A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree

Tanja Stadler[1] & James H. Degnan[2,3]
[1]Institute of Integrative Biology Universitätsstrasse 16, 8092, Zürich, Switzerland
[2]Dept. of Mathematics and Statistics, Private Bag 4800, University of Canterbury
Christchurch 8140 New Zealand
[3]National Institute of Mathematical and Biological Synthesis, Knoxville, Tennessee, USA

March 2, 2012

Abstract

In this paper, we provide a polynomial time algorithm to calculate the probability of a *ranked* gene tree topology for a given species tree, where a ranked tree topology is a tree topology with the internal vertices being ordered. The probability of a gene tree topology can thus be calculated in polynomial time if the number of orderings of the internal vertices is a polynomial number. However, the complexity of calculating the probability of a gene tree topology with an exponential number of rankings for a given species tree remains unknown.

1 Introduction

Phylogenetic reconstruction methods aim to infer the species phylogeny which gave rise to a group of extant species. Typically, this species phylogeny is obtained based on genetic data from representative individuals of each extant species. The ancestries of genes at different loci form gene trees which do not necessarily have the same topology as the species tree. Gene tree topologies and species tree topologies might be different due to such phenomena as incomplete lineage sorting, gene duplication, recombination within gene loci, and horizontal gene transfer [4]. In this paper, we focus on incomplete lineage sorting as the mechanism for incongruence of gene tree and species tree topologies, in which two gene lineages do not coalesce in the most recent population ancestral to the individuals from which the genes were sampled. As an example, the lineages sampled from species A and B in Figure 1b do not coalesce until the population ancestral to species A, B, and C, thus allowing the B and C lineages in the gene tree to have a more recent common ancestor than lineages A and B.

Given a fixed species tree, and assuming the gene tree evolved under the multi-species coalescent [4], the most probable gene tree topology can have a different topology from that of the species tree. Such a gene tree topology is called an anomalous gene tree. In fact, for every species tree topology with at least 5 leaves, we can choose edge lengths in the species tree topology such that anomalous gene trees exist [3]. This implies that the gene tree topology appearing most often when considering different genes might not agree with the species tree topology, thus we cannot use a simple majority-heuristic to infer the species tree from a collection of gene trees. Instead we need statistical tools rather than majority rule heuristics for inferring the species tree based on gene trees.

Current methods for inferring species trees from gene trees in this setting can be divided into topology-based and genealogy-based methods, in which the input for a reconstruction algorithm accepts either gene tree topologies or genealogies, i.e., gene trees with branch lengths (coalescence times). Topology-based methods include Minimize Deep Coalescence (MDC) [19, 28], STAR [18], STELLS [33], rooted triple consensus [9] and other consensus and supertree methods [2, 32]. Genealogy-based methods include Bayesian and likelihood methods such as BEST, *BEAST, and STEM [11, 13, 14] and clustering and distance-based methods [15, 17, 18, 20]. Possible pros and cons of the two approaches are that topology-based methods can be computationally faster and less sensitive to errors in estimating gene trees (and gene tree branch lengths) from sequence data [12], while methods that use coalescence times, particularly using Bayesian modelling, can be the most accurate when model assumptions are correct [16].

Another possibility that has been so far unexplored in methods for inferring species trees from gene trees is to use *ranked* gene trees, in which the temporal order of the nodes of the gene tree (the coalescence times) is used, but not the continuous-valued branch lengths. This approach might therefore be intermediate between purely topology-based methods and genealogy-based methods. By preserving more of the temporal information in the gene tree nodes, the hope is to develop methods that

