rate of substitution per site. Also, the Jukes-Cantor model supposes that the initial distribution is such that each site follows the stationary distribution of the chain, i.e. for the transition probability matrix P_t, the distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ is stationary if $\pi = \pi P_t$ for all $t \geq 0$. When $\pi = (1/4, 1/4, 1/4, 1/4)$, the chain satisfies the *detailed balance condition* $\pi Q = 0$ (the chain is then said to be *time-reversible*). Finally, the model supposes *neutral* evolution, meaning under the mutation dynamics, no site is favored more than any other. Many generalizations of the Jukes-Cantor model exist that are independent-sites, time-reversible, and neutral on sequences of fixed length L on $\Sigma = \{A, B, C, D\}$. The most general model of these is the GTR model described in [Tav86].

Identifiability of the tree with edge lengths is proven in [SHP98] for some classes of substitution-only models described in the previous paragraph, in particular for the Jukes-Cantor model. Correspondingly little theoretical work has focused on models that allow *indels* (insertions or deletions) of sites in addition to substitutions, in spite of the models being around for a while [TKF91], [TKF92], [DR13], [ARS15], [FLR20]. For fixed length, substitution-only models there is an unambiguous sequence alignment given by associating the first sites of each sequence, associating the second sites of each sequence, etc. For the rest of the chapter, we focus on the Cavender-Farris-Neyman model that

considers a two-letter alphabet $\Sigma = \{0, 1\}$.

The TKF91 (indel) model for two states. We adapt the model from [TKF91], changing their alphabet from four letters to $\Sigma = \{0, 1\}$. If substitutions happen at rate ν, then applied to the distribution (π_0, π_1) of zeros and ones for a single site, the transition rate matrix is

$$Q = \nu \begin{pmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{pmatrix}.$$

Then the transition probability is of going from character $i \in \{0, 1\}$ to $j \in \{0, 1\}$ in time t is

$$p_{ij}(t) = \begin{cases} e^{-\nu t} + \pi_j(1 - e^{-\nu t}) & i = j \\ \pi_j(1 - e^{-\nu t}) & i \neq j. \end{cases}$$

This substitution process applies to each existing site independently of all others. For the insertion-deletion component, we let the length of the sequence evolve according to a modified linear-birth death process along the tree. Starting with the usual linear birth-death process, each site gives birth with exponential rate λ and dies with exponential rate μ independently of all other sites, and this process is independent of the substitutions. When an insertion occurs, its label starts at 0 with probability π_0 and at 1 with probability π_1. At insertion, the new site then evolves like the existing sites. We will take $\lambda \leq \mu$ so that the sequences are finite almost surely. However, to ensure the sequences have non-trivial length, we make the following modification. There exists a single source that inserts new sites at the *beginning* of the sequence with exponential rate λ like any existing site independently of all sites, except this single source never dies.

To keep track of indels and substitutions, we introduce *mortal* and immortal links. To every site in the sequence, we attach a mortal link to its right. In the birth-death process, when we say that a site gives birth to another, we mean that a mortal link gives birth to a new site and mortal link, which is placed to the right of the parent link. When a site dies, we remove the site and its mortal link and concatenate the neighboring surviving site-mortal-link pairs. The immortal link, denoted by $\bullet$, acts as the source of new sites at the beginning of the sequence. When we say that the source inserts a new site at the beginning of the sequence, we mean that the immortal link gives birth to a new site and mortal link, which is placed to the right of the immortal link. Then the TKF91 Markov process $\mathcal{I} = \{\mathcal{I}_t\}_{t\geq 0}$ is continuous-time and on the space

$$\mathcal{S} := \bullet \otimes \bigcup_{M\geq 1} \{0,1\}^M.$$

This indel process is time-reversible with respect to the measure

$$\Pi(\vec{x}) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^M \prod_{i=1}^{M} x_i$$

for each $\vec{x} = (x_1, ..., x_M) \in \{0,1\}^M$ and $M \geq 1$, and $\Pi(\bullet) = 1 - \frac{\lambda}{\mu}$. Since $\lambda < \mu$, this implies Π is the stationary distribution of $\mathcal{I}$. We assume the initial distribution of the indel process on the tree T is the stationary distribution Π. Given sequences generated under this model from a model phylogenetic tree T with edge lengths, we are interested in identifiability of the tree with edge lengths. That is, for two trees with different topologies or for two trees with the same topologies but different edge lengths, are the distributions of the sequences at the leaves different?

A difficulty for indel models is their lack of mathematical tractability, see e.g. [Tha06, DR13, ARS15, FR20]. One suggestion to attain some tractability is to consider only

the evolution of the lengths of the sequences in the tree in [Tha06]. The length of the sequence is an extra piece of information under indel models versus substitution-only models, so one might guess it is enough to establish species relationships. In particular, [Tha06] shows that leaf sequence lengths *alone* are enough to reconstruct phylogenies. It is a distance-based approach. First, letting $\overline{L}$ be the expected length of the sequence at stationary, we have

$$\overline{L} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}. \tag{4.1}$$

To see why (4.1) is true, we observe the distribution of sequence lengths is

$$\gamma_M = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^M, \ M \geq 0$$

by summing $\Pi(\vec{x})$ over all $\vec{x}$ of length M. More specifically, suppose t_{uv} is the height of the subtree underneath the most recent common ancestral (MRCA) vertex between two leaf vertices u and v. Then for any $t \in [0, t_{uv}]$, [Tha06] shows the expectation $\mathcal{N}_v(t)$ of the sequence length N_v at any leaf v conditioned on the sequence length N_u is

$$\mathcal{N}_v(t) = \overline{L} + \left(N_u(t) - \overline{L}\right) e^{-\mu t_{uv}\left(1 - \frac{\lambda}{\mu}\right)}. \tag{4.2}$$

Intuitively, at $t = 0$ the expected length is N_u because u and v would be the same vertex. For positive t, the second term of (4.2) reflects the change in the number of sites from the ancestral root to the leaf v from the expected sequence length $\overline{L}$ at the root. The exponential $e^{-\mu t_{uv}(1-\lambda/\mu)}$ factors in the net loss of sites in time t_{uv}. The second term of (4.2) is then added to the initial expected length of $\overline{L}$ to obtain $\mathcal{N}_v(t)$. Solving (4.2) then implies the full distribution of sequence lengths is sufficient to recover λ/μ and μt_{uv} for all pairs of leaves u, v.

