JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 30, Number 11, 2023

Pp. 1146–1181
DOI: 10.1089/cmb.2023.0185

Research Articles

Open camera or QR reader and scan code to access this article and other resources online.

QR-STAR: A Polynomial-Time Statistically Consistent Method for Rooting Species Trees Under the Coalescent

YASAMIN TABATABAEE,[1] SEBASTIEN ROCH,[2] and TANDY WARNOW[1]

ABSTRACT

We address the problem of rooting an unrooted species tree given a set of unrooted gene trees, under the assumption that gene trees evolve within the model species tree under the multispecies coalescent (MSC) model. Quintet Rooting (QR) is a polynomial time algorithm that was recently proposed for this problem, which is based on the theory developed by Allman, Degnan, and Rhodes that proves the identifiability of rooted 5-taxon trees from unrooted gene trees under the MSC. However, although QR had good accuracy in simulations, its statistical consistency was left as an open problem. We present QR-STAR, a variant of QR with an additional step and a different cost function, and prove that it is statistically consistent under the MSC. Moreover, we derive sample complexity bounds for QR-STAR and show that a particular variant of it based on "short quintets" has polynomial sample complexity. Finally, our simulation study under a variety of model conditions shows that QR-STAR matches or improves on the accuracy of QR. QR-STAR is available in open-source form on github.

Keywords: multispecies coalescent, rooting, species tree estimation, statistical consistency.

1. INTRODUCTION

ROOTED SPECIES TREES are needed for many biological research problems, including comparative genomics (Jun et al., 2015; Skarp-de Haan et al., 2014) and dating (Renner et al., 2008). The availability of genome-wide sequencing data for many species has made it possible to estimate species trees using different loci from across the genome, thus enabling ''multi-locus'' species tree estimation. Typically, rooted species trees are estimated in two steps: first the unrooted topology of the species tree is inferred using a multi-locus species tree estimation method, and then that unrooted species tree is rooted.

Alternatively, rooted gene trees can be inferred and then combined into a rooted species tree, using methods such as MP-EST (Liu et al., 2010), STAR (Liu et al., 2009), and GLASS (Mossel and Roch,

[1]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.
[2]Department of Mathematics, University of Wisconsin–Madison, Madison, Wisconsin, USA.
An early version of this article was published as part of the 2023 Annual International Conference on Research in Computational Molecular Biology (RECOMB), held in Istanbul, Turkey on April 16–19, 2023, and was deposited to bioRxiv at https://doi.org/10.1101/2022.10.26.513897.

2008). However, the estimation of rooted gene trees is itself challenging, making this approach less reliable than methods that operate in the two-step procedure where the unrooted species tree is estimated first and then rooted (Simmons and Gatesy, 2015; Simmons et al., 2022).

The problem of estimating an unrooted species tree has been actively investigated over the past several decades. The classical approach is ''concatenation,'' where the alignments for the different loci are concatenated into one large ''super-alignment,'' which is then given to a tree estimation method, such as RAxML (Stamatakis, 2014).

However, evolutionary processes, such as incomplete lineage sorting (ILS) or gene duplication and loss (GDL), can result in different genomic regions (referred to as ''loci'') having different evolutionary histories, so that gene trees and species trees can have different topologies (Maddison, 1997). Moreover, when ILS is high, then standard concatenation analyses can have poor accuracy (Kubatko and Degnan, 2007; Molloy and Warnow, 2018), and may even be statistically inconsistent (Roch and Steel, 2015; Roch et al., 2019).

Therefore, to estimate highly accurate species trees in the presence of ILS or GDL, new methods have been developed that take the source of heterogeneity into consideration (Chifman and Kubatko, 2014; Mirarab et al., 2014b; Ogilvie et al., 2017). Species tree estimation in the presence of ILS, as modeled by the multispecies coalescent (MSC) model (Hudson, 1983), is the most well studied, and many methods have been developed for this problem; see Posada (2016) for a survey.

Given an estimate of the unrooted species tree, various methods can be used to infer the root location. Perhaps the most commonly used approach is the use of one or more outgroup species (e.g., the addition of a lizard within a collection of bird species), which allows the unrooted tree on this enlarged set of species to be rooted on the edge leading to the outgroup (Maddison et al., 1984).

Although this approach is natural, there are many challenges in selecting an appropriate outgroup species: if the outgroup is too distantly related to the other species, then it may be attached fairly randomly to the tree containing the remaining species, and if it is too closely related, it may even be an ingroup taxon rather than an outgroup (Felsenstein, 1978; Graham et al., 2002; Holland et al., 2003; Li et al., 2012).

It is also possible to use estimated branch lengths on the species tree to find the root based on specific optimization criteria, often using molecular clock analysis (Drummond et al., 2006); however, these approaches may only be highly accurate when evolutionary rates are close to following the strict molecular clock (which assumes that all sites along the genome evolve under a constant rate) (Hess and De Moraes Russo, 2007; Mai et al., 2017; Tria et al., 2017).

There are also recent developments that seek to find the root based on non-reversible models of DNA substitution (Bettisworth and Stamatakis, 2021; Naser-Khdour et al., 2022). However, none of the methods mentioned so far consider biological processes that cause discord between species trees and gene trees, and they are mainly used and evaluated for rooting gene trees (Wade et al., 2020).

Recently, a few methods have been developed that are specifically designed for rooting species trees under the MSC; these include a rooting method by Tian and Kubatko (2017) that uses site pattern probabilities, and a method that uses approximate Bayesian computation by Alanzi and Degnan (2017). The first method assumes a strict molecular clock and degrades in accuracy when there is deviation from the clock (Tian and Kubatko, 2017), and the second approach relies on a large number of calculations and may not be scalable (Alanzi and Degnan, 2017). Further, the software for these methods is not publicly available, and their performance compared with other methods is not explored in the literature.

We recently introduced Quintet Rooting (QR) (Tabatabaee et al., 2022), a polynomial-time method for rooting an unrooted species tree with at least five leaves given a set of unrooted gene trees, which is designed for use when the gene trees differ from the species tree due to ILS. QR is based on the mathematical theory by Allman, Degnan, and Rhodes (ADR) (Allman et al., 2011) that established that the rooted topology of every 5-leaf species tree is identifiable from the distribution of the unrooted 5-leaf gene tree topologies; a trivial extension to any number $n \geq 5$ species then follows.

The experimental study in Tabatabaee et al. (2022) showed that QR had good accuracy on simulated ILS datasets in comparison to alternative methods. However, we did not establish whether it was statistically consistent under the MSC. That is, we did not establish whether QR would return the correct root location with probability converging to 1 as the number of true gene trees in the input increases, when given the true unrooted species tree as input.

Although there has been much focus on proving statistical consistency for species tree estimation methods and several methods such as ASTRAL (Mirarab et al., 2014b), SVDQuartets (Wascher and

Kubatko, 2021) and BUCKy (Larget et al., 2010) have been proven statistically consistent estimators of the *unrooted* species tree under the MSC, to the best of our knowledge, no prior study has addressed the statistical consistency properties of methods for rooting species trees.

In this article, we argue that QR is not guaranteed to be statistically consistent under the MSC. We introduce a variant of QR called QR-STAR that is also polynomial-time and uses much of the same algorithmic structure of QR, but with some important changes that enable us to prove statistical consistency under the MSC. We also analyze the sample complexity for QR-STAR, and we provide a variant that achieves polynomial sample complexity. Finally, our simulation study evaluating QR and QR-STAR under a range of model conditions shows that QR-STAR matches or improves on the accuracy of QR, and its error is close to the error of the optimal rooting under many conditions.

The rest of this article is organized as follows. We provide background information on QR in Section 2, as well as the theory established by ADR (Allman et al., 2011). We introduce QR-STAR in Section 3. The theoretical results are provided in Section 4. In Section 5, we report on the results of a simulation study, including the design of QR-STAR and the evaluation of QR-STAR in comparison to QR. We conclude in Section 6 with a discussion of future research.

2. BACKGROUND

We present the theory from Allman et al. (2011) first, which establishes identifiability of the rooted species tree from unrooted quintet trees, and then we describe QR, our earlier method for rooting species trees. Together, these form the basis for deriving our new method, QR-STAR, which we present in the next section.

2.1. ADR theory

ADR (Allman et al., 2011) established that the unrooted topology of the species tree is identifiable from four-leaf unrooted gene trees under the MSC, a result that is well known and used in several ''quartet-based'' methods for estimating species trees under the MSC (Larget et al., 2010; Mahbub et al., 2021; Mirarab et al., 2014b). ADR also proved that the rooted species tree topology is identifiable from unrooted five-leaf gene tree topologies; this result is much less well known, but it was recently used in the development of QR for rooting species trees.

ADR have described the probability distribution of unrooted gene tree topologies under each 5-taxon MSC model species tree. On a given set of 5 taxa, there exist 105 different rooted binary trees, labeled with $R_1, \ldots, R_{105}$,* that can be categorized into 3 groups based on their (unlabeled) rooted shapes: caterpillar, balanced, and pseudo-caterpillar (Rosenberg, 2007). An example of a tree from each category is shown in Figure 1. Each 5-taxon model species tree defines a specific probability distribution over the 15 different unrooted gene tree topologies on the same leafset, shown with $T_1, \ldots, T_{15}$ (see Appendix Fig. A1 in Appendix A). Theorem 9 in Allman et al. (2011) states that this distribution uniquely determines the rooted tree topology and its internal branch lengths for trees with at least five taxa.

To prove this identifiability result, the ADR theory specifies a set of linear invariants (i.e., equalities) and inequalities that must hold between the probabilities of unrooted 5-taxon gene trees, for any choice of the parameters of the model species tree. These linear invariants and inequalities define a partial order on the probabilities of 5-taxon unrooted gene tree topologies. In other words, two gene tree probabilities $u_i = \mathbb{P}(T_i)$ and $u_j = \mathbb{P}(T_j)$ can have one of four possible relationships: $u_i > u_j$, $u_j > u_i$, $u_i = u_j$, or u_i and u_j are not comparable.

Figure 1 shows examples of these partial orders, described using Hasse diagrams, for a particular leaf labeling of trees from each rooted shape. Note that some probabilities are members of the same set (e.g., for R_1, set c_4 contains both u_4 and u_{13}, indicating that $u_4 = u_{13}$), and so we refer to the sets c_i as equivalence classes on these probabilities. Further, we will denote the set of equivalence classes associated with a 5-taxon rooted tree R with C_R.

*The labeling of rooted and unrooted trees in this article is consistent with the notations and leaf-labeling used in tables 4 and 5 in Allman et al. (2011) as well as in Tabatabaee et al. (2022).

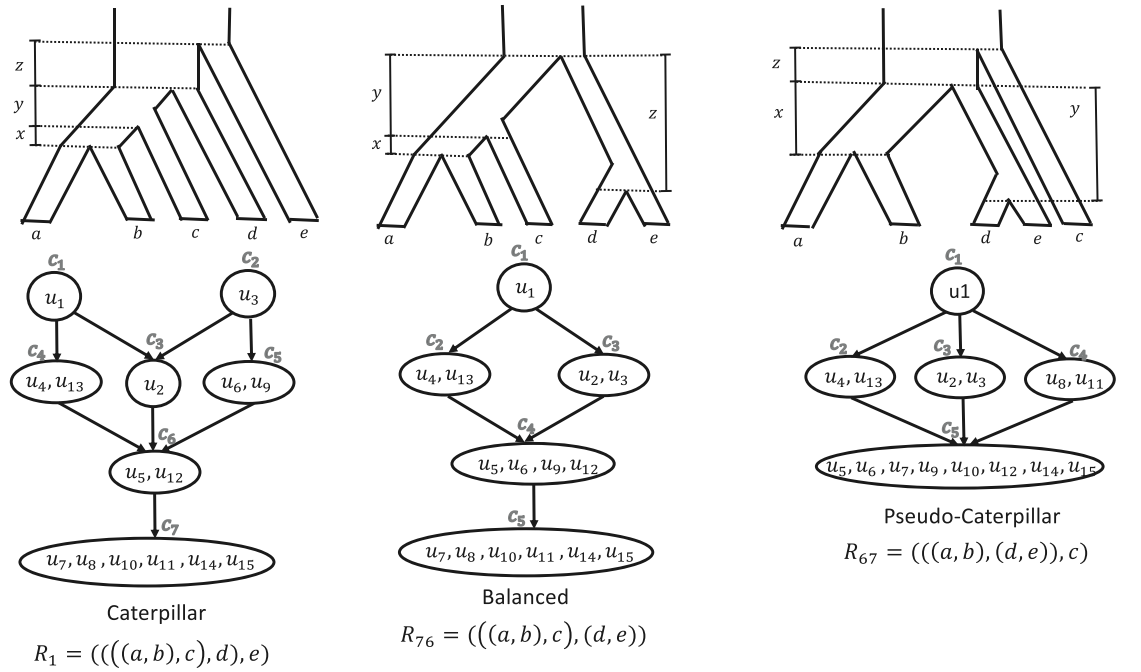


FIG. 1. ADR invariants and inequalities for different rooted topological shapes. The invariants (i.e., equalities) and inequalities found by ADR define a partial order on the probabilities of unrooted 5-taxon gene tree topologies for rooted 5-taxon model species trees with different rooted shapes (caterpillar, balanced, and pseudo-caterpillar). There are 15 unrooted binary trees on a given set of 5 leaves. Each of the 105 5-taxon rooted species trees define a specific distribution on the probabilities of these unrooted trees. The topology of the rooted binary species tree can be determined from this distribution (i.e., it is identifiable, as established by ADR). While the branch lengths of the rooted species tree depend on the actual probabilities, the linear invariants and inequalities that hold for these distributions are enough to determine the rooted topology of the model species tree. ADR, Allman, Degnan, and Rhodes.

As can be seen in Figure 1, the number of equivalence classes for caterpillar, balanced, and pseudo-caterpillar trees is 7, 5, and 5, respectively. Each directed edge between two equivalence classes in these Hasse diagrams defines an inequality, so that all gene tree probabilities in class c_a at the source of an edge are greater than all gene tree probabilities in class c_b at the target, and we show this by $c_a > c_b$.

The exact values of the unrooted gene tree probabilities depend on the internal branch lengths of the model tree, and ADR provide a set of formulas that relate the model tree parameters to the probability distribution of the unrooted gene trees in appendix B of Allman et al. (2011), which will be used in our proofs.

2.2. *Quintet Rooting*

The input to QR is an unrooted species tree T with n leaves and a set $\mathcal{G}$ of k single-copy unrooted gene trees where the gene trees draw their leaves from the leafset of T, denoted by $\mathcal{L}(T)$. Given this input, QR searches over all possible rootings of T and returns a tree most consistent with the distribution of quintets (i.e., 5-taxon trees) in the input gene trees.

QR approaches this problem by selecting a set Q of quintets of taxa from $\mathcal{L}(T)$ (called the ''quintet sampling'' step), and scoring all rooted versions of T based on their induced trees on these quintets. The subtree $T_{|q}$, that is, T restricted to taxa in quintet set q, can be rooted on any of its seven edges. In a preprocessing step, QR computes a score for each of these seven different rootings for all trees induced on the quintets in set Q, based on a cost function (described below).

This results in $7 \times |Q|$ computations, and therefore the preprocessing step takes $O(k(|Q| + n))$. Next, for every rooted version of T, QR sums up the costs of all its induced rooted trees on quintets in Q using the scores computed in the preprocessing step, and it returns the rooting with the minimum overall cost. Since T can be rooted on any of its $2n-3$ edges, the scoring step takes $O(n + |Q|)$ time.

Thus, QR provides an exact solution to the optimization problem with the following input and output:

- **Input**: An unrooted tree topology T, a set of k unrooted gene tree topologies $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$, a set Q containing quintets of taxa from leafset $\mathcal{L}(T)$, and a cost function $Cost(r, \vec{u})$.
- **Output**: Rooted tree R with topology T such that $\sum_{q \in Q} \text{Cost}(R|_q, \vec{\hat{u}}_q)$ is minimized, where $\vec{\hat{u}}_q$ is the distribution of unrooted quintet trees in $\mathcal{G}|_q = \{g_1|_q, g_2|_q, \ldots, g_k|_q\}$.

2.2.1. Cost function. The cost function $\text{Cost}(R|_q, \vec{\hat{u}}_q)$ measures the fitness of the rooted quintet tree $R|_q$ with the distribution of the unrooted gene trees restricted to q (i.e., $\vec{\hat{u}}_q$), according to the linear invariants and inequalities derived from the ADR theory. In particular, this cost function is designed to penalize a rooted tree $R|_q$ if the estimated quintet distribution $\vec{\hat{u}}_q$ violates some of the inequalities or invariants in its partial order. To this end, a penalty term was considered for each invariant and inequality in the partial order of a 5-taxon rooted tree that is violated in a quintet distribution. The cost function was defined based on a linear combination of these penalty terms, and had the following form, where r is a 5-taxon rooted tree and $\vec{\hat{u}}$ is an estimated quintet distribution:

$$\text{Cost}(r, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_r} \frac{1}{|c|} \sum_{u_a, u_b \in c} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_r} \frac{1}{|c'|} \sum_{u_a \in c,\, u_b \in c'} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}}. \quad (1)$$

The normalization factors $\frac{1}{|c|}$ and $\frac{1}{|c'|}$ were used to reduce a topological bias that arose from differences in the sizes of the equivalence classes for each tree shape.

2.3. Quintet sampling

The set Q of quintets in the QR algorithm can be selected in different ways, and here we consider sampling strategies that lead to statistical consistency. These sampling strategies differ in the number of quintets they sample and therefore their runtime. A straightforward sampling strategy is to use all $\Theta(n^5)$ quintets, which was the main approach used in Tabatabaee et al. (2022).

Alternatively, an $O(n)$ sampling method called ''Linear Encoding'' was proposed, as we now describe. The set Q_{LE} contains one quintet for each edge in the tree T, with the quintets computed as follows. If the edge e is incident to a leaf x, we note that deleting e partitions the set of taxa into three sets $\{x\}, A, B$; for this edge, the quintet q_e is formed by selecting $\{x\}$ and at least one leaf from each of the sets A and B, with the other two leaves of q_e being randomly selected.

For any edge e that is internal in the tree, the removal of e partitions the set of taxa into four subsets A_1, A_2, B_1 and B_2, and a quintet q_e is formed by picking one taxon from each of these four sets, and then picking a fifth taxon arbitrarily. Since a tree with n leaves has $2n-3$ edges, $|Q_{LE}| = 2n-3$ and therefore the runtime of the preprocessing step is $O(nk)$ when using the linear encoding. Note also that $Q_{LE}(T)$ may not be unique for trees with $n > 5$ taxa.

2.4. Lack of consistency for QR

QR uses the cost function in Equation (1) to select between different rootings of a 5-taxon unrooted species tree, given the estimated quintet distribution $\vec{\hat{u}}$. Lemma 3 shows that for each balanced tree, there are two caterpillar trees for which the set of violated inequalities becomes empty. Consider the caterpillar tree R_1 and balanced tree R_{76} shown in Figure 1.

Note that class c_3 in R_{76} is the result of merging the classes c_2 and c_3 in R_1. Moreover, class c_4 in R_{76} is the result of merging classes c_5 and c_6 in R_1. Assume that the model tree is R_{76}, and we have estimated $\vec{\hat{u}}$ given a set of k unrooted quintet gene trees. We now argue informally that even as k increases, there is no guarantee that eventually $\text{Cost}(R_{76}, \vec{\hat{u}}) < \text{Cost}(R_1, \vec{\hat{u}})$ for all large enough k.

According to the proof of Lemma 6 in Section 4, as k increases, for the model tree R_{76}, all inequality penalty terms in the form of $\max(0, \hat{u}_b - \hat{u}_a)$ will converge to zero in probability. Therefore, roughly speaking, the cost of R_{76} eventually consists primarily of the invariant penalty terms.