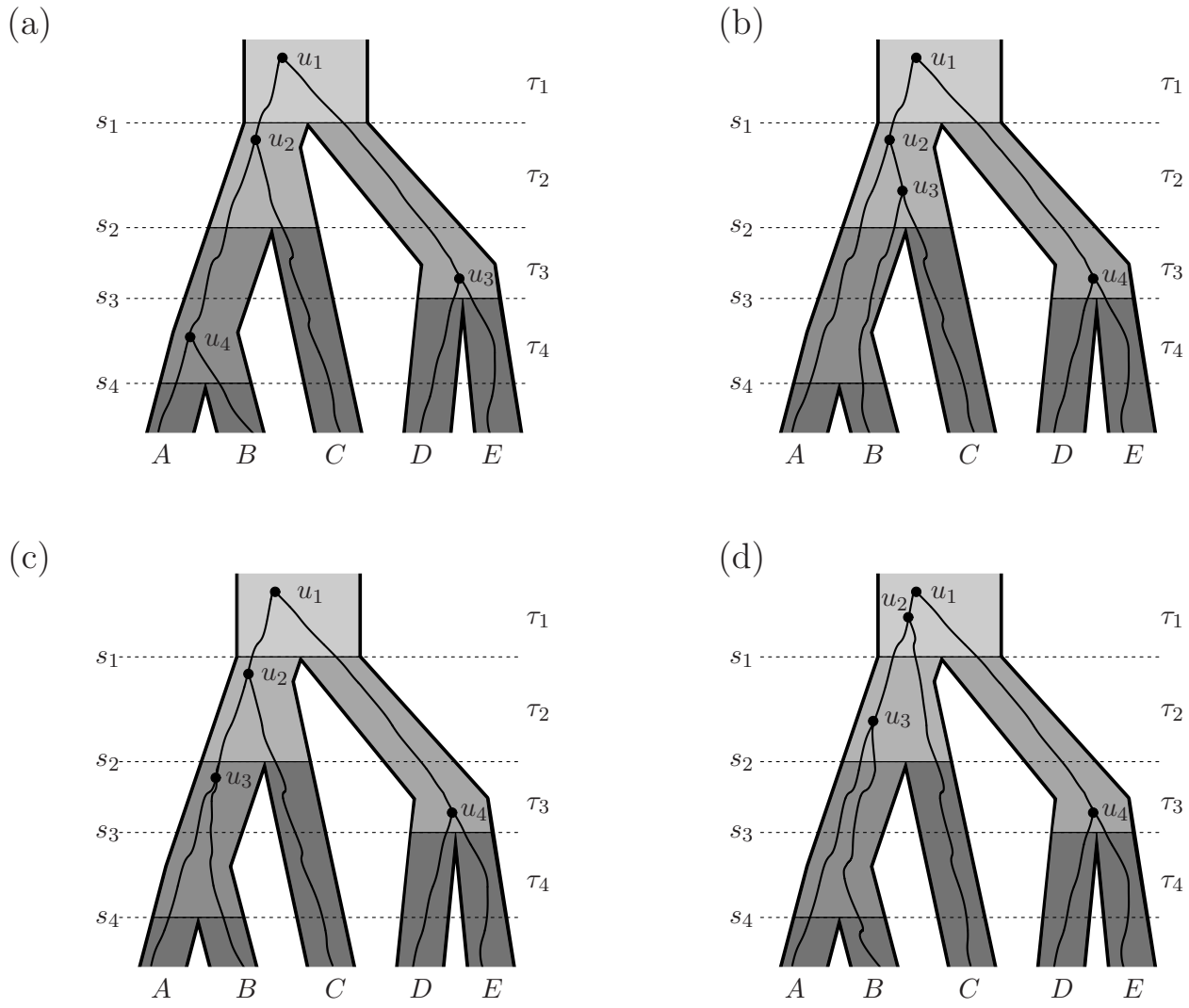

Figure 1: In (a)–(d) the ranked species tree topology is $(((A, B)_4, C)_2, (D, E)_3)_1$. (a) The ranked gene tree matches the ranked species tree. (b) The (ranked or unranked) gene tree does not match the species tree, and there is an incomplete lineage sorting event (a deep coalescence) because the lineages from species A and B fail to coalesce more recently than s_2. (c) The gene tree and species tree have the same unranked topology but have different ranked topologies, as D and E coalesce in the gene tree more recently than A and B, while A and B is the most recent divergence in the species tree. The gene tree in (c) has ranked topology $(((A, B)_3, C)_2, (D, E)_4)_1$. In (c), there are no incomplete lineage sorting events (no deep coalescences); however, there is an extra lineage at time s_3 which leads to the gene tree and species tree having different rankings. In (c), all coalescences occur in the most recent possible interval consistent with the ranked gene tree, and we have $\ell_1 = 2, \ell_2 = 3, \ell_3 = 5, \ell_4 = 5$, and $g_1 = 2$, $g_2 = 3$, $g_3 = 5$, $g_4 = 5$. (d) A gene tree with the same ranked topology as the gene tree in (c) but with coalescences occurring in different intervals.

are more powerful than purely topology-based methods and that are still computationally efficient and robust to errors in estimating gene trees and gene tree branch lengths from sequence data.

In [5], a first step toward developing methods that use ranked gene trees for inferring species trees was taken by providing formulae to calculate the probability of a ranked gene tree given a species tree. The previous work, however, was based on an exponential enumeration of what were called *ranked coalescent histories* and did not provide an algorithm for computing some of the key terms in the probability of individual ranked histories. In this paper, we improve this previous (computationally inefficient) approach, by providing a method for computing probabilities of ranked gene trees given species trees which is polynomial in the number of leaves using a dynamic programming approach.

Methods for computing probabilities of ranked gene trees efficiently may also be of interest in the context of computing probabilities of unranked gene trees, partic-

Table 1: Notation used in the paper

Symbol	meaning
$\mathcal{T}$	species tree with real-valued divergence times
$\mathscr{G}$	ranked gene tree (real-valued coalescence times not specified)
n	the number of leaves of $\mathcal{T}$ and $\mathscr{G}$
s_i	speciation times, with $s_1 > \cdots > s_{n-1}$, let $s_0 = \infty$
τ_i	intervals between speciation times, $\tau_i = [s_i, s_{i-1})$
ℓ_i	the number of gene tree lineages at time s_i
m_i	the number of coalescence events in interval τ_i
$\mathscr{G}_{i,\ell_i}$	the ranked gene tree observed from time 0 to time s_i
g_i	the minimum number of gene tree lineages at time s_i
$y_{i,z}$	population z in interval τ_i in beaded tree
u_i	internal node (coalescence) with rank i in the gene tree, u_1 is most ancient, u_{n-1} is the most recent
$k_{i,j,z}$	the number of lineages available for coalescence in population $y_{i,z}$ just after the jth coalescence (considered forward in time) in interval τ_i; $k_{i,0,z}$ is the number of lineages "exiting" at time s_{i-1}
$\delta(y), \delta(u)$	the set of leaves descended from a node of the species tree or gene tree, respectively
$\mathrm{lca}(u)$	for a node u of the gene tree, the node y of the species tree with largest rank such that $\delta(u) \subset \delta(y)$
$\tau(y)$	for a node y with rank i on the species tree, we denote $\tau(y) = \tau_i$ (the interval immediately above y)
$\lambda_{i,j}$	the overall coalescence rate in interval τ_i immediately preceding (backwards in time) the jth coalescence
h_k^1	number of sequences of coalescences above the root of the species tree starting with k lineages
f_i	the joint density of coalescence times in interval τ_i

ularly because no polynomial time algorithm has been found for calculating the probability of a gene tree topology given a species tree under the multispecies coalescent [6, 23, 29, 33]. The probability of an unranked gene tree topology can be obtained by summing over all ranked gene tree topologies with the same topology. Thus, for unranked gene trees with particular shapes where the number of rankings increases in polynomial time, using ranked gene trees can potentially increase the speed of computing probabilities of unranked gene trees as well. We note that a completely unbalanced gene tree has only one ranking, while the number of rankings can be exponential in the number of leaves when gene trees become more balanced. Thus, our approach for calculating unranked gene tree probabilities will be most useful for less balanced ranked gene trees.