Then the tree topology can be recovered from the μt_{uv} using metric properties of phylogenies, see [Bun71], [Ste16]. The following lemma was proven for positive edge weights.

Lemma 23 (Tree reconstruction from edge weights)**.** *Let T be a tree with vertices V and f be a nonzero real-valued function on the subsets of size 2 in V such that*

$$f(\{v,w\}) = \sum_{i=0}^{\gamma(v,w)} f\left(\{v_i, v_{i+1}\}\right),$$

where $v_0, ..., v_{\gamma(v,w)}$ is the unique path connecting v and w an $\gamma(v,w)$ is the number of edges between v and w. Then the topology of T and values of f on the entire domain is uniquely determined by the values of f when restricted to subsets involving leaf vertices only.

To apply the lemma, we let $t_{uv} = f(\{u,v\})$ for any pair of leaf vertices u, v. The lemma implies the tree T is identifiable from the sequence lengths in the sense that two distinct trees $T_1 \neq T_2$ produce distinct joint distributions of sequences at the leaves.

However, [Tha06] suggests that this method could be used to reconstruct the edge-weighted tree from a single sample of sequence lengths at the leaves in the limit where $\lambda \nearrow \mu$. This notion of statistical consistency suggests that as $\lambda \nearrow \mu$, the expected length of each sequence tends to ∞. Also, since the sites are independent, one might view the sites as iid random variables, where taking the expected number of variables to infinity implies that the tree can be reconstructed with probability 1.

We show in Section 2 that no such estimator exists in this limit; in particular, no consistent *distance* estimator exists in this limit. It is expected that the techniques established in Section 2 will be useful to analyze other bioinformatic methods for indel

processes, such as k-mer statistics (see [YZ08],[Hau13]). Relevant observations about k-mers will be discussed in Section 3.

4.2 Impossibility of tree reconstruction from length

Recalling the definition of the TKF91 indel process $\mathcal{I} = \{\mathcal{I}_t\}_{t\geq 0}$, we set more notation. Throughout this section, we let $\mathbb{P}_{\vec{x}}$ be the probability measure when the root state is $\vec{x}$. If the root state is chosen according to a distribution ν, then we denote the probability measure by $\mathbb{P}_\nu$. We also denote by $\mathbb{P}_M$ the conditional probability measure for the event that the root state has length M.

For our purposes, it will suffice to focus on the space $\mathcal{T}_2$ of star trees with two leaves that have the same finite distance $h \in (0, \infty)$ from the root and are labeled as $\{1, 2\}$. This distance h is the height of the tree. To prove consistency fails for sequence lengths, it suffices to show it fails on a subset of potential model trees. The indel process on a tree $T \in \mathcal{T}_2$ reduces to a pair of indel processes $(\mathcal{I}_t^1, \mathcal{I}_t^2)_{t\geq 0}$ that are independent upon conditioning on the root state $\mathcal{I}_\rho = \mathcal{I}_0^1 = \mathcal{I}_0^2$. We always assume the root state is chosen according to the equilibrium distribution Π. That is, the distribution of $(\mathcal{I}_0^1, \mathcal{I}_0^2) \in \mathcal{S} \times \mathcal{S}$ is

$$\widehat{\nu}_0(\vec{x}, \vec{y}) := \begin{cases} \Pi(\vec{x}) & \text{if } \vec{x} = \vec{y}, \\ 0 & \text{otherwise.} \end{cases}$$

4.2.1 Main result

The impossibility result we prove uses the total variation distance of distributions: the distributions of pairs of sequence lengths at different distances cannot be distinguished

with probability going to 1 as $\lambda \nearrow \mu$. Following [Tha06], we consider the asymptotic regime where $\lambda \nearrow \mu$, which we recall implies that the expected sequence length at stationarity tends to $+\infty$. Recall that the total variation distance between two probability measures τ_1 and τ_2 on a countable measure space E is

$$\|\tau_1 - \tau_2\|_{TV} = \frac{1}{2} \sum_{\sigma \in E} |\tau_1(\sigma) - \tau_2(\sigma)| .$$

Theorem 13 (Impossibility of distance estimation from sequence lengths)**.** *Let T^1 and T^2 be two trees in $\mathcal{T}_2$ with heights $h_1 > h_2 > 0$ respectively. For $i \in \{1, 2\}$, we consider a TKF91 process on tree T^i and let $\vec{N}^{(i)} = (N_1^{(i)}, N_2^{(i)}) \in \mathbb{Z}_+^2$ be the pair of sequence lengths at the leaves ∂T^i. Let*

$$\mathcal{L}_i = \mathbb{P}_\Pi(\vec{N}^{(i)} \in \cdot)$$

be the distribution of $\vec{N}^{(i)}$ under $\mathbb{P}_\Pi$. Then for any fixed deletion rate $\mu \in (0, \infty)$,

$$\limsup_{\lambda \nearrow \mu} \|\mathcal{L}_1 - \mathcal{L}_2\|_{TV} < 1. \tag{4.3}$$

Recall that the total variation distance can be written as

$$\|\tau_1 - \tau_2\|_{TV} = \sup_{A \subseteq E} |\tau_1(A) - \tau_2(A)| .$$

So (4.3) implies that there is no test that can distinguish between $\mathcal{L}_1$ and $\mathcal{L}_2$ with probability going to 1 as $\lambda \nearrow \mu$. So no consistent estimator based on sequence lengths can exist under this view.

