Bayesian-Weighted Triplet and Quartet Methods for Species Tree Inference

Andrew Richards1 · Laura Kubatko1,2

Received: 13 October 2020 / Accepted: 3 June 2021 / Published online: 23 July 2021
© The Author(s), under exclusive licence to Society for Mathematical Biology 2021

Abstract
Inference of the evolutionary histories of species, commonly represented by a species tree, is complicated by the divergent evolutionary history of different parts of the genome. Different loci on the genome can have different histories from the underlying species tree (and each other) due to processes such as incomplete lineage sorting (ILS), gene duplication and loss, and horizontal gene transfer. The multispecies coalescent is a commonly used model for performing inference on species and gene trees in the presence of ILS. This paper introduces Lily-T and Lily-Q, two new methods for species tree inference under the multispecies coalescent. We then compare them to two frequently used methods, SVDQuartets and ASTRAL, using simulated and empirical data. Both methods generally showed improvement over SVDQuartets, and Lily-Q was superior to Lily-T for most simulation settings. The comparison to ASTRAL was more mixed—Lily-Q tended to be better than ASTRAL when the length of recombination-free loci was short, when the coalescent population parameter $\theta$ was small, or when the internal branch lengths were longer.

Keywords Species phylogeny · Multispecies coalescent · Quartet assembly · Phylogenetic inference

1 Introduction
The phylogenetic inference problem is concerned with using data, including but not limited to DNA sequences, to understand the evolutionary history of a collection of
species. Consider the collection of mammals shown in Fig. 1. We are concerned with three aspects shown in the figure. First, is the unlabeled topology correct? In other words, for each node, how many descendants are there on the left and right branches? Second, is the labeled topology correct, or do the labels need to be permuted? Third, when do the speciation events occur?

We usually begin with the assumption that when a speciation event occurs each ancestral species divides into exactly two daughter species. Thus, the evolutionary history can be represented as a binary tree known as a species tree.

**Definition 1** A *species tree* is an acyclic graph \( S = (V(S), E(S), \tau) \) where \( V(S) \) is the vertex set of \( S \), \( E(S) \) is the edge set of \( S \), and \( \tau \) is a set of node times.

Biologically, internal nodes represent speciation events while the leaves represent extant species. We call this leaf set \( L_S \). The leaves will be represented by lower case letters \( a, b, c \), etc., and internal nodes by letters later in the alphabet. Internal branches represent a continuum of ancestral species. If we know the common ancestor of all the species under consideration, then the tree is *rooted*, and the tree becomes a directed graph from the root outward. If the root is unknown, then we only know the direction of the branches that connect to external nodes. The *degree* of a node is the number of other nodes a node is connected to. For a species tree, the leaves are of degree one, and due to the binary tree assumption, all internal nodes are of degree three (except the root, if known, which is of degree two).

We assume a *strict molecular clock*, so the mutation rate is constant over time (or more precisely, if \( \lambda \) represents the mutation rate there is a common \( \lambda(t) \) for all branches at time \( t \)). Therefore, the rooted tree is *ultrametric*, meaning that each leaf will be equidistant from the root. Note that because the leaves are all at time \( t = 0 \) (the present day) and the tree is ultrametric, our parameterization using node times is equivalent to using branch lengths to define distances on the tree.
Fig. 2 The fifteen rooted quartets. We can label these $S^{(4)}_1, S^{(4)}_2, \ldots S^{(4)}_{15}$. If we are unconcerned with the labels, the first column is $S^{(4)}_{u1}$ and the other columns $S^{(4)}_{u2}$. Note that each row corresponds to one of the three unrooted quartets.

Let $S^{(n)}$ refer to the set of all possible trees with $n$ taxa. The superscript is in parentheses to highlight that it refers to the number of taxa rather than as an exponent. Then, for example, $S^{(4)}_1$ and $S^{(4)}_2$ can indicate the first and second 4-taxon species trees from Fig. 2. If the labeling is unimportant, we can also use $S^{(4)}_{u1}$ and $S^{(4)}_{u2}$ to indicate the two unlabeled 4-taxon topologies.

Species tree inference is complicated by the possibility for divergence between the evolutionary history of species and individual elements of their genomes. Causes for this divergence include incomplete lineage sorting (ILS), gene duplication and loss (GDL) and horizontal gene transfer (HGT). ILS is commonly modeled by the coalescent process (Kingman 1982). The history of individual loci on the genome is represented by a gene tree.

**Definition 2** A gene tree is an acyclic graph $G = (V(G), E(G), t)$ where $V(G)$ is the vertex set of $G$, $E(G)$ is the edge set of $G$, and $t$ is a set of node times. These node times are subject to the constraint that the coalescent events in question must occur prior to the divergence time of the species in question.

The gene tree is embedded within the species tree (see Fig. 3) and usually both trees have the same leaf set. Exceptions can occur if the gene has not been sampled for all species under consideration or if there are multiple sampled individuals per species. Thus, unless otherwise noted, we will drop the subscript and just refer to the set of species under consideration as $L$. When we need to distinguish the leaves of $G$ from the leaves of $S$ we use capital letters $A, B, C, \ldots$. Internal nodes represent coalescent events, which identify the most recent common ancestor of two gene lineages.

It is worth noting here that our definition of a gene, as is common in the phylogenetic literature, refers to a recombination-free region of the genome. Thus, the
Fig. 3 Species trees with four taxa under the coalescent process. The green lines show example gene trees evolving within the underlying species tree (Color figure online)

common assumption is that no recombination occurs within a gene, and all sites in a
gene share a common evolutionary history, while sites on separate genes are indepen-
dent conditional only on the underlying species tree. This differs from the biological
definition of a gene as a segment of DNA coding for a polypeptide. There is, of course,
no underlying reason why a biological gene can or should share a common history
in its entirety. To avoid this confusion, locus is also sometimes used to describe a
recombination-free region, however, it is still more common to refer to “gene trees”
rather than “locus trees”.

A number of different approaches have been taken with regard to species tree
inference in the presence of ILS. The first is essentially to ignore the problem: perform
gene tree inference on concatenated data using methods such as RAxML (Stamatakis
2014) or FastTree (Price et al. 2009), treating all sites as if they share a single,
common evolutionary history. This can be fast and accurate for estimating $S$. But, there
are some concerns: concatenation has been shown to be statistically inconsistent for
some values of $(S, \tau)$ (Degnan and Salter 2005; Roch and Steel 2015), and speciation
time estimates are biased since the coalescent event must naturally occur before the
speciation time. Another approach is the use of summary statistics that first estimate
the gene trees independently for each gene, and then use the gene tree estimates as
inputs for species tree estimates. Examples of this approach include STEM (Kubatko
et al. 2009), ASTRAL (Mirarab et al. 2014; Mirarab and Warnow 2015; Zhang et al.
2018), and MP-EST (Liu et al. 2010). These methods can be computationally efficient,
however they depend on the accuracy of the gene tree estimates that are used as inputs
as well as proper delineation of recombination-free segments of the genome (Gatesy
and Springer 2014; Springer and Gatesy 2016). A third approach uses the full data
to coestimate the species tree and each of the gene trees, generally using Markov
chain Monte Carlo (MCMC) methods. Examples of this approach include BEST (Liu
and Pearl 2007), *BEAST (Drummond and Rambaut 2007; Ogilvie et al. 2017), and
BPP (Yang 2015). These methods can be quite accurate but are very computationally
intensive when the number of loci and/or leaf set is large. Assessment of convergence
is also a challenge, especially due to the multi-modal nature of the likelihood in the
tree space (Salter 2001).

A fourth approach, and the one we take in this paper, is to treat the gene trees as a
nuisance parameter that can be integrated over. A previous example of this approach is
SVDQuartets (Chifman and Kubatko 2014), which uses a rank-based methodology to
infer the proper unrooted species tree for each quartet of species under consideration, and then uses an assembly algorithm to infer the final n-taxon species tree estimate taking the set of unrooted quartet trees as input. The theory behind the method assumes unlinked coalescent-independent sites (CIS) data such that the gene tree underlying each site can be treated as a random draw from the distribution of all possible gene trees given the species tree, however Wascher and Kubatko (2021) recently proved that SVDQuartets is statistically consistent for multilocus data as well.

Our method, (Li)kelihood-based assembly (Lily), also assumes unlinked CIS data. From (Chifman and Kubatko 2015) we have the site pattern probabilities given a species tree topology and branching times. Then, a prior on these topologies and branching times is assumed, and posterior probabilities for each set of rooted triplets (Lily-T) or unrooted quartets (Lily-Q) are calculated. These posterior probabilities are then used as weights in an assembly algorithm to infer the final n-taxon estimated topology \( \hat{S} \). To distinguish whether we are referring to inference on the rooted triplet/unrooted quartet inputs or the final n-taxon tree, we will use \( S^{(3)} \) for a rooted triplet, \( S^{(4)} \) for an unrooted quartet, and \( S^{(n)} \) for the final n-taxon tree. We will simply use \( S \) when the inference step could apply to either \( S^{(3)} \) or \( S^{(4)} \) depending on if we are using Lily-T or Lily-Q. The details of the procedures are described in the next section.

2 Method

The outline for the Lily-T and Lily-Q procedures are laid out in Algorithms 1 and 2. First, an \( n \times n \) matrix of pairwise distances between sequences is found. Then, for each triplet (Lily-T) or quartet (Lily-Q) of species, the site pattern frequencies are found, and the likelihood for each rooted topology is calculated using a prior on the branching times. Using Bayes’s Theorem, the posterior probability of each rooted triplet or unrooted quartet is then determined. In the case of unrooted quartets, this is found by summing over the probabilities of all rooted quartets consistent with the unrooted topology. Finally, given these posteriors as inputs, the n-taxon topology is estimated using supertree assembly methods. Each step is discussed in detail in the following sections.