The bulk of the paper consists of the derivation of the polynomial time method for computing ranked gene tree probabilities. The algorithm is summarized in section 2.2. This is followed by a discussion of applications to computing probabilities of unranked gene tree topologies and to inferring ranked species trees under maximum likelihood and a modification to the MDC criterion.

2 Calculating the probability of a ranked gene tree topology

In the following, we will derive the probability of a ranked gene tree topology given a species tree, $\mathbb{P}[\mathscr{G} \mid \mathcal{T}]$. Equations (1,2,3,4,8,10) allow the calculation of $\mathbb{P}[\mathscr{G} \mid \mathcal{T}]$ in time $O(n^5)$. The model giving rise to the gene tree is the multi-species coalescent with constant population sizes [4]. Each species consists of a population of constant size where lineages merge according to the coalescent. Thus, lineages from two different species may coalesce any time previous to the split of the two species.

We begin with some notation, which is also summarized in Table 1. Let time be 0 today and increasing going into the past. Let $\mathcal{T}$ be a species tree with n species, and thus $n-1$ speciation events (denoted by $1, \ldots, n-1$) occurring at times $s_1 > \cdots > s_{n-1}$. Denote the interval between speciation event $i-1$ and speciation event i by τ_i, see Figure 1.

Let $\mathscr{G}$ be a ranked gene tree topology. It is convenient to use the same labels for the leaves of $\mathscr{G}$ and of $\mathcal{T}$. This is a slight abuse of notation, as leaf A of $\mathcal{T}$ refers to a population (or species), and A of $\mathscr{G}$ refers to a gene sampled from population A. We denote the nodes of $\mathscr{G}$ (which are coalescence events) by $u_1, \ldots, u_{n-1}$, where node u_j has rank j, and where higher rank indicates a more recent coalescence. A ranked tree topology can be

notated similarly to Newick notation, putting the rank as a subscript for each node, see also Figure 1.

Let $\mathscr{G}_{i,\ell_i}$ be part of a ranked gene tree evolving on a species tree between time s_i and time 0 (i.e. the present). $\mathscr{G}_{i,\ell_i}$ consists of ℓ_i gene tree lineages at speciation time s_i and the coalescent history of $\mathscr{G}_{i,\ell_i}$ in time interval $(0, s_i)$ is consistent with the ranked gene tree $\mathscr{G}$. Let g_i be the minimum number of lineages required in the ranked gene tree at time s_i such that $\mathscr{G}$ can be embedded into the species tree $\mathcal{T}$. Note that $n \geq \ell_i \geq g_i > i$. Next we provide a dynamic programming approach for calculating the probability of a ranked gene tree given a species tree. An efficient way to determine the required quantities $g_1, \ldots, g_{n-1}$ is provided in Section 2.1.

Essentially, in our approach, we traverse the intervals between speciation events going back in time, $\tau_{n-1}, \ldots, \tau_2$ (formalized in Theorem 2), and calculate the probability of the appropriate coalescent events occuring in interval τ_i based on how many coalescent events happened in the later intervals $\tau_{i+1}, \ldots, \tau_{n-1}$ (Theorem 3). Finally with Theorem 1, we account for the most ancetral time interval τ_1.

Theorem 1. *The probability of a ranked gene tree given a species tree is,*

$$\mathbb{P}[\mathscr{G} \,|\, \mathcal{T}] = \sum_{\ell_1 = g_1}^{n} \mathbb{P}[\mathscr{G}_{1,\ell_1} \,|\, \mathcal{T}]/H_{\ell_1} \tag{1}$$

where

$$H_{\ell_1} = \ell_1!(\ell_1 - 1)!/2^{\ell_1 - 1} \tag{2}$$

is the probability for the coalescences above the root appearing in the right order [8].

For precalculated $\mathbb{P}[\mathscr{G}_{1,\ell_1}|\mathcal{T}]$ $(\ell_1 = 2, \ldots, n)$ the complexity of calculating $\mathbb{P}[\mathscr{G} \,|\, \mathcal{T}]$ is thus $O(n)$. Next, we will provide a recursive way to calculate $\mathbb{P}[\mathscr{G}_{1,\ell_1}|\mathcal{T}]$ for $\ell_1 = 2, \ldots, n$ in polynomial time, thus $\mathbb{P}[\mathscr{G} \,|\, \mathcal{T}]$ can be calculated in polynomial time.

Theorem 2. *The probability $\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathcal{T}]$ can be calculated for all i recursively (with $l_i \geq g_i$),*

$$\begin{aligned} &\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathcal{T}] \\ &= \sum_{\ell_{i+1}=\max(\ell_i, g_{i+1})}^{n} \mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathscr{G}_{i+1,\ell_{i+1}}, \mathcal{T}]\mathbb{P}[\mathscr{G}_{i+1,\ell_{i+1}}|\mathcal{T}] \end{aligned} \tag{3}$$

with

$$\mathbb{P}[\mathscr{G}_{n-1,n} \,|\, \mathcal{T}] = 1.$$

The complexity of calculating $\mathbb{P}[\mathscr{G}_{1,\ell_1} \,|\, \mathcal{T}]$ for $\ell_1 = 2, \ldots, n$ is $O(n^3)$, given we know $\mathbb{P}[\mathscr{G}_{i,\ell_i} \,|\, \mathscr{G}_{i+1,\ell_{i+1}}, \mathcal{T}]$ for all i, ℓ_i, ℓ_{i+1}.