Proof idea. We first give a heuristic argument that underlies our formal proof. We can rescale time units by a positive constant, so without loss of generality, assume that the deletion rate is $\mu = 1$. Then the stationary sequence length distribution reduces to

$$\gamma_M = (1 - \lambda)\lambda^M \ M \geq 0. \tag{4.4}$$

The stationary length M at the root is geometric with mean and standard deviation both of order $1/(1-\lambda)$. So we can think of the root length as roughly $M \approx C/(1-\lambda)$ with significant probability. Ignoring the small effect of the immortal link and conditioning on M, the lengths at the leaves are sums of independent random variables, specifically the progenies of the M mortal links of the root. The mean and variance of these variables can be computed explicitly from continuous-time Markov chain theory (see (4.11) below; see also [Tha06, (27), (31)]). As $\lambda \nearrow 1$, the difference in expectation between heights h_1 and h_2 turns out to be

$$Me^{-(1-\lambda)h_1} - Me^{-(1-\lambda)h_2} \approx \frac{C}{1-\lambda}[-(1-\lambda)h_1 + (1-\lambda)h_2] \approx C(h_2 - h_1), \tag{4.5}$$

while the variance is of order

$$M\frac{e^{-(1-\lambda)h_i}(1-e^{-(1-\lambda)h_i})}{1-\lambda} \approx \frac{C}{1-\lambda}\frac{(1-\lambda)h_i}{1-\lambda} \approx \frac{Ch_i}{1-\lambda}. \tag{4.6}$$

The key observation is that the variance (4.6) is much greater than the square of the expectation difference (4.5). Hence, by the central limit theorem, one can expect significant overlap between the length distributions under h_1 and h_2, making them hard to distinguish even as $\lambda \nearrow 1$. We formalize this argument next.

We observe that (4.3) is equivalent to

$$\liminf_{\lambda \nearrow 1} \sum_{\vec{y} \in \mathbb{Z}_+^2} \mathbb{P}_\Pi(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_\Pi(\vec{N}^{(2)} = \vec{y}) > 0. \tag{4.7}$$

Indeed the total variation distance between two probability measures τ_1 and τ_2 on a countable space E can also be written as

$$\|\tau_1 - \tau_2\|_{TV} = 1 - \sum_{\sigma \in E} \tau_1(\sigma) \wedge \tau_2(\sigma).$$

The rest of the proof is to establish (4.7). It involves a series of steps:

1. We first reduce the problem to a sum of independent random variables by conditioning on the root sequence length and ignoring the immortal link. In particular, we use the fact that there is a fairly uniform probability that M is in an interval of size $1/(1-\lambda)$ around 1. And we remove the effect of the immortal link by conditioning on its having no descendant, an event of positive probability.

2. The central limit theorem (CLT) implies that there is a significant overlap between the two sums. More precisely, we need a local CLT (see e.g. [Dur10]) to derive the sort of pointwise lower bound needed in (4.7). However the bound we require must be uniform in λ and we did not find in the literature a result of quite this form. Instead, we use an argument based on the Berry-Esséen theorem (again see e.g. [Dur10]). We first establish overlap over $\Omega(\sqrt{M})$ constant size intervals for the sum of the first $M-1$ mortal links, and then we use the final mortal link to match the probabilities on common point values under heights h_1 and h_2.

3. Finally we bound the sum in (4.7).

4.2.2 Proof

In this section we give the details of the proof of Theorem 13. We follow the steps described in the previous section.

Step 1. Reducing the problem to a sum of independent random variables. We first show that $\mathbb{P}_\Pi$ in (4.7) can be replaced by $\mathbb{P}_M$ where M is of the order of the expected sequence length $1/(1-\lambda)$ under Π. That is, we condition on the length of the ancestral sequence. After that we further ignore the progenies of the immortal link

so that each leave sequence consists of i.i.d. progenies of the M sites in the ancestral sequence. These two simplifications are achieved in (4.8) and (4.9) below respectively.

For any $\lambda_* \in (0,1)$ and $0 < c_1 < 1 < c_2 < +\infty$,

$$\begin{aligned}
&\liminf_{\lambda \nearrow 1} \sum_{\vec{y} \in \mathbb{Z}_+^2} \mathbb{P}_\Pi(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_\Pi(\vec{N}^{(2)} = \vec{y}) \\
&\geq \inf_{\lambda \in (\lambda_*, 1)} \sum_{\vec{y} \in \mathbb{Z}_+^2} \mathbb{P}_\Pi(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_\Pi(\vec{N}^{(2)} = \vec{y}).
\end{aligned}$$

Conditioning on the sequence length and using the law of total probability and (4.4), the right-hand side equals

$$\begin{aligned}
&\inf_{\lambda \in (\lambda_*, 1)} \sum_{\vec{y} \in \mathbb{Z}_+^2} \left[\sum_{M \in \mathbb{Z}_+} \gamma_M^{(\lambda)} \, \mathbb{P}_M(\vec{N}^{(1)} = \vec{y}) \right] \wedge \left[\sum_{M \in \mathbb{Z}_+} \gamma_M^{(\lambda)} \, \mathbb{P}_M(\vec{N}^{(2)} = \vec{y}) \right] \\
&= \inf_{\lambda \in (\lambda_*, 1)} \sum_{\vec{y} \in \mathbb{Z}_+^2} \sum_{M \in \mathbb{Z}_+} (1-\lambda)\lambda^M \left[\mathbb{P}_M(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_M(\vec{N}^{(2)} = \vec{y}) \right].
\end{aligned}$$