For the caterpillar tree R_1, most of its inequality penalty terms are also penalty terms in the cost of R_{76}, but it also has additional penalty terms between classes c_2 and c_3 as well as classes c_5 and c_6 that are merged in R_{76}. By simplifying the penalty terms that are eventually zero with high probability or are included in the cost of both trees, we get:

$$\begin{aligned}&\text{Cost}(R_1, \vec{\hat{u}}) \; - \; \text{Cost}(R_{76}, \vec{\hat{u}}) \approx \\ &\max(0, \hat{u}_2 - \hat{u}_3) + \tfrac{1}{2}\max(0, \hat{u}_5 - \hat{u}_6) + \tfrac{1}{2}\max(0, \hat{u}_5 - \hat{u}_9) + \tfrac{1}{2}\max(0, \hat{u}_{12} - \hat{u}_6) + \tfrac{1}{2}\max(0, \hat{u}_{12} - \hat{u}_9) \\ &\quad + \tfrac{1}{2}|\hat{u}_6 - \hat{u}_9| + \tfrac{1}{2}|\hat{u}_5 - \hat{u}_{12}| \\ &- (\tfrac{1}{2}|\hat{u}_2 - \hat{u}_3| + \tfrac{1}{4}|\hat{u}_5 - \hat{u}_6| + \tfrac{1}{4}|\hat{u}_5 - \hat{u}_9| + \tfrac{1}{4}|\hat{u}_5 - \hat{u}_{12}| + \tfrac{1}{4}|\hat{u}_6 - \hat{u}_9| + \tfrac{1}{4}|\hat{u}_6 - \hat{u}_{12}| + \tfrac{1}{4}|\hat{u}_9 - \hat{u}_{12}|)\end{aligned} \tag{2}$$

In the limit as $k \to +\infty$, all remaining terms in that difference also go to 0, since each term corresponds to a difference between two probabilities that are in the same equivalence class under the model tree. Hence, intuitively, there is no guarantee that R_{76} will be selected by QR. Based on this informal argument, we conjecture that QR is not statistically consistent.

3. QR-STAR

QR-STAR is an extension to QR that has an additional step for determining the rooted shape (i.e., the rooted topology without the leaf labels) of each quintet tree, as well as an associated penalty term in its cost function. This penalty term compares the rooted shape of the 5-taxon tree, denoted by $S(r)$, with the rooted shape inferred by QR-STAR from the given quintet distribution, denoted by $\hat{S}(\hat{u})$. The motivation for this additional preprocessing step is that, as argued in the previous section, the cost function of QR does not guarantee statistical consistency. The cost function of QR-STAR takes the following general form:

$$\text{Cost}^*(r, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_r} \sum_{u_a, u_b \in c} \alpha_{a,b} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_r} \sum_{u_a \in c,\, u_b \in c'} \beta_{a,b} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}} + \underbrace{C1|S(r) \neq \hat{S}(\hat{u})|}_{\text{Shape Penalty}} \tag{3}$$

where for all a, b, we require $\alpha_{a,b} \geq 0$ and $\beta_{a,b}, C > 0$.[†] Let $\alpha_{\max} = \max_{a,b}(\alpha_{a,b})$ and $\beta_{\min} = \min_{a,b}(\beta_{a,b})$ where a, b ranges over all pairs of indices a, b used in the penalty terms in Equation (3).

Each of the 105 rooted binary trees on a given set of 5 leaves have a unique set of inequalities and invariants that can be derived from the ADR theory. The cost function in Equation (3) considers a penalty term for these inequalities and invariants as well as the shape of the tree, so that $\text{Cost}^*(r, \vec{\hat{u}})$ is minimized for a rooted 5-taxon tree r that best describes the given estimated quintet distribution $\vec{\hat{u}}$.

3.1. Determining the rooted shape

The different rooted shapes (i.e., caterpillar, balanced, pseudo-caterpillar) of model 5-taxon species trees define equivalence classes with different class sizes on the unrooted gene tree probability distribution. These class sizes can be used to determine the unlabeled shape of a rooted tree, when given the *true* gene tree probability distribution.

For example, the size of the equivalence class with the smallest gene tree probabilities is 8 for the pseudo-caterpillar trees and 6 for balanced or caterpillar trees. Therefore, the size of the equivalence class corresponding to the minimal element in the partial order can differentiate a pseudo-caterpillar tree from other tree shapes. Moreover, both caterpillar and balanced trees have a unique class with the second smallest probability, which is of size 2 for caterpillar trees and size 4 for balanced trees, and this can be used to differentiate a caterpillar tree from a balanced tree. This approach is used in theorem 9 in Allman et al. (2011) for establishing the identifiability of rooted 5-taxon trees from unrooted gene trees.

However, given an *estimated* gene tree distribution, it is likely that none of the invariants derived from the ADR theory exactly hold, and so the class sizes cannot be directly determined and the approach cited earlier cannot be used as is to infer the shape of a rooted quintet. Here, we propose a simple modification for determining the rooted shape of a tree from the estimated distribution of unrooted gene trees, by looking for significant gaps between quintet gene tree probabilities.

Let T be the unrooted species tree with $n \geq 5$ leaves given to QR-STAR and q be a quintet of taxa from $\mathcal{L}(T)$. Let $\vec{\hat{u}}$ be the quintet distribution estimated from input gene trees induced on taxa in set q. QR-STAR first sorts $\vec{\hat{u}}$ in ascending order to get $\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \ldots \leq \hat{u}_{\sigma_{15}}$.

[†]See Remark 1 for why $\alpha_{a,b}$ does not need to be strictly positive.

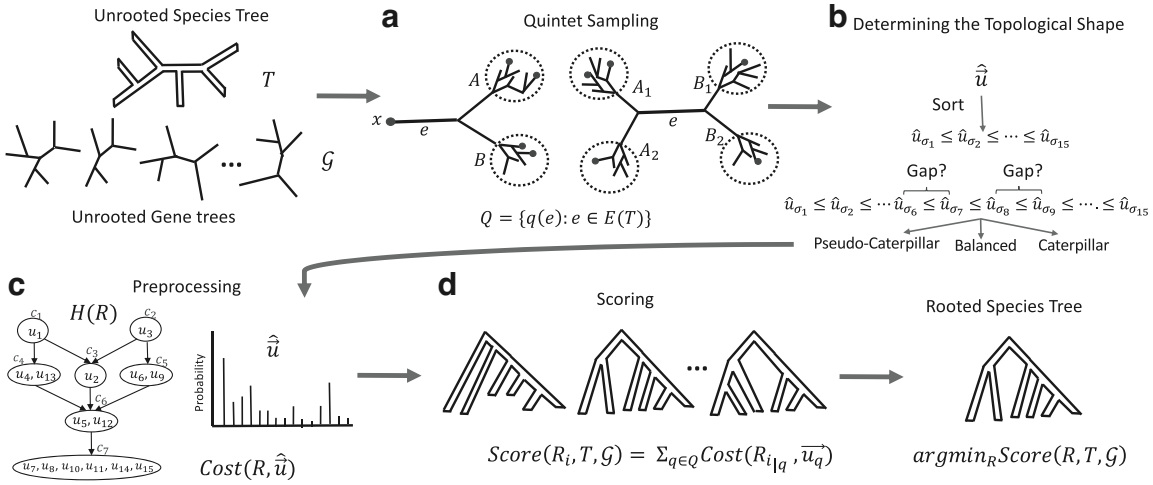


FIG. 2. QR-STAR Pipeline. The input is an unrooted species tree T and a set of unrooted gene trees $\mathcal{G}$ on the same leafset. **(a)** The sampling step selects a set Q of quintets from the leafset of T (shown is the linear encoding sampling). **(b)** The step that determines the rooted shape for each selected quintet. **(c)** The preprocessing step computes a cost for each of the seven possible rootings of each selected quintet. **(d)** The scoring step computes a score for each rooted tree in the search space based on the costs computed in the preprocessing step, and it returns a rooting of T with minimum score. The QR pipeline skips the step that determines the rooted shape for each selected quintet, and it has a simpler cost function. QR, Quintet Rooting.

We propose a general design for QR-STAR based on a given error probability $\delta > 0$, so that the algorithm returns the true rooted tree with probability at least $1-\delta$. Given δ, we define $A_{Q,\delta}(k) = \sqrt{\frac{2}{k}\ln\left(\frac{30|Q|}{\delta}\right)}$ (refer to Lemma 4 for the derivation), where k is the number of input gene trees and Q is the set of sampled quintets, which depends on the number n of taxa and is assumed fixed. The first step of QR-STAR computes an estimate of the rooted shape of a quintet q, denoted by $\hat{S}(\hat{u})$ in Equation (3), as follows:

- estimate the rooted shape $\hat{S}(\hat{u})$ as pseudo-caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A_{Q,\delta}(k)$;
- estimate the rooted shape $\hat{S}(\hat{u})$ as balanced if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \geq A_{Q,\delta}(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A_{Q,\delta}(k)$;
- estimate the rooted shape $\hat{S}(\hat{u})$ as caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \geq A_{Q,\delta}(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} \geq A_{Q,\delta}(k)$.

The runtime of QR-STAR is the same as QR, as determining the topological shape for each quintet is done in constant time, and the overall runtime remains $O(nk)$, when a linear sampling of quintets is used. Figure 2 shows the pipeline of QR-STAR and its individual steps.

4. THEORETICAL RESULTS

In this section, we provide the main theoretical results, starting with a series of lemmas and theorems that will be used in the proof of statistical consistency of QR-STAR in Theorem 2. Throughout this article, we assume that discordance between species trees and gene trees is solely due to ILS. In establishing statistical consistency, we assume that input gene trees are true gene trees and, thus, have no gene tree estimation error (GTEE). All trees are assumed to be fully resolved (i.e., binary).

4.1. Preliminaries

We begin with some definitions and key observations.

Definition 1 (Path length parameter). Let R be an MSC model species tree. Let $f(R)$ be the length of the shortest internal branch of R and $g(R)$ be the length of the longest internal path (i.e., a path formed from only the internal branches) of R. We define the path length parameter of R as:

$$h(R) = \frac{1}{18} e^{-3g(R)} (1 - e^{-f(R)})^2 \tag{4}$$

Note that $h(R) \in (0, \frac{1}{18})$ since $\exp(-x) \in (0, 1)$ for all $x > 0$ and the branch lengths have positive values. The formula for Equation (4) is derived from the proof of Lemma 2.

Lemma 1. Let R be an MSC model species tree with $n \geq 5$ leaves and q be an arbitrary set of 5 leaves from $\mathcal{L}(R)$. Then, $h(R|_q) \geq h(R)$ where $R|_q$ is the rooted tree R restricted to taxa in set q.

Proof. Every internal path of $R|_q$ is also an internal path in R, and therefore $g(R|_q) \leq g(R)$. Also, every branch in $R|_q$ is formed from one or more branches in R, so the shortest branch in $R|_q$ is at least as long as the shortest branch in R, and therefore $f(R|_q) \geq f(R)$. Hence:

$$h(R|_q) = \tfrac{1}{18} e^{-3g(R|_q)} (1 - e^{-f(R|_q)})^2 \geq \tfrac{1}{18} e^{-3g(R)} (1 - e^{-f(R)})^2 = h(R) \tag{5}$$

□

Lemma 2. Let R be an MSC model species tree with 5 leaves and internal branch lengths x, y, and z. Let $\vec{u}$ be the probability distribution that R defines on the unrooted 5-taxon gene tree topologies. If $\hat{\vec{u}}$ is an estimate of $\vec{u}$ such that given $\varepsilon > 0$, we have $|\hat{u}_i - u_i| < \varepsilon$ for all $1 \leq i \leq 15$, then the following inequality holds:

$$\forall_{c > c' \in C_R} \forall_{u_a \in c,\, u_b \in c'} : \hat{u}_a - \hat{u}_b > h(R) - 2\epsilon. \tag{6}$$

Proof. Let $X = e^{-x}$, $Y = e^{-y}$, and $Z = e^{-z}$. According to the explicit formulas for the probability distribution of unrooted gene trees $\vec{u}$ under a 5-taxon model species tree provided in appendix B of Allman et al. (2011), the exact value of each u_i can be expressed as a polynomial with variables X, Y, and Z.

We show that the lemma holds for all pairs u_a, u_b from different equivalence classes, for each tree category. For each category, we only show that the lemma holds for one example tree with that rooted shape [trees in table 1 in Tabatabaee et al. (2022)], as the rest of the trees have $\vec{u}$ distributions that are only permutations of the distributions of these three example trees (see supplementary section S2 in Tabatabaee et al., 2022) and the explicit formulas remain the same. The following equations can be derived using elementary algebraic arguments, and the fact that $X, Y, Z \in (0, 1)$.

Caterpillar trees. For a caterpillar tree $R = ((((a, b) : x, c) : y, d) : z, e)$ with $C_R = \{c_1 : \{u_1\}, c_2 : \{u_3\}, c_3 : \{u_2\}, c_4 : \{u_4, u_{13}\}, c_5 : \{u_6, u_9\}, c_6 : \{u_5, u_{12}\}, c_7 : \{u_7, u_8, u_{10}, u_{11}, u_{14}, u_{15}\}\}$, we have $c_1 > c_3, c_4 > c_6 > c_7$ and $c_2 > c_3, c_5 > c_6 > c_7$. Therefore:

- $u_a \in c_1, u_b \in c_3$

$$\begin{gathered} u_a - u_b = (1 - \tfrac{2}{3}X - \tfrac{2}{3}Y + \tfrac{1}{3}XY + \tfrac{1}{18}XY^3 + \tfrac{1}{90}XY^3Z^6) - (\tfrac{1}{3}Y - \tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) = \\ 1 - \tfrac{2}{3}X - Y + \tfrac{1}{2}XY + \tfrac{1}{6}XY^3 = (1 - Y) - \tfrac{1}{6}X(4 - 3Y - Y^3) = \\ (1 - Y) - \tfrac{1}{6}X(1 - Y)(4 + Y + Y^2) \geq (1 - Y) - \tfrac{1}{6}X(1 - Y)6 = (1 - Y)(1 - X) \\ \Rightarrow u_a - u_b \geq (1 - Y)(1 - X) \geq (1 - e^{-f(R)})^2 > h(R) \end{gathered} \tag{7}$$

where $(1 - Y)(1 - X) \geq (1 - e^{-f(R)})^2$ follows from the fact that $(1 - X) = 1 - e^{-x} \geq 1 - e^{-f(R)}$ as $x \geq f(R)$, and the same is true for y.

To save space, the rest of the derivations are provided in Appendix B. Therefore, we have:

$$\forall_{c > c' \in C_R} \forall_{u_a \in c,\, u_b \in c'} : u_a - u_b > \frac{1}{18} e^{-3g(R)} (1 - e^{-f(R)})^2 = h(R) \tag{8}$$

Since $|u_i - \hat{u}_i| < \epsilon$, we have $-\epsilon < \hat{u}_i - u_i < \epsilon$ and $-\epsilon < u_i - \hat{u}_i < \epsilon$. According to Equation (8):

$$\begin{gathered} \hat{u}_a - \hat{u}_b = (u_a - u_b) + (\hat{u}_a - u_a) + (u_b - \hat{u}_b) \Rightarrow \hat{u}_a - \hat{u}_b > (u_a - u_b) - \epsilon - \epsilon \\ \Rightarrow \hat{u}_a - \hat{u}_b > h(R) - 2\epsilon \end{gathered} \tag{9}$$

□

Definition 2. For a 5-taxon rooted tree R, we define I_R as the set of ordered pairs (i, j), $1 \leq i \neq j \leq 15$, corresponding to inequalities in the form $u_i > u_j$ defined according to the partial order of R. The inequalities that are a result of transitivity (i.e., $u_i > u_j$ and $u_j > u_k$ implies $u_i > u_k$) are not included in I_R.

Definition 3. Let $V(R, R')$ be the set of violated inequalities of two rooted 5-taxon trees R and R', that is, all pairs $\{i, j\}$ such that $(i, j) \in I_R$ and $(j, i) \in I_{R'}$.

Figure 3a shows an example of $V(R, R')$ computed for caterpillar trees, and Figure 3b is a heatmap showing the function $|V(R, R')|$ computed for the seven possible rootings of an unrooted quintet tree. The set $V(R, R')$ can be easily computed from I_R and $I_{R'}$ for all pairs of rooted 5-taxon trees, and I_R is derived from the ADR theory for all 105 5-taxon rooted trees in the supplementary section S2 in Tabatabaee et al. (2022).

Lemma 3. (a) For 5-taxon binary rooted trees R and R' with the same rooted shape, the set $V(R, R')$ is always non-empty. (b) For each balanced tree B, there exist two caterpillar trees C_1 and C_2 such that $V(B, C_i) = \emptyset$ for $i = 1, 2$.

Proof. (a) In Appendix A, we provide Appendix Figures A2–A4, which show the function $|V(R, R')|$ (number of violated inequalities) for all rooted quintet tree pairs R and R' with the same unlabeled topological shape (i.e., caterpillar, balanced, and pseudo-caterpillar), computed using the invariants and inequalities derived from the ADR theory [for details on how these are computed, refer to Tabatabaee et al. (2022), supplementary section S2]. It is clear that, except for the numbers on the main diagonal, all other values are non-zero. Therefore, $V(R, R')$ is always non-empty when R and R' have the same rooted topological shape.

(b) W.L.O.G. (Without loss of generality) assume we have a particular unrooted quintet tree T_1 (see table 5 in Allman et al., 2011) so that its seven possible rootings are caterpillar trees R_1, R_2, R_{59}, R_{60}, pseudo-caterpillar tree R_{67}, and balanced trees R_{76} and R_{105}. Figure 3b shows the function $|V(R, R')|$ for all these trees, and it is evident that for the balanced trees R_{76} and R_{105}, there are two caterpillar trees (R_1 and R_2 for R_{76}, and R_{59} and R_{60} for R_{105}) for which $|V(R, R')|$ becomes zero. The same can be observed for trees with other unrooted topologies in Appendix Figure A5 in Appendix A. □

Lemma 4. Let R be an MSC model species tree with $n \geq 5$ leaves and Q be a set of quintets of taxa from $\mathcal{L}(R)$. Given $\delta > 0$ and $k > 0$ unrooted gene tree topologies, the following inequality holds, where $A_{Q,\delta}(k) = \sqrt{\frac{2}{k} \ln\left(\frac{30|Q|}{\delta}\right)}$

$$\mathbb{P}\left(\forall_{q \in Q} \forall_{1 \leq i \leq 15} |(\hat{u}_q)_i - (u_q)_i| \; < \frac{A_{Q,\delta}(k)}{2}\right) \geq 1 - \delta. \tag{10}$$

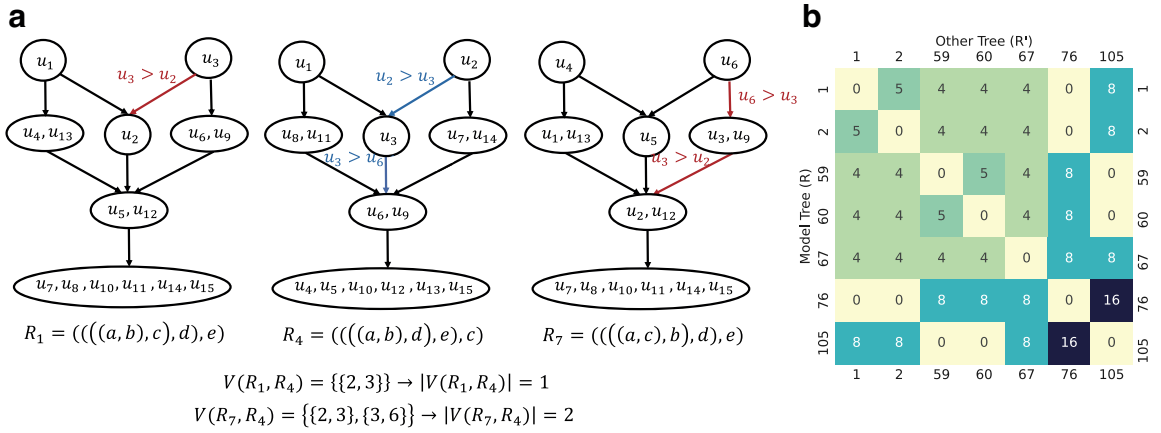


FIG. 3. Conflicting inequality penalty terms between rooted 5-taxon species trees. **(a)** Set of violated inequality penalty terms in the partial orders of R_1 and R_7 with respect to R_4, which are all caterpillar trees. The red edges show violations of inequalities in tree R_4, highlighted in *blue*. **(b)** Heatmap showing the number of pairwise violated penalty terms [function $|V(R, R')|$] of seven possible rooted trees having unrooted topology with bipartitions $ab|cde$ and $abc|de$. The *dark colors* indicate more violations, and the *lightest color* corresponds to no violations ($|V(R, R')| = 0$).