2.1 Data Structure

Our data structure and data reduction method for Lily-Q are summarized in Fig. 4. (The data reduction process for Lily-T is largely similar and will be discussed at the end of this section.) We begin by assuming a matrix of aligned sequence data \( D_{raw} \) where the rows represent the species under consideration and the columns each represent an aligned site. We assume in the sequel that this alignment has been performed without error. Even with a correct alignment, there may also be sites present in one sequence and not another, either due to the data truly being missing or because of an insertion or deletion. Thus, \( D_{ij} \in \{ A, C, G, T, - \} \). Here, \( i \) is an index for the taxon, \( j \) is the index for the site, the letters represent the four nucleotides, and the dash represents missing data.
Algorithm 1: Lily-T procedure

Calculate an $n \times n$ matrix of distances between species (Sect. 2.2);

for $l \leftarrow 1$ to $\binom{n}{3}$ do
  Find site pattern frequencies $D_{JC}$ for the $l$th triplet of species (Sect. 2.1);
  Estimate speciation rate $\beta$ using procedures in Sect. 2.4;
  Find the site pattern probabilities $\delta(S^{(3)}, \tau)$ for each of the three rooted triplet topologies from (Chifman and Kubatko 2015) (Sect. 2.3 and appendix 6.2);
  Integrate over $\tau$ using $\theta = 0.003$ and $f(\tau|\beta)$ as estimated from Eq. 1 to find $\delta(S^{(3)})$ and then $L(S^{(3)}|D_{JC})$ for each rooted triplet (Sects. 2.4 and 4.1);
  From Bayes’s Theorem find the posterior probability of each of the three rooted triplets (Sect. 2.4);
end

Using the $3\binom{n}{3}$ posterior probabilities as input, estimate $S^{(n)}$ using the Triplet MaxCut algorithm (Sevillya et al. 2016) (Sect. 2.5);

Algorithm 2: Lily-Q procedure

Calculate an $n \times n$ matrix of distances between species (Sect. 2.2);

for $l \leftarrow 1$ to $\binom{n}{4}$ do
  Find site pattern frequencies $D_{JC}$ for the $l$th quartet of species (Sect. 2.1);
  Estimate speciation rate $\beta$ using procedures in Sect. 2.4;
  Find the site pattern probabilities $\delta(S^{(4)}, \tau)$ for each of the fifteen rooted quartet topologies from (Chifman and Kubatko 2015) (Sect. 2.3);
  Integrate over $\tau$ using $\theta = 0.003$ and $f(\tau|\beta, S^{(4)})$ as estimated from Eq. 1 to find $\delta(S^{(4)})$ and then $L(S^{(4)}|D_{JC})$ for each rooted quartet (Sects. 2.4 and 4.1);
  From Bayes’s Theorem find the posterior probability of each of the fifteen rooted quartets (Sect. 2.4);
  Calculate the posterior probability of the three unrooted quartet topologies as the sum of the corresponding rooted quartets (see Fig. 5);
end

Given the $3\binom{n}{4}$ posterior probabilities as input, estimate $S^{(n)}$ using the Weighted QMC algorithm (Avni et al. 2015) (Sect. 2.5);

If the data are iid, or if the possibility of varying rates across sites is treated as a random effect that can be integrated over, then the columns are exchangeable. The following is repeated for each subset of four of the $n$ taxa. Let $\delta_{TCCG}$ represent the probability that at a certain site the first species has $T$, the next two have $C$, and the fourth species has $G$, i.e., $\delta_{TCCG} = P(i_A = T, i_B = C, i_C = C, i_D = G)$ where $i_x$ is the state for species $x$. Under this assumption of exchangeability, $\delta_{TCCG}$ will be the same at every site. Then, the number of sites where the first taxon has character $i_A$, the second has character $i_B$, etc., which we label $d_{i_A,i_B,i_C,i_D}$ will follow a binomial distribution with probability $\delta_{i_A,i_B,i_C,i_D}$, and the joint probability of all possible site patterns is a multinomial distribution. Since there are $4^n$ possibilities (or $5^n$ with missing data), the numbers of sites that follow each pattern $d_{i_A,i_B,i_C,i_D}$ is a sufficient statistic. Then we can map $D_{raw}$ down to a $4^n \times 1$ vector $D_{ind} \sim Multinom(J, \delta_{ind})$, where each element of $\delta$ represents one of the site pattern probabilities and $J$ is the total number of sites. In the sequel, we will only consider sites where all four species in the quartet
Fig. 4 (Color figure online) Data reduction (Lily-Q): For each set of four species, the raw aligned sequence data can be reduced down to $D_{ind}$ and then $D_{JC}$ under the Jukes–Cantor assumptions have a nucleotide present—i.e., sites where $D_{ij} \in \{A, C, G, T\}$, $i = A, B, C, D$, which will not affect inference if sites are missing at random.

If we further assume the Jukes–Cantor (JC69) substitution model (Jukes and Cantor 1969), all nucleotides have the same limiting frequency of $1/4$ and all substitution rates between nucleotides are the same. As a result, for determining probabilities, if we use the JC69 model we don’t need to keep track of what nucleotides are present where; we only need to note whether or not they are the same. For example with two taxa, $P(i_A = C, i_B = C) = P(i_A = G, i_B = G)$, as is the probability of any two different nucleotides at the same site. So, we can call two identical nucleotides at the same site an XX pattern regardless of whether they represent AA, CC, GG, or TT. Similarly, XY represents the case where the two nucleotides are different. Again, the distribution of the site pattern frequencies follows a multinomial distribution, but we need to keep track of fewer cases. The number of different cases required for $n$ taxa is $\sum_{i=1}^{4} Z_{n,i}$ where $Z_{n,i}$ is the Sterling number of the second kind. Most relevant for our later discussion, for a four-taxon tree, we can map $D_{ind}$ down to a $15 \times 1$ vector $D_{JC}$.

We repeat this mapping for each quartet of species, creating an $\binom{4}{4}$ set of site pattern frequency vectors $D_{JC}$.

This data mapping process is similar for Lily-T, except that it is repeated of each of the $\binom{4}{3}$ set of triplets. $D_{ind}$ then is a $4^3 \times 1$ vector and $D_{JC}$ is a $5 \times 1$ vector.

### 2.2 Distance Estimation Between Sequences

We estimate the distance between sequences as follows. From the JC69 model we can infer the well-known pairwise distance estimate for each pair of species

$$\hat{\lambda}_t = \frac{-3 \log(1 - \frac{4\hat{p}^3}{3})}{4}$$
where \( t \) is the time measured in years, \( \lambda \) is the per-year mutation rate, and \( \hat{p} \) is the fraction of discordant sites between the two species. Then, assuming a value of 0.003 (we will defer the justification for this assumption to Sect. 4.1) for \( \theta \) and generation time \( g \), we have:

\[
0.003 = \theta = 4N_e \mu = 4\lambda N_e g
\]

After a quick algebraic manipulation we can estimate time in coalescent units (\( 2N_e \) generations) as:

\[
\frac{\hat{t}}{2N_e g} = \frac{-6 \log(1 - 4 \hat{p})}{0.012}
\]

Then, to get node times, we divide this distance by 1/2 since the distance measures time “there and back again” to the node. If we have multiple individuals per species, we simply take the average of the pairwise distances. Note that since this estimate uses concatenated data, it estimates the expected coalescent time between taxa \( a \) and \( b \), not the speciation time. Since the expected coalescent time must necessarily be greater than the speciation time, this estimate is then biased to overestimate the speciation time. We will show in Sect. 4.1, however, that the Lily procedure is more sensitive to overestimating the speciation rate (which is inversely proportional to speciation times) and so a biased estimator may in this case prove useful.

### 2.3 Derivation of \( L((S, \tau)|D_{JC}) \) for 3- or 4-Taxon Trees

Chifman and Kubatko (2015) derived the site pattern probabilities \( \delta_k|(S^{(4)}, \tau) \) for a 4-taxon tree where \( \delta_k|(S^{(4)}, \tau) \) is the probability of the \( k \)th site pattern, \( k \in \{XXXX, XXXY, \ldots XZW\} \), occurring at a site given the species tree topology and branching times under the JC69 model and the strict molecular clock. Then since \( D_{JC} \sim Multinom(J, \delta|(S^{(4)}, \tau)) \) the likelihood is given by

\[
L((S^{(4)}, \tau)|D_{JC}) \propto \prod_{k=1}^{K} [\delta_k|(S^{(4)}, \tau)]^{d_k}
\]

To find \( P(D_{JC} = d_{JC}|(S^{(4)}, \tau)) \) we first recognize that this can be factored into two processes: the coalescent process and the substitution process:

\[
P(D_{JC} = d_{JC}|(S^{(4)}, \tau)) = \int_{(G,t)} P(D_{JC} = d_{JC}|(G, t), (S^{(4)}, \tau)) f((G, t)|(S^{(4)}, \tau)),
\]

\[
= \int_{(G,t)} P(D_{JC} = d_{JC}|(G, t)) f((G, t)|(S^{(4)}, \tau)),
\]
where the second equality is true under the assumption of neutral selection, whereby the substitution process and coalescent process are independent and then the first term depends directly only on the gene tree (Wakeley 2009).

We begin by noting that:

$$\delta_{XXXY}|((a, (b, (c, d))), \tau) = \delta_{XYXX}|((a, (d, (c, b))), \tau).$$

A similar argument can be made for all fifteen site pattern frequencies and all fifteen rooted 4-taxon topologies. Thus, we need only derive the site pattern probabilities for the two topologies shown in Fig. 3, ((a, b), (c, d)) and (a, (b, (c, d))) and the site pattern probabilities for the other 13 topologies is a permutation of one of these two cases. The likelihood of any other 4-taxon tree is then given by Eq. 2 with the data permuted as necessary by the permutation function $\sigma(\cdot)$:

$$L((S^{(4)}, \tau)|D_{JC}) \propto \prod_{k=1}^{K} [\delta_k|(S^{(4)}, \tau)]^{\sigma(d_k)} \quad (3)$$

$\delta|(S^{(3)}, \tau)$ for the lone unlabeled 3-taxon tree topology can be found by marginalizing any one of the four taxa. The details are shown in the Appendix, where we marginalize over taxon $a$ in the asymmetric 4-taxon species tree (see Fig. 3.)