Restricting the sum over M, the right-hand side is greater than or equal to

$$\begin{aligned}
&\inf_{\lambda \in (\lambda_*, 1)} \sum_{M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]} (1-\lambda)\lambda^M \sum_{\vec{y} \in \mathbb{Z}_+^2} \mathbb{P}_M(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_M(\vec{N}^{(2)} = \vec{y}) \\
&\geq c_3(c_2 - c_1) \inf_{\lambda \in (\lambda_*, 1)} \inf_{M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]} \sum_{\vec{y} \in \mathbb{Z}_+^2} \mathbb{P}_M(\vec{N}^{(1)} = \vec{y}) \wedge \mathbb{P}_M(\vec{N}^{(2)} = \vec{y}), \qquad (4.8)
\end{aligned}$$

where c_3 is a lower bound on λ^M for $M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]$ and $\lambda \in (\lambda_*, 1)$.

Let $\mathcal{Z}_0$ be the event that the immortal link of the root sequence produces no mortal link in either leaf sequences. Let $\mathbb{P}_{M,\bullet}$ be the probability conditioned on that event, and c_4 be a lower bound on the probability of $\mathcal{Z}_0$ uniform in $\lambda \in (\lambda_*, 1)$. Under $\mathbb{P}_{M,\bullet}$, the two components of $\vec{N}^{(1)}$ are conditionally independent and each is a sum of M i.i.d. random

variables corresponding to the progenies of mortal links. Hence (4.8) is at least

$$c_4c_3(c_2-c_1)\inf_{\lambda\in(\lambda_*,1)}\inf_{M\in\left[\frac{c_1}{1-\lambda},\frac{c_2}{1-\lambda}\right]}\sum_{\vec{y}\in\mathbb{Z}_+^2}\mathbb{P}_{M,\bullet}(\vec{N}^{(1)}=\vec{y})\wedge\mathbb{P}_{M,\bullet}(\vec{N}^{(2)}=\vec{y})$$
$$\geq c_4c_3(c_2-c_1)\inf_{\lambda\in(\lambda_*,1)}\inf_{M\in\left[\frac{c_1}{1-\lambda},\frac{c_2}{1-\lambda}\right]}\sum_{\vec{y}\in\mathbb{Z}_+^2}\left[p^{(\lambda)}_{M,y_1}(h_1)\,p^{(\lambda)}_{M,y_2}(h_1)\right]\wedge\left[p^{(\lambda)}_{M,y_1}(h_2)\,p^{(\lambda)}_{M,y_2}(h_2)\right]. \tag{4.9}$$

where we let $p^{(\lambda)}_{y_1,y_2}(t)=\mathbb{P}_{y_1,\bullet}(|\mathcal{I}_t|=y_2)$ be the transition probability of the length process *without* the immortal link.

The sum in (4.9) leads us to study the overlap between the probability distributions $p^{(\lambda)}_{M,\cdot}(t):=\{p^{(\lambda)}_{M,j}(t)\}_{j\in\mathbb{Z}_+}$ for $t=h_1,h_2$ and $M\in\left[\frac{c_1}{1-\lambda},\frac{c_2}{1-\lambda}\right]$. The central limit theorem is what we need. However, because of our need for a bound that is uniform in λ, we shall apply the Berry-Esséen theorem. Specifically, we apply the latter bound to the progenies of the first $M-1$ mortal links of the root sequence. The idea is to show that $\Omega(\sqrt{M})$ summands in (4.9) have value $\Omega(1/\sqrt{M})$, for each of h_1 and h_2 separately, and then use the last mortal link to "match" all these values between h_1 and h_2.

Step 2a. Establishing a uniform bound for $p^{(\lambda)}_{M-1,\cdot}(t)$. Note that $p^{(\lambda)}_{M,\cdot}(t)$ is the distribution of $S_M(t):=\sum_{i=1}^M L^i_t$, where $\{L^i_t\}_{i\geq1}$ are i.i.d. random variables having the distribution of the progeny length of a single mortal link at time $t>0$.

Let the mean and the variance of L^i_t be

$$\beta:=\beta(\lambda,t):=\mathbb{E}[L^i_t]\quad\text{and}\quad\sigma^2:=\sigma^2(\lambda,t):=\mathbb{E}|L^i_t-\beta|^2. \tag{4.10}$$

As is expected, *the distribution $p^{(\lambda)}_{M,\cdot}(t)$ is approximately Gaussian with mean βM and variance σ^2M*. We quantify this statement in the bound (4.12) below, after proving some moment bounds.

Lemma 24. *Let $\beta(\lambda,t)$ and $\sigma(\lambda,t)$ be the mean and the standard deviation of L_t^i defined in (4.10) and consider the absolute third moment $\rho(\lambda,t) := \mathbb{E}|L_t^i - \beta|^3$. For any $t \in (0,\infty)$,*

$$\beta(\lambda,t) = e^{-(1-\lambda)t} \quad \textit{and} \quad \sigma^2(\lambda,t) = \frac{1+\lambda}{1-\lambda} e^{-(1-\lambda)t}(1 - e^{-(1-\lambda)t}). \tag{4.11}$$

Furthermore,

$$0 < \inf_{\lambda\in[\lambda_*,1]} \sigma(\lambda,t) < \sup_{\lambda\in[\lambda_*,1]} \sigma(\lambda,t) < \infty \quad \textit{and} \quad \sup_{\lambda\in[\lambda_*,1]} \rho(\lambda,t) < \infty.$$

Proof. For (4.11), see e.g. [DR13, (3), (4)] and [FR20].

