Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes

Julia Chifman∗

Department of Cancer Biology
Wake Forest School of Medicine, Winston-Salem, NC, 27157

Laura Kubatko†

Department of Statistics
Department of Evolution, Ecology, and Organismal Biology
Mathematical Biosciences Institute
The Ohio State University, Columbus, OH 43210

Abstract

The inference of the evolutionary history of a collection of organisms is a problem of fundamental importance in evolutionary biology. The abundance of DNA sequence data arising from genome sequencing projects has led to significant challenges in the inference of these phylogenetic relationships. Among these challenges is the inference of the evolutionary history of a collection of species based on sequence information from several distinct genes sampled throughout the genome. It is widely accepted that each individual gene has its own phylogeny, which may not agree with the species tree. Many possible causes of this gene tree incongruence are known. The best studied is incomplete lineage sorting, which is commonly modeled by the coalescent process. Numerous methods based on the coalescent process have been proposed for estimation of the phylogenetic species tree given multi-locus DNA sequence data. However, use of these methods assumes that the phylogenetic species tree can be identified from DNA sequence data at the leaves of the tree, although this has not been formally established. We prove that the unrooted topology of the n-leaf phylogenetic species tree is generically identifiable given observed data at the leaves of the tree that are assumed to have arisen from the coalescent process with time-reversible substitution.

∗Electronic address: jchifman@wakehealth.edu
†Electronic address: lkubatko@stat.osu.edu; Corresponding author
1 Introduction

The field of evolutionary genetics has benefitted enormously from recent advances in sequencing technology that have led to the availability of DNA sequence information for hundreds or thousands of species. This data is commonly used to study evolutionary patterns and processes. A fundamental problem in this area is the inference of a phylogenetic species tree that describes the evolutionary relationships among a collection of species for which data have been collected. Formally, a species phylogeny is a directed, acyclic graph in which all internal nodes have degree three, except for the root node, which has degree two. The tree represents the biological process of speciation, in which one population splits into two populations which then evolve independently, with no subsequent exchange of genetic material. The root node represents the most ancestral population to all sampled species, while the leaves represent present-day populations. An example of a phylogenetic species tree for four species, $a$, $b$, $c$, and $d$, is shown by the outlined tree in Figure 1.

Although the goal is generally to estimate the species phylogeny from available DNA sequence data, these sequence data are only directly informative about the gene tree – the phylogenetic tree underlying the gene for which the DNA sequences are available. It is well-accepted that gene trees and species trees may not agree with one another (see, e.g., [27, 28]), with many evolutionary processes known to give rise to variability in gene phylogenies within a fixed species phylogeny. Examples of such processes are incomplete lineage sorting (ILS), hybridization, horizontal gene transfer, and gene duplication and loss [27]. The best studied of these processes is ILS, which results when two lineages fail to share a most recent common ancestor (MRCA; represented by an internal node in the gene tree) until further back in time than the immediately ancestral population. For example, in Figure 1a, the gene tree embedded within the species tree represents the phylogenetic history of the lineages sampled from species $a$, $b$, $c$, and $d$, which are denoted by $A$, $B$, $C$, and $D$, respectively. Throughout the text, we use uppercase letters to refer to gene lineages, and the corresponding lowercase letters to refer to the species from which these lineages are sampled. Although it is possible for lineages $C$ and $D$ to share their most recent common ancestor in the population labeled $P_1$, they remain distinct in this population, and instead share their MRCA in population $P_2$, thus providing an example of ILS. Note that the topology (branching pattern) of the gene tree in Figure 1a matches the topology of the species tree. Figure 1b gives another example of ILS, but in this case lineage $A$ coalesces with lineage $C$ in their ancestral population, and the gene tree topology does not match the species tree topology.

One of the reasons that ILS has been well-studied is that it can be modeled by the coalescent process. The coalescent process can be derived as the large sample limit (as the population size goes to $\infty$) of the Wright-Fisher and other common population genetics models [20, 21, 34]. The key property of the coalescent model is that the waiting time back into the past for a pair of lineages to find their MRCA follows an exponential distribution, with a parameter that depends on the sample size. The coalescent model thus provides a link between the phylogenetic species tree and the
set of gene trees embedded within the species tree that give rise to the actual data. For this reason, numerous methods based on the coalescent process have recently been proposed for estimation of the phylogenetic species tree given multi-locus DNA sequence data (e.g., BEST [25], *BEAST [17], STEM [22], MP-EST [26], SNAPP [8]). Use of these methods assumes that the phylogenetic species tree can be identified from DNA sequence data at the leaves of the tree, but this has not formally been established (note, however, that Allman et al. (2011) [2] have established identifiability given a collection of gene tree topologies and Allman et al. (2011) [7] have considered identifiability given clade probabilities).

Here, we prove that the unrooted topology of the phylogenetic species tree is identifiable given observed data at the leaves of the tree that are assumed to have arisen from the coalescent process. Our results hold for data for which a single observation corresponds to recording which of \( \kappa \) possible states occurs at each leaf. These data are modeled by continuous-time Markov processes that specify the rates of transitions between states along the phylogeny and that satisfy the condition of time-reversibility. For DNA sequence data, there are four states (i.e., \( \kappa = 4 \)) corresponding to the four nucleotides A, C, G, and T. In this case, our results hold for the General Time Reversible (GTR; [35]) model and all associated sub-models.

In the next section, we give the necessary background on the coalescent process and on the process of mutation for general \( \kappa \)-state models, pointing out the application to DNA sequence data where relevant. We then present our main results and show how they are used to establish identifiability in the general case. Based on these results, we propose a method for inferring species-level relationships for empirical data sets consisting of multi-locus DNA sequences. We conclude by suggesting extensions of our current work.