Proof. For an arbitrary $\epsilon > 0$, we have:

$$\begin{aligned}\mathbb{P}(\forall_{q\in Q}\forall_{1\le i\le 15}|(\hat{u}_q)_i-(u_q)_i|<\epsilon) &= 1-\mathbb{P}(\exists_{q\in Q,\,1\le i\le 15}|(\hat{u}_q)_i-(u_q)_i|\ge\epsilon)\\ &\ge 1-\sum_{q\in Q}\sum_{i=1}^{15}\mathbb{P}(|(\hat{u}_q)_i-(u_q)_i|\ge\epsilon)\end{aligned} \tag{11}$$

according to the union bound. Using the Hoeffding inequality (Mitzenmacher and Upfal, 2017) for each of the 15 unrooted 5-taxon tree topologies, we get:

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{j=1}^{k}X_{q,j,i}-\mu\right|\ge\epsilon\right)\le 2e^{\frac{-2k\epsilon^2}{(b-a)^2}}\Rightarrow\mathbb{P}(|(\hat{u}_q)_i-(u_q)_i|\ge\epsilon)\le 2e^{-2k\epsilon^2} \tag{12}$$

where $X_{q,j,i}$ is a binary random variable that is 1 when the quintet gene tree $g|_{qj}$ has the unrooted topology T_i and is zero otherwise, and so $0\le X_{q,j,i}\le 1$ almost surely. Substituting Equation (12) in Equation (11), we obtain:

$$\mathbb{P}(\forall_{q\in Q}\forall_{1\le i\le 15}|(\hat{u}_q)_i-(u_q)_i|<\epsilon)\ge 1-\sum_{q\in Q}\sum_{i=1}^{15}\mathbb{P}(|(\hat{u}_q)_i-(u_q)_i|\ge\epsilon)\ge 1-30|Q|e^{-2k\epsilon^2} \tag{13}$$

Setting $\epsilon=\sqrt{\frac{1}{2k}\ln(\frac{30|Q|}{\delta})}=\frac{A_{Q,\delta}(k)}{2}$ in the equation cited earlier proves the lemma:

$$\mathbb{P}(\forall_{q\in Q}\forall_{1\le i\le 15}|(\hat{u}_q)_i-(u_q)_i|<\epsilon)\ge 1-30|Q|e^{-2k\epsilon^2}=1-\delta. \tag{14}$$

□

4.2. Statistical consistency for 5-leaf trees

We now establish statistical consistency for QR-STAR under the MSC and provide a sufficient condition for a set of sampled quintets that leads to consistency. That is, we prove that as the number of input true gene trees increases, the probability that QR-STAR and its variants correctly root the given unrooted species tree converges to 1. We first prove statistical consistency for QR-STAR when the model tree has only five taxa in Theorem 1 and then extend the proofs to trees with arbitrary numbers of taxa in Theorem 2.

The main idea of the proof of consistency for 5-taxon trees is that we show as the number of input gene trees increases, the cost of the true rooted tree becomes arbitrarily close to zero, but the cost of any other rooted tree is bounded away from zero, where the bound depends on the path length parameter of the model tree $h(R)$ (see Definition 1).

To establish statistical consistency in Theorems 1 and 2, we assume that δ (now seen as a sequence depending on k) is such that $\lim_{k\to\infty}\delta=0$ and $\lim_{k\to\infty}A_{Q,\delta}(k)=0$. For instance, the choice $\delta=1/k$ satisfies these assumptions. Throughout this section, we will write $A(k)$ instead of $A_{Q,\delta}(k)$ since the species tree has only five leaves and the δ-sequence is fixed.

Lemma 5 (Correct determination of rooted shape). Let R be a 5-taxon model species tree and $\vec{u}$ be the probability distribution that it defines on the unrooted 5-taxon gene tree topologies. There is an integer $k>0$ such that if we are given at least k unrooted gene trees drawn i.i.d. (independent and identically distributed) from the distribution $\vec{u}$, the first step of QR-STAR will correctly determine the rooted shape of R with probability at least $1-\delta$.

Proof. Let $\mathcal{E}$ be the event that $|\hat{u}_i-u_i|<\frac{A(k)}{2}$ for all $1\le i\le 15$. Assume k is large enough so that $A(k)<\frac{1}{2}h(R)$. This is, indeed, possible under our assumption that $\lim_{k\to\infty}A(k)=0$. According to Lemma 4, the probability that $\mathcal{E}$ occurs is at least $1-\delta$. We assume that $\mathcal{E}$ holds in the rest of this proof. In this case, according to Lemma 2, we have

$$\forall_{c>c'\in C_R}\forall_{u_a\in c,\,u_b\in c'}:\hat{u}_a-\hat{u}_b>h(R)-A(k)>A(k) \tag{15}$$

where the last inequality is a result of the assumption $A(k)<\frac{1}{2}h(R)$. Therefore, the minimum distance between elements of any two equivalence classes that are related by an inequality in the partial order, that is, $c>c'\in C_R$, is greater than $A(k)$. Moreover, we now show that the maximum distance between elements inside an equivalence class is less than $A(k)$. Since $u_a=u_b$ and according to the triangle inequality, we obtain:

$$u_a,u_b\in c:|\hat{u}_a-\hat{u}_b|<|\hat{u}_a-u_a|+|u_a-u_b|+|u_b-\hat{u}_b|<\frac{A(k)}{2}+0+\frac{A(k)}{2}=A(k) \tag{16}$$

The partial orders on unrooted gene trees defined for each of the three topological shapes have a unique equivalence class whose members have the minimum probability, and for the caterpillar and balanced shapes, there is a unique class whose members have the second smallest probability. Since the distance between elements in different equivalence classes related by an inequality is greater than $A(k)$, and the distance between elements inside an equivalence class is less than $A(k)$, after sorting $\vec{\hat{u}}$ in ascending order, the elements of the equivalence class with the smallest probability appear at the beginning, followed by the elements of the second smallest class (for caterpillar and balanced shapes).

Let $\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \ldots \leq \hat{u}_{\sigma_{15}}$ be the result of sorting $\vec{\hat{u}}$ in ascending order. For a pseudo-caterpillar tree, the class with the minimum probability has eight elements, and for caterpillar or balanced trees it has six elements.

- The first step of QR-STAR determines the tree shape as pseudo-caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$ and else it will determine the shape as either caterpillar or balanced. When $\hat{u}_{\sigma_7}$ and $\hat{u}_{\sigma_6}$ belong to different classes, their distance must be greater than $A(k)$. Therefore, when $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$ holds, $\hat{u}_{\sigma_7}$ and $\hat{u}_{\sigma_6}$ must belong to the same equivalence class, and this only happens when the model tree is a pseudo-caterpillar tree. On the other hand, if R is a pseudo-caterpillar tree, then $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$, as the first eight elements in the sorted list must belong to the same class. Therefore, condition $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$ holds if and only if R has a pseudo-caterpillar shape and QR-STAR determines the correct unlabeled shape in this case.
- If $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} > A(k)$, then R is either a balanced or a caterpillar tree. The equivalence class with the second smallest probability values, for both caterpillar and balanced trees, is unique and has size 2 for caterpillar trees and size 4 for balanced trees. The first step of QR-STAR determines the tree shape as balanced if condition $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$ holds and else it would determine the tree shape as caterpillar. Similar to the explanation cited earlier for the case of pseudo-caterpillar trees, when $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$, $\hat{u}_{\sigma_8}$ and $\hat{u}_{\sigma_9}$ must belong to the same equivalence class and this only happens for balanced trees. Moreover, when R is balanced, $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$. Therefore, conditions $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} > A(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$ hold if and only if R is a balanced tree, so that QR-STAR correctly determines the tree shape in this case as well.
- Finally, when R is a caterpillar tree, $\hat{u}_{\sigma_8}$ and $\hat{u}_{\sigma_9}$ belong to different equivalence classes; therefore, $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} > A(k)$, and the other side can be shown similarly. Therefore, by comparing $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6}$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8}$ against $A(k)$ when k is large enough so that $A(k) < \frac{1}{2}h(R)$, QR-STAR will correctly determine the rooted shape of the model tree with probability at least $1-\delta$.

The argument is summarized next:

$$\begin{aligned}
\text{Pseudo}-\text{caterpillar}: \quad & \hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \ldots \leq \underbrace{\hat{u}_{\sigma_6} \leq \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}<A(k)} \leq \underbrace{\hat{u}_{\sigma_8} < \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}>A(k)} \leq \hat{u}_{\sigma_{10}} \leq \ldots \leq \hat{u}_{\sigma_{15}} \\
\text{Balanced}: \quad & \hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \ldots \leq \underbrace{\hat{u}_{\sigma_6} < \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}>A(k)} \leq \underbrace{\hat{u}_{\sigma_8} \leq \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}<A(k)} \leq \hat{u}_{\sigma_{10}} \leq \ldots \leq \hat{u}_{\sigma_{15}} \\
\text{Caterpillar}: \quad & \hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \ldots \leq \underbrace{\hat{u}_{\sigma_6} < \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}>A(k)} \leq \underbrace{\hat{u}_{\sigma_8} < \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}>A(k)} \leq \hat{u}_{\sigma_{10}} \leq \ldots \leq \hat{u}_{\sigma_{15}}
\end{aligned} \tag{17}$$

□

Lemma 6 (Upper bound on the cost of the model tree). Let R be a 5-taxon model species tree and $\vec{u}$ be the probability distribution that it defines on the unrooted 5-taxon gene tree topologies. There is an integer $k > 0$ such that if we are given at least k unrooted gene trees drawn i.i.d. from distribution $\vec{u}$, then $Cost^*(R, \vec{\hat{u}})$ is less than $31\alpha_{\max}A(k)$ with probability at least $1-\delta$.

Proof. Let $\mathcal{E}$ be the event that $|\hat{u}_i - u_i| < \frac{A(k)}{2}$ for all $1 \leq i \leq 15$. According to Lemma 4, the probability that $\mathcal{E}$ holds is at least $1-\delta$. When $\mathcal{E}$ holds, according to Lemma 2, the following inequality is true for $\vec{\hat{u}}$ and the path length parameter $h(R)$ of the model tree R:

$$\forall_{c>c' \in C_R} \forall_{u_a \in c,\, u_b \in c'} : \hat{u}_a - \hat{u}_b > h(R) - A(k). \tag{18}$$

when k is sufficiently large that $A(k) < \frac{1}{2}h(R)$ (which is possible under our assumption that $\lim_{k\to\infty} A(k) = 0$), then $\hat{u}_a - \hat{u}_b$ will be positive. Therefore, all inequality penalty terms that are defined as $\max(0, \hat{u}_b - \hat{u}_a)$ in $Cost^*(R, \vec{\hat{u}})$ become zero, since $\hat{u}_b - \hat{u}_a$ is a negative term.

Therefore, the total sum of the inequality penalty terms in $\text{Cost}^*(R, \vec{\hat{u}})$ will be zero for large enough k. Moreover, Lemma 5 states that the topological shape of R can be correctly determined when $A(k) < \frac{1}{2}h(R)$ and $\mathcal{E}$ holds, and hence the shape penalty term $1|S(R) \neq \hat{S}(\hat{u})|$ also becomes zero. Therefore, all elements of $\text{Cost}^*(R, \vec{\hat{u}})$ except the invariant penalty terms become zero.

According to Equation (16), for each invariant penalty term, we have $|\hat{u}_a - \hat{u}_b| < A(k)$. Hence:

$$\text{Cost}^*(R, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_R} \sum_{u_a,\, u_b \in c} \alpha_{a,b} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} < \alpha_{\max} \sum_{c \in C_R} \sum_{u_a,\, u_b \in c} A(k) = \alpha_{\max} A(k) \sum_{c \in C_R} \binom{|c|}{2} \leq 31 \alpha_{\max} A(k). \tag{19}$$

The last inequality holds since (1) caterpillar trees have seven equivalence classes with class sizes 1, 1, 1, 2, 2, 2, 6 and therefore $\sum_{c \in C_R} \binom{|c|}{2} = 18$, (2) balanced trees have five equivalence classes with sizes 1, 2, 2, 4, 6 and $\sum_{c \in C_R} \binom{|c|}{2} = 23$, and (3) pseudo-caterpillar trees have five equivalence classes with sizes 1, 2, 2, 2, 8 and $\sum_{c \in C_R} \binom{|c|}{2} = 31$. Hence, in all cases, we have $\sum_{c \in C_R} \binom{|c|}{2} \leq 31$ and the inequality follows. □

Theorem 1 (Statistical Consistency of QR-STAR for 5-taxon trees). Let R be a rooted 5-taxon model species tree and $\vec{u}$ be the distribution that it defines on the unrooted 5-taxon gene tree topologies. Given a set $\mathcal{G}$ of unrooted true quintet gene trees drawn i.i.d. from $\vec{u}$, QR-STAR is a statistically consistent estimator of R under the MSC.

Proof. We will show that we can find k large enough so that QR-STAR will correctly return the rooted version of R with probability at least $1 - \delta$ when given at least k true gene trees. Hence, QR-STAR is statistically consistent for rooting R, since $\lim_{k \to \infty} \delta = 0$ by assumption.

According to Lemma 6, when k is large enough so that $A(k) < \frac{1}{2}h(R)$, if $\vec{\hat{u}}$ is the distribution estimated from $\mathcal{G}$, then $\text{Cost}^*(R, \vec{\hat{u}})$ is at most $31\alpha_{\max}A(k)$ with probability at least $1 - \delta$. We now prove that for every other rooted 5-taxon tree R', $\text{Cost}^*(R', \vec{\hat{u}})$ is bounded away from zero. Note that according to Lemma 3, for every rooted 5-taxon tree $R' \neq R$ with the same rooted shape as R, we have $V(R, R') \neq \emptyset$ and therefore, there exists $1 \leq x \neq y \leq 15$ such that $(x, y) \in I_R, (y, x) \in I_{R'}$.

Let $\mathcal{E}$ be the event that $|\hat{u}_i - u_i| < \frac{A(k)}{2}$ for all $1 \leq i \leq 15$. According to Lemma 2, when $\mathcal{E}$ holds, then $\hat{u}_x - \hat{u}_y > h(R) - A(k)$ as $(x, y) \in I_R$. However, since $(y, x) \in I_{R'}$, an inequality penalty term in the form of $\max(0, \hat{u}_x - \hat{u}_y)$ is added to $Cost^*(R', \vec{\hat{u}})$ when R' has the same shape as R.

Moreover, according to Lemma 5, when $\mathcal{E}$ holds and $A(k) < \frac{1}{2}h(R)$, the first step of QR-STAR correctly determines the rooted shape of R. Therefore, if R' has a different rooted shape than R, then the penalty $1|S(R') \neq \hat{S}(\hat{u})|$ becomes 1 and a positive cost C is added to the cost of R'. Therefore, both cases (i.e., whether R' has a different topology from R or not) lead to a positive penalty in the cost function for R' that is bounded away from zero. Hence:

$$\begin{aligned} \text{Cost}^*(R', \vec{\hat{u}}) &= \underbrace{\sum_{c \in C_{R'}} \sum_{u_a,\, u_b \in c} \alpha_{a,b} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_{R'}} \sum_{u_a \in c,\, u_b \in c'} \beta_{a,b} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}} + \underbrace{C 1|S(R') \neq \hat{S}(\hat{u})|}_{\text{Shape Penalty}} \\ &\Rightarrow \forall_{R' \neq R} \text{Cost}^*(R', \vec{\hat{u}}) \geq \min(\beta_{\min}(\hat{u}_x - \hat{u}_y), C) > \min(\beta_{\min}(h(R) - A(k)), C) \end{aligned} \tag{20}$$

Equation (20) defines a lower bound for the cost of any tree other than the true rooted tree and Lemma 6 gives an upper bound for the cost of the true tree, both with respect to the estimated quintet distribution $\vec{\hat{u}}$. Therefore, when k is large enough so that:

$$A(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{\alpha_{\max}}{\beta_{\min}} + 1}, \frac{1}{2}h(R) \right) \tag{21}$$

(which, once again, is possible under our assumption that $\lim_{k\to\infty} A(k)=0$), we will have:

$$\text{Cost}^*(R,\vec{\tilde{u}}) < 31\alpha_{\max}A(k) < \min(\beta_{\min}(h(R)-A(k)), C) < \text{Cost}^*(R',\vec{\tilde{u}}) \tag{22}$$

which means that the cost of the true rooted tree will be less than the cost of any other rooted tree on the same leafset with probability at least $1-\delta$. Precisely, $\forall_{R'\neq R}\text{Cost}^*(R,\vec{\tilde{u}}) < \text{Cost}^*(R',\vec{\tilde{u}})$ when Equation (21) holds, where $\beta_{\min}, C, h(R) > 0$ and $\alpha_{\max} \geq 0$ are constants.

As a result, QR-STAR will return the true rooted species tree topology with probability converging to 1 as the number of gene trees grows large, proving the statistical consistency for 5-leaf trees.

Remark 1. Note that when $\alpha_{\max}=0$, meaning that the invariant penalty terms are removed from the cost function, the cost of the true tree is exactly zero according to the proof of Lemma 6, and the cost of any other tree is positive when k is large enough. Hence in this case, the condition in Equation (21) reduces to $A(k) < \frac{1}{2}h(R)$.

Remark 2. Note that Lemma 3(a) holds for *all* pairs of 5-taxon rooted trees with the same rooted shape and with different permutations of the leaf-labeling, regardless of whether they have the same (leaf-labeled) unrooted topology or not. Due to this property, it is possible to differentiate all pairs of 5-taxon rooted trees in a statistically consistent manner with the cost function of QR-STAR without prior knowledge about the unrooted tree topology, and hence Theorem 1 does not need to assume that the unrooted topology is given as input.

4.3. Extending to larger trees

The next lemma and theorem extend the proof of statistical consistency to trees with $n > 5$ taxa. Recall that linear encodings of T were defined in Section 2.3.

Lemma 7 (Identifiability of the root from the linear encoding). Let R and R' be rooted trees with unrooted topology T and distinct roots. Let $Q_{LE}(T)$ be the set of quintets of leaves in a linear encoding of T. There is at least one quintet of taxa $q \in Q_{LE}(T)$ so that $R_{|q}$ and $R'_{|q}$ have different rooted topologies.