Moreover, from [TKF91, (8)–(10)], the probability that a normal link as n descendants including itself is

$$\mathbb{P}(L_t^i = n) = \begin{cases} (1-\eta(\lambda,t))(1-\lambda\eta(t))[\lambda\eta(\lambda,t)]^{n-1} & \text{for } n \geq 1 \\ \eta(\lambda,t) & \text{for } n = 0 \end{cases},$$

where $\eta(\lambda,t) = \frac{1-e^{-(1-\lambda)t}}{1-\lambda e^{-(1-\lambda)t}}$. It can be seen from L'Hôpital's rule that $\eta(\lambda,t)$ is continuous as a function of λ around 1 and that $\eta(\lambda,t) = \frac{t}{1+t} + O(|1-\lambda|)$ as $\lambda \to 1$. From this explicit formula for the probability mass function of L_t^i, which we note is a geometric sequence, it follows that all moments of L_t^i are bounded from above uniformly in $\lambda \in [\lambda_*, 1]$.

To show that the variance is bounded from below uniformly in $\lambda \in [\lambda_*, 1]$, we note (again using L'Hôpital's rule) that $\sigma^2(\lambda,t)$ is continuous in λ around 1, strictly positive and tends to $2t$ as $\lambda \to 1$. Hence the variance is bounded from below, uniformly in $\lambda \in [\lambda_*, 1]$ □

Let $F_M^{(\lambda)}(t)$ be the cumulative distribution function (CDF) of the probability distribution $p_{M,\cdot}^{(\lambda)}(t)$. That is,

$$F_M^{(\lambda)}(t)(x) = \sum_{j:j\leq x} p_{M,j}^{(\lambda)}(t) = \mathbb{P}(S_M(t) \leq x).$$

Lemma 25 (Uniform bound for $p^{(\lambda)}_{M-1,\cdot}(t)$)**.** *For each $t > 0$, there exists a constant $C > 0$ such that*

$$\sup_{\lambda\in[\lambda_*,1]}\sup_{x\in\mathbb{R}}\left|F^{(\lambda)}_M(t)\Big(M\beta(\lambda,t)\,+\,x\,\sigma(\lambda,t)\,\sqrt{M}\Big)-\mathcal{N}(x)\right|\le\frac{C}{\sqrt{M}},\tag{4.12}$$

for all $M \in Z_+$, where $\mathcal{N}$ is the CDF of the standard normal distribution.

Proof. Since $\beta(\lambda,t), \sigma^2(\lambda,t), \rho(\lambda,t) \in (0,\infty)$, the Berry-Esséen theorem applies and asserts that

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_{M-1}-(M-1)\beta(\lambda,t)}{\sigma(\lambda,t)\sqrt{M-1}}\le x\right)\,-\,\mathcal{N}(x)\right|\le\frac{3\rho(\lambda,t)}{\sigma^3(\lambda,t)\sqrt{M-1}}$$

for all $\lambda \in [0,1]$. By Lemma 24, for each $t > 0$, the right hand side is bounded from above uniformly for $\lambda \in [\lambda_*, 1)$. □

Step 2b. Controlling the overlap of $p^{(\lambda)}_{M-1,\cdot}(h_1)$ and $p^{(\lambda)}_{M-1,\cdot}(h_2)$ in (4.9). To quantify the overlap between $p^{(\lambda)}_{M-1,\cdot}(h_1)$ and $p^{(\lambda)}_{M-1,\cdot}(h_2)$, we first compare their expectations. From the formula of β in (4.11), we have

$$\beta(\lambda,h_1)-\beta(\lambda,h_2)\le(1-\lambda)(h_1-h_2)$$

and so, for $M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]$, the means of S_{M-1} for h_1 and h_2 are close in the sense that

$$\beta(\lambda,h_1)(M-1)-\beta(\lambda,h_2)(M-1)\le c_6\tag{4.13}$$

for some $c_6 > 0$ not depending on λ.

Now consider the interval with length roughly the standard deviation and centered at around one of the means, $\beta(\lambda,h_1)(M-1)$. Then consider an equi-partition of this interval

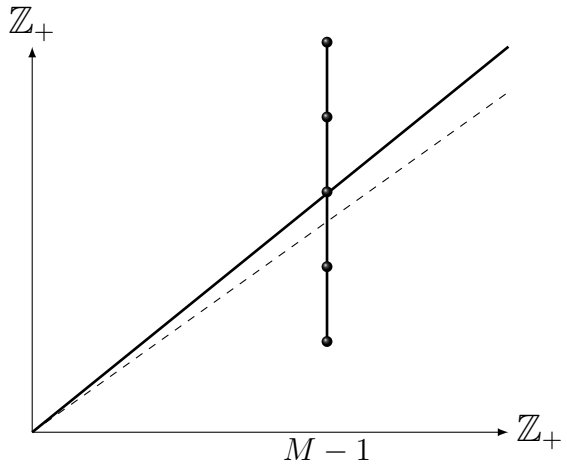

Figure 6: The solid straight line $j = \beta_1 M$ has slope β_1 and the dotted line $j = \beta_2 M$ has slope β_2 where $\beta_i = \beta(\lambda, h_i)$ for $i = 1, 2$. The vertical line has length $2\sigma_1\sqrt{M-1}$ where $\sigma_1 = \sigma(\lambda, h_1)$ and represents the union of sub-intervals $\bigcup_{r\in\Lambda_K^M} \mathcal{J}_r^M(c)$. Lemma 26 says that for each $M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]$, both probability measures $p^{(\lambda)}_{M-1,\cdot}(h_1)$ and $p^{(\lambda)}_{M-1,\cdot}(h_2)$ have mass at least $c_8/\sqrt{M-1}$ on $\mathcal{J}_r^M(c)$, uniformly for all $r \in \Lambda_K^M$ and $\lambda \in [\lambda_*, 1)$.