2 Background

In this section, we review the models used for both the coalescent process and the mutation process along a phylogenetic tree. Let \( \sigma = (S, \tau) \) represent a phylogenetic species tree with \( n \) leaves with topology \( S \) and vector of speciation times \( \tau = (\tau_1, \tau_2, \ldots, \tau_{n-1}) \). Let \( g = (G, t) \) denote a gene tree with topology \( G \) and vector of coalescent times \( t = (t_1, t_2, \ldots, t_{n-1}) \). In particular, \( t_j \) is the time from the \( j \)th coalescent event to the next speciation event (looking forward in time), \( 0 < t_j < \infty \), for \( j = 1, 2, \ldots, n - 1 \). Figure 1 shows examples of gene trees nested within species trees with all of these quantities labeled.

2.1 The Coalescent Process

The coalescent is a model for the evolutionary history (i.e., sequence of coalescent events) of a sample of lineages within a population backward in time from the present to the past [20] [21] [34]. In particular, given a sample of \( j \) gene lineages, the coalescent model specifies that the time \( t_j \) until the next pair of lineages coalesces follows an
exponential distribution with rate \((\frac{j}{2})^{-1}\), i.e. \(t_j\) has probability density

\[
f(t_j) = \frac{j(j-1)2}{\theta} \exp\left(-\frac{j(j-1)2}{\theta}t_j\right), \quad t_j > 0,
\]

where the parameter \(\theta\) is the effective population size \((\theta = 2N\mu\), where \(N\) is the population size and \(\mu\) is the mutation rate).

Now consider a population within a species tree, e.g., \(P_1\) in Figure 1a, which corresponds to a time interval of length \(\tau = \tau_3 - \tau_1\), and let \(b\) represent this branch. Following Rannala and Yang (2003), let \(u\) denote the number of lineages “entering” the population (e.g., the number of lineages in the population closest to the present time) and let \(v\) be the number of lineages “leaving” the population, \(v \leq u\). The number of coalescent events within the population is \(u - v\), and the density of the time of each coalescent event can be determined via Equation 1 by setting \(j\) to be the number of current lineages in the population and following an assumption of the coalescent model that each pair of lineages within a population is equally likely to be the next to coalesce. This means that when there are \(j\) lineages available to coalesce in the population, there are \(\binom{j}{2}\) possible pairs that might be the next to coalesce, and the density for the next event should be weighted by the probability of a particular pair coalescing, which is \(1/(\binom{j}{2})\). Let \(t^b_j\) denote the time from the speciation event that immediately precedes branch \(b\) (looking backward in time) to the coalescent event that reduces the number of
lineages on branch \( b \) from \( j \) to \( j - 1 \). Define \( t_{u+1}^b \) to be 0, and let \( \tau_b \) refer to the length of branch \( b \). Then, we can write the joint density of coalescent times \( t_u^b, t_{u-1}^b, \ldots, t_{v+1}^b \) within population \( P \) on branch \( b \) in the case in which \( u > v \) as

\[
f_{P_b}(t_u^b, t_{u-1}^b, \ldots, t_{v+1}^b) = \prod_{j=v+1}^u \left[ \frac{2}{\theta} \exp \left( -\frac{j(j-1)}{\theta} (t_j^b - t_{j+1}^b) \right) \right] \\
\times \exp \left( -\frac{v(v-1)}{\theta} (\tau_b - t_{v+1}^b) \right),
\]

\[0 < t_j^b < \infty, \ j = u + 1, u, u - 1, \ldots, v + 1.\]

The terms inside the product in Equation 2 correspond to observed coalescent events, while the final term reflects the fact that when \( v \neq 1 \), the coalescent event that decreases the number of lineages from \( v \) to \( v - 1 \) does not occur within the time remaining in population \( P_b \), which is \( \tau_b - t_{v+1}^b \).

When \( u = v \) for branch \( b \), no coalescent events happen on that branch. The probability of this occurring under the model is

\[
P_{P_b}(\text{no coalescence among } v \text{ lineages}) = \exp \left( -\frac{v(v-1)}{\theta} \tau_b \right),
\]

where \( \tau_b \) is defined as above to be the length of branch \( b \). Note that when \( u = v = 1 \), as is the case for the branches at the tips of the tree, \( P_{P_b} = 1 \).

As an example, refer again to Figure 1a. For population \( P_1 \), we have \( u = v = 2 \) and the probability of this is given by Equation 3 as

\[
P_{P_1}(\text{no coalescence among 2 lineages}) = \exp \left( -\frac{2}{\theta} (\tau_3 - \tau_1) \right).
\]

For population \( P_2 \), \( u = 3, v = 1, t_{u+1}^{P_2} = t_2, t_{u-1}^{P_2} = t_{v+1}^{P_2} = t_3 \), and the joint density is

\[
f_{P_2}(t_u^{P_2}, t_{v+1}^{P_2}) = f_{P_2}(t_2, t_3) = \left\{ \frac{2}{\theta} \exp \left( -\frac{2}{\theta} (t_3 - t_2) \right) \right\} \left\{ \frac{2}{\theta} \exp \left( -\frac{6}{\theta} t_2 \right) \right\}. \tag{5}
\]

For population \( P_3 \), \( u = 2, v = 1, \) and \( t_{u+1}^{P_3} = t_{v+1}^{P_3} = t_1 \), and the joint density is

\[
f_{P_3}(t_u^{P_3}) = f_{P_3}(t_1) = \frac{2}{\theta} \exp \left( -\frac{2}{\theta} t_1 \right). \tag{6}
\]

Equation 2 and Equation 3 describe the coalescent process within a population. We now wish to apply the coalescent model to the entire phylogenetic species tree in order to derive the probability density for gene trees nested within the species tree. Again following Rannala and Yang (2003), we note that once the number of lineages entering and leaving a population on a species phylogeny is specified, the coalescent processes within each of the populations are conditionally independent of one another. Thus, the
densities within individual populations can be multiplied to give the overall gene tree density given a particular species tree with speciation time vector $\tau$, 

$$f_{(G,t)|(S,\tau)}((G,t)) = \prod_{b=1}^{n-1} f_{P_b}(t_{u_b}^b, t_{u_b-1}^b, \ldots, t_{v_b}^b),$$  \hspace{1cm} (7)$$

where the index $b$ is over populations (e.g., branches) in the species phylogeny $(S,\tau)$, $u_b$ is the number of lineages entering population $P_b$, and $v_b$ is the number of lineages leaving population $P_b$. Note that the collection of $t_j^b$ terms across all branches is equivalent to the vector $t$ in $(G,t)$. The notation $f_{(G,t)|(S,\tau)}$ for this joint density is chosen to reflect the fact that this is the joint density of the topology and branch lengths of the gene tree $(G,t)$, conditional on the species tree $(S,\tau)$.