Proof. Let e be the edge in T corresponding to the root of R. Let $q(e)$ be the quintet of leaves corresponding to edge e in $Q_{LE}(T)$. It is clear that $R_{|q(e)}$ is also rooted at edge e. The following cases can happen:

- When edge e is not incident to a leaf, it partitions the set of leaves of T into four subsets. Let A_1, A_2, B_1 and B_2 be the subsets resulting from deleting edge e from T, where A_1 and A_2 are incident to one endpoint of e and B_1 and B_2 are incident to the other. According to the definition of quintets in the linear encoding, $q(e)$ must have at least one leaf in each subset, which we call $a_1, a_2, b_1,$ and b_2 respectively.
  For every other rooted tree R' with topology T, the root edge e' of R' will fall into one of the four subsets A_1, A_2, B_1, B_2, including the edges sharing an endpoint with e. W.L.O.G. assume that e' falls into A_1. Then when T is rooted at edge e' (resulting in R'), the leaves a_2, b_1 and b_2 in $q(e)$ fall into one side of the root and a_1 falls into another.
  Therefore, the leaves a_1, a_2, and b_1 form the rooted triplet $((a_2, b_1), a_1)$. However, when T is rooted at e (producing R), these leaves form the rooted triplet $((a_1, a_2), b_1)$, as edge e separates a_1 and a_2 from b_1. Hence, in this case, $R_{|q(e)}$ and $R'_{|q(e)}$ are topologically different because they induce different rooted triplets.
- When edge e is adjacent to a leaf x, it partitions the set of taxa into three subsets, where one of them contains the single node $\{x\}$. Let A, B be the two other subsets resulting from deleting the edge e, where e separates x from the sets A and B. According to the definition of linear encoding, $q(e)$ must contain x and at least one leaf in A and B, which we call a and b respectively.
  For every other rooted tree R' with topology T, the root edge e' of R' will fall into A or B (including the edges directly adjacent to e). W.L.O.G. assume that it falls in A. Then when T is rooted at e (producing R), the nodes a, b, x form the rooted triplet $((a, b), x)$, but when T is rooted at e' (producing R'), they form the rooted triplet $((b, x), a)$. Therefore, this case also leads to different rooted topologies for $R_{|q(e)}$ and $R'_{|q(e)}$. Therefore, in both cases, R and R' restricted to the leaves in quintet $q(e)$ produce topologically different rooted quintet trees, completing the proof. □

Lemma 7 states that no two distinct rooted trees with topology T induce the same set of rooted quintet trees on quintets of taxa in a linear encoding $Q_{LE}(T)$. Clearly, the same is true for any superset Q such that $Q_{LE}(T) \subseteq Q$, including the set Q_5 of all quintets of taxa on the leafset of T.

There are also other quintet sets that are not a superset of $Q_{LE}(T)$, but have the property that no two rooted versions of T define the same set of rooted quintets on their elements. We generalize the proof of consistency to all sets of sampled quintets with this property.

Definition 4. Let T be an unrooted tree and Q be a set of quintets of taxa from $\mathcal{L}(T)$. We say Q is ''root-identifying'' if every rooted tree R with topology T is identifiable from T and the set of rooted quintet trees in $\{R|_q : q \in Q\}$, that is, no two rooted trees with topology T induce the same set of rooted quintet trees on Q.

Theorem 2 (Statistical Consistency of QR-STAR). Let R be an MSC model species tree with $n \geq 5$ leaves and let T denote its unrooted topology. Given T and a set $\mathcal{G}$ of unrooted true gene trees on the leafset $\mathcal{L}(T)$, QR-STAR is a statistically consistent estimator of the rooted version of T under the MSC, if the set of sampled quintets Q is root-identifying.

Proof. QR-STAR computes the score of each rooted tree R with topology T as $\sum_{q\in Q} \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$. Let $\mathcal{E}_q$ be the event that $|(\hat{u}_q)_i - (u_q)_i| < \frac{A_{Q,\delta}(k)}{2}$ for all $1 \leq i \leq 15$ for a quintet $q \in Q$. Arguing as in the proof of Theorem 1, when $\mathcal{E}_q$ holds and k is large enough so that $A_{Q,\delta}(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{\alpha_{\max}}{\beta_{\min}}+1}, \frac{1}{2}h(R)\right)$ (that guarantees $A_{Q,\delta}(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R|_q)}{31\frac{\alpha_{\max}}{\beta_{\min}}+1}, \frac{1}{2}h(R|_q)\right)$ according to Lemma 1), $\text{Cost}(R|_q, \hat{\vec{u}}_q)$ will be less than the cost of each of the six alternative rooted trees on the five taxa in q with high probability, as $R|_q$ is the *true* rooting of the unrooted quintet tree $T|_q$.

According to Lemma 4, $\mathcal{E}_q$ simultaneously holds for all $q \in Q$ with probability at least $1-\delta$. In this event, for every other rooted tree $R' \neq R$ and each $q \in Q$, we have $\text{Cost}^*(R'|_q, \hat{\vec{u}}_q) \geq \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$, according to Theorem 1. This means that for every rooted tree R' with topology T, $\sum_{q\in Q} \text{Cost}^*(R'|_q, \hat{\vec{u}}_q) \geq \sum_{q\in Q} \text{Cost}^*(R|_q, \hat{\vec{u}}_q)$.

Also, according to Definition 4, Q has the property that no other rooted tree R' induces the exact same set of rooted quintet trees as R on Q. Hence, there exists $q^* \in Q$ such that $\text{Cost}^*(R'|_{q^*}, \hat{\vec{u}}_{q^*}) > \text{Cost}^*(R|_{q^*}, \hat{\vec{u}}_{q^*})$, as the cost of $R|_{q^*}$ is strictly less than the cost of each of the six alternative rooted trees on q^* when the conditions in Theorem 1 hold. Therefore, the function $\sum_{q\in Q} \text{Cost}^*(r|_q, \hat{\vec{u}}_q)$ obtains its *unique* minimum for the rooted tree R.

As a result, QR-STAR returns the true rooted topology of T with probability converging to 1 as the number of input gene trees increases, establishing the statistical consistency given trees with an arbitrary number of taxa. □

4.4. Sample complexity of QR-STAR

Having established statistical consistency, we now discuss sample complexity—that is, the number of genes that suffice for QR-STAR to correctly root the model species tree with probability at least $1-\delta$, for an arbitrary $\delta > 0$.

Theorem 3 (Sample Complexity of QR-STAR). Let R be an MSC model species tree with $n \geq 5$ leaves and let T denote its unrooted topology. Given T, Q (a root-identifying set of sampled quintet trees), $\delta > 0$, and k true unrooted gene trees on the leafset of T, QR-STAR returns the true tree R with probability at least $1-\delta$, when the number of gene trees satisfies:

$$k > 2 \times 36^2 \ln\left(\frac{30|Q|}{\delta}\right) \frac{e^{6g}}{(1-e^{-f})^4} \tag{23}$$

where f and g are the lengths of the shortest internal branch and the longest internal path in R, respectively. When the linear encoding is used so that $|Q| = 2n-3$ and in the limit of small f, QR-STAR returns the true rooted tree with probability at least $1-\delta$ when the number k of gene trees satisfies:

$$k = \Omega(f^{-4} e^{6g} (\ln(n) - \ln(\delta))). \tag{24}$$

Proof. According to the proof of Theorem 2, when k is large enough so that $A_{Q,\delta}(k) < \min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{2\max}{\beta_{\min}}+1}, \frac{1}{2}h(R)\right)$, the function $\sum_{q\in Q}\text{Cost}^*(r|_q, \hat{\tilde{u}}_q)$ will be minimized for rooted tree R and QR-STAR will return the true tree with probability at least $1-\delta$. For simplicity, we consider the case where the weights α_i, β_i, C in the cost function of QR-STAR are set such that $\min\left(\frac{C}{31\alpha_{\max}}, \frac{h(R)}{31\frac{2\max}{\beta_{\min}}+1}, \frac{1}{2}h(R)\right)$ reduces to $\frac{1}{2}h(R)$ (also see Remark 1). Substituting the definitions of $A_{Q,\delta}(k)$ and $h(R)$, we get:

$$\begin{gathered} A_{Q,\delta}(k) < \tfrac{1}{2}h(R) \Leftrightarrow \\ \sqrt{\tfrac{2}{k}\ln\left(\tfrac{30|Q|}{\delta}\right)} < \tfrac{1}{36}e^{-3g}(1-e^{-f})^2 \Leftrightarrow \\ \tfrac{2}{k}\ln\left(\tfrac{30|Q|}{\delta}\right) < \left(\tfrac{1}{36}\right)^2 e^{-6g}(1-e^{-f})^4 \Leftrightarrow \\ k > 2\times(36)^2\ln\left(\tfrac{30|Q|}{\delta}\right)\tfrac{e^{6g}}{(1-e^{-f})^4} \end{gathered} \tag{25}$$

When the linear encoding is used so that $|Q| = 2n-3$ and in the limit of small f, we have $\lim_{f\to 0}\frac{1-e^{-f}}{f} = 1$, and hence QR-STAR returns the true rooted tree with probability at least $1-\delta$ when the number of gene trees satisfy:

$$k = \Omega(f^{-4}e^{6g}(\ln(n) - \ln(\delta))). \tag{26}$$

□

Theorem 3 yields a sample complexity that is exponential in g, the length of the longest internal path in R. However, an improved sample complexity can be obtained through a more nuanced analysis as well as using a modified version of the linear encoding, as we now show.

Definition 5 (Q-Restricted path length parameter). Let R be an MSC model species tree. Let Q be a set of quintets of taxa from $\mathcal{L}(R)$, and let $R|_Q = \{R|_q : q \in Q\}$ be the corresponding set of rooted quintet trees. Let f_Q be the length of the shortest internal branch in any (rooted) quintet tree in $R|_Q$ and g_Q be the length of the longest internal path of any quintet tree in $R|_Q$. We define the path length parameter of R restricted to Q as:

$$h_Q(R) = \frac{1}{18}e^{-3g_Q}(1-e^{-f_Q})^2 \tag{27}$$

Note that $f_Q \geq f(R)$ and $g_Q \leq g(R)$, where $f(R)$ and $g(R)$ are defined in Definition 1. However, if Q is the set of all quintets of taxa on the leafset of R, then $f_Q = f(R)$ and $g_Q = g(R)$, so that the Q-restricted path length parameter of R is identical to the path length parameter of R, as given in Definition 1. Finally, note that we can replace $h(R)$ by $h_Q(R)$, without any change in the theoretical results.

We now define a variant of the linear encoding, which we refer to as a ''short quintet encoding'' of the tree. This variant is motivated by the concept of ''short quartets'' (which are quartets of leaves sampled around each internal edge in a tree so that they are the closest leaves to that edge) and the strong theoretical properties of the ''short quartet methods,'' which estimate phylogenetic trees from aligned sequences by first estimating quartet trees that seem likely to be short quartets and then combining the quartet trees into a tree on the full dataset (Erdős et al., 1999a; Erdős et al. 1999b; Roch, 2019; Warnow et al., 2001).

As proven in Erdős et al. (1999a, 1999b), even simple versions of these methods have provably polynomial sample complexity under standard models of sequence evolution down trees [e.g., the Generalized Time Reversible model (Tavaré, 1986)], after bounding (arbitrarily) the length of the shortest and longest internal edges in the model tree.

Definition 6 (Short Quintet Encoding). Let T be an unrooted tree and e an internal edge in T, so that deleting e and its endpoints produces four subtrees. A short quintet around edge e contains a nearest leaf (in topological distance) in each of the four subtrees around e, and one other leaf that is chosen so that it is either tied for nearest in its subtree or second nearest within its subtree. For the case where e is incident with a leaf x, removing e and its endpoints splits the tree into two subtrees, A and B. A short quintet around e will include x and then four other leaves. If each of A and B has at least two leaves, then we pick the two nearest leaves in each of them. Otherwise, we pick the single leaf from one subtree and the three nearest leaves from the other subtree. We modify the linear encoding algorithm to ensure that each of the sampled quintets is a short quintet, and we refer to this as a Short Quintet Encoding of T.

Note that there can be more than one short quintet encoding of a tree, and that every short quintet encoding is root-identifying (since each is a linear encoding).

Theorem 4. Let R be an MSC model species tree with $n \geq 5$ leaves with unrooted topology T, let $f(R)$ be the length of the shortest internal edge in R, and let z_T be the length of the longest internal edge in the unrooted topology T for R. Given $\delta > 0$, k true unrooted gene trees on the leafset of T, short quintet encoding Q of T, and tree T, in the limit of small f with probability at least $1-\delta$, QR-STAR returns the correct rooting of T (i.e., true tree R), when the number k of gene trees satisfies:

$$k = \Omega\left(\frac{n^{O(z_T)}(\ln(n) - \ln(\delta))}{f(R)^4}\right). \tag{28}$$

Proof. Recall that for any set Q of quintets, $f_Q \geq f(R)$. Note also that any short quintet encoding is root-identifying since this is just a special case of a linear encoding.

By arguments similar to the proof provided for Theorem 3 and by substituting $h(R)$ with $h_Q(R)$, in the limit of small $f(R)$, QR-STAR returns R with probability at least $1-\delta$ if:

$$k = \Omega(f(R)^{-4} e^{6g_Q}(\ln(n) - \ln(\delta))). \tag{29}$$

Note also that g_Q (the length of the longest internal path of the quintet trees using the short quintets in Q) is $O(z_T \log n)$, since the topological diameter within T of any short quintet is $O(\log n)$ [based on the same arguments that short quartets have topological diameters that are $O(\log n)$ (Erdős et al., 1999a)]. Hence, $e^{6g_Q} = e^{O(z_t \log n)} = O(n^{O(z_t)})$. The result follows. □

This means that QR-STAR has a polynomial sample complexity when we fix $f(R)$ and z_T, the length of the shortest internal edge in R and longest internal edge in T, respectively.

5. EXPERIMENTAL STUDY

5.1. Overview

We performed four experiments in this study. Experiment 0 was used for the design of QR-STAR, where we used a training dataset with 101-taxon species trees to set the numeric parameters in its cost function. Experiments 1–3 are on test datasets, which are separate from the training data. Experiments 1 and 2 examine rooting of the true or estimated species trees, respectively, on a dataset with 201-taxon trees generated using SimPhy (Mallo et al., 2016) under different model conditions.

Experiment 3 examines rooting of estimated species trees on two simulated datasets with model trees resembling real biological datasets [a 48-taxon avian species tree from Jarvis et al. (2014) and a 37-taxon mammalian tree from Song et al. (2012)]. Overall, the model conditions in the test datasets vary in terms of the number of taxa, number of genes, GTEE, level of ILS, and topological shape of the species tree.

For each model condition (both in training and in test datasets), we report the level of ILS using the average normalized RF [i.e., Robinson-Foulds (Robinson and Foulds, 1981)] distance between the model species tree and true gene trees, and denote this value by AD, or average distance. We also report the average GTEE using normalized RF distance between true and estimated gene trees.

We evaluated rooting error using normalized clade distance (nCD) (Tabatabaee et al., 2022), which is a rooted version of the normalized RF distance. For the training experiments, we also report the proportion of the trees that are correctly rooted.

For the training experiment, we only rooted the true species tree topology to directly observe the rooting error. In the test experiments, we rooted both the model species tree and estimated species tree, as produced by ASTRAL, using both true and estimated gene trees. Throughout these experiments, we set $\delta = \frac{1}{k}$ in QR-STAR, where k is the number of gene trees in the input.

All datasets, along with the estimated gene trees, are from prior studies (Mirarab and Warnow, 2015; Mirarab et al., 2014a; Zhang et al., 2018) and are available online. Additional information about the simulation study is provided in Appendix C.

5.2. Designing QR-STAR

We used the 101-taxon simulated datasets from Zhang et al. (2018) as our training data, which had model conditions characterized by four GTEE levels, ranging from 0.23 to 0.55 for 1000 genes. The normalized RF distance between the model species tree and true gene trees (denoted AD) in this dataset was 0.46, which indicates moderate ILS.

We explored a range of values for the shape coefficient (parameter C) and the relative weight of inequalities and invariants (the ratio $\frac{\alpha_{max}}{\beta_{min}}$) in the cost function of QR-STAR on the training dataset. When $\alpha_{max} > 0$, these two values can impact the sample complexity of QR-STAR as Equation (21) suggests. We report the proportion of the trees (from the 50 replicates in each condition) that are correctly rooted, as well as the rooting error (nCD values) for rooting the true species tree topology.

Figure 4 shows the impact of shape coefficient on the accuracy of QR-STAR, where the weights of invariant and inequality penalty terms are fixed to the weights in the original cost function of QR. For small C values (i.e., less than 1E-02), the accuracy of QR-STAR does not seem to be affected by the shape coefficient, but as C gets larger, the accuracy degrades until it reaches a stationary point again.

This suggests that the shape coefficient should be kept relatively small compared with the invariant and inequality penalty weights, as they may better capture the difference between two rooted quintet trees. Since Equation (21) suggests that larger C values are theoretically preferred, on the experiments on the test dataset, we set the value of C as 1E-02 (the largest value before accuracy degrades).

Figure 5 shows the impact of the ratio $\frac{\alpha_{max}}{\beta_{min}}$ on QR-STAR. Here, all α and β values are set as equal. The results suggest that when the inequalities are weighed more than the invariants (and so $\frac{\alpha_{max}}{\beta_{min}}$ is <1), QR-STAR has its optimal accuracy, and the accuracy degrades when the invariants are weighed more.

For both figures, the trends for different sequence lengths are similar, and the degradation in accuracy starts almost at the same point, but the accuracy is higher for longer sequence lengths, which is expected as shorter sequence lengths correspond to higher levels of GTEE. In general, these experiments show that optimal accuracy could be achieved for a wide range of parameters in QR-STAR. For experiments on the test dataset, we set C as 1E-02 and $\frac{\alpha_{max}}{\beta_{min}}$ as 0 (essentially removing invariants from the cost function), although we note that the optimal values could be dataset-dependant, and better training procedures might be needed to find robust parameter values that work well across different datasets.

5.3. Evaluating QR-STAR

Using the numeric parameters selected in Experiment 0, we compared QR-STAR to QR in two basic experiments on the test datasets. Experiment 1 compares QR and QR-STAR when rooting the true (model) species tree, given true or estimated gene trees, where the final error solely shows the rooting error.

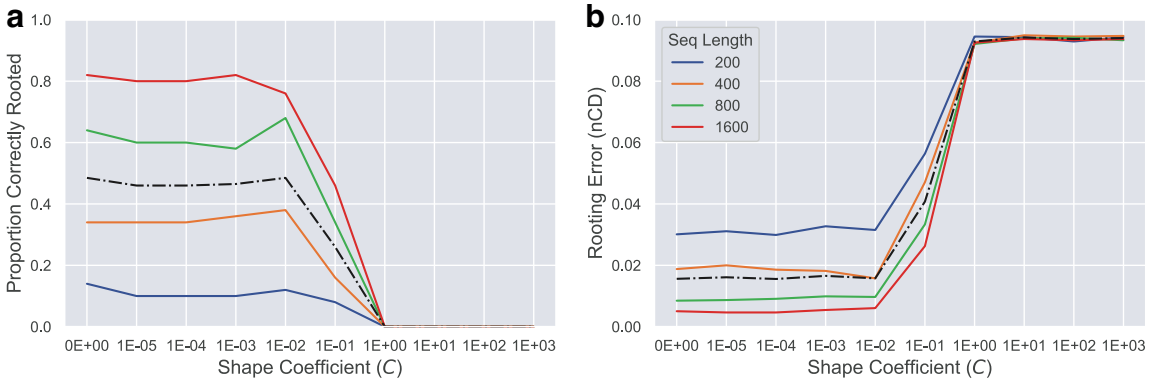


FIG. 4. Impact of shape coefficient (C) on QR-STAR. **(a)** Proportion of the trees correctly rooted and **(b)** rooting error (nCD) is shown for the 101-taxon dataset from Zhang et al. (2018) averaged over 50 replicates. The number of genes is 1000, and the average AD level is 0.46. The sequence length used to produce estimated gene trees varies between 200 and 1600 bp. The black dashed line corresponds to the average among sequence lengths. The weights of invariant and inequality penalty terms are set as in the cost function of QR. The value of C varies between 0 and 10^3, with $C = 0$ corresponding to the cost function of QR that does not guarantee consistency. AD, average distance; nCD, normalized clade distance.