into roughly $\sqrt{M-1}$ many pieces of constant length. Precisely, we write $\beta_1 := \beta(\lambda, h_1)$ and $\sigma_1 := \sigma(\lambda, h_1)$ for simplicity. Then for an arbitrary constant $K > 0$,

$$\Big(\beta_1(M-1) - \sigma_1\sqrt{M-1},\ \beta_1(M-1) + \sigma_1\sqrt{M-1}\Big) = \bigcup_{r\in\Lambda_K^M} \mathcal{J}_r^M(c)$$

where the sub-intervals

$$\mathcal{J}_r^M(K) := \big(\beta_1(M-1) + r\sigma_1 K,\ \beta_1(M-1) + (r+1)\sigma_1 K\big)$$

have constant width $\sigma_1 K$ for $r \in \Lambda_K^M$, and

$$\Lambda_K^M := \Big\{-\sigma_1\sqrt{M-1},\ -\sigma_1(\sqrt{M-1} - K)\ \ldots, 0,\ \sigma_1 K,\ \ldots, \sigma_1(\sqrt{M-1} - K)\Big\}.$$

Lemma 26 below says that there exists constants $c = c_7$ and K large enough (depending on c_5 and c_6 but not on λ) such that each of these intervals contains mass at least $\frac{c_8}{\sqrt{M-1}}$ under both probability distributions $p^{(\lambda)}_{M-1,\cdot}(h_1)$ and $p^{(\lambda)}_{M-1,\cdot}(h_2)$. See Figure 6. Write $p^{(\lambda)}_{M-1,\,A}(t) = \sum_{j\in A} p^{(\lambda)}_{M-1,\,j}(t)$ for simplicity.

Lemma 26. *There exist positive constants c_7, c_8 such that, with $\mathcal{J}_r = \mathcal{J}_r^M(c_7)$ and $\Lambda^M = \Lambda_{c_7}^M$,*

$$p^{(\lambda)}_{M-1,\,\mathcal{J}_r}(h_1) \wedge p^{(\lambda)}_{M-1,\,\mathcal{J}_r}(h_2) \geq \frac{c_8}{\sqrt{M-1}}$$

for all $r \in \Lambda^M$, $M \in \left[\frac{c_1}{1-\lambda}, \frac{c_2}{1-\lambda}\right]$ and $\lambda \in [\lambda_*, 1)$.

Proof. The Berry-Esséen theorem (4.12) implies that

$$\sup_{r\in\Lambda^M} \left| p^{(\lambda)}_{M-1,\,\mathcal{J}_r}(h_1) - \int_{\widetilde{\mathcal{J}_r}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \right| \leq \frac{6\rho}{\sigma^3\sqrt{M-1}}$$

for all $\lambda \in [\lambda_*, 1]$ and $M \geq 2$, where

$$\widetilde{\mathcal{J}_r} := \left(\frac{rK}{\sqrt{M-1}}, \frac{(r+1)K}{\sqrt{M-1}} \right).$$

Then $\{\widetilde{\mathcal{J}_r}\}_{r\in\Lambda^M}$ is roughly an equi-partition of the interval $(-1,1)$ into $\frac{2\sqrt{M-1}}{K}$ sub-intervals of length $\frac{K}{\sqrt{M-1}}$. Furthermore,

$$\int_{\widetilde{\mathcal{J}_r}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \geq \frac{K}{\sqrt{M-1}} \frac{e^{-1/2}}{\sqrt{2\pi}}.$$

Pick K large enough (call it c_7), we obtain from the first display in this proof that

$$\inf_{r\in\Lambda^M,\,\lambda\in[\lambda_*,1)} p^{(\lambda)}_{M-1,\,\mathcal{J}_r}(h_1) \geq \frac{c}{\sqrt{M-1}}$$

for some constant $c > 0$ that does not depend on M. By the same argument and using (4.13), we have

$$\inf_{r\in\Lambda^M,\,\lambda\in[\lambda_*,1)} p^{(\lambda)}_{M-1,\,\mathcal{J}_r}(h_2) \geq \frac{c'}{\sqrt{M-1}}$$

for some constant $c' > 0$ that does not depend on M, even though $\mathcal{J}_r$ is constructed using h_1. The proof is complete by taking $c_8 = \min\{c, c'\}$. □

Step 2c. Matching $p^{(\lambda)}_{M,\cdot}(h_1)$ and $p^{(\lambda)}_{M,\cdot}(h_2)$ by the last mortal link. Lemma 26 establishes overlap of $p^{(\lambda)}_{M-1,\cdot}(h_1)$ and $p^{(\lambda)}_{M-1,\cdot}(h_2)$ over constant size intervals. The next lemma uses the final mortal link to establish overlap of $p^{(\lambda)}_{M,\cdot}(h_1)$ and $p^{(\lambda)}_{M,\cdot}(h_2)$ over specific values.

Lemma 27. *There exists a positive constant c_9 such that*

$$\inf_{j^*_{r+1}\in\mathcal{J}_{r+1}\cap\mathbb{Z}_+} p^{(\lambda)}_{M,j^*_{r+1}}(h_1)\wedge p^{(\lambda)}_{M,j^*_{r+1}}(h_2) > \frac{c_8c_9}{c_7\sqrt{M-1}}.$$

for all $r\in\Lambda^M$, $M\in\left[\frac{c_1}{1-\lambda},\frac{c_2}{1-\lambda}\right]$ and $\lambda\in[\lambda_*,1)$.