### 2.2 The Mutation Process

The process of evolutionary change along a phylogeny is commonly modeled by a continuous-time Markov process. In this section, we describe the general case of a Markov mutation process on $\kappa$ states, and then consider the commonly used models for DNA sequence data for which $\kappa = 4$. We begin by specifying a general $\kappa \times \kappa$ instantaneous rate matrix $Q$ such that entry $q_{ij}$ gives the instantaneous rate of change from state $i$ to state $j$.

$$Q = \begin{pmatrix}
-\pi_1\mu_1 & \pi_2\mu_2 & \cdots & \pi_{\kappa}\mu_{\kappa-1} \\
\pi_1\mu_1 & - & \pi_3\mu_3 & \cdots & \pi_{\kappa}\mu_{2\kappa-3} \\
\pi_1\mu_2 & \pi_2\mu_\kappa & - & \cdots & \pi_{\kappa}\mu_{3\kappa-6} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\pi_1\mu_{\kappa-1} & \pi_2\mu_{2\kappa-3} & \pi_3\mu_{3\kappa-6} & \cdots & -
\end{pmatrix}. \hspace{1cm} (8)$$

In the matrix above, each $\pi_j$ term represent the frequency of state $j$ at equilibrium, with the constraints that $\pi_j > 0$ for $j \in [\kappa] := \{1, 2, \ldots, \kappa\}$, and $\sum_{j=1}^{\kappa} \pi_j = 1$. The model contains $\binom{\kappa}{2}$ additional parameters $\mu_1, \ldots, \mu_{\frac{1}{2}\kappa(\kappa-1)}$ that specify the rates of mutation between states, with the assumption that $\mu_j > 0$ for $j = 1, 2, \ldots, \kappa(\kappa-1)/2$. The diagonal entries of $Q$ are set to the negative sum of the corresponding row. Note that the $\mu_i$ term in the $(i,j)^{th}$ entry is the same as in the $(j,i)^{th}$ entry, so that the model satisfies the condition of time-reversibility, i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$. The intuition of the model is that a pair of states $i$ and $j$ have the same basic rate of mutating from one to the other, represented by $\mu_i$, but the rate of moving to state $j$ depends also on the empirical frequency of state $j$, given by $\pi_j$.

The instantaneous rate matrix is used to compute the transition probability matrix $P(t)$ such that the $(i,j)^{th}$ entry of $P(t)$ gives the probability that state $i$ mutates to state $j$ in an interval of time of length $t$. This is done by solving the matrix differential equation $P'(t) = QP(t)$ with initial condition $P(0) = I$, to give $P(t) = e^{Qt}$. In phylogenetics, it is common to refer to the matrix $P(t)$ as the substitution matrix rather
than the transition probability matrix, because the word “transition” has a particular meaning in the process of DNA sequence mutation.

The matrix $Q$ can be thought of as the $\kappa$-state analog of the GTR model for sequence data in the sense that it allows a separate parameter for mutations between each pair of states. All of the models we consider here can then be thought of as sub-models of the model given by $Q$. For example, if we specify $\mu_1 = \mu_2 = \ldots = \mu_\frac{1}{2}\kappa(\kappa-1)$ and $\pi_j = \frac{1}{\kappa}$ for $j \in [\kappa]$, the resulting model is the $\kappa$-state analog of the Jukes-Cantor model \cite{JC69}, which has been called the Mk model by Lewis (2001) \cite{Lewis2001}.

When DNA sequence data are available at the tips of the tree, $\kappa = 4$ and we let $j = 1, 2, 3, 4$ correspond to the four nucleotides $A, C, G, \text{and } T$, respectively. Below we list the restrictions on the parameters in $Q$ that lead to many of the commonly-used substitution models in empirical phylogenetics.

\textbf{[JC69] Jukes-Cantor model \cite{JC69}}: $\pi_i = \frac{1}{4}$ for $i = 1, 2, 3, 4$; $\mu_j = \mu$ for all $j = 1, 2, \ldots, 6$.

\textbf{[K2P] Kimura’s 2-parameter model \cite{K2P}}: $\pi_i = \frac{1}{4}$ for $i = 1, 2, 3, 4$; $\mu_1 = \mu_6 > 0$; $\mu_2 = \mu_3 = \mu_4 = \mu_5 > 0$.

\textbf{[F81] Felsenstein’s 1981 model \cite{Felsenstein1981}}: $\mu_j = \mu$ for $j = 1, 2, \ldots, 6$; $\pi_i \in (0, 1)$ for $i = 1, 2, 3, 4$ with $\sum_{i=1}^{4} \pi_i = 1$.

\textbf{[HKY85] Hasegawa, Kishino, and Yano’s 1985 model \cite{HKY85}}: $\mu_1 = \mu_6 > 0$; $\mu_2 = \mu_3 = \mu_4 = \mu_5 > 0$; $\pi_i \in (0, 1)$ for $i = 1, 2, 3, 4$ with $\sum_{i=1}^{4} \pi_i = 1$.

\textbf{[TN93] Tamura and Nei’s 1993 model \cite{TN93}}: $\mu_1 > 0$; $\mu_6 > 0$; $\mu_2 = \mu_3 = \mu_4 = \mu_5 > 0$; $\pi_i \in (0, 1)$ for $i = 1, 2, 3, 4$ with $\sum_{i=1}^{4} \pi_i = 1$.

\textbf{[GTR] General time-reversible model \cite{GTR}}: $\mu_j > 0$ for $j = 1, 2, \ldots, 6$; $\pi_i \in (0, 1)$ for $i = 1, 2, 3, 4$ with $\sum_{i=1}^{4} \pi_i = 1$.