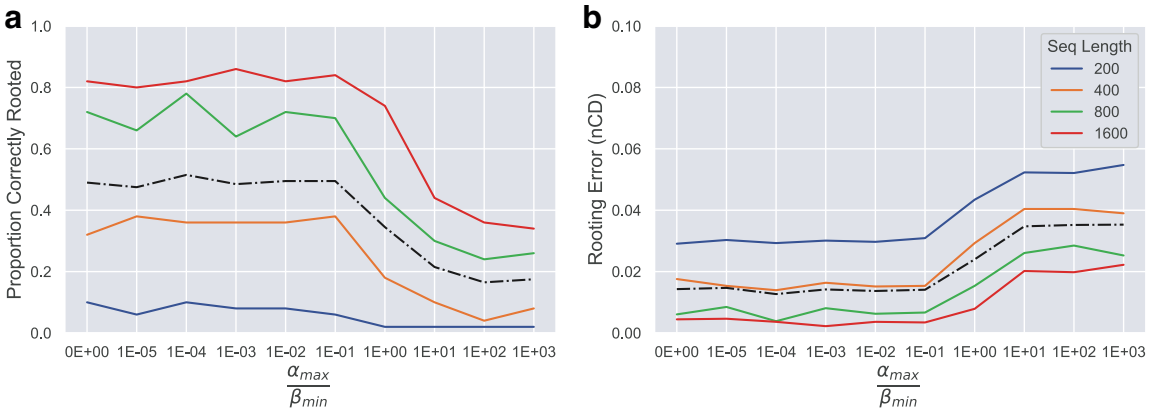


FIG. 5. Impact of $\frac{\alpha_{max}}{\beta_{min}}$ on QR-STAR. **(a)** Proportion of the trees correctly rooted and **(b)** rooting error (nCD) is shown for the 101-taxon dataset from Zhang et al. (2018) averaged over 50 replicates. The number of genes is 1000, and the average AD level is 0.46. The sequence length used to produce estimated gene trees varies between 200 and 1600 bp. The black dashed line corresponds to the average among sequence lengths. The value of $\frac{\alpha_{max}}{\beta_{min}}$ varies between 10^{-5} to 10^{3}, in addition to 0. All α and β values are set as equal for all invariant or inequality penalty terms.

Experiment 2 compares these methods when rooting an estimated species tree produced by ASTRAL, given true or estimated gene trees, where the final error is a combination of species tree estimation and rooting error.

For this second experiment, as the clade distance from the rooted version of the ASTRAL tree is a combination of RF distance between the estimated species tree and the true species tree as well as the error produced by the rooting method, we also report the optimal rooted species tree error, which is the lowest nCD error rate achieved across all possible rootings of the ASTRAL tree (see Appendix D for additional comments).

5.3.1. Datasets. We used a set of 201-taxon simulated datasets from Mirarab and Warnow (2015) as our test data; these are characterized by two different speciation rates and three tree heights (thus six tree shapes), and three number of genes for each tree shape. The AD levels for this dataset for 1000 genes ranged from 0.09 (for the 10M, 1e-07 condition) to 0.69 (for the 500K, 1e-06 condition).

The estimated gene trees were inferred using FastTree 2 (Price et al., 2010). The GTEE levels on the test data varied from 0.22 (for the 10M, 1e-06 condition) to 0.49 (for the 500K, 1e-06 condition). Appendix Table C1 in the Appendix C summarizes these statistics. The number of replicates for each model condition in this dataset was 50.

We also performed experiments on the 48-taxon avian-like and 37-taxon mammalian-like simulated datasets from Mirarab et al. (2014a), which had model species trees based on biological datasets from Jarvis et al. (2014) and Song et al. (2012), respectively. The default model condition in these datasets (shown with 1X ILS) had an ILS level that resembled the gene tree discordance in the corresponding biological data, but additional model conditions were created by multiplying or dividing branch lengths by two, thus decreasing or increasing the level of ILS, respectively (i.e., the highest ILS level we test for each biological dataset is indicated by 0.5X).

True gene trees were simulated within the model species trees under the MSC, and then sequences with varying lengths were evolved under each gene tree. Finally, RAxML (Stamatakis, 2014) was used to estimate gene trees from these sequence alignments, creating conditions with varying GTEE levels. These datasets had 20 replicates in each model condition, but the model tree in all replicates was the same tree from the corresponding biological study. Appendix Tables C2 and C3 in the Appendix C summarize the statistics for these two datasets.

5.3.2. Results for experiment 1: rooting the true species tree. Figure 6 (left) shows the result of rooting the model species tree with true gene trees on the test datasets. These results show that rooting error for both QR and QR-STAR decreases with the number of genes, as expected. We also see that rooting error is lowest for the highest ILS level (left-most column), and it increases as the ILS level decreases.

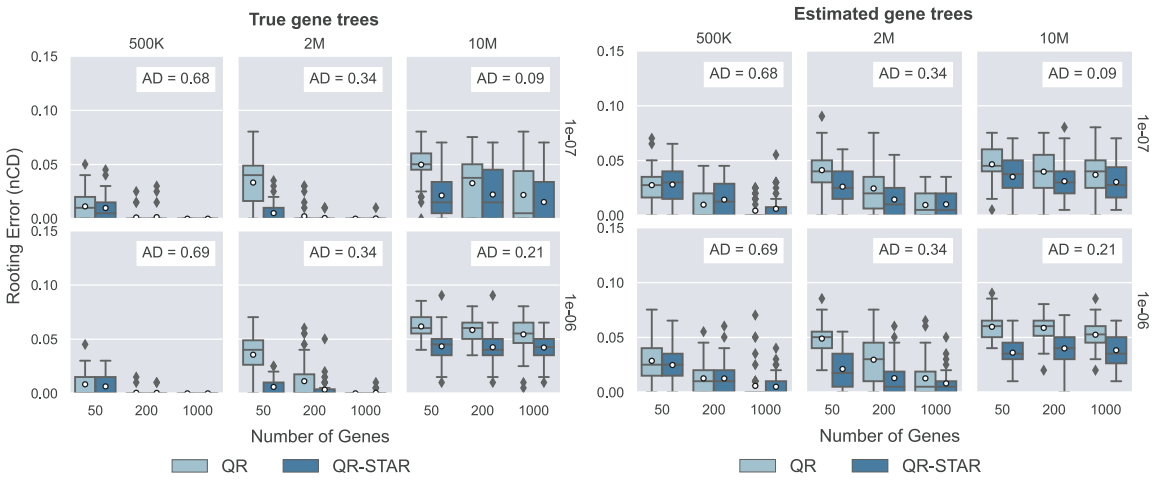


FIG. 6. Rooting the model species tree on 201-taxon simulated datasets. Comparison between QR and QR-STAR in terms of rooting error (nCD) for rooting the true unrooted species tree topology using true or estimated gene trees on the 201-taxon datasets, with 50 replicates in each model condition. The columns show tree height (500K for high ILS, 2M for moderate ILS, and 10M for low ILS), and the rows show speciation rate (1e-06 for recent speciation, 1e-07 for deep speciation). ILS, incomplete lineage sorting.

The impact of speciation rate on rooting error is seen by comparing the top and bottom rows; this impact is small except for the lowest ILS case, where deep speciation (1e-07) generally leads to lower error than recent speciation (1e-06). A comparison of nCD error rates for QR and QR-STAR shows that the two methods are close in accuracy for some conditions (notably for high or moderately high ILS with a sufficient number of genes) but when there are differences, QR-STAR has lower rooting error.

The advantage for QR-STAR over QR is largest for conditions with moderate to low ILS, and a few genes. However, under the lowest ILS condition and with speciation rate 1E-06 (bottom right subfigure), there is a consistent advantage to QR-STAR across all numbers of genes.

Figure 6 (right) shows the same comparison with estimated gene trees. As with true gene trees, increasing the ILS level (by reducing tree height) decreases the rooting error, increasing the number of genes also generally reduces rooting error (although much less under the lowest ILS level where tree height is 10M), and changing the speciation rate has a small impact (even on the low ILS condition).

A comparison between QR and QR-STAR shows that the relative accuracy depends on the ILS level. For the highest ILS condition (leftmost column), QR and QR-STAR are very close but with possibly a small advantage to QR. However, for moderate to low ILS conditions, QR-STAR matched or improved on QR.

5.3.3. Results for experiment 2: rooting an estimated species tree. Figure 7 show results on the test dataset, when rooting species trees estimated using ASTRAL with true or estimated gene trees. For all three methods (QR, QR-STAR, and optimal rooting), and using both true and estimated gene trees, increasing the number of genes improves accuracy, but increasing the ILS level reduces accuracy (in contrast to Experiment 1).

We also see that under high ILS, the error in the rooted species tree is high (on average 14.6% when using only 50 true gene trees, and 21.4% when using 50 estimated gene trees), but decreases rapidly as the number of genes increases. Error is higher for speciation close to the leaves (1e-06) than for speciation closer to the root (1e-07), a pattern that was also observed when rooting the model species tree.

The trends relating QR-STAR and QR are interesting to discuss. When using true gene trees, the relative accuracy of QR and QR-STAR depends on the ILS level, with an essentially identical error for the high ILS condition, but then an advantage to QR-STAR for the moderate or low ILS conditions (except when there is a sufficient number of genes).

When used with estimated gene trees, the relative accuracy between QR and QR-STAR depends on the ILS level, but the gap between QR and QR-STAR is smaller. There is essentially no difference for the high ILS condition, a very small difference for the moderate ILS condition (but only if the number of genes is small), and a small difference for the low ILS condition that holds across both low and moderate numbers of genes.

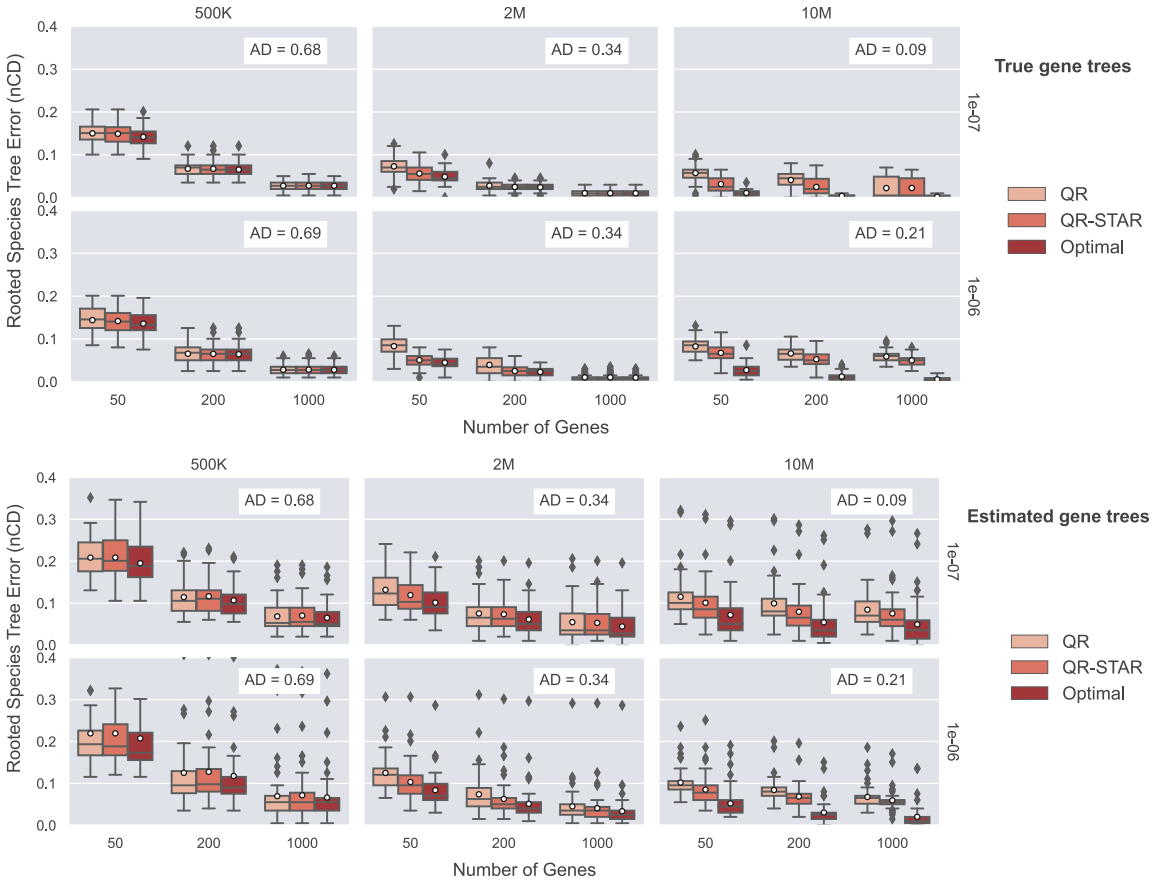


FIG. 7. Rooting the ASTRAL species tree on 201-taxon datasets. Comparison between QR, QR-STAR, and optimal rooting in terms of rooted species tree error (nCD) for rooting the species trees estimated by ASTRAL, using true or estimated gene trees on the 201-taxon datasets across 50 replicates. The columns show tree height (500K for high ILS, 2M for moderate ILS, and 10M for low ILS), and the rows show speciation rate (1e-06 for recent speciation, 1e-07 for deep speciation). The *y*-axes are cut at 0.4 to improve clarity, removing five outliers in the bottom figure from all methods in the 500K, 1e-06 model condition (see Appendix Fig. A6 in Appendix A for the full-scale figure).

Thus, the trends for rooting ASTRAL species trees are somewhat different in terms of absolute rooting error (which increases, compared with rooting the true species tree), but the relative performance of QR-STAR and QR shows similar results as for rooting true species trees. The main difference is that the difference between the methods seems to have decreased.

Finally, a comparison between QR-STAR and the optimal rooting provides some noteworthy trends. Specifically, for both true and estimated gene trees and under both high and moderate ILS, QR-STAR and optimal rooting are extremely close in terms of rooting error (with no detectable differences under high ILS and only a small difference under moderate ILS with only 50 genes).

Thus, under these conditions, there is little room for improvement over QR-STAR. Interestingly, there is a bigger gap between QR-STAR and optimal rooting for low ILS than under the higher ILS conditions, especially when the speciation rate is 1e-06 (i.e., speciation toward the leaves). We also see that there is a slightly bigger gap between QR-STAR and the optimal rooting when QR-STAR is using estimated gene trees than when using true gene trees; as expected.

5.3.4. Results for experiment 3: rooting an estimated species tree on biological model trees. Figure 8 shows results when using QR or QR-STAR to root the ASTRAL species trees on the avian or mammalian simulated datasets. Most trends are similar to the trends seen on 201-taxon datasets: accuracy with true gene trees is better than with estimated gene trees, and using more genes improve the results, as expected. The accuracy advantage of QR-STAR over QR can be seen in these two datasets as well, especially with true gene trees or low ILS.

FIG. 8. Rooting the ASTRAL species tree on biological simulations. Comparison between QR, QR-STAR, and optimal rooting on (left) 48-taxon avian simulated datasets, (right) 37-taxon mammalian simulated datasets, both from Mirarab et al. (2014a). The columns show the ILS level, and the rows show whether true or estimated gene trees (based on 500 bp sequences) were used. For both datasets, the number of replicates in each model condition is 20, but the model species tree is fixed across all replicates.

However, unlike the 201-taxon datasets where the accuracy of QR-STAR was close to optimal in most cases, here we see a bigger gap between the optimal rooting and QR-STAR in general. On the mammalian simulated datasets, this gap is small under the highest ILS condition (0.54 AD) and the medium ILS (0.33 AD) with true gene trees, but there is a larger gap in the other three conditions.

On the avian simulations, this gap is visible in all model conditions, but it becomes smaller as ILS level increases. For the avian simulations, the overall rooted species tree error is higher under high ILS conditions, suggesting that the error is dominated by species tree estimation error (Appendix Table C6 in Appendix C), and is consistent with the trend seen on the 201-taxon datasets.

However, this trend is reversed for the mammalian simulations, and the final rooted tree error is higher under the lowest ILS (0.18 AD) condition when rooting with QR-STAR (also see Appendix Table C5 in the Appendix C for ASTRAL RF rates).

5.4. Discussion of experimental results

Some trends seen here are as expected: for example, accuracy generally improves for both QR and QR-STAR with the number of genes, and using true gene trees produces better accuracy than using estimated gene trees. These trends can be explained by noting that more data and better-quality data improve accuracy. On the other hand, we also see that the combination of low ILS and deep speciation (toward the root) makes for easier conditions for both QR and QR-STAR, whereas low ILS and recent speciation (toward the leaves) makes for more challenging conditions for QR and QR-STAR; it is not clear why this is true.

An interesting trend seen in our experimental study is that rooting with QR and QR-STAR is more accurate under higher levels of discordance due to ILS, and becomes less accurate as the ILS level decreases. An explanation for this is that for a fixed number of gene trees, with less discordance due to ILS, it is likely that many gene trees that have low probability of appearing will not appear in the input, or will appear with very low frequencies, thus leading to higher error in the estimated probability distribution on quintet trees.

This will increase error in the rooting performed by QR and QR-STAR. Further, when enough gene trees fail to appear in the distribution, some estimates of quintet probabilities would become zero, and it may not be possible to differentiate some of the rooted quintets using the inequalities and invariants derived from the ADR theory.

In the extreme case where there is no discordance due to ILS (and so all true gene trees are identical to the species tree), there will be only one quintet gene tree with non-zero probability: when this happens, the identifiability theorem in Allman et al. (2011) would not hold and it becomes impossible to find the root. In contrast, the accuracy of ASTRAL and other species tree estimation methods decreases under higher levels of ILS (Mirarab and Warnow, 2015; Mirarab et al., 2014b; Molloy and Warnow, 2018).

Inference of the rooted species tree depends on both accurate estimation of the unrooted species tree topology as well as correct rooting of that tree. However, the level of ILS has a very different impact on these two steps. In most cases in our study, the overall error was dominated by species tree estimation error, and hence increased as the ILS level increased.

However, we saw a different trend on the mammalian dataset, in which the species tree estimation error was very low in the lowest ILS condition, but the rooting error was high such that the overall error was dominated by rooting error. In general, for the purposes of estimating a rooted species tree using this approach, moderate levels of ILS may make for a better overall outcome than very low or very high levels of ILS.

Another important trend is that QR-STAR is nearly always at least as accurate as QR, and it is more accurate under most conditions. In general, there is a clear advantage to QR-STAR over QR for the low ILS condition that holds across the different conditions (varying number of genes, using true or estimated gene trees, and rooting true or estimated species trees), and this advantage is also seen for the moderate ILS condition when the number of genes is small.

In contrast, for the high ILS condition, there is typically no or very little difference between the two methods, and in some cases QR can be somewhat more accurate. We also note that QR-STAR's advantage over QR is largest when using true gene trees, even under high ILS (Fig. 6), which suggests that QR may be somewhat more robust to GTEE than QR-STAR. Thus, QR-STAR has a theoretical advantage over QR but not always an accuracy advantage.

For many conditions, we observed a small gap between optimal rooting and QR-STAR. For example, on the 201-taxon dataset in moderate or high ILS conditions (Experiment 2), there was a very small difference in rooting error between QR-STAR and the optimal rooting, even using estimated gene trees, suggesting that QR-STAR is doing very well in these conditions.

Under the low ILS conditions of the 201-taxon data, however, there is a larger gap between QR-STAR and optimal rooting, especially when using only a small to moderate number of estimated gene trees. We also saw a larger gap between QR-STAR and the optimal rooting in Experiment 3 where the model trees were based on the avian and mammalian datasets, although the gap was less under the high ILS conditions than for the low ILS conditions.