Proof. At the time of the most recent speciation event, s_{n-1}, we have n lineages with probability 1, which is the initial value of the recursion. Calculating $\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathcal{T}]$ for $i < n-1$ can be done in the following way,

$$\begin{aligned} &\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathcal{T}] \\ &= \sum_{\ell_{i+1}=\max(\ell_i, g_{i+1})}^{n} \mathbb{P}[\mathscr{G}_{i,\ell_i}, \mathscr{G}_{i+1,\ell_{i+1}}|\mathcal{T}] \\ &= \sum_{\ell_{i+1}=\max(\ell_i, g_{i+1})}^{n} \mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathscr{G}_{i+1,\ell_{i+1}}, \mathcal{T}]\mathbb{P}[\mathscr{G}_{i+1,\ell_{i+1}}|\mathcal{T}]. \end{aligned}$$

Suppose $\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathscr{G}_{i+1,\ell_{i+1}}, \mathcal{T}]$ is known. Given we calculated the probability $\mathbb{P}[\mathscr{G}_{i+1,\ell_{i+1}}|\mathcal{T}]$ for $\ell_{i+1} = i+2, \ldots, n$, then calculating $\mathbb{P}[\mathscr{G}_{i,\ell_i}|\mathcal{T}]$ for $\ell_i = i+1, \ldots, n$ requires $O(\sum_{j=1}^{n-i} j) = O(\binom{n-i+1}{2})$ calculations. Summing up over $i = 1, \ldots, n-1$ yields a complexity of $O(\sum_{i=2}^{n} \binom{i}{2}) = O(\binom{n+1}{3}) = O(n^3)$. □

It remains to determine $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$. Note that during the interval τ_i, we have i branches in the species tree. Let m_i be the number of coalescent events in τ_i, so $m_i = \ell_i - \ell_{i-1}$. Let the number of lineages on branch z just after the jth coalescent event (going forward in time) in τ_i be $k_{i,j,z}$. Calculation of $k_{i,j,z}$ can be done efficiently as shown in Section 2.1.

Theorem 3. *We have,*

$$\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}] = \sum_{j=0}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1} - s_i)}}{\prod_{k=0, k\neq j}^{m_i}(\lambda_{i,k} - \lambda_{i,j})} \tag{4}$$

where $\lambda_{i,j} = \sum_{z=1}^{i} \binom{k_{i,j,z}}{2}$ and $\binom{1}{2} := 0$.

Proof. The density for the coalescence events in interval τ_i can be obtained by considering the waiting time to the "next" coalescent event (going backwards in time) as being due to competing exponentials in the different branches, where the coalescence rate within branch z is $\binom{k_{i,j,z}}{2}$. Thus, the waiting time until the next coalescent event has rate $\lambda_{i,j} = \sum_{z=1}^{i} \binom{k_{i,j,z}}{2}$.

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

The density for the coalescent events in the interval τ_i is [5],

$$\begin{aligned} f_i(v_0, v_1, \ldots, v_{m_i}) &= e^{-\sum_{j=0}^{m_i} \sum_{z=1}^{i} \binom{k_{i,j,z}}{2} v_j} \\ &= e^{-\sum_{j=0}^{m_i} \lambda_{i,j} v_j}. \end{aligned}$$

It remains to integrate over v, for which we distinguish between case (i) $\lambda_{i,0} = 0$, and case (ii) $\lambda_{i,0} > 0$.

Case (i): If $\lambda_{i,0} = 0$ (which occurs if $\ell_{i-1} = i$, i.e., all lineages within each population coalesce), then we rewrite f_i as,

$$f_i(v_0, v_1, \ldots, v_{m_i}) = \frac{\prod_{j=1}^{m_i} \lambda_{i,j} e^{-\lambda_{i,j} v_j}}{\prod_{j=1}^{m_i} \lambda_{i,j}}. \tag{5}$$

Using the fact that the integral of the numerator of Equation (5) is a hypoexponential distribution based on the sum of m_i exponential random variables [24] (with density functions $\lambda_{i,j}e^{-\lambda_{i,j}v_j}$, $j = 1, \ldots, m_i$), the probability of the coalescent events in the interval is the *cumulative distribution function* of the hypoexponential distribution evaluated at $s_{i-1} - s_i = \sum_{j=0}^{m_i} v_i$. Thus, with $\lambda_{i,j} < \lambda_{i,j+1}$,

$$\begin{aligned}
&\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}] \\
&\quad = \frac{1}{\prod_{j=1}^{m_i}\lambda_{i,j}} - \sum_{j=1}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\lambda_{i,j}\prod_{k=1,k\neq j}^{m_i}(\lambda_{i,k}-\lambda_{i,j})} \\
&\quad = \frac{1}{\prod_{j=1}^{m_i}\lambda_{i,j}} + \sum_{j=1}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i}(\lambda_{i,k}-\lambda_{i,j})} \\
&\quad = \sum_{j=0}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i}(\lambda_{i,k}-\lambda_{i,j})}. \qquad (6)
\end{aligned}$$

where the second line follows because $-\lambda_{i,j} = \lambda_{i,0} - \lambda_{i,j}$.