### 2.4 Derivation of $P(S|D_{JC})$ for 3- or 4-Taxon Trees

From the Law of Total Probability, it is immediate that:

$$\delta*S = \int_{\tau} \delta|(S, \tau) f(\tau|S) d\tau$$

A wide variety of forms for the density $f(\tau|S)$ can be chosen. We choose priors to be uninformative and to allow for analytic solutions to $\delta|S$. Refer to Fig. 3 for a description of $\tau$ and note that the 3-taxon case can be viewed as the asymmetric case with taxon $a$ removed. We base our prior on the Yule model: given a speciation rate $\beta$, the time to the first node below the root follows $Exp(2\beta)$, the time to the next node follows $Exp(3\beta)$ and so on. So, for three taxa:

$$\tau_1 \sim Exp(3\beta) \quad \text{and} \quad (\tau_2 - \tau_1) \sim Exp(2\beta)$$

For the four-taxon asymmetric case:

$$\tau_1 \sim Exp(4\beta) \quad \text{and} \quad (\tau_2 - \tau_1) \sim Exp(3\beta) \quad \text{and} \quad (\tau_3 - \tau_2) \sim Exp(2\beta)$$

In the four-taxon symmetric case, either $\tau_1$ or $\tau_2$ can occur first with equal probability, so one can draw $\eta \sim Bern(.5)$ and then:

$$\tau_1 \sim Exp(4\beta) \quad \text{and} \quad (\tau_2 - \tau_1) \sim Exp(3\beta) \quad \text{and} \quad (\tau_3 - \tau_2) \sim Exp(2\beta), \eta = 1$$
\( \tau_2 \sim \text{Exp}(4\beta) \) and \( (\tau_1 - \tau_2) \sim \text{Exp}(3\beta) \) and \( (\tau_3 - \tau_1) \sim \text{Exp}(2\beta) \), \( \eta = 0 \)

To estimate \( \beta \), we take an empirical Bayes approach. From (Stadler and Steel 2011), we have that the total tree distance (the sum of all branch lengths) under the Yule model follows a \( \frac{\Gamma(n - 1, \frac{1}{\beta})}{\Gamma_1(n - 1, 1)} \) distribution. This tree distance is equal to \( 2\tau_2 + \tau_1 \) for three taxa. Consider the set of pairwise distances estimated as described in Sect. 2.2, and let \( a(j) \) be the \( j \)th smallest distance. For example, in the case of three taxa, we can order the three pairs of distances from smallest to largest as \( a(1) \), \( a(2) \), and \( a(3) \). Then a natural estimator for \( \tau_1 \) is \( a(1) \) and \( a(2) + a(3) \) for \( \tau_2 \), giving

\[
\hat{\beta} = \frac{2}{a(1) + a(2) + a(3)},
\]

in the case of three taxa. Similarly, the tree distance for four taxa is \( 2\tau_3 + \tau_1 + \tau_2 \) leading to an estimator of

\[
\hat{\beta} = \begin{cases} 
\frac{3}{a(1) + a(2) + (1/2)*a(3) + a(4) + a(5) + a(6)}, & \text{4-taxon symmetric tree} \\
\frac{3}{a(1) + (1/2)*a(2) + a(3) + (2/3)*a(4) + a(5) + a(6)}, & \text{4-taxon asymmetric tree}
\end{cases}
\]

Yang (2014) suggests putting an exponential prior on the root age, and applying a Dirichlet prior for the other node times given the root age, arguing that this reduces the coefficient of variation in total tree distance. We did not find much difference in performance between the Yang method and the Yule model, however it is more intellectually consistent to use the Yule model for both \( \tau \) and \( S \). There were two other practical benefits—integrating over the prior is simpler and therefore the implementation is more likely to be error-free, and we found calculations using Yang’s method required long double numerical precision to avoid floating-point errors. For extremely short branching times, these errors could lead to calculating \( \delta_{X\ Y\ Z\ W} < 0 \) without 128-bit precision. We thank an anonymous reviewer for suggesting the Yule model for \( \tau \).

Given these priors on the branching times, \( \delta | S \) is calculated as derived in the appendix and the likelihood of any tree can be found as before by taking these site pattern probabilities and applying the standard multinomial likelihood to a permutation of the data:

\[
L(S | D_{JC}) \propto \prod_{k=1}^{K} \left[ \delta_k | s \right]^{\sigma(d_k)}
\]

An unfortunate side effect of this choice of prior is that for an asymmetric four-taxon tree and a given assumed speciation rate \( E(\tau_1) = \frac{1}{4\beta} \) and \( E(\tau_2) = E(\tau_2 - \tau_1) + E(\tau_1) = \frac{1}{4\beta} + \frac{1}{4\beta} = \frac{7}{12\beta} \). Meanwhile, for the symmetric topology either \( \tau_1 \) or \( \tau_2 \) can occur first with equal probability so \( E(\tau_1) = E(\tau_2) = (\frac{1}{2}) \frac{1}{4\beta} + (\frac{1}{2}) \frac{7}{12\beta} = \frac{5}{12\beta} \). In other words, if \( (C,D) \) is a cherry, inferring an asymmetric topology automatically implies the prior assumption that this divergence time between species C and D occurs more
Fig. 5  Each unrooted 4-taxon tree corresponds to five different rooted trees, arising from placing the root on one of the five branches in the unrooted tree

recently. This does not appear to be a problem at first glance, since we are not concerned with inferring species divergence times here, and are treating $\tau$ as a nuisance parameter. But, it turns out that this makes accurate inference of the root location of a 4-taxon tree impossible. Given our choice of model, if the true tree is symmetric—for example $((a, b), (c, d))$—then various pairs of rooted asymmetric trees have the same likelihood with speciation times integrated out: $L[((a, b), (c, d))] = L[(b, (a, (c, d)))]$, $L[(c, (d, (a, b)))] = L[(d, (c, (a, b)))]$, etc. Worse, for some true values of $\tau_1$, $\tau_2$, $\tau_3$, $L[(a, (b, (c, d)))] > L[((a, b), (c, d))]$. This was also a problem for the prior using Yang’s method. These inference errors all concern the location of the root rather than the unrooted topology. As a result, we limit our inference on 4-taxon trees to unrooted topologies.

For the prior on topologies, we assume the tree generation follows a Yule model: there is a constant rate of species divergence over time, and the rate of species divergence is equal for all branches. This is meant to be as uninformative as possible. For 3-taxon trees, this intuitively assigns a 1/3 probability to each of the 3 rooted 3-taxon trees. For 4-taxon trees, each of the three symmetric topologies has a 1/9 prior probability and each of the twelve asymmetric topologies has a 1/18 prior probability. Interestingly, while any individual symmetric topology is twice as probable as any individual asymmetric topology, in the prior it is twice as probable that the unlabeled topology will be asymmetric rather than symmetric. The details of the prior calculation are given by (Harding 1971).

Taking both the prior on the topologies and each topology’s likelihood, it is a simple application of Bayes’s Theorem in the 3-taxon case to show:

$$P(S = s|D_{JC}) = \frac{L(S|D_{JC})P(S = s)}{\sum_{i=1}^{3} L(S_i|D_{JC})P(S_i = s_i)}$$

For the 4-taxon case, the summation is performed over 15 rather than 3 topologies, and the final probability of the unrooted topology is the sum of the five rooted topologies compatible with it (see Fig. 5).

For the 3-taxon case, we can show that given a sufficient number of unlinked sites under the molecular clock, we can infer the correct topology with probability 1. For the unrooted 4-taxon case, we have demonstrated this in simulation studies for a wide variety of true trees, but it remains unproven.
Theorem 1. Given CIS data for three taxa evolving under the multispecies coalescent with the JC69 model and the molecular clock, if $\hat{S}^{(3)}$ is the maximum a posteriori tree, $P(\hat{S}^{(3)} = S^{(3)}) \to 1$ as the number of sites $J \to \infty$ for any prior for which $0 < \tau_1 < \tau_2$ holds with probability 1.

Then, using a Bonferroni adjustment, we can ensure that given a sufficient number of sites, $P(\hat{S}^{(3)} = S^{(3)}) > 1 - \epsilon$ uniformly for all the triplets. It then follows that given proposition four from (Steel and Penny 1993), we can estimate $S^{(n)}$ with probability $> 1 - \epsilon$.

2.5 Estimating the n-Taxon Species Tree

There is no theoretical impediment to prevent extending this procedure to infer full posterior probabilities for 5-, 6-, or even n-taxon trees. But, to see the practical difficulties, consider the challenges to extend this result to just five taxa. First, for 5 taxa $\delta_{JC}$ is now a $51 \times 1$ vector for each of three unlabeled topologies. Then, a $51 \times 105$ permutation matrix is needed to find the site pattern probabilities for all 105 5-taxon labeled gene tree topologies. A general recursive formula for the number of unordered gene histories embedded in an n-taxon species tree is provided by (Rosenberg 2007). We calculated this total to be 379, 313, and 208 for $S_1$, $S_2$, and $S_3$ from Fig. 9, respectively. The probabilities of these histories need to be calculated and then integrated over to find the site pattern probabilities of the three unlabeled species tree topologies, then the same $51 \times 105$ permutation matrix must again be applied to get the full site pattern probabilities for all species tree topologies. As a result, the full set of site pattern probabilities $\delta | (S^{(5)}, \tau)$ remains to be calculated, and an even more complex process would be needed to extend to six or more taxa.

Instead, to infer $S^{(n)}$, we use an assembly algorithm that takes the posterior probabilities of the rooted triplets or unrooted quartets as inputs to infer the final estimate $\hat{S}^{(n)}$. We repeat the process of Sects. 2.3 and 2.4 for each of the $\binom{n}{3}$ sets of triplets (Lily-T) or $\binom{n}{4}$ quartets (Lily-Q). Then we have a set of $3 \binom{n}{3}$ posterior probabilities for each possible rooted triplet or $3 \binom{n}{4}$ posterior probabilities for each possible unrooted quartet from the leaf set. These probabilities are then used in an assembly algorithm to find $\hat{S}^{(n)}$. There are many options for assembly methods. The methods we choose allow us to use the posterior probabilities as weights in the subtree inputs. Both SVDQuartets and ASTRAL are unweighted methods and thus treat all inputs equally regardless of the inference uncertainty (see Fig. 6). For Lily-T, we chose the Triplet MaxCut (TMC) method of Sevillya et al. (2016) due to its speed, accuracy, ease of implementation, and its ability to work with rooted trees as inputs, and similarly chose the Weighted Quartet Max Cut (weighted QMC) algorithm of Avni et al. (2015) for Lily-Q. Further, the implementation of these algorithms are very similar to the algorithm used by SVDQuartets, and so using it reduces, but does not eliminate, one source of variation in comparing the two methods.

Triplet MaxCut works as follows: first we obtain posterior probabilities on all $\binom{n}{3}$ subsets from the leaf set. When we have multiple individuals from a species, posterior probabilities for each combination of individuals from each triplet of species are
Bayesian-Weighted Triplet and Quartet Methods… Page 13 of 39

93

Fig. 6 Unweighted tree inputs: an unweighted method treats the two cases the same, even though the right side contains far more information about the true topology.

calculated. Then two symmetric $n \times n$ matrices are formed, $G$ and $B$, called the “good” and “bad” matrices with $i, j = 1, 2, 3 \ldots n$ corresponding to the taxa in alphabetical order: $a, b, c, \ldots$. For each triplet input, we have a set of two taxa that form a cherry and a third that is not part of the cherry, as well as a weighting for the triplet. For each of the $\binom{n}{3}$ pairs in the triplet, if taxa $i$ and $j$ form a cherry, the weight of the triplet is added to $B_{ij}$ and $B_{ji}$ and if $i$ and $j$ do not form a cherry, the weight of the triplet is added to $G_{ij}$ and $G_{ji}$. As an example, consider the case where we are trying to infer $S$ for $L = \{a, b, c, d, e, f\}$. One of our input triplets is $(a, (c, d))$, which is given the weight 0.5. We would add 0.5 to $B_{34}$, $B_{43}$, $G_{13}$, $G_{31}$, $G_{14}$, and $G_{41}$. This process is repeated for all sets of input triplets. Thus, large entries in $G_{ij}$ indicate taxa $i$ and $j$ should, all else equal, be separated, and large entries in $B_{ij}$ indicate taxa $i$ and $j$ should, all else equal, be grouped together in the final tree $S$.