These differences indicate that there are conditions where improvements to QR-STAR for its empirical performance should be sought, especially when the ILS level in the data is low. There are at least two ways to improve empirical performance, without sacrificing statistical consistency—modifying the cost function and changing the quintet sampling strategy—and both of these should be explored in future work.

6. CONCLUSIONS

We have presented QR-STAR, a polynomial-time statistically consistent method for rooting species trees under the MSC model. QR-STAR is an extension to QR, a method for rooting species trees introduced in Tabatabaee et al. (2022). QR-STAR differs from QR in that it has an additional step for determining the topological shape of each unrooted quintet selected in the QR algorithm, and incorporates the knowledge of this shape in its cost function, alongside the invariants and inequalities previously used in QR.

We also showed that the statistical consistency for QR-STAR holds for a larger family of optimization problems based on cost functions and sampling methods, and that modifying the linear encoding to be based on short quintets enables QR-STAR to have polynomial sample complexity.

To the best of our knowledge, this is the first work that established the statistical consistency of any method for rooting species trees under a model that incorporates gene tree heterogeneity. It remains to be investigated whether other rooting methods can also be proven statistically consistent under models of gene evolution inside species trees, such as the MSC or models of GDL.

For example, STRIDE (Emms and Kelly, 2017) and DISCO+QR (Willson et al., 2023) are methods that have been developed for rooting species trees from gene family trees, where genes evolve under GDL; however, it is not known whether these methods are statistically consistent under any GDL model.

This study suggests several directions for future research. For example, we proved statistical consistency for one class of cost functions, which was a linear combination of the invariant, inequality, and shape penalty terms; however, cost functions in other forms could also be explored and proven statistically consistent.

Theorem 3 shows that the sample complexity of QR-STAR depends on both the length of the shortest branch and the longest path in the model tree. This suggests that having very short or very long branches

can both confound rooting under ILS, which is also suggested in previous studies (Alanzi and Degnan, 2017; Allman et al., 2011). This is unlike what is known for species tree estimation methods such as ASTRAL, where the sample complexity is only affected by the shortest branch of the model tree (Chan et al., 2022; Shekhar et al., 2017), and trees with long branches are easier to estimate.

Another theoretical direction is the construction of the rooted species tree directly from the unrooted gene trees. As explained in Remark 2, the proof of consistency of QR-STAR for 5-taxon trees does not depend upon the knowledge of the unrooted tree topology; this suggests that it is possible to estimate the rooted topology of the species tree in a statistically consistency manner *directly* from unrooted gene tree topologies. Future work could focus on developing statistically consistent methods for this problem, which is significantly harder than the problem of rooting a given tree.

There are also directions for improving empirical results. An important consideration in designing a good cost function is its empirical performance, as many cost functions can lead to statistical consistency but may not provide accurate estimations of the rooted tree in practice (Figs. 4 and 5).

One potential direction is to incorporate estimated branch lengths, whether of the gene trees or of the unrooted species tree, into the rooting procedure. These improvements can especially be useful for datasets with low levels of ILS, which create the most difficult conditions for QR-STAR and where there is a gap between the accuracy of QR-STAR and the optimal rooting.

Finally, the experiments in this study were limited to comparisons between QR, QR-STAR, and the optimal rooting of the ASTRAL species trees. In our prior study presenting QR (Tabatabaee et al., 2022), we showed that QR had good accuracy compared with many prior rooting methods. That study, however, was restricted to a small number of model conditions. Hence, future work should also include a comparison of QR-STAR to a larger number of rooting methods, including outgroup rooting, and under a wider range of model conditions.

CODE AND DATA AVAILABILITY

QR-STAR is available at https://github.com/ytabatabaee/Quintet-Rooting. The scripts and data used in this study are available at https://github.com/ytabatabaee/QR-STAR-paper.

ACKNOWLEDGMENTS

S.R. thanks Cécile Ané and her group for helpful discussions. Y.T. thanks Mohammed El-Kebir for helpful suggestions on an earlier version of this work.

AUTHORS' CONTRIBUTIONS

The authors all contributed to conceptualization, formal analysis, writing, review, and editing. Y.T.: software, visualization, investigation, methodology, data curation, and original draft preparation. S.R.: methodology, investigation, and funding acquisition. T.W.: investigation, project administration, supervision, funding acquisition, and resources.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

S.R. was supported by NSF grants DMS-1902892, DMS1916378, and DMS-2023239 (TRIPODS Phase II), as well as a Vilas Associates Award. T.W. was supported by the Grainger Foundation. Y.T. was supported in part by UIUC C.L. and Jane W-S. Liu Award.

REFERENCES

Alanzi AR, Degnan JH. Inferring rooted species trees from unrooted gene trees using approximate Bayesian computation. Mol Phylogenet Evol 2017;116:13–24.

Allman ES, Degnan JH, Rhodes JA. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 2011;62(6):833–862.

Bettisworth B, Stamatakis A. Root Digger: A root placement program for phylogenetic trees. BMC Bioinform 2021;22(1):225.

Chan Y-B, Li Q, Scornavacca C. The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference. J Math Biol 2022;85(3):1–22.

Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics 2014;30(23): 3317–3324.

Drummond AJ, Ho SYW, Phillips MJ, et al. Relaxed phylogenetics and dating with confidence. PLoS Biol 2006; 4(5):e88.

Emms DM, Kelly S. STRIDE: Species tree root inference from gene duplication events. Mol Biol Evol 2017;34(12): 3267–3278.

Erdős PL, Steel MA, Székely LA, et al. A few logs suffice to build (almost) all trees (I). Random Struct Algorithms 1999a;14(2):153–184.

Erdős PL, Steel MA, Székely L, et al. A few logs suffice to build (almost) all trees: Part II. Theor Comput Sci 1999b;221(1–2):77–118.

Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 1978;27(4): 401–410.

Graham SW, Olmstead RG, Barrett SC. Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. Mol Biol Evol 2002;19(10):1769–1781.

Hess PN, De Moraes Russo CA. An empirical test of the midpoint rooting method. Biol J Linn Soc Lond 2007; 92(4):669–674.

Holland B, Penny D, Hendy M. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—A simulation study. Syst Biol 2003;52(2):229–238.

Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution 1983;37:203–217.

Jarvis ED, Mirarab S, Aberer AJ, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 2014;346(6215):1320–1331.

Jun S-R, Leuze MR, Nookaew I, et al. Ebolavirus comparative genomics. FEMS Microbiol Rev 2015;39(5):764–778.

Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 2007;56(1):17–24.

Larget BR, Kotha SK, Dewey CN, et al. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 2010;26(22):2910–2911.

Li C, Matthes-Rosana KA, Garcia M, et al. Phylogenetics of Chondrichthyes and the problem of rooting phylogenies with distant outgroups. Mol Phylogenet Evol 2012;63(2):365–373.

Liu L, Yu L, Edwards SV. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 2010;10(1):1–18.

Liu L, Yu L, Pearl DK, et al. Estimating species phylogenies using coalescence times among sequences. Syst Biol 2009;58(5):468–477.

Maddison WP. Gene trees in species trees. Syst Biol 1997;46(3):523–536.

Maddison WP, Donoghue MJ, Maddison DR. Outgroup analysis and parsimony. Syst Biol 1984;33(1):83–103.

Mahbub M, Wahab Z, Reaz R, et al. wQFM: Highly accurate genome-scale species tree estimation from weighted quartets. Bioinformatics 2021;37(21):3734–3743.

Mai U, Sayyari E, Mirarab S. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. PLoS One 2017;12(8):e0182238.

Mallo D, de Oliveira Martins L, Posada D. SimPhy: Phylogenomic simulation of gene, locus, and species trees. Syst Biol 2016;65(2):334–344.

Mirarab S, Bayzid MS, Boussau B, et al. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 2014a;346(6215):1250463.

Mirarab S, Reaz R, Bayzid MS, et al. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 2014b;30(17):i541–i548.

Mirarab S, Warnow T. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 2015;31(12):i44–i52.

Mitzenmacher M, Upfal E. Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis. Cambridge University Press; 2017.

Molloy EK, Warnow T. To include or not to include: The impact of gene filtering on species tree estimation methods. Syst Biol 2018;67(2):285–303.
Mossel E, Roch S. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. IEEE/ACM Trans Comput Biol Bioinform 2008;7(1):166–171.
Naser-Khdour S, Quang Minh B, Lanfear R. Assessing confidence in root placement on phylogenies: An empirical study using nonreversible models for mammals. Syst Biol 2022;71(4):959–972.
Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol 2017;34(8):2101–2114.
Posada D. Phylogenomics for systematic biology. Syst Biol 2016;65(3):353–356.
Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 2010;5(3):e9490.
Renner SS, Grimm GW, Schneeweiss GM, et al. Rooting and dating maples (Acer) with an uncorrelated-rates molecular clock: Implications for North American/Asian disjunctions. Syst Biol 2008;57(5):795–808.
Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci 1981;53(1–2):131–147.
Roch S. Hands-on Introduction to Sequence-Length Requirements in Phylogenetics. In: Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret. (Warnow T. ed.) Springer, 2019; pp. 47–86.
Roch S, Nute M, Warnow T. Long-branch attraction in species tree estimation: Inconsistency of partitioned likelihood and topology-based summary methods. Syst Biol 2019;68(2):281–297.
Roch S, Steel M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol 2015;100:56–62.
Rosenberg NA. Counting coalescent histories. J Comput Biol 2007;14(3):360–377.
Shekhar S, Roch S, Mirarab S. Species tree estimation using ASTRAL: How many genes are enough? IEEE/ACM Trans Comput Biol Bioinform 2017;15(5):1738–1747.
Simmons MP, Gatesy J. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol Phylogenet Evol 2015;91:98–122.
Simmons MP, Springer MS, Gatesy J. Gene-tree misrooting drives conflicts in phylogenomic coalescent analyses of palaeognath birds. Mol Phylogenet Evol 2022;167:107344.
Skarp-de Haan C, Culebro A, Schott T, et al. Comparative genomics of unintrogressed *Campylobacter coli* clades 2 and 3. BMC Genomics 2014;15(1):1–14.
Song S, Liu L, Edwards SV, et al. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A 2012;109(37):14942–14947.
Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–1313.
Tabatabaee Y, Sarker K, Warnow T. Quintet Rooting: Rooting species trees under the multi-species coalescent model. Bioinformatics 2022;38(Suppl 1):i109–i117.
Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 1986;17(2):57–86.
Tian Y, Kubatko L. Rooting phylogenetic trees under the coalescent model using site pattern probabilities. BMC Evol Biol 2017;17(1):1–11.
Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol 2017;1(1):1–7.
Wade T, Rangel LT, Kundu S, et al. Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. PLoS One 2020;15(5):e0232950.
Warnow T, Moret BM, St. John K. Absolute convergence: True trees from short sequences. Proc Annu ACM-SIAM Symp Discrete Algorithms 2001;7:186–195.
Wascher M, Kubatko L. Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. Syst Biol 2021;70(1):33–48.
Willson J, Tabatabaee Y, Liu B, et al. DISCO+QR: Rooting species trees in the presence of GDL and ILS. Bioinform Adv 2023;3(1):vbad015.
Zhang C, Rabiee M, Sayyari E, et al. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform 2018;19(6):15–30.

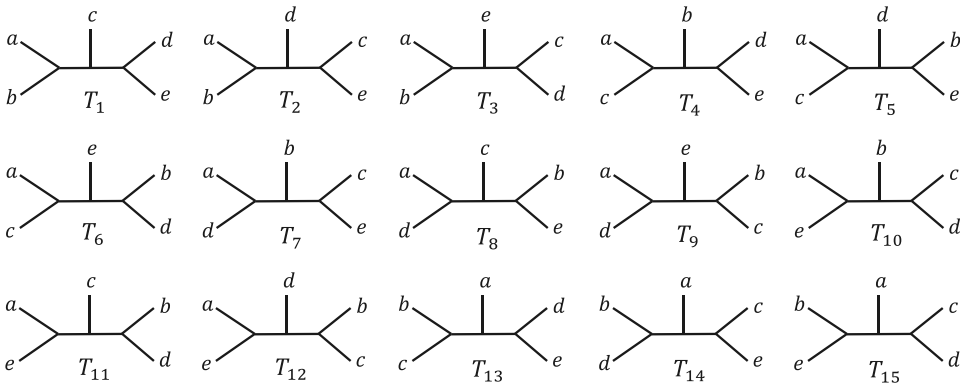
Address correspondence to:
Dr. Tandy Warnow
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois 61801
USA

E-mail: warnow@illinois.edu

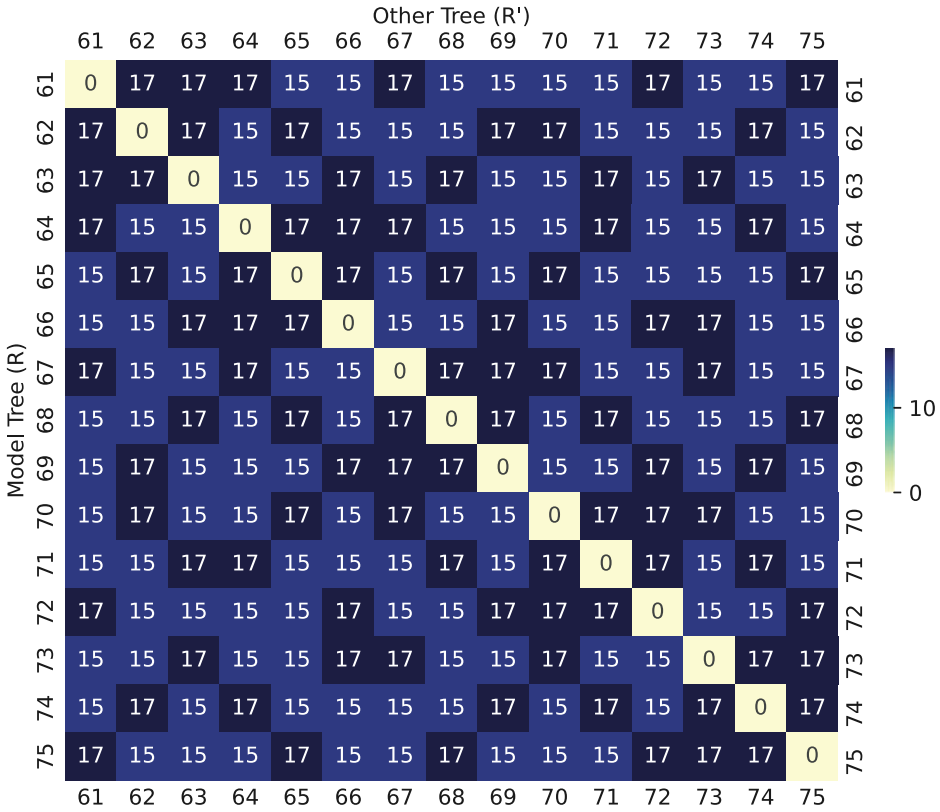
(*Appendix follows* →)

7. Appendix

A. ADDITIONAL FIGURES

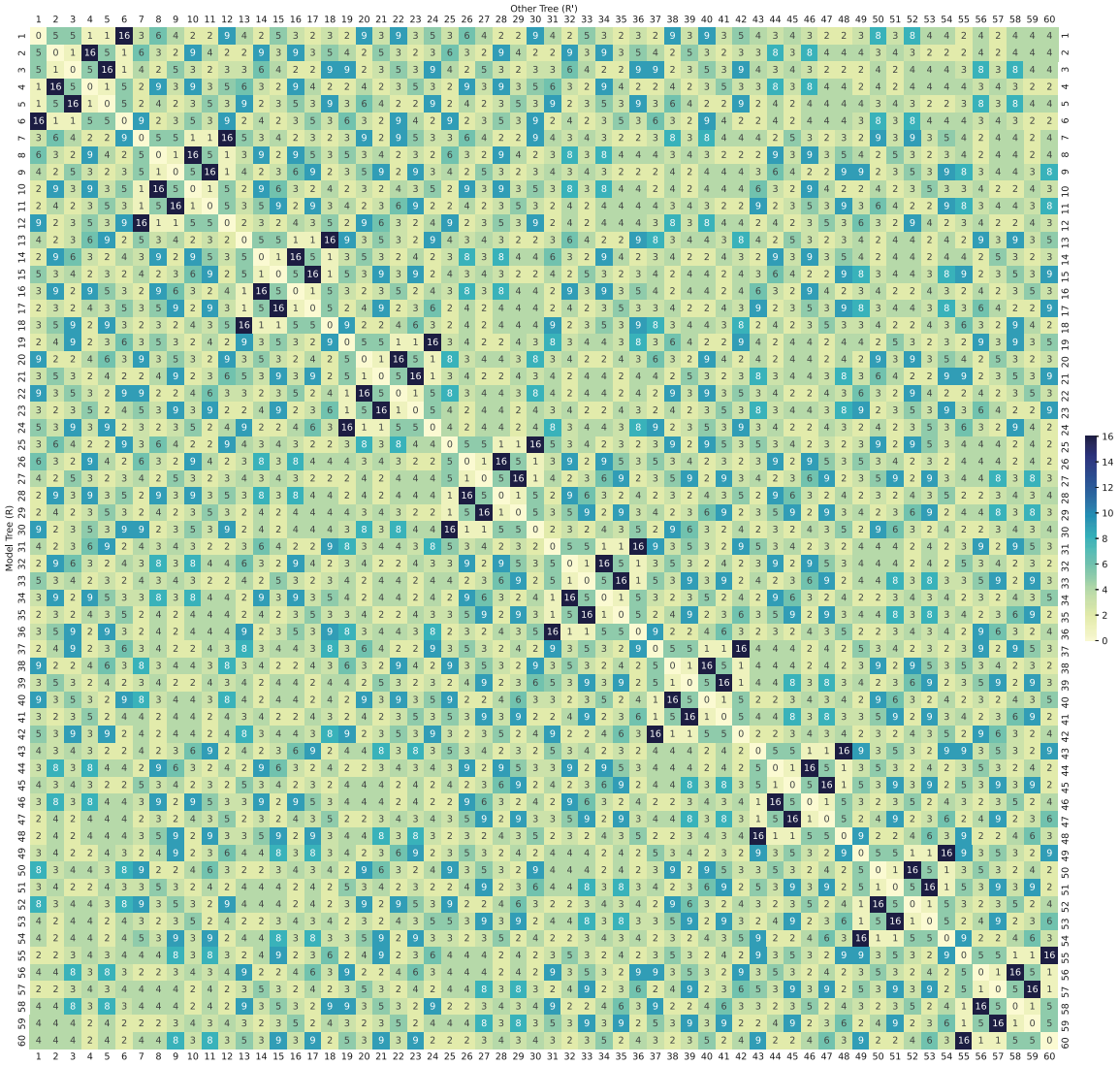


APPENDIX FIG. A1. Topologies of the 15 unrooted 5-taxon gene trees labeled according to Allman et al. (2011).