Thus far, we have described the evolutionary model for mutations between states along specific branches of a gene tree. We now describe how these models are used to compute the probability distribution of the data at the leaves of a phylogenetic gene tree.

Consider the gene trees labeled as in Figure \ref{fig:gene_tree}. Let $X_y$ be the state observed for lineage $y$, $y = A, B, C, D$. Considering the tree to be rooted at node $r$ (see \cite{Felsenstein1981}), the site pattern probability on gene tree $(G, t)$ in Figure \ref{fig:gene_tree} for a particular observation $i_1i_2i_3i_4$, $i_j \in [\kappa]$, at the tips of the tree is given by

$$p_{i_1i_2i_3i_4}(G, t) = P(X_A = i_1, X_B = i_2, X_C = i_3, X_D = i_4)$$

$$= \sum_{r=1}^{\kappa} \sum_{s=1}^{\kappa} \pi_{rs} P_{i_1}(v_1) P_{i_2}(v_2) P_{i_3}(v_3) P_{i_4}(v_4) P_{i_4}(v_5), \quad (9)$$

where $v_1 = \tau_2 + t_1, v_2 = \tau_2 + t_1, v_3 = 2t_3 - t_2 + (\tau_3 - t_1 - \tau_2), v_4 = \tau_3 + t_2$, and $v_5 = \tau_3 + t_2$. These values for the $v_i$ are obtained by considering the gene tree to be
rooted at node \( r \) (rather than rooted as shown in the top portion of Figure 1a). The set \( P_{(G,t)} = \{ p_{i_1i_2i_3i_4|(G,t)} : i_j \in [k] \} \) gives the probability distribution on all possible site patterns for the gene tree in Figure 1a.

A similar calculation can be carried out on the gene tree in Figure 1b—the key idea is that the states are not observed at the internal nodes of the tree, and so the probability must be computed by summing over all possible states for these nodes. Additionally, probabilities along each branch are multiplied together because we assume that the mutation process proceeds independently along each branch. Finally, we point out that we are utilizing what have been called rate matrix models \(^4\) in phylogenetics, since we assume that the Markov mutation process is homogeneous across branches of the phylogeny.

3 The Site Pattern Probability Distribution Under the Coalescent Model

Our goal is to establish identifiability of the \( n \)-leaf phylogenetic species tree given data at the leaves of the tree. To achieve this, we first prove identifiability of the 4-taxon species tree, and thus dedicate this section to the description of the coalescent model for a fixed species tree \( S \) on four leaves. We would like to point out that everything in this section is easily extendable to the \( n \)-taxon case.

To compute the site pattern probabilities on the species tree, we will use gene tree site pattern probabilities along with the gene tree density given in Equation 7. To distinguish the probability distribution for data arising on the species tree under the coalescent model from the probability distribution of data arising on a gene tree, we will use the notation \( P^*_{(G)} = \{ p^*_{i_1i_2i_3i_4|(G,S,\tau)} : i_j \in [k] \} \) for data that come from the species phylogeny \( S \) with vector of speciation times \( \tau \) and with embedded gene tree \( G \).

Consider first the case in which \( G = S \), i.e., the gene tree and species tree have the same topology (see, for example, Figure 1a). In that case, for a fixed \( i_1, i_2, i_3, \) and \( i_4 \), we find

\[
p^*_{i_1i_2i_3i_4|(G,S,\tau)} = P(X_a = i_1, X_b = i_2, X_c = i_3, X_d = i_4) = \int_{t} p_{i_1i_2i_3i_4|(G,t)} f_{(G,t)|S,\tau}(G,t) dt.
\]

When \( G \neq S \), the labels at the leaves of the species tree may correspond to a different ordering of the labels of the lineages in the gene tree. For example, in Figure 1b species \( a \) and \( b \) are sister leaves in the species tree, but lineages \( A \) and \( C \) are sisters in the gene tree. Hence, for a gene tree \( G \) and species tree \( (S,\tau) \) in Figure 1b we compute

\[
p^*_{i_1i_2i_3i_4|(G,S,\tau)} = P(X_a = i_1, X_b = i_2, X_c = i_3, X_d = i_4) = \int_{t} p_{i_3i_1i_2i_4|(G,t)} f_{(G,t)|(S,\tau)}(G,t) dt.
\]
Figure 2: Example of two different histories for the same gene tree topology (red) nested within an asymmetric species tree (blue). In (a), the second coalescent event happens prior to the root of the species tree, while in (b) it occurs in the first population ancestral to both species.

It is clear that many observations at the leaves of the species tree will not always result in the same observations at the tips of the embedded gene tree. We will use $\sigma(i_1, i_2, i_3, i_4)$ to represent the observation at the leaves of the gene tree, where $\sigma$ is a permutation of $i_1, i_2, i_3, \text{and } i_4$ for a fixed gene tree topology nested within a species tree. In general we write

$$p^*_i G, S, \tau = \int p_{\sigma(i_1, i_2, i_3, i_4)}(G, t) f(G, t) | (S, \tau) (G, t) dt. \quad (10)$$

Equation 10 gives the probability of the event $\{X_a = i_1, X_b = i_2, X_c = i_3, X_d = i_4\}$ for a fixed gene tree topology $G$ and species tree $(S, \tau)$. Our goal is to obtain these probabilities conditioning only on the species tree $(S, \tau)$. Since the true gene tree is unobserved, we must consider all possible gene trees that are consistent with the given species tree, and weight the probability of the site pattern of interest appropriately by the probability of each gene tree under the coalescent model. In addition, we have to consider histories for each gene tree, since for example, the same gene tree might have coalescent events happening above all the speciation events, as in Figure 2a, or in between two speciation events, as in Figure 2b. We will denote a gene tree for a specific history by $G_h$. It is worth pointing out that the gene tree in Figure 1a will have additional histories but the gene tree in Figure 1b has only one history. Also, note that the limits of integration will depend on where the coalescent events happen. This leads to the following expression for the probability that $\{X_a = i_1, X_b = i_2, X_c = i_3, X_d = i_4\}$ for species tree $(S, \tau)$,

$$p^*_i G, S, \tau = \sum_{G_h} \int p_{\sigma(i_1i_2i_3i_4)}(G_h, t) f(G_h, t) | (S, \tau) (G_h, t) dt. \quad (11)$$

We denote the probability distribution on all possible site patterns given species tree $(S, \tau)$ by

$$P^*_i (S, \tau) = \{p^*_i G, S, \tau : i_j \in [\kappa]\}.$$
We note that in general each $p^*_{11i2i3i4}((S, \tau))$ is not a polynomial but an analytic function.