Next, for the set of $n$ taxa $L = \{1, 2, 3 \ldots n\}$ we obtain all subsets of the taxa of cardinality $\leq \lfloor \frac{n}{2} \rfloor$. For each subset, we can arbitrarily label the taxa in the subset $L_1$ and call the remainder $L_2 = L \setminus L_1$, resulting in a bipartitioning of $L$. Then, for each bipartition, we score the bipartition by the ratio $\frac{\sum_{i \in L_1} \sum_{j \in L_2} G_{ij}}{\sum_{i \in L_1} \sum_{j \in L_2} B_{ij}}$. Note that this score can be undefined if there are entries in $B = 0$. To avoid this problem, a very small number such as $10^{-200}$ can be added to each element of $B$. The bipartition with the highest score is accepted. If the cardinality of either $L_1$ or $L_2$ is greater than two, the process is repeated recursively on $L_1$ and/or $L_2$ as needed. In essence, each step of the procedure creates a node and assigns taxa to the left and right branches of the node until the final tree is resolved.

The operation of the weighted QMC is largely similar, with posteriors first being calculated for all $\binom{n}{4}$ subsets of quartets from the leaf set to use as inputs in the assembly algorithm. The major difference in the assembly is that care must be taken if one of $L_1$ or $L_2$ is a set of 3 elements. Then the set is augmented with an additional artificial taxon and the procedure is performed on this augmented set to evaluate which two elements of the set constitute a cherry.

2.6 Uncertainty Quantification

An immediate concern with using an assembly procedure is that one of the key advantages of our likelihood-based approach—the ability to produce posterior
probabilities—is lost. The assembly procedures we chose are based on heuristics that make sense—grouping species together that have a high probability of being cherries. But, unfortunately we have no distributional results to assess. The output tree is a point estimate only, and while we can generate simulation data to say that it is reasonably accurate, once we apply it to real data we no longer have a measure of uncertainty. The difficulty lies in the same factors that led us to pursue the assembly procedure in the first place—the inability to calculate a joint $n$-taxon set of site pattern probabilities.

A standard method for measuring uncertainties of estimated phylogenies is the bootstrapping method of Efron (1979). We implemented a nonparametric bootstrap: for CIS data we resampled with replacement from all sites, and for multilocus data we resampled the genes with replacement and then for each gene we resampled with replacement the sites within each gene. An advantage of the bootstrap is that it can give unbiased estimates of uncertainty without any distributional assumptions. But, a number of cautions are in order. First, we only have asymptotic guarantees about the approximation of $\frac{d_k}{\delta_k}$ to $\delta_k$ and we have no finite-sample knowledge of how good the approximation is. As a result, we don’t know how much uncertainty we have about our uncertainty. Second, without parallelization, the time required to estimate the bootstrap samples increases linearly with the number of bootstrap samples taken. Last, bootstrap support values are not probabilities, and should not be treated as such. That said, we can compare the bootstrap support to the actual proportion of times we infer the correct tree to see if the bootstrap support reasonably “mimics” the true probability of being correct for reasonable parameter values.

3 Implementation

We have written source code in C++ that implements Lily-T and Lily-Q. The programs take as inputs alignments in PHYLIP format and output either the final output tree (Lily-T) or a properly formatted input file for the weighted QMC program (Lily-Q). We also wrote programs for calculating the number of gene tree histories in Sect. 2.5 and a Perl wrapper for executing the simulation runs. The Robinson-Foulds (RF) distance (Robinson and Foulds 1981) from the true tree was calculated using the R packages ape (Paradis and Schliep 2019) and phangorn (Schliep 2011). These programs, as well as a file summarizing the results are available at https://github.com/richards-1227/Lily.

To check against coding errors, we performed chi-squared goodness of fit tests comparing the site pattern probabilities calculated by our C++ programs against 1,000,000 CIS simulated at various settings of $\tau$ and $\theta$ with four taxa. There was no evidence that the simulated frequencies differed from that expected (only one individual test $p$ value was below 0.05 (0.006), and it was no longer significant after making a Bonferroni adjustment).
4 Results

Data were simulated using the *ms* (Hudson 2003) and Seq-Gen (Rambaut and Grassly 1997) programs in C++ using a Unix HPC platform. The *ms* program takes the node times (in coalescent units) and population parameter $\theta$ as input and simulates a set of gene trees under the multispecies coalescent model. Seq-Gen then takes these gene trees and the mutation model parameters as input and generates the aligned sequence data $D_{raw}$. One point of caution is that *ms* measures time in $4N_e$ generations, so to use the units described here, the input the node times (measured in coalescent units) must divided by 2, and then scaled for the Seq-Gen calculations using the --s switch by $\theta$ rather than $\theta/2$.

4.1 Testing Robustness to Model Assumptions and Prior Specification

We next relaxed various assumptions of the substitution model to verify that the true tree was still estimated with high probability. One thousand runs were performed with 10,000 CIS generated with the Jukes–Cantor assumptions being progressively relaxed, first allowing for different stationary probabilities, then different relative substitution rates, then between-site rate heterogeneity, and finally allowing for invariant sites. The settings are are listed below with results in Table 1.

- **Setting one:** `seq-gen < infile > outfile -q -mHKY -s 0.003 -l 10000 -p 10000`
- **Setting two:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -l 10000 -p 10000 -s 0.003`
- **Setting three:** `seq-gen < infile > outfile -q -mGTR -f 0.05 0.45 0.45 0.05 -r 1.0 3.0 1.0 1.0 3.0 1.0 -l 10000 -p 10000 -s 0.003`
- **Setting four:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -r 1.0 5.0 1.0 1.0 3.0 1.0 -l 10000 -p 10000 -s 0.003`
- **Setting five:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -r 1.0 2.0 0.75 0.5 1.5 0.25 -g 3 -a 3.0 -l 10000 -s 0.003`
- **Setting six:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -r 1.0 2.0 0.75 0.5 1.5 0.25 -g 9 -a 9.0 -l 10000 -s 0.003`
- **Setting seven:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -r 1.0 2.0 0.75 0.5 1.5 0.25 -g 3 -a 3.0 -l 10000 -s 0.003`
- **Setting eight:** `seq-gen < infile > outfile -q -mGTR -f 0.15 0.35 0.35 0.15 -r 1.0 2.0 0.75 0.5 1.5 0.25 -g 3 -a 3.0 -l 10000 -s 0.003`

The results shown in Table 1 are typical and are presented for Lily-Q and each of the eight substitution model settings using $\tau_1 = 0.2$, $\tau_2 = 0.4$, $\tau_3 = 0.6$, and $\theta_0 = 0.003$. There does not appear to be a large impact from relaxing the JC69 assumptions, and for
Table 1  Average posterior probabilities assigned to the true tree as the JC69 assumptions are progressively relaxed

| Setting | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
|---------|------|------|------|------|------|------|------|------|
| Posterior probability | 0.925 | 0.941 | 0.873 | 0.891 | 0.886 | 0.910 | 0.873 | 0.869 |

Fig. 7  (Color figure online) Mean posterior probability assigned by Lily-Q to the true tree for varying numbers of sites and true population parameter θ for a symmetric 4-taxon with branch lengths of 1.0 coalescent unit. The left panel a uses θ = 0.001 and the right panel b uses θ = 0.01

the remainder of our simulations all the data were generated under simulation setting 8, so performance of our methods is measured against a deliberately misspecified model.

Here we should note that the site pattern probabilities δ are also conditional on θ as well as (S, τ), but keeping with the notation of (Chifman and Kubatko 2015) we have ignored this conditioning. In the sequel, we will denote the true but unknown population parameter as θ₀, whereas θ represents the value we use in our likelihood calculations. For the remainder of our work, we used θ = 0.003 as an input to both Lily-T and Lily-Q, which we justify as follows. Most empirical values of θ₀ fall between 0.001 and 0.01 (see (Jennings and Edwards 2005; Kopp and Barmina 2005; Kubatko and Degnan 2007; Rannala and Yang 2003) for estimates on species ranging from primates to finches). So, we simulated data using θ₀ = {0.0003, 0.001, 0.003, 0.01, 0.03}, extending beyond the empirical range in both directions, and calculated posteriors using θ = 0.01 and θ = 0.001. The results for Lily-T and Lily-Q are shown in Fig. 7. Since there is no apparent difference in performance for assuming a θ different from the actual data-generating θ₀, we simply use a θ = 0.003 as it is in the middle of the empirical range on the log scale.

We cannot, however, simply assume a value for the speciation rate β. We simulated data for a symmetric species tree with τ₁ = τ₂ = 1.0, τ₃ = 2.0 with the posterior probability assigned by Lily-Q to the true tree shown in Fig. 8. We performed inference using a β = 13/24 and then two orders of magnitude above and below this value. β = 13/24 was chosen since E(τ₃) = E(τ₃ − τ₂) + E(τ₂ − τ₁) + E(τ₁) = 1/2β + 1/3β + 1/4β = 13/12β so β = 13/24 corresponds to an expected root age of 2 coalescent units. If the speciation rate is overestimated, however, there can be a large loss of accuracy, even with 9000 sites. This is to be expected given the asymmetric nature of
Fig. 8 (Color figure online) Mean posterior probability assigned by Lily-Q to the true tree for varying numbers of sites and assumed speciation rate $\beta$ for a 4-taxon symmetric tree with branch lengths of 1.0 coalescent units. ($\beta = 0.5417$ gives an expected root age of 2.0, matching true tree)

the exponential distribution—a small $\beta$ flattens the prior and can still have adequate prior mass near the true value, while a large $\beta$ puts most prior mass near zero.

4.2 Application to Simulated Data

Data were simulated for 5, 8, and 12-taxon trees for a total of 12 different topologies ranging from fully symmetric to fully asymmetric. The full set of topologies simulated is displayed in Fig. 9. For each topology, we used four different values of the true population parameter: $\theta_0 = \{0.0008, 0.0024, 0.0072, 0.0216\}$. These were chosen to extend slightly above and below the empirical range of 0.001 to 0.01 and the values are linear on the log scale. We used five different settings for the minimum branch length (MBL): 0.1, 0.2, 0.5, 1.0, and 2.0 coalescent units. Multilocus data were simulated with 10, 50, and 500 genes and 50, 200, and 500 sites per gene. For CIS data, we simulated either 5,000 or 50,000 sites. 100 runs were performed at each combination of settings. Not all settings were used in all combinations, as with low branch lengths and/or small values of $\theta_0$, there arose cases where there was no mutation between two of the sequences. Although there are ways to handle zero sequence distance and infer non-binary trees, it was not practical to handle these cases given the large number of simulation runs and different settings, so we excluded settings where we did not get a full run of 100 simulations with at least some distance between all sequences.