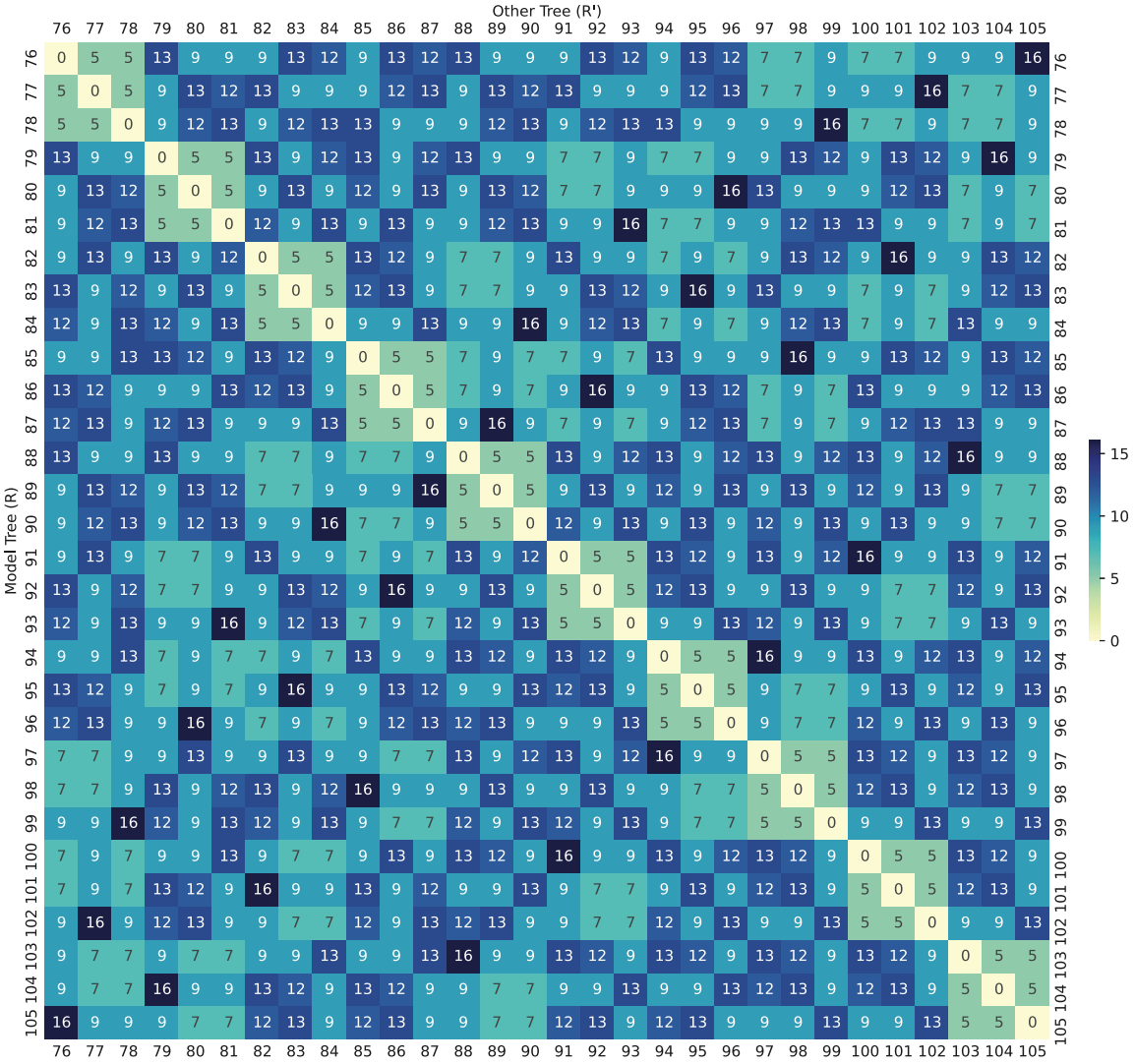
APPENDIX FIG. A2. Conflicts between 5-taxon pseudo-caterpillar trees. Heatmap showing the number of conflicting inequality penalty terms [the function $|V(R, R')|$] for pairs of pseudo-caterpillar 5-taxon rooted trees.

(*Appendix continues* →)



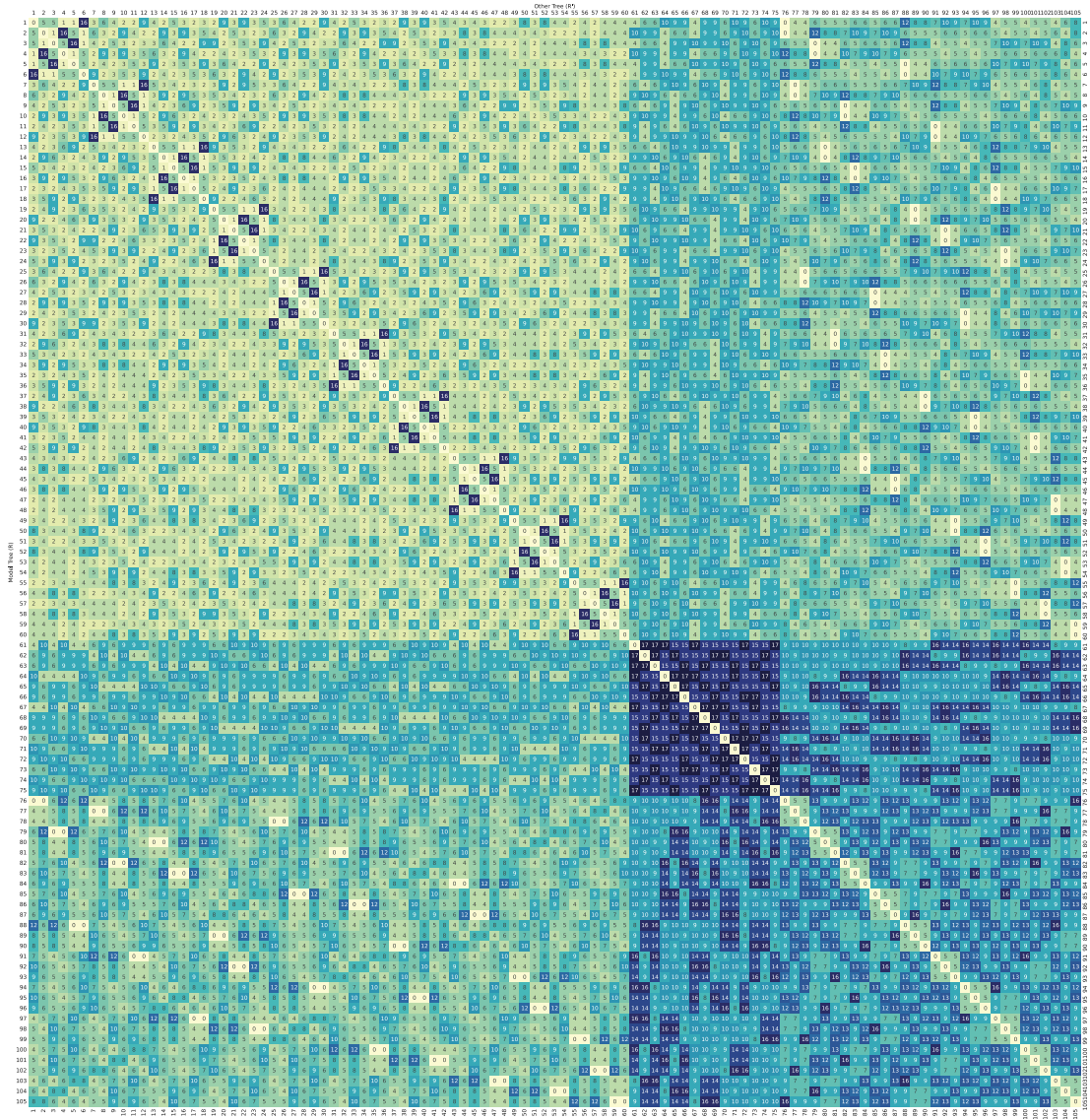
APPENDIX FIG. A3. Conflicts between 5-taxon caterpillar trees. Heatmap showing the number of conflicting inequality penalty terms [the function $|V(R, R')|$] for pairs of caterpillar 5-taxon rooted trees.

(*Appendix continues* →)



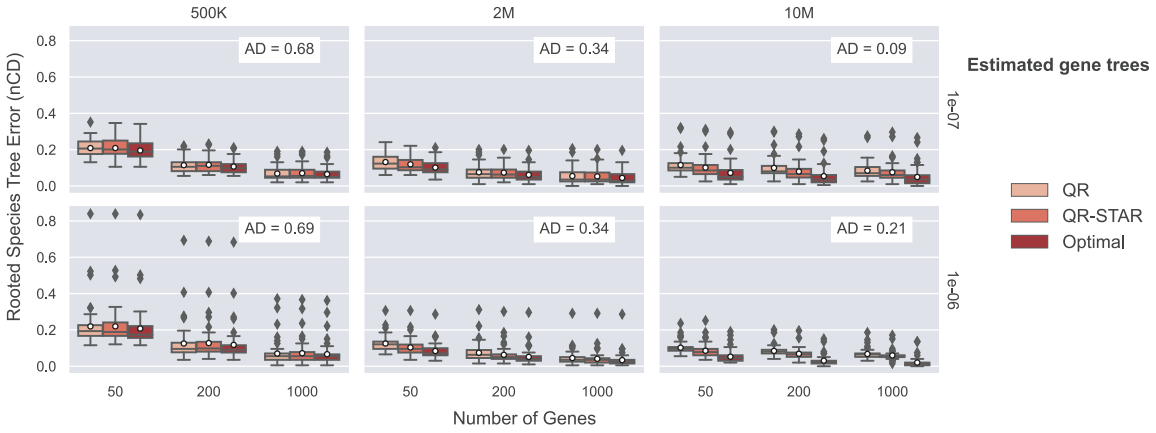
APPENDIX FIG. A4. Conflicts between 5-taxon balanced trees. Heatmap showing the number of conflicting inequality penalty terms [the function $|V(R, R')|$] for pairs of balanced 5-taxon rooted trees.

(*Appendix continues* →)

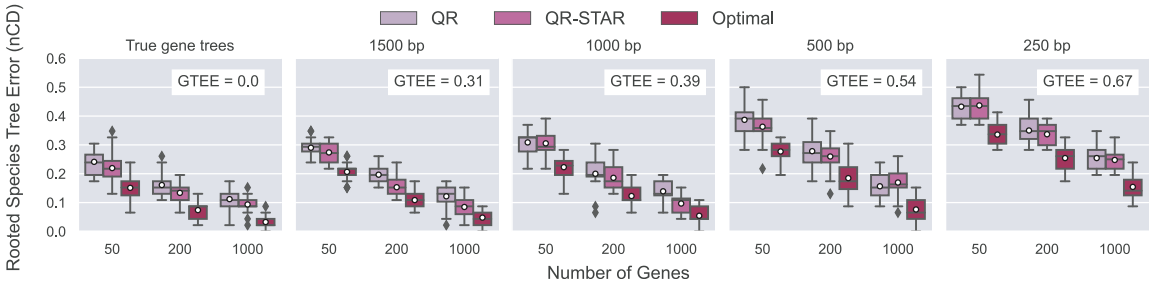


APPENDIX FIG. A5. Conflicts between all 5-taxon rooted trees. Heatmap showing the number of conflicting inequality penalty terms [the function $|V(R, R')|$] for all pairs of binary 5-taxon rooted trees.

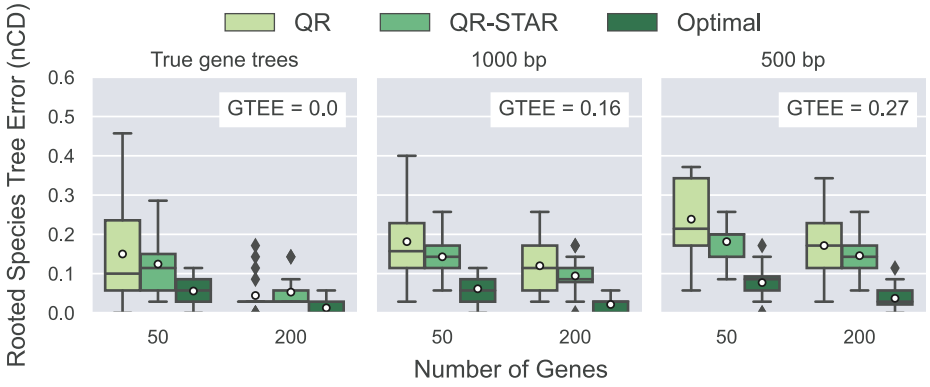
(*Appendix continues* →)



APPENDIX FIG. A6. Rooting the ASTRAL species tree on 201-taxon datasets with estimated gene trees. Full version of Figure 7 for estimated gene trees (including outliers). The results are shown across 50 replicates. The columns show tree height (500K for high ILS, 2M for moderate ILS, and 10M for low ILS) and the rows show speciation rate (1e-06 for recent speciation, 1e-07 for deep speciation). ILS, incomplete lineage sorting.



APPENDIX FIG. A7. Rooting the ASTRAL species tree on avian simulated dataset. Comparison between QR, QR-STAR, and optimal rooting on 48-taxon avian simulated dataset for 1X ILS model condition. The columns show whether true or estimated gene trees were used, and the sequence length used to produce the gene trees. The number of replicates in each model condition is 20.



APPENDIX FIG. A8. Rooting the ASTRAL species tree on mammalian simulated dataset. Comparison between QR, QR-STAR, and optimal rooting on 37-taxon mammalian simulated dataset for 1X ILS model condition. The columns show whether true or estimated gene trees were used, and the sequence length used to produce the gene trees. The number of replicates in each model condition is 20.

(*Appendix continues* →)

B. PROOF OF LEMMA 2 CONTINUED

Proof. *Caterpillar Trees*

- $u_a \in c_1, u_b \in c_4$

$$u_a - u_b = (1 - \tfrac{2}{3}X - \tfrac{2}{3}Y + \tfrac{1}{3}XY + \tfrac{1}{18}XY^3 + \tfrac{1}{90}XY^3Z^6) - (\tfrac{1}{3}X - \tfrac{1}{3}XY + \tfrac{1}{18}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$1 - X - \tfrac{2}{3}Y + \tfrac{2}{3}XY = (1-X)(1-\tfrac{2}{3}Y) > \tfrac{1}{3}(1-X)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{3}(1-X) \geq \tfrac{1}{3}(1-e^{-f(R)}) > h(R)$$

- $u_a \in c_2, u_b \in c_3$

$$u_a - u_b = (\tfrac{1}{3}Y - \tfrac{1}{6}XY - \tfrac{1}{18}XY^3 - \tfrac{2}{45}XY^3Z^6) - (\tfrac{1}{3}Y - \tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$\tfrac{1}{18}XY^3 - \tfrac{1}{18}XY^3Z^6 = \tfrac{1}{18}XY^3(1-Z^6) = \tfrac{1}{18}XY^3(1-Z)(1+Z+Z^2)(1+Z^3)$$
$$> \tfrac{1}{18}XY^3(1-Z)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{18}XY^3(1-Z) \geq \tfrac{1}{18}e^{-3g(R)}(1-e^{-f(R)}) > h(R)$$

where $XY^3 \geq e^{-3g(R)}$ follows from the fact that $XY = e^{-x}e^{-y} = e^{-(x+y)} \geq e^{-g(R)}$ as $x+y$ correspond to a (sub)-length of an internal path in R and hence $x+y \leq g(R)$ and $Y^2 \geq e^{-2g(R)}$ as $y \leq g(R)$.

- $u_a \in c_2, u_b \in c_5$

$$u_a - u_b = (\tfrac{1}{3}Y - \tfrac{1}{6}XY - \tfrac{1}{18}XY^3 - \tfrac{2}{45}XY^3Z^6) - (\tfrac{1}{6}XY - \tfrac{1}{18}XY^3 - \tfrac{2}{45}XY^3Z^6) =$$
$$\tfrac{1}{3}Y - \tfrac{1}{3}XY = \tfrac{1}{3}Y(1-X)$$
$$\Rightarrow u_a - u_b = \tfrac{1}{3}Y(1-X) \geq \tfrac{1}{3}e^{-g(R)}(1-e^{-f(R)}) > h(R)$$

- $u_a \in c_3, u_b \in c_6$

$$u_a - u_b = (\tfrac{1}{3}Y - \tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) - (\tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$\tfrac{1}{3}Y - \tfrac{1}{3}XY = \tfrac{1}{3}Y(1-X)$$
$$\Rightarrow u_a - u_b = \tfrac{1}{3}Y(1-X) \geq \tfrac{1}{3}e^{-g(R)}(1-e^{-f(R)}) > h(R)$$

- $u_a \in c_4, u_b \in c_6$

$$u_a - u_b = (\tfrac{1}{3}X - \tfrac{1}{3}XY + \tfrac{1}{18}XY^3 + \tfrac{1}{90}XY^3Z^6) - (\tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$\tfrac{1}{3}X - \tfrac{1}{2}XY + \tfrac{1}{6}XY^3 = \tfrac{1}{6}X(2-3Y+Y^3) = \tfrac{1}{6}X(1-Y)^2(Y+2) > \tfrac{1}{3}X(1-Y)^2$$
$$\Rightarrow u_a - u_b > \tfrac{1}{3}X(1-Y)^2 \geq \tfrac{1}{3}e^{-g(R)}(1-e^{-f(R)})^2 > h(R)$$

- $u_a \in c_5, u_b \in c_6$

$$u_a - u_b = (\tfrac{1}{6}XY - \tfrac{1}{18}XY^3 - \tfrac{2}{45}XY^3Z^6) - (\tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$\tfrac{1}{18}XY^3 - \tfrac{1}{18}XY^3Z^6 = \tfrac{1}{18}XY^3(1-Z^6) = \tfrac{1}{18}XY^3(1-Z)(1+Z+Z^2)(1+Z^3)$$
$$> \tfrac{1}{18}XY^3(1-Z)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{18}XY^3(1-Z) \geq \tfrac{1}{18}e^{-3g(R)}(1-e^{-f(R)}) > h(R)$$

(*Appendix continues* →)

- $u_a \in c_6, u_b \in c_7$

$$u_a - u_b = (\tfrac{1}{6}XY - \tfrac{1}{9}XY^3 + \tfrac{1}{90}XY^3Z^6) - (\tfrac{1}{18}XY^3 + \tfrac{1}{90}XY^3Z^6) =$$
$$\tfrac{1}{6}XY - \tfrac{1}{6}XY^3 = \tfrac{1}{6}XY(1-Y^2) = \tfrac{1}{6}XY(1-Y)(1+Y) > \tfrac{1}{6}XY(1-Y)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{6}XY(1-Y) \geq \tfrac{1}{6}e^{-g(R)}(1-e^{-f(R)}) > h(R)$$

Balanced Trees

For a balanced model species tree $R = (((a, b) : x, c) : y, (d, e) : z)$, with $C_R = \{c_1 : \{u_1\}, c_2 : \{u_4, u_{13}\}, c_3 : \{u_2, u_3\}, c_4 : \{u_5, u_6, u_9, u_{12}\}, c_5 : \{u_7, u_8, u_{10}, u_{11}, u_{14}, u_{15}\}\}$, we have $c_1 > c_2, c_3 > c_4 > c_5$. Therefore,

- $u_a \in c_1, u_b \in c_3$

$$u_a - u_b = (1 - \tfrac{2}{3}X - \tfrac{2}{3}YZ + \tfrac{1}{3}XYZ + \tfrac{1}{15}XY^3Z) - (\tfrac{1}{3}YZ - \tfrac{1}{6}XYZ - \tfrac{1}{10}XY^3Z) =$$
$$1 - \tfrac{2}{3}X - YZ + \tfrac{1}{2}XYZ + \tfrac{1}{6}XY^3Z = 1 - YZ - \tfrac{1}{6}X(4 - 3YZ - Y^3Z) >$$
$$1 - YZ - \tfrac{1}{6}(4 - 3YZ - Y^3Z) = \tfrac{1}{3} - \tfrac{1}{6}YZ(3 - Y^2) >$$
$$\tfrac{1}{3} - \tfrac{1}{6}Y(3 - Y^2) = \tfrac{1}{6}(2 - 3Y + Y^3) = \tfrac{1}{6}(1-Y)^2(2+Y) > \tfrac{1}{3}(1-Y)^2$$
$$\Rightarrow u_a - u_b > \tfrac{1}{3}(1-Y)^2 \geq \tfrac{1}{3}(1 - e^{-f(R)})^2 > h(R)$$