Let $C_{\text{GTR}(\kappa)}$ be the $\kappa$-state analytic GTR model under the coalescent for a 4-taxon species tree. The model parameters are the topology of the species tree $S$, the matrix $Q$ as described by (8), the vector of speciation times $\tau = (\tau_1, \tau_2, \tau_3)$ for $S$ and the effective population size $\theta$. For a fixed species tree $S$ we will denote the continuous parameter space of $C_{\text{GTR}(\kappa)}$ by an open set $U_S \subseteq \mathbb{R}^M$. Then the parameterization map $\psi_S$, for the analytic model, giving the probability distribution of the variables at the leaves of the 4-taxon species tree $S$ is

$$\psi_S : U_S \rightarrow \Delta^{\kappa^4-1}$$

$$u \mapsto P^*_u(S, \tau),$$

where $\Delta^{\kappa^4-1} \subseteq [0, 1]^{\kappa^4}$ is the probability simplex

$$\Delta^{\kappa^4-1} := \{(p_1, p_2, \ldots, p_{\kappa^4}) \in \mathbb{R}^{\kappa^4} \mid \sum_{i=1}^{\kappa^4} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\}.$$

The image of the map, $C_S := \psi_S(U) \subseteq \Delta^{\kappa^4-1}$, is the coalescent phylogenetic model. One can easily see that the model can be extended to any $n$-taxon species tree.

### 3.1 A few remarks about embedded gene trees within a species tree

Let $S$ be a four-taxon symmetric $((a, b), (c, d))$ or asymmetric $(a, (b, (c, d)))$ species tree with a cherry $(c, d)$. Recalling that the gene trees arising under the coalescent model will all satisfy the molecular clock (i.e., the distance of each tip to the root is identical), it might be obvious to some readers that for any observation $i_1i_2i_3i_4$ at the leaves of $S$, $i_1, i_2, i_3, i_4 \in [\kappa]$

$$p^*_{i_1i_2i_3i_4}((S, \tau)) = p^*_{i_1i_2i_4i_3}((S, \tau)).$$

At the same time for any gene tree for which $(C, D)$ is not a cherry, the above Equation [13] is false.

**Example 3.1.** Let gene tree $G$ arising from the species tree $((a, b), (c, d))$ be $[((A, C), B), D)$, as in Figure 7b. Then according to Equation [10] for any observation $i_1i_2i_3i_4$ at the leaves of $S$ we get

$$p^*_{i_1i_2i_3i_4}((G, S, \tau)) = \int_{t} p_{i_1i_3i_2i_4}((G, t))f(G, t)|(S, \tau)(G, t)dt,$$

$$p^*_{i_1i_2i_4i_3}((G, S, \tau)) = \int_{t} p_{i_1i_4i_2i_3}((G, t))f(G, t)|(S, \tau)(G, t)dt.$$

Clearly, $p^*_{i_1i_2i_3i_4}((G, S, \tau)) \neq p^*_{i_1i_2i_4i_3}((G, S, \tau))$ for this particular gene tree, since $p_{i_1i_3i_2i_4}((G, t)) \neq p_{i_1i_4i_2i_3}((G, t))$.

This example demonstrates that in general for many individual gene trees Equation [13] does not hold yet when the sum is taken over all possible gene trees and their
respective histories the result does hold. In this section we would like to clarify why and how the above Equation 13 is true under the coalescent model.

**Case 1:** Consider a collection of gene trees with \((C, D)\) as a cherry, denoted by \(G_{CD}\). Then for each \(G_{CD} \in G_{CD}\) it is true that

\[
p_{i_1 i_2 i_3 i_4 | (G_{CD}, S, \tau)}^* = p_{i_1 i_2 i_4 i_3 | (G_{CD}, S, \tau)},
\]

for all \(i_1, i_2, i_3, i_4 \in [k]\).

**Case 2:** Now, consider the set of gene trees for which \((A, B)\) is a cherry and \((C, D)\) is not a cherry. There are only two gene trees of this form: \(G^1 = (((A, B), C), D)\) and \(G^2 = (((A, B), D), C)\). Regardless of whether \(S\) is symmetric or asymmetric, the number of histories associated with the two gene trees above is the same (two histories are possible in the case that \(S\) is symmetric, while only a single history is possible when \(S\) is asymmetric). Furthermore, the probability densities of these histories are identical under the coalescent model; we denote them below by \(f_{(G_h^1, t)}((S, \tau))\) and \(f_{(G_h^2, t)}((S, \tau))\), but emphasize that they are equivalent. Then, for all \(i_1, i_2, i_3, i_4 \in [k]\)

\[
p_{i_1 i_2 i_3 i_4 | (G_h^1, S, \tau)}^* = \int t \sigma(i_1 i_2 i_3 i_4) f_{(G_h^1, t)}((S, \tau)) (G_h^1, t) dt,
\]

\[
p_{i_1 i_2 i_4 i_3 | (G_h^2, S, \tau)}^* = \int t \sigma(i_1 i_2 i_4 i_3) f_{(G_h^2, t)}((S, \tau)) (G_h^2, t) dt.
\]

Now, note that

\[
P_{\sigma(i_1 i_2 i_3 i_4)}((G_h^u, S, \tau)) = P_{\sigma(i_1 i_2 i_3 i_4)}((G_h^1, S, \tau)),
\]

for \(u, v \in \{1, 2\}, u \neq v\), which gives

\[
p_{i_1 i_2 i_3 i_4 | (G_h^1, S, \tau)}^* + p_{i_1 i_2 i_4 i_3 | (G_h^2, S, \tau)}^* = p_{i_1 i_2 i_3 i_4 | (G_h^1, S, \tau)}^* + p_{i_1 i_2 i_4 i_3 | (G_h^2, S, \tau)}^*.
\]

**Case 3:** Consider the set of gene trees for which neither \((A, B)\) nor \((C, D)\) are cherries. Then at least one of the following pairs are cherries:

\[\{(A, C), (A, D), (B, C), (B, D)\}\.