The tree topologies in Fig. 9 are drawn to scale. Comparing trees $S_4$ and $S_5$ we note that for the same MBL, the root for the asymmetric tree $S_5$ will occur much further in the past than for the symmetric tree $S_4$. Trees with any degree of asymmetry will naturally have some branch lengths longer than the minimum due to the structure imposed by the tree. For example, $S_1$ has each of the pendant branches of increasing length. In $S_3$, the branches from the root to two of the species are all at the MBL, while
Fig. 9  The twelve topologies used in the simulation study. 5 taxa: \( S_1 \) through \( S_3 \), 8 taxa: \( S_4 \) through \( S_7 \), 12 taxa: \( S_8 \) through \( S_{12} \). The red bar indicates scale of the MBL, which was set at 0.1, 0.2, 0.5, 1.0, and 2.0 coalescent units.

one pendant branch is two times MBL and the other pendant branches are 1.5 times MBL. Full tree descriptions are given in the supplementary material.

In order to evaluate how well each method performed, we first need a proper metric of distance between trees. The most common is the RF distance (Robinson and Foulds 1981). The distance between the two trees \( T_1 \) and \( T_2 \) is defined as the sum of the number of bipartitions in \( T_1 \) not contained in \( T_2 \) and vice versa. For binary trees, the presence of a bipartition is symmetric, since the number of bipartitions is equal to the number of internal branches, which is in turn equal to \( n - 2 \) for a rooted tree and \( n - 3 \) for an unrooted tree. Thus, if there is a bipartition of \( T_1 \) not in \( T_2 \), there must be a bipartition in \( T_2 \) not in \( T_1 \). As a result, the RF metric can take on any even value from zero to \( 2(n - 2) \) (or \( 2(n - 3) \) if unrooted). Since the metric is also a function of whether or not the two trees are rooted, to compare Lily-T to the other methods (which produce unrooted trees), we must first unroot both the estimated and true tree. For multilocus data, we compared SVDQuartets [as implemented in PAUP* (Swofford 2003)], ASTRAL [using FastTree for gene tree estimation as per the example in Chou et al. (2015)], Lily-T, and Lily-Q for each run and calculated the mean RF distance and standard error for the 100 runs. For CIS data, we excluded ASTRAL from comparison since a gene tree cannot be estimated for a single site. For some cases with difficult settings (low MBL and/or \( \theta_0 \)) SVDQuartets did not fully resolve some nodes, estimating non-binary trees. Although the RF metric is still defined in this case, it tended to produce smaller RF values when SVDQuartets estimated a non-binary tree. Thus, the improved performance of SVDQuartets versus other methods for difficult problems is partly illusory. (See Fig. 10 panel (d) where SVDQuartets seems to improve with less data.)
A representative selection of the simulation results are shown in Figs. 10 through 14 (the full plots are in the supplementary material) corresponding to topologies $S_4$ and $S_8$ from Fig. 9. A number of results become evident from the RF plots. First, all estimation is better with a larger $\theta$, as there is more mutation along each branch allowing us to pick up more of the phylogenetic signal. Second, with enough data, all methods perform well. Third, by comparing the 50 genes/500 sites per gene and the 500 genes/50 sites per gene case, we can see that all methods do better with more genes even when the
Fig. 11 (Color figure online) Mean normalized RF distance for true tree (((a, b), (c, d)), ((e, f), (g, h))), minimum branch length 0.5 and 1.0. Note that ASTRAL cannot be used for genes of length 1. Groups on the horizontal axes indicate number of genes and sites per gene total number of sites is held constant. Also, comparing to the 25,000 CIS setting, we generally see great improvement with Lily-T and Lily-Q compared to multilocus data, which is to be expected given the theoretical basis of the method. At smaller sample sizes, SVDQuartets is generally outperformed by all other methods, and Lily-T is in turn outperformed by Lily-Q under most simulation settings.

Therefore, we focused on the comparison between Lily-Q and ASTRAL. For MBL of 1.0 or 2.0 coalescent units, Lily-Q generally performs no worse, and in many cases,
Fig. 12 (Color figure online) Mean normalized RF distance for true tree (((a, b), (c, d)), ((e, f), (g, h))), minimum branch length 2.0, and ((a, (b, (c, (d, (e, (f, (g, (h, (i, (j, (k, l))))))))))), minimum branch length 0.1. Note that ASTRAL cannot be used for genes of length 1. Groups on the horizontal axes indicate number of genes and sites per gene.

better than ASTRAL. For an MBL of 0.5 coalescent units, Lily-Q usually performed as well or better than ASTRAL, except with 10 genes, 500 sites per gene, and \( \theta_0 \) of 0.0072 or 0.0216. For 0.1 and 0.2 coalescent units, ASTRAL performs better with more sites per gene and larger values of \( \theta_0 \), with the results more even with a smaller \( \theta_0 \) and fewer sites per gene. With regard to topologies, we could not discern a clear pattern of which topologies favored one method or the other. ASTRAL outperforms
Lily-T for most settings, except when the branch lengths are large and $\theta$ is smaller. These results match our expectations, as ASTRAL depends on accurate gene tree inputs and those gene trees are easier to resolve when the genes are long (many sites per gene) and there is more mutation along each branch (higher $\theta$).
Fig. 14 (Color figure online) Mean normalized RF distance for true tree (a, (b, (c, (d, (e, (f, (g, (h, (i, (j, (k, l)))))))))), minimum branch lengths 1.0 and 2.0. Note that ASTRAL cannot be used for genes of length 1. Groups on the horizontal axes indicate number of genes and sites per gene.

4.3 Bootstrapping Results

For six different combinations of topology, $\theta$, MBL, gene length, and number of genes for each of 5, 8 and 12 taxa, we generated then estimated 100 trees. We then drew 100 bootstrap samples on each tree by resampling both the genes and then resampling the sites within each gene. For both Lily-T and Lily-Q, we compared the bootstrap
support for each clade and whether the estimated clade was present in the true tree. Figure 15 shows the bootstrap support for correctly identified clades versus incorrectly estimated clades. We had a total of 11,400 clade estimations for Lily-T and 9,600 for Lily-Q. This shows that the bootstrap support tends to be higher when the clade has been correctly identified. Then, we checked the predictive power of the bootstrap support (see Table 2). The positive predictive power was excellent, as the frequency of estimating the correct clade is as high or higher than the bootstrap support. The negative predictive power, however, is lower as the frequency of estimating the correct clade may still be high when the bootstrap support is low.

We looked at different settings to see if these general results were influenced by aspects of the tree. We did not see a difference for topology or number of taxa. The difficulty of the problem did tend to impact how well bootstrap support tracks estimation accuracy. Figure 16 and Table 3 show results for estimating the true tree $(a, (b, (c, (d, e))))$ with 200 sites per gene. In the easier case, we have 10 genes, $\theta_0 = 0.0216$, and MBL of 1.0 coalescent units. In the harder case, we have 50 genes, $\theta_0 = 0.0008$, and MBL of 0.2 coalescent units (the additional genes are more than offset by the branch length and mutation rate). In the easier case, bootstrap support generally underestimates accuracy, and clades may be accurate even when the support is low. For the harder case, the frequency of correctly estimating a clade appears to rise linearly with the bootstrap support.

These simulation studies suggest that the bootstrap support may be useful as a guide to triaging larger phylogenetic questions—if the bootstrap support for all clades is high, our evidence indicates that there is a high probability that we have inferred the correct tree. Conversely, if the support is low, it serves as an indication that a more computationally intensive MCMC-based method may be necessary for an accurate assessment of the uncertainty associated with specific nodes.

**4.4 Application to Empirical Sequence Data**

We applied Lily-T and Lily-Q to four empirical datasets. For all the datasets we estimated the topology and then drew 100 bootstrap samples and calculated the bootstrap support at each node. All analysis was run on a UNIX HPC with 15 parallel proces-
Table 2 Percentage of correct nodes for different bootstrap support values

| Bootstrap support | Correct | Incorrect | Percent correct |
|-------------------|---------|-----------|-----------------|
| [0, 0.2]          | 63      | 147       | 0.300           |
| (0.2, 0.4]        | 657     | 747       | 0.468           |
| (0.4, 0.6]        | 1848    | 695       | 0.727           |
| (0.6, 0.8]        | 2664    | 183       | 0.935           |
| (0.8, 1]         | 4374    | 22        | 0.995           |
| [0, 0.2]          | 9       | 29        | 0.237           |

Data are taken from the 18 different simulation settings: Lily-T (top) and Lily-Q (bottom).

Fig. 16 Bootstrap support for correctly and incorrectly estimated clades—comparing an easier case (a, left) with a more difficult case (b, right). In both cases, the true tree is \((a, (b, (c, (d, e))))\) and 200 sites per gene were simulated.

The second dataset consisted of approximately 127,000 sites over 106 genes from eight yeast species: *Saccharomyces cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae*...
Table 3  Percentage of correct nodes for different bootstrap support values for an easier case (top) and a more difficult case (bottom)

| Bootstrap support | Correct | Incorrect | Percent correct |
|-------------------|---------|-----------|-----------------|
| [0, 0.2]          | 0       | 0         | -               |
| (0.2, 0.4]        | 2       | 0         | 1               |
| (0.4, 0.6]        | 20      | 7         | 0.741           |
| (0.6, 0.8]        | 52      | 8         | 0.867           |
| (0.8, 1]          | 211     | 0         | 1               |

| Bootstrap support | Correct | Incorrect | Percent correct |
|-------------------|---------|-----------|-----------------|
| (0.2, 0.4]        | 43      | 85        | 0.335           |
| (0.4, 0.6]        | 60      | 60        | 0.500           |
| (0.6, 0.8]        | 25      | 12        | 0.676           |
| (0.8, 1]          | 3       | 0         | 1               |

The true tree is (a, (b, (c, (d, e)))) and 200 sites per gene were simulated.

Fig. 17  Estimates for the 9-taxon primate dataset with bootstrap support for each clade

Fig. 18  Estimates for 8-taxon yeast dataset with bootstrap support for each clade. Full species names in text

(Smik), S. kudriavzevii (Skud), S. bayanus (Sbay), S. castellii (Scas), S. kluyveri (Sklu), and the outgroup Candidas albicans (Calb) (Rokas et al. 2003; Wen and Nakhleh 2018). Both methods matched the results from (Wen and Nakhleh 2018). Estimation took under a minute. All but one node had 100% bootstrap support. The exception was some uncertainty over whether Skud and Sbay represent a clade or if Sbay is an outgroup to all of Scer, Spar, and Smik (see Fig. 18).