Proof. By Lemma 26, $\mathcal{J}_r$ contains at least one integer, say $j^{(1)}_r$, with mass at least $\frac{c_8}{c_7\sqrt{M-1}}$ under the probability measure $p^{(\lambda)}_{M-1,\cdot}(h_1)$. This is because there are c_7 integers in $\mathcal{J}_r$. Similarly, there exists $j^{(2)}_r$ with mass at least $\frac{c_8}{c_7\sqrt{M-1}}$ under $m_{F^{(\lambda)}_{M-1}(h_2)}$. Hence

$$p^{(\lambda)}_{M-1,\,j^{(1)}_r}(h_1)\wedge p^{(\lambda)}_{M-1,\,j^{(2)}_r}(h_2)\geq\frac{c_8}{c_7\sqrt{M-1}}.$$

Let j^*_{r+1} be an arbitrary integer in $\mathcal{J}_{r+1}$. The progeny of the M-th mortal link has positive probability, say c_9, over integers in $[0,2c_7]$, uniformly over $\lambda\in[0,1]$. It follows that

$$p^{(\lambda)}_{M,j^*_{r+1}}(h_1)=\sum_{k=0}^{j^*_{r+1}}p^{(\lambda)}_{M-1,k}(h_1)\,p^{(\lambda)}_{1,j^*_{r+1}-k}(h_1)>p^{(\lambda)}_{M-1,j^{(1)}_r}(h_1)\,p^{(\lambda)}_{1,j^*_{r+1}-j^{(1)}_r}(h_1)\geq\frac{c_8c_9}{c_7\sqrt{M-1}}$$

and similar for h_2. The proof is complete. □

Step 3. Putting everything together. Lemma 27 implies the sum in (4.9) is at least a positive constant, uniformly in $M\in\left[\frac{c_1}{1-\lambda},\frac{c_2}{1-\lambda}\right]$ and $\lambda\in(\lambda_*,1)$, because that sum

is

$$\begin{aligned}
&\sum_{\vec{y}\in\mathbb{Z}_+^2} \left[p_{M,y_1}^{(\lambda)}(h_1)\, p_{M,y_2}^{(\lambda)}(h_1)\right] \wedge \left[p_{M,y_1}^{(\lambda)}(h_2)\, p_{M,y_2}^{(\lambda)}(h_2)\right] \\
&\geq \sum_{y_1\in\cup_{r\in\Lambda^M}\mathcal{J}_{r+1}\cap\mathbb{Z}_+,\, y_2\in\cup_{r\in\Lambda^M}\mathcal{J}_{r+1}\cap\mathbb{Z}_+} \left[p_{M,y_1}^{(\lambda)}(h_1)\wedge p_{M,y_1}^{(\lambda)}(h_2)\right]\cdot\left[p_{M,y_2}^{(\lambda)}(h_1)\wedge p_{M,y_2}^{(\lambda)}(h_2)\right] \\
&\geq \left(\frac{c_8c_9}{c_7\sqrt{M-1}}\right)^2 \left|\{y_1\in\cup_{r\in\Lambda^M}\mathcal{J}_{r+1}\cap\mathbb{Z}_+,\ y_2\in\cup_{r\in\Lambda^M}\mathcal{J}_{r+1}\cap\mathbb{Z}_+\}\right| \\
&\geq \left(\frac{c_8c_9}{c_7}\right)^2.
\end{aligned}$$

The proof of (4.7) and hence that of Theorem 13 are complete.

Chapter 5

Discussion and future work

In this chapter, we discuss how to follow-up on the work in this thesis.

In Chapter 2, we introduced ASTRAL-one and ASTRAL-multi to adapt the method of [MRB+14] to multi-copy gene trees under the GDL model of [ALS09] and showed they are statistically consistent. Already we have a follow-up to this question in Section 2.3 by proving parallel results for the more complicated DLCoal model in [RK12]. It is interesting that the methods of Chapter 2 hold for any Markovian birth-death process on the edges of the species tree T_S. In particular, any forward-in-time processes like lateral gene transfer (LGT) implemented in [Gal07] could be combined with the GDL process and it would be relatively simple to replicate the results of this chapter. Developing results for DLCoal was particularly challenging because of the backward-in-time coalescent process being performed after locus trees are obtained.

In Chapter 3, of ASTRAL-one we answered the question of how many multi-copy genes are enough to estimate the species tree T_S with high probability. While we give an upper bound on a sufficient number of gene trees in terms of the constant birth rate and death rate and tree depth and minimum branch length, there may still be more to say about the birth and death rates. In particular, a follow-up to this project is to develop tight bounds involving the net insertion rate $\lambda - \mu$ and in particular establishing whether there is a difference between the subcritical and supercritical regimes. Another

follow-up is to consider the effects of focusing on the number of copies at the vertices in the species tree, whose effect is expressed by α. A method might be considered better than ASTRAL-one if it had a reasonable sample complexity independent of α. Finally, we turn to ASTRAL-multi. We did not perform sample complexity analysis on ASTRAL-multi because the difference in expectations between $\mathcal{N}_{AB|CD}$ and $\mathcal{N}_{AC|BD}$ are not enough to grasp the concentration of mass of these random variables. A follow-up to Chapters 2 and 3 that may even be more interesting than a specific sample complexity result is characterizing the moments of $\mathcal{N}_{AB|CD}$ in hopes of deriving subtle concentration bounds, utilizing [Lug04] and [Ver18].

In Chapter 4, we showed that it is impossible to recover an edge-weighted rooted species tree of even two leaves from the lengths of the sequences of the extant species. To do this, we used the Berry-Esséen theorem to characterize the overlap between distributions of sequence based on different tree depths. Going forward, it might be possible to derive similar results based on finer statistics like k-mer counts for integers $k \geq 1$, see [YZ08] and [ARS15], but such work will be challenging even with the new methods presented here because of a lack of results from local limit theory. We would look to Doeblin's method in [Cul61] to count k-mers using a discrete-time Markov chain, but local limit theorems appear elusive. However, with length-based data, it is interesting to explore new methods of ancestral state reconstruction on INDEL sequences in dense regimes, following up on [FR20].

Bibliography

[ADR11] Elizabeth S Allman, James H Degnan, and John A Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6):833–862, 2011.