Without loss of generality, suppose that \((A, C)\) is a cherry.

(i). First, suppose that \((B, D)\) is also a cherry, and denote this gene tree by \(G_1\). Regardless of whether \(S\) is symmetric or asymmetric, there are two possible histories consistent with this gene tree. Now consider a second gene tree, \(G_2\), in which \((A, D)\) and \((B, C)\) are cherries. This gene tree also has two possible histories. For both \(G_1\) and \(G_2\), define \(h_1\) to be the history in which the first coalescent event occurs below the root of the species tree, and \(h_2\) to be the history in which the first coalescent event occurs above the root of the species tree. As in Case 2, the probability densities of \(h_i\) in \(G_1\) and \(G_2\) are identical for \(i = 1, 2\). Then the relationships in Equations 14, 15, and 16 (and thus, Equation 17) hold.

(ii). Now, suppose that \((B, D)\) is not a cherry. Then two gene trees are possible: \((B, (D, (A, C)))\) and \((D, (B, (A, C)))\). Let \(G_1\) be the tree \((B, (D, (A, C)))\) (the other
case is analogous). Now consider the tree \( G_2 = (B, (C, (A, D))) \). \( G_1 \) and \( G_2 \) have the same number of histories and the same associated probability densities for a given species trees with \((c, d)\) as a cherry. The same arguments in the previous cases can be used to show that Equation 17 holds in this case as well. This concludes Case 3.

It is straightforward to check that all of the possible histories for both the symmetric and asymmetric species trees are included in exactly one of the cases described above. Recall that the probability of any observation \( i_1 i_2 i_3 i_4 \) at the leaves of \( S \) is the sum over all possible gene trees and their associated histories. Thus, Equation 13 is true under the coalescent model. If the species tree \( S \) is symmetric then in addition we also get
\[
p^*_{i_1 i_2 i_3 i_4}(S, \tau) = p^*_{i_2 i_1 i_3 i_4}(S, \tau),
\]
for any observation \( i_1 i_2 i_3 i_4 \) at the leaves of \( S, i_1, i_2, i_3, i_4 \in [\kappa] \).

4 Splits, Flattenings, Invariants and Identifiability Background

In this section we provide definitions and results that we will use in subsequent sections. Let \( L \) denote a set of taxa for a tree (species or gene), e.g. if \( S \) is a species tree in Figure 1, then \( L = \{a, b, c, d\} \).

**Definition 4.1.** A *split* of a set of taxa \( L \) is a bipartition of \( L \) into two non-overlapping subsets \( L_1 \) and \( L_2 \), denoted \( L_1 | L_2 \). A split \( L_1 | L_2 \) is *valid* for tree \( T \) if the subtrees containing the taxa in \( L_1 \) and in \( L_2 \) do not intersect.

In general, the definition of a split is used for unrooted trees. To avoid any ambiguity, by a split of a rooted 4-taxon species tree (symmetric or asymmetric) we will always mean a bipartition of \( L \) into two equal subsets, i.e. \(|L_1| = |L_2|\). This means that we are not considering in this article all possible splits induced by all internal edges, e.g. the split over the branch \( \tau_3 - \tau_2 \) for the asymmetric species tree in Figure 2.

In particular, for the four-leaf species trees \(((a, b), (c, d))\) (Figure 1) or \((a, (b, (c, d)))\) (Figure 2) there are three splits according to our discussion above. The split \( L_1 | L_2 = ab | cd \) is valid and splits \( ac | bd \) and \( ad | bc \) are not valid. For the gene tree in Figure 1b the split \( AC | BD \) is valid, while \( AB | CD \) and \( AD | BC \) are not. This also demonstrates that a valid split for a species tree will not always result in the same valid split for an embedded gene tree.

For a particular \( n \)-leaf species or gene tree, the probability distribution on all possible site patterns can be viewed as an \( n \)-dimensional \( \kappa \times \cdots \times \kappa \) array \( P \). Splits of a set of taxa provide a natural way to rearrange a tensor \( P \) as a matrix.

**Definition 4.2.** Let \( L_1 | L_2 \) be any split of a set of taxa \( L \). A *flattening* of \( P \), denoted by \( \text{Flat}_{L_1 | L_2}(P) \), is an \( \kappa^{|L_1|} \times \kappa^{|L_2|} \) matrix, whose rows are indexed by possible states for the leaves in \( L_1 \) and columns by possible states in \( L_2 \). The entries of \( \text{Flat}_{L_1 | L_2}(P) \) correspond to the probability of the site pattern specified by the row and column indices.
For example, for $\kappa = 4$ and a species tree $S$ as in Figure 1, the $16 \times 16$ flattening of $P^*_S(S,\tau)$ for a split $L_1 \mid L_2 = ad \mid bc$ is

$$\text{Flat}_{ad \mid bc}(P^*_S(S,\tau)) = \begin{pmatrix}
p^*_{AAAA} & p^*_{AAC} & p^*_{AGA} & p^*_{ATA} & p^*_{ACAA} & \cdots & p^*_{ATT} \\
p^*_{AAAC} & p^*_{AC} & p^*_{AAC} & p^*_{ACT} & p^*_{ACAC} & \cdots & p^*_{ATTC} \\
p^*_{AAAG} & p^*_{AACG} & p^*_{ACAG} & p^*_{ACAT} & p^*_{ACAG} & \cdots & p^*_{ATTG} \\
p^*_{CAAA} & p^*_{CACA} & p^*_{CAGA} & p^*_{CATA} & p^*_{CCAA} & \cdots & p^*_{CTTA} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
p^*_{TAAT} & p^*_{TACT} & p^*_{TAGT} & p^*_{TATT} & p^*_{TCAT} & \cdots & p^*_{TTTT}
\end{pmatrix}.$$  