Next, we applied Lily-T and Lily-Q to a dataset of 52 individuals from seven North American snake species consisting of aligned sequences from 18 nuclear and
Fig. 19 Estimates for 7-taxon rattlesnake dataset with bootstrap support for each clade. Agk.—*Agkistrodon pscivorus* and *A. contortrix*, S.c.e.—*Sistrurus catenatus edwardsii*, S.c.t.—*S. c. tergeminus*, S.c.c.—*S. c. catenatus*, S.m.m.—*S. miliarius miliarius*, S.m.s.—*S. m. streckeri*, S.m.b.—*S. m. barbouri*

Fig. 20 Estimates for 4-taxon mosquito dataset with bootstrap support (Lily-T). Lily-Q gives 100% posterior probability to this quartet

1 mitochondrial genes with 190-850 characters per gene. Runtime was 34 min (the increase in run time was due to the many different combination of individuals within each subset of three or four species). This is much less than the runtimes reported by (Chifman and Kubatko 2014): <1 day for SVDQuartets and ∼10 days for *BEAST*, although we acknowledge runtimes for these methods have improved in the time since those results were published. Our results are shown in Fig. 19. Bootstrap support was high (≥ 88%) for all clades except the (*S.m.miliarius*, *S.m.barbouri*) clade. For that clade, bootstrap support was 43% for Lily-T and 70% for Lily-Q. SVDQuartets also exhibited low bootstrap support for this clade, so the weak support may be a function of the weak phylogenetic signal present in the data.

Finally, we estimated the topology for four mosquito species using aligned sequences consisting of over 25 million sites from around 80,000 different genes. For Lily-T, we ran 100 bootstrap samples and there was 100% bootstrap support for both clades, with total computation time of fourteen min. Because we did not have the delineation of the different genes, we resampled the sites for bootstrapping in a single stage. The estimated species tree is shown in Fig. 20 and matched the results from (Thawornwattana et al. 2018). Since there were only four taxa, Lily-Q could calculate a posterior probability of 100% using a single run which took around 15 seconds.
5 Conclusions and Further Work

Lily-T and Lily-Q are two new methods for fast, accurate species tree estimation under the multispecies coalescent model along with the assumption of the molecular clock. The methods are insensitive to the value of the coalescent population parameter $\theta$, and sensitivity to the prior on the speciation rate can be controlled through our method for estimating the rate from the data. Lily-Q is more accurate than Lily-T, and we would recommend it over Lily-T unless one needs to estimate the root location or practitioners are concerned about Lily-Q lacking a theoretical guarantee, in which case Lily-T is preferred. We suspect the advantage of Lily-Q over Lily-T comes from the fact it better captures the joint distribution of the data whereas Lily-T in essence only looks at which pattern $d_{XXY}$, $d_{XYX}$, or $d_{YXX}$ is more common.

The comparison between Lily-Q and ASTRAL follows our expectations in that ASTRAL is able to perform well under conditions when estimation of the individual genes is likely to be accurate—when $\theta$ is large and when there are a large number of sites per gene. It is worth noting that our comparisons are of a correctly specified ASTRAL model against an incorrectly specified Lily-Q model: we assume that we have delineated the genes correctly and that sites within each gene have a common history whereas Lily-Q assumes that all sites are unlinked. Even with this disadvantage, Lily-Q outperformed ASTRAL for many parameter settings. There is also reason to believe that the number of sites per gene needed for ASTRAL to do better than Lily-Q may not be realistic—(Hobolth et al. 2011) estimated 75% of loci in a primate dataset have a recombination-free length of between 17 and 93bp and (Springer and Gatesy 2016) reported a locus length of around 12bp for a large mammalian dataset. An avenue for further research would be to test Lily-Q against ASTRAL when the genes are not properly delimited to measure how much this degrades ASTRAL’s performance.

We also did not compare directly to any concatenation methods. These comparisons may indicate the limits of concatenation methods, especially for the highly asymmetric trees with short MBL. One comparison in particular that may be of interest would be to compare Lily-T to SMRT-ML (DeGiorgio and Degnan 2020), which estimates rooted triplets using concatenated data. We should highlight that, unlike concatenation or coestimation methods, we can only estimate $S^{(n)}$, and not any other parameters of interest such as $\tau$. Once $S^{(n)}$ is estimated, we could use our distance matrix to estimate $\tau$, but those estimates are biased and highly sensitive to $\theta$. Peng et al. (2021) have recently proposed a coalescent-based method for estimation of branch lengths within a species tree using a composite likelihood approach, and demonstrated that their method is both accurate and computationally efficient in comparison to Bayesian approaches.

While both the Lily-T and Lily-Q assembly steps generate errors in the face of three or more identical sequences, there are ways to work around this problem. For example, consider the case with six taxa where taxa $a$, $b$, and $c$ have identical sequences. As a first step $a$ and $b$ could be simply removed from consideration, inference applied to the remaining taxa—perhaps inferring the split $(c, d, (e, f))$—which results in the final inferred tree in Fig. 21. While such a workaround would be unwieldy for the large-scale simulation study performed here, it could be reasonably implemented on empirical data.
We have presented some data showing that high bootstrap support is indicative of high tree estimation accuracy, but we wish to highlight that we need to explore this claim further in future work. First, to minimize simulation time, we only performed the bootstrap analysis on eighteen different parameter settings. Second, we would like to perform further simulation studies to compare the usefulness of bootstrap support values across different estimation methods.

This work should scale well to higher numbers of taxa. For one individual per species, the number of quartet estimates is \( \binom{n}{4} \), which requires \( O(n^4) \) operations. Our point estimate for the snake data required over 68,000 quartet estimates and was completed in under a minute. At the same level of parallelization, for one individual per species, we could estimate \( \binom{100}{4} \) quartets in under an hour and could evaluate more species at a higher level of parallelization. Of course, the ability to scale our method also depends on the assembly method, and we note that our approach of providing weights for quartets or triplets can be paired with any assembly method that takes weighted input. We have tested wQMC for over 30 taxa and found it to be reasonably efficient. However, development of computationally efficient methods for building supertrees from triples and quartets remains an active area of current research [see, e.g., (Mahbub et al. 2020)].

Finally, this work has all been done under the strict molecular clock assumption, and the method can be sensitive to the assumption. To see why, consider the true tree \( (a : 1, (b : x, c : x) : (1 - x)) \), where \( x \in (0, 1) \) and \( \theta_0 = 0.003 \). Looking at the proof of theorem 6.4, the maximum a posteriori tree will not be the true tree if either \( \delta_{XXY} \) or \( \delta_{XYX} \) is greater than \( \delta_{YXX} \). In work in review, we have generalized the work of (Chifman and Kubatko 2015) to allow for different relative rates on different branches of the tree, which we represent by the vector \( \gamma \).

Without loss of generality, we can focus on species \( c \) as the result for \( b \) will be similar. Let \( \gamma_C \) represent the relative rate on the pendant branch to \( c \) and \( \gamma_{BC} \) the relative rate on the common branch leading to the root. Increasing \( \gamma_C \) and holding other rates constant, we have that \( \delta_{XXY} > \delta_{YXX} \) when \( \gamma_C > 9 \) for \( x = 0.2 \), \( \gamma_C > 3 \) for \( x = 0.5 \), and \( \gamma_C > 1.5 \) for \( x = 0.8 \). For the interior branch, \( \delta_{XXY} > \delta_{YXX} \) when \( \gamma_{BC} > 6.5 \) for \( x = 0.2 \), \( \gamma_{BC} > 4.5 \) for \( x = 0.5 \), and \( \gamma_{BC} > 3.0 \) for \( x = 0.8 \). This shows that Lily-T is most sensitive to violations of the clock for what are already the hardest problems—those with short internal branches. It also suggests that clock violations are more serious on a pendant than on an internal branch. Although we do

---

**Fig. 21** Handling identical sequences: After first treating the identical sequences as coming from a single species, the rest of the tree is resolved as \((c, d, (e, f))\). Then \( c \) is replaced by the polytomy \((a, b, c)\).
not have clear criteria for when Lily-Q, or any assembly procedure, would produce maximum a posteriori trees that differ from the true tree in the limit, at a minimum it would occur if any 3-taxon subtree met these conditions.

Correcting for sensitivity to violations of the clock would require either estimating \( \hat{\gamma} \) from the data or applying a prior to \( \gamma \) and integrating over it to get \( \delta|S, \tau \). Whether these steps can be performed without substantially slowing down the methodology remains an open question.

Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s11538-021-00918-z.

Acknowledgements   The authors thank the Associate Editor and three anonymous reviewers for helpful comments on an earlier draft of the manuscript.

6 Appendix

6.1 Calculation of \( \delta|S \) for 4-Taxon Trees

From Chifman and Kubatko (2015), as in the 3-taxon case \( P(\delta|S, \tau) \) is the inner product \( c^Tb \) where \( c \) is a vector of constants times a function of \( \tau \) that is of the form \( e^{-a\tau_1-b\tau_2-g\tau_3} \).

6.1.1 Asymmetric Case

From (Chifman and Kubatko 2015) we have the following, converting to coalescent units from mutation units and using 4 characters \( \{A, C, G, T\} \):

\[
\delta_k|S, \tau = \frac{1}{256} \left( c_0 + c_1 e^{-a\theta\tau_1} \frac{e^{-a\theta\tau_2}}{(1+a\theta)} + c_2 e^{-a\theta(\tau_1/2+\tau_2)} \right) + c_4 e^{-a\theta\tau_3} \frac{e^{-a\theta(\tau_1/2+\tau_3)}}{(1+a\theta)(1+2a\theta)} + c_5 e^{-a\theta(\tau_1/2+\tau_2/2+\tau_3)} \frac{e^{-a\theta(\tau_1+\tau_3)}}{(1+a\theta)(1+2a\theta)^2} + c_6 e^{-a\theta(\tau_1/2+\tau_2+\tau_3)} \frac{e^{-a\theta(\tau_1+\tau_3)}}{(1+a\theta)(1+2a\theta)^2(1+3a\theta)} + c_7 e^{-a\theta(\tau_2+\tau_3)-(\tau_2-\tau_1)} \frac{e^{-a\theta(\tau_2+\tau_3)}}{(1+a\theta)^2(1+2a\theta)^2(1+3a\theta)}
\]

where the coefficients are given by Table 4.