Case (ii): If $\lambda_{i,0} > 0$, then we rewrite f_i as,

$$f_i(v_0, v_1, \ldots, v_{m_i}) = \frac{\prod_{j=0}^{m_i}\lambda_{i,j}e^{-\lambda_{i,j}v_j}}{\prod_{j=0}^{m_i}\lambda_{i,j}} \qquad (7)$$

For integrating f_i, we use the fact that the integral of the numerator in Equation (7) is the convolution of $m_i + 1$ exponential random variables with parameters $\lambda_{i,0}, \ldots, \lambda_{i,m_i}$, which is the hypoexponential distribution. Now, since $\lambda_{i,j} < \lambda_{i,j+1}$, we observe, using the *probability density function* of the hypoexponential distribution,

$$\begin{aligned}
&\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}] \\
&\quad = \int_v f_i(v_0, v_1, \ldots, v_{m_i})\, dv \\
&\quad = \sum_{j=0}^{m_i} \frac{e^{-\lambda_{i,j}(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i}(\lambda_{i,k}-\lambda_{i,j})},
\end{aligned}$$

which is the same expression as for the $\lambda_{i,0} = 0$ case (6). Note that for case (i) we made use of the cumulative distribution function of the hypoexponential distribution, while for case (ii) we made use of the density function of the hypoexponential distribution. Both cases yield the same final expression for $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$, which establishes the proof. □

Corollary 4. *The probabilities $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$ for all possible i, m_i and ℓ_i (recall that $m_i = \ell_i - \ell_{i-1}$) are calculated in $O(n^5)$, given all $\lambda_{i,j}$.*

Proof. For a fixed i, m_i and ℓ_i, we require $O(m_i^2)$ calculations to evaluate $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$. We need to determine $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$ for all possible i, m_i and ℓ_i. First, we observe that $i \leq \ell_{i-1} \leq n$, and thus for a fixed ℓ_i, we have, $0 \leq m_i \leq \ell_i - i$. Second, $i < \ell_i \leq n$. And third, $2 \leq i \leq n-1$. Thus, the number of calculations needed to calculate $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$ for all possible i, m_i and ℓ_i is,

$$\begin{aligned}
O\left(\sum_{i=2}^{n-1}\sum_{\ell_i=i+1}^{n}\sum_{m_i=0}^{\ell_i-i} m_i^2\right) &= O\left(\sum_{i=2}^{n-1}\sum_{\ell_i=i+1}^{n}(\ell_i - i)^3\right) \\
&= O\left(\sum_{i=2}^{n-1}(n-i)^4\right) \\
&= O\left(n^5\right).
\end{aligned}$$

□

Corollary 5. *The quantities $\lambda_{i,j}$ can be calculated for all possible i, m_i, ℓ_i and j in $O(n^5)$, given all $k_{i,j,z}$.*

Proof. For a fixed i, m_i, ℓ_i and j, we require $O(i)$ calculations to evaluate $\lambda_{i,j}$. As $j = 0, \ldots, m_i$, with the same arguments as in Corollary 4, we obtain,

$$\begin{aligned}
O\left(\sum_{i=2}^{n-1}\sum_{\ell_i=i+1}^{n}\sum_{m_i=0}^{\ell_i-i}\sum_{j=0}^{m_i} i\right) &= O\left(\sum_{i=2}^{n-1} i \sum_{\ell_i=i+1}^{n}\sum_{m_i=0}^{\ell_i-i} m_i\right) \\
&= O\left(\sum_{i=2}^{n-1} i \sum_{\ell_i=i+1}^{n}(l_i-1)^2\right) \\
&= O\left(\sum_{i=2}^{n-1} i(n-i)^3\right) \\
&= O\left(n^5\right).
\end{aligned}$$

□

We note that the terms $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$ are analogous to the functions $g_{i,j}$ defined in [27, 31], which give the probability that i lineages coalesce into j within time t in a single population and are used extensively in computing probabilities related to unranked gene trees [6, 21, 22, 33]. In particular, if only one population, say z^*, has coalescence events, then we have

$$\begin{aligned}
&\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}] \\
&\quad = \frac{g_{\ell_{i+1},\ell_i}(s_i - s_{i+1})\prod_{z\neq z^*} g_{k_{i,0,z},k_{i,0,z}}(s_i - s_{i+1})}{\prod_{k=1}^{\ell_{i+1}-\ell_i}\binom{\ell_{i+1}-k+1}{2}},
\end{aligned}$$

a product of $g_{i,j}$ functions with the denominator counting the number of sequences in which m_i coalescences could have occurred. The terms $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}}|\mathscr{G}_{i,\ell_i}, \mathcal{T}]$ allow for the coalescences to occur in separate populations, however, and are constrained by the ranking of the gene tree. For example, in interval τ_3 of Figure 1c, there are two coalescences which occur in different populations. If the ranking of the gene tree were

not important, the branches could be considered independent, and the probability of this event would be $g_{2,1}(s_2-s_3)g_{2,1}(s_2-s_3)$. However, the gene tree ranking constrains the coalescence of A and B to be less recent than that of D and E, so the probability for events in this interval is,

$$\mathbb{P}[\mathscr{G}_{3,2}|\mathscr{G}_{4,3},\mathcal{T}] = [g_{2,1}(s_2-s_3)]^2/2.$$

We illustrate that we get the same result from Theorem 3: there are two coalescence events in interval τ_3, so we use $j=0,1,2$, and calculate

$$\lambda_{3,0} = \binom{1}{2}+\binom{1}{2}+\binom{1}{2} = 0,$$
$$\lambda_{3,1} = \binom{2}{2}+\binom{1}{2}+\binom{1}{2} = 1,$$
$$\lambda_{3,2} = \binom{2}{2}+\binom{1}{2}+\binom{2}{2} = 2.$$

Thus, Equation (4) from Theorem 3 evaluates to

$$\begin{aligned}
&\frac{e^{-0(s_2-s_3)}}{(2-0)(1-0)} + \frac{e^{-1(s_2-s_3)}}{(0-1)(2-1)} + \frac{e^{-2(s_2-s_3)}}{(0-2)(1-2)} \\
&= \quad \frac{1}{2} - e^{-(s_2-s_3)} + \frac{1}{2}e^{-2(s_2-s_3)} \\
&= \quad \frac{1}{2}\left(1-e^{-(s_2-s_3)}\right)^2 \\
&= \quad [g_{2,1}(s_2-s_3)]^2/2.
\end{aligned}$$