A very important concept that will be tied with the minors of the matrix $\text{Flat}_{L_1 \mid L_2}(P)$ is that of invariants. Consider an asymmetric species tree $S = (a, (b, (c, d)))$, where $c$ and $d$ are sister leaves. Then,

$$p^*_{sij}(S,\tau) - p^*_{sji}(S,\tau) = 0,$$

for all $i, j \in [\kappa]$, where $*$ indicates any value in $[\kappa]$ (see Section 3.1). This is an example of a linear invariant. It is termed linear since it is a linear function in the site pattern probabilities. More precisely, an invariant is a function in the site pattern probabilities that vanishes when evaluated on any distribution arising from the model. Linear invariants for gene trees have been studied by several authors [23, 9, 13, 14]. In general, many linear invariants for a gene tree are not invariants for the same species tree. Intuitively this makes sense, since when the sum is taken over all gene trees embedded within a species tree it is quite obvious that some linear invariants will vanish on one gene tree topology but not the other.

Let $\mathcal{R}$ be a collection of analytic functions $f_1, f_2, \ldots, f_n$ defined on the common domain $D \subseteq \mathbb{R}^m$. An analytic variety $V(\mathcal{R})$ is a simultaneous zero-set of functions in $\mathcal{R}$,

$$V(\mathcal{R}) = \{ a \in D | f_i(a) = 0, 1 \leq i \leq n \}.$$  

Fix a coalescent phylogenetic model $C_S \subseteq \Delta^{\kappa^4-1}$ as defined in Section 3. Then an analytic function $f$ is called a coalescent phylogenetic invariant if $f(a) = 0$ for all $a \in C_S$.

A very powerful and important concept about the zero-sets of analytic varieties will play an important role in our arguments about identifiability. It is well-known that if a real function $f$ is analytic in the connected open set $U \subseteq \mathbb{R}^m$ and its zero set is of positive measure, then $f \equiv 0$. Let a set $U \subseteq \mathbb{R}^m$ be of full dimension and set $W := U \cap V(\mathcal{R})$, where $\mathcal{R}$ is a finite set of nonconstant analytic functions defined on $U$. If $W$ is a proper subvariety of $U$ then it must be of dimension strictly less than $m$. In particular, $W$ is necessarily of Lebesgue measure zero.

A key issue when formulating any statistical model is whether the parameters of that model are identifiable. Classically, identifiability means the following: suppose
that $\mathcal{M}(\Theta) = \{P_\theta : \theta \in \Theta\}$ defines a family of probability distributions with parameters $\theta$. The model $\mathcal{M}(\Theta)$ is said to be identifiable if the mapping $\theta \mapsto P_\theta$ is injective. This definition is rather too strict as for many models this map will not be injective. Describing a set of parameters on which a model is non-identifiable and excluding it from a domain is not always possible, especially for identifiability of numerical parameters. Nonetheless, identifiability for parametric models, both algebraic and analytic, can be proved by demonstrating that the set of parameters on which the model is non-identifiable is a subset of a proper subvariety of measure zero. In this case we say that model parameters are generically identifiable.

Generic identifiability of a gene tree topology and numerical parameters for many evolutionary models is well-studied and established. A series of papers by E. Allman and J. Rhodes and collaborators \[1, 5, 6, 3\] have used an algebraic framework to establish identifiability of the unrooted gene tree and associated model parameters for substitution models as general as the General Markov model with a proportion of sites invariant (GM+I) as well as for the general time reversible model with rate variation following the gamma distribution (GTR+Γ).

For the general $\kappa$-state Markov model $\mathcal{M}$ a flattening of a tensor $P$ is used to prove generic identifiability of a gene tree topology. Briefly, let $P$ be a joint distribution arising from $\mathcal{M}_T$ on an $n$-leaf gene tree $T$ then: (i) if $L_1|L_2$ is a valid split for $T$, the rank$(\text{Flat}_{L_1|L_2}(P)) \leq \kappa$, and (ii) if $L_1|L_2$ is not a valid split for $T$, generically the rank$(\text{Flat}_{L_1|L_2}(P)) > \kappa$. In particular, for a valid split on $T$ the $(\kappa + 1)$-minors of a matrix $\text{Flat}_{L_1|L_2}(P)$, called edge invariants, all vanish on $\mathcal{M}_T$.

## 5 Main Results

To prove identifiability of a 4-leaf species tree from site pattern probabilities we will use precise formulas for the generalized Jukes-Cantor $k$-state model under the coalescent for symmetric and asymmetric 4-leaf species trees, which are described fully in Section \[6\]. In addition our proof only applies to the unrooted 4-leaf species tree (note that the rooted symmetric and the rooted asymmetric species trees yield a single unrooted species tree topology) although we make a distinction between the two trees (symmetric and asymmetric) when applicable.

In the next theorem we are going to identify the parameter space $U_S$ with a full dimensional subset of $\mathbb{R}^M$. In particular, $U_S$ is an open connected subset of $\mathbb{R}^M$. Also, recall that the parameterization map $\psi_S$ defined by (12) is given by analytic functions. In Section \[4\] we have stated a result from the theory of analytic functions of several complex variables that will play an important role in our next argument. For clarity we rephrase this fact as follows: if a real function $f$ is analytic in the connected open set $U \subseteq \mathbb{R}^n$ and $f$ is not identically zero then the set $Z(f) = \{x \in U | f(x) = 0\}$ has $n$-dimensional Lebesgue measure zero. For more information about analytic functions of several complex variables and the proof of the above fact, the reader is encouraged to consult \[15\].
Theorem 5.1. Let $S$ be a four-taxon symmetric $((a, b), (c, d))$ or asymmetric $(a, (b, (c, d)))$ species tree with a cherry $(c, d)$ and let $L_1|L_2$ be one of the splits of taxa, $abcd, ac|bd$ or $ad|bc$. Consider the $C_{GTR(\kappa)}$ $\kappa$-state analytic GTR model under the coalescent for a species tree $S$. Identify the parameter space $U_S$ with a full dimensional subset of $\mathbb{R}^M$.