We have from the Yule model that \( \tau_1 \sim \text{Exp}(4\beta) \), \( \tau_2 - \tau_1 \sim \text{Exp}(3\beta) \), and \( \tau_3 - \tau_2 \sim \text{Exp}(2\beta) \). Then:

\[
f(\tau) = (4\beta)e^{-4\beta\tau_1} (3\beta)e^{-3\beta(\tau_2-\tau_1)} (2\beta)e^{-2\beta(\tau_3-\tau_2)} = (24\beta^3)e^{-\beta(\tau_1-\beta\tau_2-2\beta\tau_3)}
\]
### Table 4: Coefficients for each site pattern probability—asymmetric

| Site pattern | \(c_0\) | \(c_1\) | \(c_2\) | \(c_3\) | \(c_4\) | \(c_5\) | \(c_6\) | \(c_7\) | \(c_8\) | \(c_9\) |
|--------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| XXXX         | 1       | 3       | 6       | 12      | 9       | 12      | 9       | 24      | 48      | \((6\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XXXY         | 1       | -1      | 2       | -4      | 5       | -4      | -3      | 8       | -16     | \((-2\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XYXX         | 1       | 3       | -2      | -4      | 5       | 12      | -3      | -8      | -16     | \((-2\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| YXXX         | 1       | 3       | 6       | 12      | -3      | -4      | -3      | -8      | -16     | \((-2\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XYYY         | 1       | 3       | -2      | -4      | 1       | -4      | 9       | -8      | -16     | \(2\alpha\theta(4 + \alpha\theta)^2\) |
| XYXY         | 1       | -1      | 2       | -4      | 1       | -4      | 1       | -8      | 16      | \(\alpha\theta(32 + 40\alpha\theta + 10\alpha^2\theta^2)\) |
| XXYZ         | 1       | -1      | -2      | 4       | 1       | 4       | -3      | -8      | 16      | \(2\alpha\theta(4 + \alpha\theta)^2\) |
| YZXX         | 1       | 3       | -2      | -4      | -3      | -4      | -3      | 8       | 16      | \(2\alpha\theta(4 + \alpha\theta)^2\) |
| XYZX         | 1       | -1      | -2      | 4       | 1       | -4      | 1       | 0       | 0       | \(-\alpha\theta(4 + \alpha\theta)^2\) |
| YXXZ         | 1       | -1      | 2       | -4      | -3      | 4       | 1       | 0       | 0       | \(-\alpha\theta(4 + \alpha\theta)^2\) |
| XYZW         | 1       | -1      | -2      | 4       | -3      | 4       | 1       | 8       | -16     | \(2\alpha^2\theta^3\) |
Then, integrating over $f(\tau)$ gives:

\[
\int_0^\infty \int_0^{\tau_3} \int_0^{\tau_2} e^{-\alpha \tau_1 - b \tau_2 - \gamma \tau_3} (24\beta^3) e^{-\beta \tau_1 - \beta \tau_2 - 2\beta \tau_3} d\tau_1 d\tau_2 d\tau_3
\]

\[
= \int_0^\infty (24\beta^3) e^{-\gamma \tau_3(c+2\beta)} \left( \int_0^{\tau_3} e^{-\tau_2(b+\beta)} \left( \int_0^{\tau_2} e^{-\tau_1(a+\beta)} d\tau_1 \right) d\tau_2 \right) d\tau_3
\]

\[
= \frac{24\beta^3}{a + \beta} \int_0^\infty e^{-\gamma \tau_3(c+2\beta)} \left( \int_0^{\tau_3} e^{-\tau_2(b+\beta)} \left( 1 - e^{-\tau_2(a+\beta)} \right) d\tau_2 \right) d\tau_3
\]

\[
= \frac{24\beta^3}{a + \beta} \left( \frac{1}{(c + \beta)(b + \beta)} - \frac{1}{(b + \beta)(b + c + 3\beta)} - \frac{1}{(a + b + 2\beta)(c + 2\beta)} \frac{1}{(a + b + 2\beta)(a + b + c + 4\beta)} \right)
\]

\[
= \frac{24\beta^3}{(c + 2\beta)(b + c + 3\beta)(a + b + c + 4\beta)}
\]

Applying this formula to each term above gives:

\[
\delta_k | S = \frac{1}{256}(c_0 + c_1 \frac{4\beta}{(1 + \alpha \theta)(4\beta + \alpha \theta)}) + c_2 \frac{12\beta^2}{(1 + \alpha \theta)(3\beta + \alpha \theta)(4\beta + \alpha \theta)}
\]

\[
+ c_3 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(3\beta + \alpha \theta)(4\beta + 3\alpha \theta/2)}
\]

\[
+ c_4 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(3\beta + \alpha \theta)(4\beta + \alpha \theta)}
\]

\[
+ c_5 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + \alpha \theta)(4\beta + 3\alpha \theta/2)}
\]

\[
+ c_6 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + 2\alpha \theta)}
\]

\[
+ c_7 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + 3\alpha \theta/2)(4\beta + 3\alpha \theta/2)}
\]

\[
+ c_8 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + 3\alpha \theta/2)(4\beta + 2\alpha \theta)}
\]

\[
+ c_9 \frac{24\beta^3}{(1 + \alpha \theta)(2 + \alpha \theta)^2(3 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + 2\alpha \theta + 1)(4\beta + 2\alpha \theta)}
\]
From (Chifman and Kubatko 2015) we have the following, converting to coalescent units from mutation units and using 4 characters \{A, C, G, T\}:

\[
\delta_k = \frac{1}{256}(c_0 + c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 + c_8)
\]

where the coefficients are given by Table 5.

The prior differs from the asymmetric case in that \(\tau_1 < \tau_2\) and \(\tau_2 < \tau_1\) occurs with equal probability. The form of the integrals are the same in either case, just substituting the coefficients in front of each integral. As a result, integrating over the prior gives (where the only difference between the terms is the \(a\) or \(b\) in the middle factor of the denominator):

\[
\frac{12\beta^3}{(c + 2\beta)(b + c + 3\beta)(a + b + c + 4\beta)} + \frac{12\beta^3}{(c + 2\beta)(a + c + 3\beta)(a + b + c + 4\beta)}
\]

Applying this formula to each term above gives:

\[
\delta_k|S = \frac{1}{256}\left(\frac{c_0 + c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 + c_8}{\frac{2\beta}{4\beta + \alpha\theta} + \frac{6\beta^2}{(3\beta + \alpha\theta)(4\beta + \alpha\theta)}}\right)
\]

\[
+ c_3 \left(\frac{12\beta^2}{(1 + \alpha\theta)^2(3\beta + \alpha\theta)(4\beta + 2\alpha\theta)}\right)
\]

\[
+ c_4 \left(\frac{24\beta^3}{(1 + \alpha\theta)(2\beta + \alpha\theta)(3\beta + \alpha\theta)(4\beta + \alpha\theta)}\right)
\]

---

**Table 5** Coefficients for each site pattern probability—symmetric

| Site pattern | \(c_0\) | \(c_1\) | \(c_2\) | \(c_3\) | \(c_4\) | \(c_5\) | \(c_6\) | \(c_7\) | \(c_8\) |
|-------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| XXXX        | 1      | 3      | 3      | 9      | 12     | 24     | 24     | 48     | \((6\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XXYY        | 1      | -1     | 3      | -3     | 4      | -8     | 8      | -16    | \((-2\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XYXX        | 1      | 3      | -1     | 3      | 4      | -8     | 8      | -16    | \((2\alpha\theta)(4 + \alpha\theta)(4 + 3\alpha\theta)\) |
| XYXY        | 1      | -1     | -1     | 1      | 4      | -8     | -8     | -16    | \(\alpha\theta(32 + 40\alpha\theta + 10\alpha^2\theta^2)\) |
| XXYZ        | 1      | 3      | 3      | 9      | -4     | -8     | -8     | -16    | \(2\alpha\theta(4 + \alpha\theta)^2\) |
| YXZ        | 1      | -1     | -1     | -1     | 1      | 0      | 0      | 0      | \(-\alpha\theta(4 + \alpha\theta)^2\) |
| XYZW        | 1      | -1     | -1     | -1     | -4     | 8      | 8      | -16    | \(2\alpha^2\theta^3\) |
\[
\begin{align*}
&+ \frac{(c_5 + c_6)12\beta^2}{(1 + \alpha\theta)(2 + \alpha\theta)} \left( \frac{1}{(2\beta + \alpha\theta)(3\beta + \alpha\theta)(4\beta + 3\alpha\theta/2)} \right) \\
&+ \frac{1}{(2\beta + \alpha\theta)(3\beta + 3\alpha\theta/2)(4\beta + 3\alpha\theta/2)} \\
&+ c_7 \frac{24\beta^3}{(1 + \alpha\theta)(2 + \alpha\theta)(3 + \alpha\theta)(2\beta + \alpha\theta + 2)(3\beta + 2\alpha\theta + 1)(4\beta + 2\alpha\theta)} \\
&+ c_8 \frac{24\beta^3}{(1 + \alpha\theta)^2(2 + \alpha\theta)(3 + \alpha\theta)(2\beta + \alpha\theta + 2)(3\beta + 2\alpha\theta + 1)(4\beta + 2\alpha\theta)}
\end{align*}
\]

6.2 Calculation of $\delta(S, \tau)$ for 3-Taxon Trees

The fifteen site pattern probabilities shown above for the 4-taxon case can be mapped down to the five site patterns for three taxa by summing over the fourth taxon we are not interested in. Any of the four taxa can be used, however, we found it simplest to marginalize over the outgroup of the asymmetric tree (see Fig. 3) to get the following relationships (dropping the conditioning on $(S, \tau)$ for clarity):

\[
\begin{align*}
p_0 &= \delta_{XXX} = \delta_{XYY} + 3\delta_{YXX} \\
p_1 &= \delta_{YXX} = \delta_{XYY} + \delta_{XYX} + 2\delta_{XZZ} \\
p_2 &= \delta_{XYY} = \delta_{XXY} = \delta_{XYX} + 2\delta_{XZZ} = \delta_{XYX} + \delta_{XYY} + 2\delta_{YZZ} \\
p_3 &= \delta_{XYZ} = \delta_{XXY} + \delta_{YXX} + \delta_{XYZ} + \delta_{XZZ}
\end{align*}
\]

These probabilities take the form:

\[
\begin{align*}
p_0 &= c_0 + 3c_1 + 6c_2 + 12c_3 \\
p_1 &= c_0 + 3c_1 - 2c_2 - 4c_3 \\
p_2 &= c_0 - c_1 + 2c_2 - 4c_3 \\
p_3 &= c_0 - c_1 - 2c_2 + 4c_3
\end{align*}
\]

where (measuring $\tau_1$ and $\tau_2$ in coalescent units):

\[
\begin{align*}
c_0 &= \frac{1}{64} \\
c_1 &= \frac{e^{-\tau_1\alpha\theta}}{64(1 + \alpha\theta)} \\
c_2 &= \frac{e^{-\tau_2\alpha\theta}}{64(1 + \alpha\theta)} \\
c_3 &= \frac{e^{-\tau_1\alpha\theta/2}e^{-\tau_2\alpha\theta}}{64(1 + 2\alpha\theta)(1 + \alpha\theta)}
\end{align*}
\]

It is easily verified that:

\[
4p_0 + 12p_1 + 24p_2 + 24p_3 = 1
\]
The coefficients in Eq. 8 arise as follows. $X$ can represent any of $A$, $C$, $G$, or $T$. $Y$ can represent any of the remaining three characters, and $Z$ any of the remaining two. Finally, the coefficient of $p_2$ is doubled to account for both $\delta_{XYX}$ and $\delta_{XXY}$.