Remark 6. *The probability of a gene tree topology is the sum of the probabilities of each ranked gene tree with the given topology. A given tree topology has* $(n-1)!/\prod_{i=1}^{n-1}(c_i-1)$ *rankings, where* c_i *is the number of descendant leaves of interior vertex* i*. A proof can be found in [26]. For a completely balanced tree on* $n=2^k$ *leaves, the number of rankings grows faster than polynomial: the numerator can be approximated by,*

$$n! \approx \sqrt{2\pi n}(n/e)^n,$$

and the denominator can be approximated by,

$$\prod_{i=1}^{n-1}(c_i-1) = \prod_{i=1}^{k}(2^i-1)n/2^i \approx n^k = n^{\log_2 n},$$

showing that the ratio grows faster than polynomial in n*.*

2.1 Calculation of g_i and $k_{i,j,z}$

Calculation of g_i

If $\mathcal{T}$ and $\mathscr{G}$ have the same ranked topology, then $g_i = i+1$. In general, to compute g_i, we let $\mathrm{lca}(u_j)$ be the *least common ancestor* node on the species tree for a node u_j on the ranked gene tree – i.e., the node with the largest rank on the species tree which is ancestral to all species represented in u_j. For a node y on the species tree, let $\tau(y)$ be the interval immediately above y. For example, in Figure 1c, $\tau(\mathrm{lca}(u_4)) = \tau_3$ where u_4 is the gene tree node with rank 4 — the node ancestral to D and E only. We then express g_i as

$$g_i = n - \sum_{j=i+1}^{n-1}\prod_{k=j}^{n-1} I(\tau(\mathrm{lca}(u_k)) > \tau_i) \qquad (8)$$

where $\tau_j < \tau_i$ iff $j<i$, and where $I(\cdot)$ is an indicator function taking the value 1 if the condition holds and otherwise 0. Assuming each lca() operation is $O(1)$ [10, 25]), preprocessing allows all lca terms to be computed in $O(n)$ time. Similarly all needed products and the sum in Equation (8) can each then be computed in $O(n)$ time. Thus, calculating $g_1,\ldots,g_{n-1}$ can be done in $O(n)$ time.

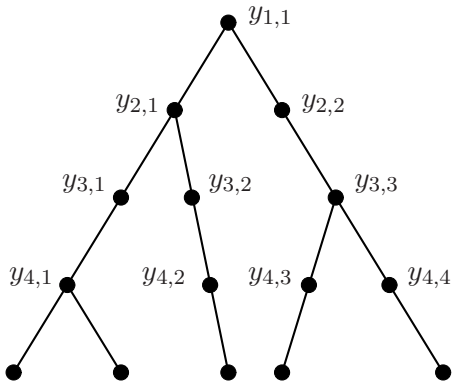

Figure 2: The beaded version of the species tree topology in Figure 1a–d.

Calculation of $k_{i,j,z}$

We let $y_{i,j}$ be the jth population (read left to right) in interval τ_i (equivalently, the jth branch or jth node subtending the branch). In order to label every population before and after a speciation time s_i uniquely, extra nodes can be added to the species tree to form a *beaded species tree* (Figure 2), so that there are i nodes at time s_i, $i=1,\ldots,n-1$. For each $i\in\{1,\ldots,n-1\}$, there is one node of outdegree 2, and $i-1$ nodes of outdegree 1. Thus, population $y_{i,j}$ corresponds to a branch (equivalently, a node) in the beaded species tree. We denote the outdegree of a node y by $outdeg(y)$.

In the remainder of this section, we compute the values $k_{i,j,z}$, i.e. the number of lineages on branch $y_{i,z}$ of the beaded species tree during the interval immediately after the jth coalescence event (going forward in time),

with $k_{i,0,z}$ being the number of lineages "exiting" the branch at time s_{i-1}. For example, in Figure 1b, we have

$$k_{2,0,1} = 1, \quad k_{2,1,1} = 2, \quad k_{2,2,1} = 3,$$
$$k_{2,0,2} = 1, \quad k_{2,1,2} = 1, \quad k_{2,2,2} = 1$$

The value of $k_{i,j,z}$ depends on the number of lineages entering branch i, ℓ_i, as well as the number of lineages exiting the branch, and not just on the number of coalescence events in the interval. For example, in Figure 1c, $k_{2,0,1} = 1$ and $k_{2,1,1} = 2$, while in Figure 1d, $k_{2,0,1} = 2$ and $k_{2,1,1} = 3$, although the two gene trees have the same ranked topology and $m_2 = 1$ for both cases.

To determine the terms $k_{i,j,z}$ we note that the number of coalescences that have occurred more recently than interval τ_i is $n - \ell_i$. In a given interval τ_i, we let $z^{(1)}$ and $z^{(2)}$ be the left and right children, respectively, of population z of outdegree 2, and let $z^{(1)} = z^{(2)}$ be the only child of a node z of outdegree 1.