- $u_a \in c_1, u_b \in c_2$

$$u_a - u_b = (1 - \tfrac{2}{3}X - \tfrac{2}{3}YZ + \tfrac{1}{3}XYZ + \tfrac{1}{15}XY^3Z) - (\tfrac{1}{3}X - \tfrac{1}{3}XYZ + \tfrac{1}{15}XY^3Z) =$$
$$1 - X - \tfrac{2}{3}YZ + \tfrac{2}{3}XYZ = (1-X)(1 - \tfrac{2}{3}YZ) > \tfrac{1}{3}(1-X)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{3}(1-X) \geq \tfrac{1}{3}(1 - e^{-f(R)}) > h(R)$$

- $u_a \in c_3, u_b \in c_4$

$$u_a - u_b = (\tfrac{1}{3}YZ - \tfrac{1}{6}XYZ - \tfrac{1}{10}XY^3Z) - (\tfrac{1}{6}XYZ - \tfrac{1}{10}XY^3Z) =$$
$$\tfrac{1}{3}YZ - \tfrac{1}{3}XYZ = \tfrac{1}{3}YZ(1-X)$$
$$\Rightarrow u_a - u_b = \tfrac{1}{3}YZ(1-X) \geq \tfrac{1}{3}e^{-g(R)}(1 - e^{-f(R)}) > h(R)$$

- $u_a \in c_2, u_b \in c_4$

$$u_a - u_b = (\tfrac{1}{3}X - \tfrac{1}{3}XYZ + \tfrac{1}{15}XY^3Z) - (\tfrac{1}{6}XYZ - \tfrac{1}{10}XY^3Z) =$$
$$\tfrac{1}{3}X - \tfrac{1}{2}XYZ + \tfrac{1}{6}XY^3Z = \tfrac{1}{6}X(2 - 3YZ + Y^3Z) > \tfrac{1}{6}X(2Z - 3YZ + Y^3Z) =$$
$$\tfrac{1}{6}XZ(2 - 3Y + Y^3) = \tfrac{1}{6}XZ(2+Y)(1-Y)^2 > \tfrac{1}{3}XZ(1-Y)^2$$
$$\Rightarrow u_a - u_b > \tfrac{1}{3}XZ(1-Y)^2 \geq \tfrac{1}{3}e^{-g(R)}(1 - e^{-f(R)})^2 > h(R)$$

- $u_a \in c_4, u_b \in c_5$

$$u_a - u_b = (\tfrac{1}{6}XYZ - \tfrac{1}{10}XY^3Z) - (\tfrac{1}{15}XY^3Z) =$$
$$\tfrac{1}{6}XYZ - \tfrac{1}{6}XY^3Z = \tfrac{1}{6}(XYZ)(1-Y^2) = \tfrac{1}{6}(XYZ)(1-Y)(1+Y) > \tfrac{1}{6}(XYZ)(1-Y)$$
$$\Rightarrow u_a - u_b > \tfrac{1}{6}(XYZ)(1-Y) \geq \tfrac{1}{6}e^{-g(R)}(1 - e^{-f(R)}) > h(R)$$

where $XYZ \geq e^{-g(R)}$ follows from $x + y + z \leq g(R)$.

(*Appendix continues* →)

Pseudo-Caterpillar Trees

For a pseudo-caterpillar model species tree $R=(((a,b):x,(d,e):y):z,c)$ with $C_R=\{c_1:\{u_1\}, c_2:\{u_4,u_{13}\}, c_3:\{u_2,u_3\}, c_4:\{u_8,u_{11}\}, c_5:\{u_5,u_6,u_7,u_9,u_{10},u_{12},u_{14}\}\}$, we have $c_1 > c_2, c_3, c_4 > c_5$. Therefore,

- $u_a \in c_1, u_b \in c_2$

$$u_a-u_b=(1-\tfrac{2}{3}X-\tfrac{2}{3}Y+\tfrac{4}{9}XY-\tfrac{2}{45}XYZ^6)-(\tfrac{1}{3}X-\tfrac{5}{18}XY+\tfrac{1}{90}XYZ^6)=$$
$$1-X-\tfrac{2}{3}Y+\tfrac{13}{18}XY-\tfrac{1}{18}XYZ^6=1-X-\tfrac{2}{3}Y+\tfrac{2}{3}XY+\tfrac{1}{18}XY-\tfrac{1}{18}XYZ^6$$
$$=(1-X)(1-\tfrac{2}{3}Y)+\tfrac{1}{18}XY(1-Z^6)>\tfrac{1}{18}XY(1-Z)(1+Z+Z^2)(1+Z^3)>\tfrac{1}{18}XY(1-Z)$$
$$\Rightarrow u_a-u_b>\tfrac{1}{18}XY(1-Z)\geq\tfrac{1}{18}e^{-g(R)}(1-e^{-f(R)})>h(R)$$

- $u_a \in c_1, u_b \in c_3$

$$u_a-u_b=(1-\tfrac{2}{3}X-\tfrac{2}{3}Y+\tfrac{4}{9}XY-\tfrac{2}{45}XYZ^6)-(\tfrac{1}{3}Y-\tfrac{5}{18}XY+\tfrac{1}{90}XYZ^6)=$$
$$1-\tfrac{2}{3}X-Y+\tfrac{13}{18}XY-\tfrac{1}{18}XYZ^6=1-\tfrac{2}{3}X-Y+\tfrac{2}{3}XY+\tfrac{1}{18}XY-\tfrac{1}{18}XYZ^6$$
$$=(1-Y)(1-\tfrac{2}{3}X)+\tfrac{1}{18}XY(1-Z^6)>\tfrac{1}{18}XY(1-Z)(1+Z+Z^2)(1+Z^3)>\tfrac{1}{18}XY(1-Z)$$
$$\Rightarrow u_a-u_b>\tfrac{1}{18}XY(1-Z)\geq\tfrac{1}{18}e^{-g(R)}(1-e^{-f(R)})>h(R)$$

- $u_a \in c_1, u_b \in c_4$

$$u_a-u_b=(1-\tfrac{2}{3}X-\tfrac{2}{3}Y+\tfrac{4}{9}XY-\tfrac{2}{45}XYZ^6)-(\tfrac{1}{9}XY-\tfrac{2}{45}XYZ^6)=$$
$$1-\tfrac{2}{3}X-\tfrac{2}{3}Y+\tfrac{1}{3}XY=(1-X)(1-Y)-\tfrac{2}{3}XY+\tfrac{1}{3}X+\tfrac{1}{3}Y=(1-X)(1-Y)+\tfrac{1}{3}(X+Y-2XY)$$
$$>(1-X)(1-Y)+\tfrac{1}{3}(X^2+Y^2-2XY)=(1-X)(1-Y)+\tfrac{1}{3}(X-Y)^2>(1-X)(1-Y)$$
$$\Rightarrow u_a-u_b>(1-X)(1-Y)\geq(1-e^{-f(R)})^2>h(R)$$

- $u_a \in c_2, u_b \in c_5$

$$u_a-u_b=(\tfrac{1}{3}X-\tfrac{5}{18}XY+\tfrac{1}{90}XYZ^6)-(\tfrac{1}{18}XY+\tfrac{1}{90}XYZ^6)=$$
$$\tfrac{1}{3}X-\tfrac{1}{3}XY=\tfrac{1}{3}X(1-Y)$$
$$\Rightarrow u_a-u_b=\tfrac{1}{3}X(1-Y)\geq\tfrac{1}{3}e^{-g(R)}(1-e^{-f(R)})>h(R)$$

- $u_a \in c_3, u_b \in c_5$

$$u_a-u_b=(\tfrac{1}{3}Y-\tfrac{5}{18}XY+\tfrac{1}{90}XYZ^6)-(\tfrac{1}{18}XY+\tfrac{1}{90}XYZ^6)=$$
$$\tfrac{1}{3}Y-\tfrac{1}{3}XY=\tfrac{1}{3}Y(1-X)$$
$$\Rightarrow u_a-u_b=\tfrac{1}{3}Y(1-X)\geq\tfrac{1}{3}e^{-g(R)}(1-e^{-f(R)})>h(R)$$

- $u_a \in c_4, u_b \in c_5$

$$u_a-u_b=(\tfrac{1}{9}XY-\tfrac{2}{45}XYZ^6)-(\tfrac{1}{18}XY+\tfrac{1}{90}XYZ^6)=$$
$$\tfrac{1}{18}XY-\tfrac{1}{18}XYZ^6=\tfrac{1}{18}XY(1-Z^6)=\tfrac{1}{18}XY(1-Z)(1+Z+Z^2)(1+Z^3)>\tfrac{1}{18}XY(1-Z)$$
$$\Rightarrow u_a-u_b>\tfrac{1}{18}XY(1-Z)\geq\tfrac{1}{18}e^{-g(R)}(1-e^{-f(R)})>h(R)$$

□

C. DETAILS OF THE EXPERIMENTAL STUDY

C.1. Details of the Datasets

For the training dataset (101-taxon dataset from Zhang et al., 2018), the incomplete lineage sorting (ILS) level (measured using the average distance between the true gene trees and the model species tree, or AD) for most replicates ranged from 0.3 to 0.6 with an average of 0.46. The mean gene tree estimation error (GTEE) values for the four sequence lengths were 0.23, 0.31, 0.42, and 0.55 for the 1600, 800, 400, and 200 bp sequences, respectively. The speciation rate for this dataset was 1e-07.

(*Appendix continues* →)

Appendix Table C1. Statistics for the 201-Taxon Simulated Datasets

Speciation rate	*Tree height*	*ILS level (AD)*	*GTEE*
1E-06	500K	0.69	0.49
1E-06	2M	0.34	0.28
1E-06	10M	0.21	0.22
1E-07	500K	0.68	0.46
1E-07	2M	0.34	0.34
1E-07	10M	0.09	0.29

ILS level (measured using the average distance between the true gene trees and the model species tree, or AD) and average GTEE of the estimated gene trees for the six model conditions of the 201-taxon datasets with 1000 gene trees. The two speciation rates used in this dataset indicate whether speciation happened close to the leaves (i.e., recent speciation for 1e-06) or close to the root (i.e., deep speciation for 1e-07). The shorter tree height (500K) indicates shorter branches and therefore higher levels of ILS.

AD, average distance; GTEE, gene tree estimation error; ILS, incomplete lineage sorting.

Appendix Table C2. Statistics for the 48-Taxon Avian Simulated Dataset

Model condition	*Sequence length (bp)*	*ILS level (AD)*	*GTEE*
0.5X	500	0.59	0.59
1X	250	0.47	0.67
	500	0.47	0.54
	1000	0.47	0.39
	1500	0.47	0.31
2X	500	0.35	0.57

ILS level (measured using the average distance between the true gene trees and the model species tree, or AD) and average GTEE of the estimated gene trees for different model conditions of the avian datasets with 1000 gene trees.

Appendix Table C3. Statistics for the 37-Taxon Mammalian Simulated Dataset

Model condition	*Sequence length (bp)*	*ILS level (AD)*	*GTEE*
0.5X	500	0.54	0.27
1X	500	0.33	0.27
	1000	0.33	0.16
2X	500	0.18	0.27

ILS level (measured using the average distance between the true gene trees and the model species tree, or AD) and average GTEE of the estimated gene trees for different model conditions of the mammalian datasets with 200 gene trees.

Appendix Table C4. ASTRAL RF Error Rates on the 201-Taxon Datasets

Speciation rate	*Tree height*	*Number of genes*	*RF (true gene trees)*	*RF (est gene trees)*
1E-06	500K	1000	0.028 ± 0.012	0.067 ± 0.065
		200	0.065 ± 0.021	0.118 ± 0.103
		50	0.137 ± 0.030	0.208 ± 0.118
1E-06	2M	1000	0.010 ± 0.008	0.034 ± 0.041
		200	0.023 ± 0.011	0.052 ± 0.045
		50	0.045 ± 0.015	0.084 ± 0.042
1E-06	10M	1000	0.006 ± 0.006	0.020 ± 0.028
		200	0.012 ± 0.010	0.031 ± 0.029
		50	0.027 ± 0.014	0.053 ± 0.037
1E-07	500K	1000	0.028 ± 0.011	0.065 ± 0.036
		200	0.066 ± 0.018	0.107 ± 0.038
		50	0.143 ± 0.024	0.197 ± 0.046
1E-07	2M	1000	0.010 ± 0.007	0.045 ± 0.040
		200	0.024 ± 0.011	0.062 ± 0.038
		50	0.049 ± 0.017	0.102 ± 0.041
1E-07	10M	1000	0.002 ± 0.003	0.050 ± 0.054
		200	0.004 ± 0.004	0.055 ± 0.055
		50	0.011 ± 0.007	0.072 ± 0.061

Species tree estimation error in terms of normalized RF distance for ASTRAL trees given true and estimated gene trees for the 201-taxon datasets. The values show mean and standard deviation of RF errors across 50 replicates for each model condition.

RF, Robinson-Foulds.

(*Appendix continues* →)

Appendix Table C5. ASTRAL RF Error Rates on the Mammalian Simulated Dataset

Model condition	*Sequence length (bp)*	*Number of genes*	*RF rate*
0.5X	NA (true gene trees)	200	0.035 ± 0.024
		50	0.094 ± 0.044
	500	200	0.049 ± 0.025
		50	0.107 ± 0.049
1X	NA (true gene trees)	200	0.013 ± 0.017
		50	0.057 ± 0.039
	1000	200	0.022 ± 0.021
		50	0.063 ± 0.031
	500	200	0.038 ± 0.032
		50	0.079 ± 0.044
2X	NA (true gene trees)	200	0.003 ± 0.009
		50	0.018 ± 0.020
	500	200	0.025 ± 0.021
		50	0.066 ± 0.031

Species tree estimation error in terms of normalized RF distance for ASTRAL trees on the 37-taxon mammalian simulated dataset. The values show mean and standard deviation of RF errors across 20 replicates for each model condition. For true gene trees, sequence length is not applicable and hence shown as NA.

Appendix Table C6. ASTRAL RF Error Rates on the Avian Simulated Dataset

Model condition	*Sequence length (bp)*	*Number of genes*	*RF rate*
0.5X	NA (true gene trees)	1000	0.048 ± 0.030
		200	0.151 ± 0.029
		50	0.242 ± 0.040
	500	1000	0.131 ± 0.040
		200	0.243 ± 0.042
		50	0.332 ± 0.053
1X	NA (true gene trees)	1000	0.033 ± 0.023
		200	0.076 ± 0.030
		50	0.154 ± 0.041
	1500	1000	0.049 ± 0.028
		200	0.111 ± 0.031
		50	0.211 ± 0.032
	1000	1000	0.056 ± 0.029
		200	0.126 ± 0.032
		50	0.228 ± 0.042
	500	1000	0.078 ± 0.038
		200	0.189 ± 0.053
		50	0.283 ± 0.035
	250	1000	0.158 ± 0.045
		200	0.260 ± 0.043
		50	0.343 ± 0.040
2X	NA (true gene trees)	1000	0.021 ± 0.016
		200	0.046 ± 0.025
		50	0.104 ± 0.039
	500	1000	0.043 ± 0.019
		200	0.126 ± 0.043
		50	0.226 ± 0.047

Species tree estimation error in terms of normalized RF distance for ASTRAL trees on the 48-taxon avian simulated dataset. The values show mean and standard deviation of RF errors across 20 replicates for each model condition. For true gene trees, sequence length is not applicable and hence shown as NA.

(*Appendix continues* →)

C.2. Software Commands and Version Numbers

- ASTRAL: We used ASTRAL (v5.7.8) to estimate unrooted species trees, with the specified number of true or estimated gene trees for each model condition. ASTRAL is available at https://github.com/smirarab/ASTRAL. We used the following command:
  `java -jar astral.5.7.8.jar -i <input-genes.tre> -o <output.tre>`
- Quintet Rooting (QR): We used QR (v1.2.4) to root unrooted species trees. QR is available at https://github.com/ytabatabaee/Quintet_Rooting. We used the following command:
  `python3 quintet_rooting.py -t <input-tree.tre> -g <input-genes.tre> -o <output.tre> -sm le`
  The -LE option specifies the quintet sampling method as ''linear encoding.''
- QR-STAR: QR-STAR is available as part of the QR software package, and we ran it using the following command in the comparisons to QR, that sets $C = 1E-02$ and $\frac{\alpha_{max}}{\beta_{min}} = 0$:
  `python3 quintet_rooting.py -t <input-tree.tre> -g <input-genes.tre> -o <output.tre> -sm le -c STAR -abratio 0 -coef 0.01`
- Optimal rooting: We used the script available at https://github.com/ytabatabaee/QR-STAR-paper/blob/main/scripts/optimal_rooting.py to find a rooting of an estimated tree that has minimum normalized clade distance (nCD) error with respect to a reference rooted tree.
  `python3 optimal_rooting.py -r <reference-rooted.tre> -t <unrooted.tre> -o <output.tre>`
- GTEE, AD and Rooting (nCD) Error: GTEE and average distance between model species trees and true gene was computed using a script for computing normalized Robinson-Foulds (RF) distance written by Erin. K. Molloy available at https://github.com/ekmolloy/njmerge/blob/master/python/compare_trees.py
  Rooting error was measured in terms of average nCD using the script available at https://github.com/ytabatabaee/Quintet-Rooting/blob/main/scripts/clade_distance.py

D. COMPUTING THE OPTIMAL ROOTING SCORE

In Experiments 2 and 3, we reported the lowest possible normalized clade distance (nCD) rate achievable across all rootings of the estimated species tree; this was performed through an exhaustive search (rooting on all possible edges and computing the nCD rate). However, here we show that this best possible nCD has a close relationship to the missing branch (false negative or FN) rate of the unrooted estimated species tree with respect to the model species tree.

Let R be the true rooted species tree and let T denote its unrooted topology. Let $\hat{T}$ be an estimate of T. We are interested in rooting $\hat{T}$ to minimize the number of its missing clades with respect to R, and we will call this the missing clade number (not a rate). We define $FN(\hat{T}, T)$ to be the number of bipartitions in T that are not found in $\hat{T}$. We will show that the missing clade number for an optimal rooting of $\hat{T}$ with respect to R is either $FN(\hat{T}, T)$ or $FN(\hat{T}, T)+1$, depending on whether the bipartition at the root of R is present in $\hat{T}$ or not.

Lemma 8. Let R be the true rooted species tree with T the unrooted version. We draw R with root r having two children v_A and v_B. Let A be the clade below v_A and B be the clade below v_B. Let $\hat{T}$ be an estimate of T. If bipartition $A|B$ is present in $\hat{T}$, induced by edge e, then the optimal rooting of $\hat{T}$ that minimizes the number of missing clades in R is achieved by rooting on edge e, and results in $FN(\hat{T}, T)$ missing clades. If $A|B$ is not present in $\hat{T}$, then the optimal rooting of $\hat{T}$ that minimizes the number of clades in R missing from the rooted version of $\hat{T}$ results in $FN(\hat{T}, T)+1$ missing clades.

Proof. Assume $A|B$ is not present in $\hat{T}$. Take R and mark all edges of T that define bipartitions that do not appear in $\hat{T}$. Since the bipartition that is defined by the edge that the root bisects is not present in $\hat{T}$, those two edges incident to r [i.e., the edges (r, v_A) and (r, v_B)] are both marked. Make r a leaf, by attaching a new leaf that is adjacent to the original root location. Collapse all marked edges, to obtain an unresolved tree. Now refine this unresolved tree so that it induces $\hat{T}$ (and so produces $\hat{T}$ when the root leaf is removed). Since this tree contains the root, this can be seen as a rooted version of $\hat{T}$. Note that the missing clade number for this rooted tree is identical to $FN(\hat{T}, T)+1$. That is, the *two* edges that define the bipartition $A|B$ in T each contribute 1 to the missing clade number, and every other missing edge also contributes 1 to the missing clade number. The proof for when $A|B$ is present in $\hat{T}$ is similar and is omitted. □