[ALS09] Lars Arvestad, Jens Lagergren, and Bengt Sennblad. The gene evolution model and computing its associated probabilities. *Journal of the ACM*, 56(2):7, 2009.

[AN72] K.B. Athreya. and P.E. Ney. *Branching processes by K. B. Athreya and P. E. Ney.* Springer-Verlag Berlin, New York, 1972.

[ARS15] Elizabeth S. Allman, John A. Rhodes, and Seth Sullivant. Statistically consistent k-mer methods for phylogenetic tree reconstruction. *Journal of computational biology : a journal of computational molecular cell biology*, 24 2:153–171, 2015.

[BSD+13] Bastien Boussau, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, 2013.

[Bun71] Peter Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the Archaeological and Historical Sciences*, pages 387–395, 1971.

[Cul61] I.V. Culanovski. *Twenty-Five Papers on Statistics and Probability.* Sel. transl. math. stat. probab. American Mathematical Society, 1961.

[Deg18] J. H. Degnan. Modeling Hybridization Under the Network Multispecies Coalescent. *Syst. Biol.*, 67(5):786–799, 09 2018.

[DHN19] Peng Du, Matthew W Hahn, and Luay Nakhleh. Species tree inference under the multispecies coalescent on data with paralogs is accurate. *bioRxiv*, 2019.

[DR13] Constantinos Daskalakis and Sebastien Roch. Alignment-free phylogenetic reconstruction: sample complexity via a branching process analysis. *Ann. Appl. Probab.*, 23(2):693–721, 2013.

[Dur10] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 4 edition, 2010.

[FLR20] Wai-Tong Louis Fan, Brandon Legried, and Sebastien Roch. Impossibility of consistent distance estimation from sequence lengths under the tkf91 model, 2020.

[FR20] Wai-Tong Louis Fan and Sebastien Roch. Statistically consistent and computationally efficient inference of ancestral dna sequences in the tkf91 model under dense taxon sampling. *Bulletin of Mathematical Biology*, 82(2):21, 2020.

[Gal07] Nicolas Galtier. A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem. *Systematic Biology*, 56(4):633–642, 08 2007.

[Hau13] Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15(3):407–418, 11 2013.

[HLR20] Max Hill, Brandon Legried, and Sebastien Roch. Species tree estimation under joint modeling of coalescence and duplication: sample complexity of quartet methods, 2020.

[HRS11] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, USA, 2011.

[JC69] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. 1969.

[Ken48] David G. Kendall. On the generalized birth-and-death process. *Ann. Math. Statist.*, 19(1):1–15, 03 1948.

[Kin82] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[KK10] L. Knowles and Laura Kubatko. Estimating species trees: an introduction to concepts and models. *Estimating Species Trees: Practical and Theoretical Aspects*, 01 2010.

[LGSC20] Qiuyi Li, Nicolas Galtier, Celine Scornavacca, and Yao-Ban Chan. The multilocus multispecies coalescent: A flexible new model of gene family evolution. *bioRxiv*, 2020.

[LKDA10] Bret R Larget, Satish K Kotha, Colin N Dewey, and Cécile Ané. BUCKy: Gene Tree/Species Tree Reconciliation with Bayesian Concordance Analysis. *Bioinformatics*, 26(22):2910–2911, 2010.

[LMWR19] Brandon Legried, Erin K. Molloy, Tandy Warnow, and Sébastien Roch. Polynomial-time statistical estimation of species trees under gene duplication and loss. *bioRxiv*, 2019.

[Lug04] Gábor Lugosi. Concentration-of-measure inequalities, 2004.

[Mad97] Wayne Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, 1997.

[MK09] Chen Meng and Laura Kubatko. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, 75:35–45, 02 2009.

[MRB+14] S. Mirarab, R. Reaz, Md. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[RK12] M. D. Rasmussen and M. Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765, 2012.

[RNW18] Sebastien Roch, Michael Nute, and Tandy Warnow. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods, 2018.

[RS15] Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, 2015.

[RSM19] Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. Multi-allele species reconstruction using ASTRAL. *Molecular Phylogenetics and Evolution*, 130:286–296, 2019.

[RY03] B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, Aug 2003.

[SHP98] Mike Steel, Michael D. Hendy, and David Penny. Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results. *Discrete Applied Mathematics*, 88(1):367 – 396, 1998. Computational Molecular Biology DAM - CMB Series.

[SRM18] Shubhanshu Shekhar, Sebastien Roch, and Siavash Mirarab. Species tree estimation using astral: How many genes are enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5):1738–1747, Sep 2018.

[SS03] C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003.

[Ste16] Mike Steel. *Phylogeny—discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[Tav86] Simon Tavaré. Some probabilistic and statistical problems in the analysis

of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[Tha06] Bhalchandra D. Thatte. Invertibility of the tkf model of sequence evolution. *Mathematical Biosciences*, 200(1):58 – 75, 2006.

[TKF91] Jeffrey L. Thorne, Hirohisa Kishino, and Joseph Felsenstein. An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, 33(2):114–124, Aug 1991.

[TKF92] Jeffrey L Thorne, Hirohisa Kishino, and Joseph Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of molecular evolution*, 34(1):3–16, 1992.

[UWR16] Lawrence H. Uricchio, Tandy J. Warnow, and Noah A. Rosenberg. An analytical upper bound on the number of loci required for all splits of a species tree to appear in a set of gene trees. *BMC Bioinformatics*, 17:241 – 250, 2016.

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[War17] Tandy Warnow. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, USA, 1st edition, 2017.

[YZ08] Kuan Yang and Liqing Zhang. Performance comparison between k -tuple distance and four model-based distances in phylogenetic tree reconstruction . *Nucleic Acids Research*, 36(5):e33–e33, 02 2008.