The number of lineages available to coalesce in population z of interval τ_i is

$$k_{i,m_i,z} = \sum_{j=1}^{outdeg(y_{i,z})} k_{i+1,0,z^{(j)}} \tag{9}$$

where the $z^{(j)}$ are the daughter populations (one or two) of z. Further, $k_{n,0,z} = 0$ for all z. Since the beaded species tree has $n^2/2$ nodes, precalculating $outdeg(y_{i,z})$ requires $O(n^2)$. For $0 \leq j < m_i$, we have

$$k_{i,j,z} = \begin{cases} k_{i,j+1,z} - 1 & j\text{th coalescence on branch } z \\ k_{i,j+1,z} & \text{otherwise} \end{cases} \tag{10}$$

Consequently, determining a particular $k_{i,j,z}$ is $O(1)$. Thus determining $k_{i,j,z}$ for all possible i, m_i and ℓ_i is (see also Corollary 4),

$$\begin{aligned} &= O\left(\sum_{i=2}^{n-1} \sum_{\ell_i=i+1}^{n} \sum_{m_i=0}^{\ell_i - i} \sum_{j=0}^{m_i} O(1)\right) \\ &= O\left(n^4\right). \end{aligned}$$

Note that taking the sum over all z is not necessary, as in all but one branch the $k_{i,j,z}$ equals the $k_{i,j+1,z}$.

2.2 An algorithm

In summary, we derived an algorithm with runtime $O(n^5)$ for calculating the probability of a ranked gene tree given a species tree on n tips:

1. Calculate $g_1, \ldots g_{n-1}$ using Equation (8).
2. Calculate $k_{i,j,z}$ (for $i, j = 1, \ldots, n; z = 1 \ldots i$), using Equations (9) and (10).
3. Calculate $\lambda_{i,j} = \sum_{z=1}^{i} \binom{k_{i,j,z}}{2}$ (for $i, j = 1, \ldots, n$).
4. Calculate $\mathbb{P}[\mathscr{G}_{i-1,\ell_{i-1}} | \mathscr{G}_{i,\ell_i}, \mathcal{T}]$ (for $i = 2, \ldots, n$; $\ell_{i-1} = g_{i-1}, \ldots, n$; $\ell_i = g_i, \ldots, n$), using Theorem 3.
5. Calculate $\mathbb{P}[\mathscr{G}_{1,\ell_1} | \mathcal{T}]$ using Theorem 2.
6. Calculate $\mathbb{P}[\mathscr{G} \,|\, \mathcal{T}]$ using Theorem 1.

3 Discussion

In this paper, we provide a polynomial-time algorithm ($O(n^5)$ where n is the number of species) to calculate the probability of a ranked gene tree topology given a species tree, summarized in Section 2.2. We now discuss applying these results to computing probabilities of unranked gene tree topologies and to inferring ranked species trees.

3.1 Computing probabilities of unranked gene tree topologies

Previous work on computing probabilities of unranked gene tree topologies used the concept of *coalescent histories*, which specify the branches in the species tree in which each node of the gene tree occurs. An unranked gene tree probability can then be computed by enumerating all coalescent histories and computing the probability of each. The number of coalescent histories grows at least exponentially when the (unranked) gene tree matches the species tree, making this approach computationally intensive. Coalescent histories can be enumerated either recursively (e.g., in PHYLONET [30] or [23]) or nonrecursively (COAL [6]).

A much faster approach using dynamic programming similar to that used in this paper is implemented in STELLS [33], which conditions on the ancestral configuration in each branch rather than the number of lineages. Here an ancestral configuration keeps track not only of the number of lineages in a branch in the species tree, but also the particular nodes of the gene tree. Different ancestral configurations can potentially have the same number of lineages within a population. Enumerating ancestral configurations turns out to have exponential running time for arbitrarily shaped trees, but the number of ancestral configurations is still much smaller than the number of coalescent histories. When computing probabilities of ranked gene tree topologies, however, the ranking specifies the sequence of coalescence events, leading to a unique ancestral configuration given the number of lineages in a time interval. This fortuitously enables probabilities of ranked gene tree topologies to be computed in polynomial time.

We note that although the number of rankings for a gene tree is not polynomial in the number of leaves in

general, the number of rankings can be small for certain tree shapes. For example, if the gene tree has a *caterpillar* shape, in which each internal node has a leaf as a descendant, then there is only one ranking, and thus computing the ranked and unranked gene tree are equivalent. For a *pseudo-caterpillar*, a tree made by replacing the subtree with four leaves of a caterpillar with a balanced tree on four leaves [23], there are only two rankings possible, and for a *bicaterpillar* [23], for which the left subtree is a caterpillar with n_L leaves and the right subtree is a caterpillar with $n-n_L$ leaves, there are $\binom{n-2}{n_L-1}$ rankings. Thus computing unranked gene tree probabilities by summing ranked gene tree probabilities can be done in polynomial time for some tree shapes. We note that for the approach used by STELLS, some tree shapes can also be computed in polynomial time, including the cases we mentioned that have a polynomial number of rankings. An open question is whether there are any classes of unranked gene trees which have a polynomial number of rankings but an exponential number of ancestral configurations, or vice versa.

3.2 Inferring species trees from ranked gene trees

Our fast calculation of the probability of ranked gene tree topologies can be used to determine the maximum likelihood species tree from a collection of known gene trees. Assume we have observed N ranked gene trees (i.e., N loci). Now the maximum likelihood species tree $\mathcal{T}_{ML}$ (with branch lengths on internal branches) is

$$\mathcal{T}_{ML} = \underset{\mathcal{T}}{\operatorname{argmax}}\ \mathbb{P}[\mathscr{G}_1, \ldots, \mathscr{G}_N | \mathcal{T}]$$

where

$$\mathbb{P}[\mathscr{G}_1, \ldots, \mathscr{G}_N | \mathcal{T}] = \prod_{k=1}^{N} \mathbb{P}[\mathscr{G}_k | \mathcal{T}] = \prod_{i=1}^{H_n} \mathbb{P}[\mathscr{G}^{(i)} | \mathcal{T}]^{n_i} \quad (11)$$

is a multinomial likelihood. Here $\mathbb{P}[\mathscr{G}_k|\mathcal{T}]$ can be determined with our polynomial-time algorithm, we let $\mathscr{G}^{(i)}$ denote the ith ranked topology, and n_i is the number of times ranked topology i is observed, with $\sum_{i=1}^{H_n} n_i = N$. Note in particular that the ranked topology of $\mathcal{T}_{ML}$ might differ from the most frequent ranked gene tree topology[5].