### 6.3 Calculation of $\delta|S$ for 3-Taxon Trees

One can see from Eq. 6 that $\delta_k(S, \tau)$ is a linear combination of terms, so $\delta_k|S$ is also linear combination of terms. Given $\tau_1 \sim Exp(3\beta)$ and $(\tau_2 - \tau_1) \sim Exp(2\beta)$, the prior is:

$$f(\tau|\beta) = 3\beta e^{-3\beta \tau_1} 2\beta e^{-2\beta(\tau_2 - \tau_1)} = 6\beta^2 e^{-\beta \tau_1 - 2\beta \tau_2}$$

Then to integrate over $f(\tau|\beta)$, each term has the form:

$$\int_0^\infty \int_0^{\tau_2} e^{-\alpha \tau_1} e^{-b \tau_2} 6\beta^2 e^{-\beta \tau_1 - 2\beta \tau_2} d\tau_1 d\tau_2$$

$$= \int_0^\infty 6\beta^2 e^{-\tau_2(b+2\beta)} (\int_0^{\tau_2} e^{-\tau_1(a+\beta)} d\tau_1) d\tau_2$$

$$= \frac{6\beta^2}{a + \beta} \int_0^\infty 6\beta^2 e^{-\tau_2(b+2\beta)} (1 - e^{-\tau_2(a+\beta)}) d\tau_2$$

$$= \frac{6\beta^2}{a + \beta} \left( \frac{1}{b + 2\beta} - \frac{1}{a + b + 2\beta} \right)$$

$$= \frac{6\beta^2}{(b + 2\beta)(a + b + 3\beta)}$$

Taken together, $\delta_k|S$ has the linear form of Eq. 6 with the following terms replacing those of Eq. 7:

$$\begin{cases}
  c_0 = 1/64 \\
  c_1 = \frac{3\beta}{(1 + \alpha \theta)(3\beta + \alpha \theta)} \\
  c_2 = \frac{6\beta^2}{(1 + \alpha \theta)(2\beta + \alpha \theta)(3\beta + \alpha \theta)} \\
  c_3 = \frac{3\beta}{(1 + \alpha \theta)(1 + 2\alpha \theta)(2\beta + \alpha \theta)(3\beta + 3\alpha \theta/2)}
\end{cases}
$$

where $\alpha = 4/3$ from the JC69 model. Comparison to the results in the supplementary material reveals the computational advantages of the Yule model approach.
6.4 Proof of Theorem 1

Proof Without loss of generality, let $S_0 = (a, (b, c))$ be the true tree and $\tau_0$ be the true value of $\tau$. We want to show that $\forall \epsilon > 0 \exists J_0 : \forall J > J_0 \ P(\hat{S} = S) > 1 - \epsilon$.

First, note that for all values of $(\tau, \theta)$, we have $\delta_{XXY} = \delta_{XYX}$.

Second, we note using the results of Eq. 6

$$\delta_{YXX}|((a, (b, c)), \tau) > \delta_{XYX}|((a, (b, c)), \tau)$$

which holds w.p.1 by assumption.

Next, note

$$\int_{\tau} \delta_{YXX}|((a, (b, c)), \tau) f(\tau)d(\tau) = \delta_{XYX}|((a, (b, c)), \tau) f(\tau)d(\tau)$$

by properties of expectations from Eq. 10 and since $f(\tau) > 0$ a.e. under the prior.

Since each topology $(a, (b, c)), (b, (a, c)),$ and $(c, (a, b))$ has a prior probability of $1/3$, the maximum posterior topology will be the maximum likelihood topology. Recall from Sect. 2.4 that we can permute any site pattern probability given one topology to find a site pattern probability given another topology—e.g., $\delta_{YXX}|((a, (b, c)) = \delta_{XYX}|((b, (a, c)))$. So, the likelihood of trees $(a, (b, c))$ and $(b, (a, c))$ given the data are:

$$L((a, (b, c))|d) = c_d (p_{xxx})^{d_{xxx}} (p_{xxy})^{d_{xxy}} (p_{xyx})^{d_{xyx}} (p_{xyz})^{d_{xyz}}$$
$$L((b, (a, c))|d) = c_d (p_{xxx})^{d_{xxx}} (p_{xxy})^{d_{xxy}} (p_{xyx})^{d_{xyx}} (p_{xyz})^{d_{xyz}}$$

By cancelling terms in common, one can see that the likelihood ratio comes down to permuting the number of sites that follow the XYX and YXX patterns:

$$\frac{L((a, (b, c))|d)}{L((b, (a, c))|d)} = (p_{xyx})^{d_{xyx} - d_{yxx}} (p_{xxy})^{d_{xyx} - d_{yxx}}$$

Together $p_{yxx} > p_{xyx}$ and $d_{yyy} > d_{yxx}$ imply that $\frac{L((a, (b, c))|d)}{L((b, (a, c))|d)} > 1$, which implies that $\hat{S} = S$. As we have shown that $p_{yxx} > p_{xyx}$, it is sufficient to show that $d_{yxx} \rightarrow p_{yxx}$ and $d_{xyx} \rightarrow p_{xyx}$ for sufficiently large number of sites.
Let $p_{y|x,0} = \delta_{y|x}|((a, (b, c)), \tau_0)$ and $p_{x|y,0} = \delta_{x|y}|((a, (b, c)), \tau_0)$. From the SLLN, we have

$$\forall \epsilon_1, \delta_1 > 0 \exists J_1 : \forall J > J_1 P\left(\frac{d_{y|x,J}}{J} > p_{y|x,0} - \delta_1\right) > 1 - \epsilon_1$$

$$\forall \epsilon_2, \delta_2 > 0 \exists J_2 : \forall J > J_2 P\left(\frac{d_{x|y,J}}{J} < p_{x|y,0} + \delta_2\right) > 1 - \epsilon_2$$

Choose $\delta_1, \delta_2$ such that $\delta_1 + \delta_2 < p_{y|x,0} - p_{x|y,0}$, $\epsilon_1 = \epsilon_2 = \epsilon / 2$, and $J_0 = \max\{J_1, J_2\}$ by applying the Bonferroni inequality we have that $P\left(\frac{d_{y|x,J}}{J} > p_{y|x,0} - \delta_1, \frac{d_{x|y,J}}{J} > \delta_2\right) > 1 - \epsilon$ and the result holds.

---

**References**

Avni E, Cohen R, Snir S (2015) Weighted Quartets Phylogenetics. Systematic Biology 64(2):233–242

Chifman J, Kubatko L (2014) Quartet Inference from SNP Data Under the Coalescent Model. Bioinformatics 30(23):3317–3324

Chifman J, Kubatko L (2015) Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. Journal of Theoretical Biology 374(1):35–47

Chou J, Gupta A, Yaduvanshi S, Davidson R, Nute M, Mirarab S, Warnow T (2015) A comparative study of SVD quartets and other coalescent-based species tree estimation methods. BMC Genomics 16(S2). https://doi.org/10.1186/1471-2164-16-S10-S2

DeGiorgio M, Degnan JH (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. Mol Biol Evol 27(3):552–569

Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution 59(1):24–37

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7(214):1. https://doi.org/10.1186/1471-2148-7-214

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26

Gatesy J, Meredith RW, Janega JE, Simmons MP, Murphy WJ, Springer MS (2017) Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. Cladistics 33:295–332

Gatesy J, Springer MS (2014) Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatenation conundrum. Mol Phylogenetics Evol 80:231–266

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab 3(1):44–77

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res 21:349–356

Hudson RR (2003) Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18(2):337–338

Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (Phepila) inferred from thirty gene trees. Evolution 59(9):2033–2047

Jukes TH, Cantor CR (1969) Evolution of protein molecules. Academic Press, New York

Kingman JFC (1982) The coalescent. Stoch Process Appl 13(3):235–248

Kopp A, Barmina O (2005) Evolutionary history of the drosophila bipectinata species complex. Genetical Res 85(1):23–46

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25(7):971–973

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56(1):14–24
Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56(3):504–514
Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. https://doi.org/10.1186/1471-2148-10-302
Mahim M, Zahir W, Rezwana R, Bayzid MS (2020) wQFM: statistically consistent genome-scale species tree estimation from weighted quartets. bioRxiv. https://www.biorxiv.org/content/early/2020/12/01/2020.11.30.403352
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):i541–i548
Mirarab S, Warnow T (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52
Ogulvie HA, Bouckaert RR, Drummond AJ (2017) StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol 34(8):2101–2114
Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528
Peng J, Swofford D, Kubatko L (2021) Estimation of speciation times under the multispecies coalescent (in review)
Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650
Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13(3):235–238
Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164(4):1645–1656
Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53(1):131–147
Roch S, Steel M (2015) Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol 100c:56–62
Rokas A, Williams BL, Carroll S (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804
Rosenberg NA (2007) Counting coalescent histories. J Comput Biol 14(3):360–377
Salter L (2001) Complexity of the likelihood surface for a large DNA data set. Syst Biol 50(6):970–978
Schliep KP (2011) Phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592–593
Sevilla G, Frenkel Z, Snir S (2016) TripletMaxCut: a new toolkit for rooted supertree. Methods Ecol Evol 7(11):1359–1365
Springer MS, Gatesy J (2016) Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatenation conundrum. Mol Phylogenetics Evol 94:1–33
Stadler T, Steel M (2011) Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J Theor Biol 297:33–40
Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313, 01
Steel M, Penny D (1993) Distribution of tree comparison metrics—some new results. Syst Biol 42(2):136–141
Swofford DL (2003) Paup*. phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates. Sunderland, Massachusetts
Thawornwattana Y, Dalquen D, Yang Z (2018) Coalescent analysis of phylogenomic data confidently resolves the species relationships in the anopheles gambiae species complex. Mol Biol Evol 35(10):2512–2527
Wakeley J (2009) Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village
Wасher M, Kubatko L (2021) Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. Syst Biol 70(1):33–48
Wen D, Nakheh L (2018) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst Biol 67(1):439–457
Whidden C, Matsen IV FA (2015) Quantifying MCMC exploration of phylogenetic tree space. Syst Biol 64(3):472–491
Yang Z (2014) Molecular evolution: a statistical approach. Oxford University Press, New York
Yang Z (2015) The BPP program for species tree estimation and species delimitation. Curr Zool 61(5):854–865
Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14:717–724
Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(Supp 6):15–30

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.