Our derivation of the ranked gene tree probability also suggests a way to infer a ranked species tree topology from ranked gene tree topologies with a similar flavor as the MDC criterion. In MDC, for an input gene tree and candidate species tree, the number of extra lineages (lineages which necessarily fail to coalesce due to topological differences between gene and species trees) on each edge of the species tree is counted. For MDC, whether the edge of the species tree is long or short does not affect the deep coalescence cost. In working with ranked gene trees, however, we can keep track of the minimum number of extra lineages within each time interval τ_i. The total number of extra lineages in this sense is

$$\sum_{i=1}^{n-1} g_i - (i+1) \quad (12)$$

Minimizing (12) as a criterion for the ranked species tree will tend to penalize long edges of the species tree which have multiple lineages persisting through multiple species divergence events. As an example, in Figure 1b, the gene tree has a MDC cost of 1 since there are two lineages exiting the population immediately ancestral to A and B; however the cost according (12) is 2 because there are two edges on the beaded version of the species tree (Figure 2) that each have an extra lineage. In Figure 1c, the gene tree has a MDC cost of 0 for the species tree since it has the matching unranked topology; however, the number of extra lineages from equation (12) is 1. We note that in Figure 1c, interval τ_3, incomplete lineage sorting (and deep coalescence) have not occurred as these concepts are normally used. To capture the idea that coalescence has nevertheless occurred in a more ancient time interval than allowed, we might refer to the coalescence of A and B in Figure 1c as an "ancient lineage sorting" event (rather than incomplete lineage sorting event) or an ancient coalescence rather than a deep coalescence. We could therefore refer to minimizing equation (12) as the Minimize Ancient Coalescence (MAC) criterion, which would provide an interesting comparison to the usual topology-based MDC criterion.

In practice, a method of inferring a species tree from ranked gene trees would require estimating the ranked gene trees. This would require clock-like gene trees, or trees with times estimated for nodes, which can also be inferred under relaxed clock models in BEAST [7]. To account for the uncertainty in the gene trees, the counts for different ranked gene trees could be weighted by their posterior probabilities obtained from Bayesian estimation of the gene trees [1]. Thus, in equation (11), we would let n_{ik} be the posterior probability of ranked topology i at locus k, and use $n_i = \sum_{k=1}^{H_n} n_{ik}$ as the estimated number of times that ranked topology i was observed. Similarly, for equation (12), the coalescence cost at a locus could be distributed over multiple topologies weighted by their posterior probabilities.

Acknowledgements

We thank David Bryant for suggesting the dynamic programming approach to this problem and two anonymous referees for valuable comments, particularly on calculating g_i and $k_{i,j,z}$. JHD was funded by the New Zealand

Marsden fund and by a Sabbatical Fellowship at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville. TS was funded by the Swiss National Science Foundation.

References

[1] E. S. Allman, J. H. Degnan, and J. A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, 62:833–862, 2011.

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

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

[4] J. H. Degnan and N. A. Rosenberg. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends Ecol. Evol.*, 24:332–340, 2009.

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

[6] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

[7] A. J. Drummond and A. Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolut. Biol.*, 7:214, 2007.

[8] A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *J. R. Stat. Soc. Ser. B*, 32:155–174, 1970.

[9] G. B. Ewing, I. Ebersberger, H. A. Schmidt, and A. von Haeseler. Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.*, 8:118, 2008.

[10] D. Harel and R. E. Tarjan. Fast algorithms for finding nearest common ancestors. *SIAM J. Comput.*, 13:338–355, 1984.

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

[12] H. Huang, Q. He, L. S. Kubatko, and L. L. Knowles. Sources of error for species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.*, 59:573–583, 2009.

[13] L. S. Kubatko, B. C. Carstens, and L. L. Knowles. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25:971–973, 2009.

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

[15] L. Liu and L. Yu. Estimating species trees from unrooted gene trees. *Syst. Biol.*, 60:661–667, 2011.

[16] L. Liu, L. Yu, L. S. Kubatko, D. K. Pearl, and S. V. Edwards. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, 53:320–328, 2009.

[17] L. Liu, L. Yu, and D. K. Pearl. Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.*, 60:95–106, 2010.

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

[19] W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, 55:21–30, 2006.

[20] E. Mossel and S. Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, 7:166–171, 2010.

[21] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol. Biol. Evol.*, 5:568–583, 1988.

[22] N. A. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theor. Pop. Biol.*, 61:225–247, 2002.

[23] N. A. Rosenberg. Counting coalescent histories. *J. Comput. Biol.*, 14:360–377, 2007.

[24] S. Ross. *Introduction to Probability Models*. Academic Press, San Diego, CA, 9th edition, 2007.

[25] B. Schiever and U. Vishkin. On finding lowest common ancestors: simplification and parallelization. *SIAM J. Comput.*, 17:1253–1262, 1988.

[26] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[27] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26:119–164, 1984.

[28] C. Than and L. Nakhleh. Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.*, 5:e1000501, 2009.

[29] C. Than, D. Ruths, H. Innan, and L. Nakhleh. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.*, 14:517–535, 2007.

[30] C. Than, D. Ruths, and L. Nakhleh. Phylonet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9:322, 2008.

[31] J. Wakeley. *Coalescent Theory*. Roberts & Company, Greenwood Village, CO, 2008.

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

[33] Y. Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, doi:10.1111/j.1558-5646.2011.01476.x, 2011.