1. If $L_1|L_2$ is a valid split for $S$, then for all distributions $P^*_{(S, \tau)}$ arising from the model

$$\text{rank}(\text{Flat}_{L_1|L_2}(P^*_{(S, \tau)})) \leq \left( \frac{\kappa + 1}{2} \right).$$

2. If $L_1|L_2$ is not a valid split for $S$, then for generic distributions $P^*_{(S, \tau)}$ arising from the model

$$\text{rank}(\text{Flat}_{L_1|L_2}(P^*_{(S, \tau)})) > \left( \frac{\kappa + 1}{2} \right).$$

Proof. Suppose that $L_1|L_2$ is a valid split for $S$, that is $L_1|L_2 = ab|cd$. From our discussion in Section 3.1 it is clear that for any distribution $P^*_{(S, \tau)}$, arising from the $\kappa$-state model $C_{GTR(\kappa)}$ the entries $((i, j), (k, l))$ and $((i, j), (l, k))$ of the $\kappa^2 \times \kappa^2$ matrix $\text{Flat}_{ab|cd}(P^*_{(S, \tau)})$ are equal, for all $k \neq l \in [\kappa]$. In particular, the columns labeled by the $cd$-indices $kl$ and $lk$ for distinct $k, l \in [\kappa]$ are identical, e.g. columns labeled by 12 and 21 are equal. Since there are $\binom{\kappa}{2}$ such pairs, we get that

$$\text{rank}(\text{Flat}_{ab|cd}(P^*_{(S, \tau)})) \leq \kappa^2 - \left( \frac{\kappa}{2} \right) = \left( \frac{\kappa + 1}{2} \right),$$

which establishes (1).

Now suppose that $L_1|L_2$ is not a valid split for $S$. Let $X_{ac|bd}$ and $X_{ad|bc}$ denote the sets of $((\kappa+1)/2)$-minors of the $\kappa^2 \times \kappa^2$ matrices $\text{Flat}_{ac|bd}$ and $\text{Flat}_{ad|bc}$ respectively. Let $V_{ac|bd}$ be the zero set of $X_{ac|bd}$ and $V_{ad|bc}$ be the zero set of $X_{ad|bc}$.

Define $W$ as a zero set of real functions $f \circ \psi_S$ that are analytic in the connected set $U_S$, where $f$ vanishes on $V_{ac|bd} \cup V_{ad|bc}$. It is clear that $W$ is an analytic subvariety and a subset of $U_S$. We are going to show that $W$ is a proper subvariety of $U_S$ by finding $u \in U_S \setminus W$ with $\psi_S(u) \notin V_{ac|bd} \cup V_{ad|bc}$. In particular, we show that the ranks of $\text{Flat}_{ac|bd}$ and $\text{Flat}_{ad|bc}$ are greater than $\binom{\kappa + 1}{2}$ for this choice of parameters.

Recall that the continuous model parameters are the vector of speciation times $\tau = (\tau_1, \tau_2, \tau_3)$ for $S$ as depicted in Figures 1 and 2, the effective population size $\theta$ and the matrix $Q$ (the rates of mutation between states $\mu = (\mu_1, \mu_2, \ldots, \mu_2^{\kappa(\kappa-1)})$ and relative frequencies of the states at equilibrium $\pi = (\pi_1, \pi_2, \ldots, \pi_\kappa)$). Note that for a fixed species tree $S$ the matrices $\text{Flat}_{ac|bd}(P^*_{(S, \tau)})$ and $\text{Flat}_{ad|bc}(P^*_{(S, \tau)})$ will be identical, since for any observation at the leaves of $S$ for $i_1, i_2, i_3, i_4 \in [\kappa]$ we have

$$P^*_{i_1i_2i_3i_4|S, \tau} = P^*_{i_1i_2i_3i_4|S, \tau}.$$

Thus, for simplicity let $F := \text{Flat}_{ac|bd}(P^*_{(S, \tau)}) = \text{Flat}_{ad|bc}(P^*_{(S, \tau)})$. Next, let $Q$ be a generalized Jukes-Cantor matrix as described by (22) and let $\tau_2 = \tau_3$ and $\tau_1 = \frac{\tau_2}{10}$.
Since, \( \tau_2 = \tau_3 \) then the distributions \( P_{(S, \tau)}^* \) will be the same for both symmetric and asymmetric species trees. Furthermore, let \( \tau_3 = 1 \), \( \theta = 0.1 \), and \( \mu_i = 0.1 \) for all \( i \in \{1, \ldots, \frac{1}{2}\kappa(\kappa - 1)\} \).

For \( \kappa \in \{2, 3\} \), by using precise formulas for the site pattern probabilities for the generalized Jukes-Cantor as described in Section 6 and Supplement A and the choice of parameters made above, we find that the determinant of \( F \) is not zero, that is, generically \( \text{rank}(F) = \kappa^2 \) (see Mathematica supplementary files for computations).

For \( \kappa \geq 4 \) we are going to select all rows and columns from the matrix \( F \) labeled by \( ac, bd \)-indices with distinct \( a \) and \( c \) (equivalently with distinct \( b \) and \( d \)). Denote this \( \kappa(\kappa - 1) \times \kappa(\kappa - 1) \) submatrix by \( F^* \). Note, that for \( \kappa \geq 4 \), \( \kappa(\kappa - 1) > \left(\frac{\kappa + 1}{2}\right) \). By the symmetry of the matrix \( Q \) the matrix \( F^* \) will have six distinct entries (see Supplement A),

\[
\begin{align*}
&\quad\quad\quad\quad p_{xxyy}^*, p_{xyxz}^*, p_{xyzx}^*, p_{yxxx}^*, p_{yxzw}^*, \text{ and } p_{xzyw}^*.
\end{align*}
\]

The diagonal \( (\kappa - 1) \times (\kappa - 1) \) blocks of \( F^* \) will be of the following form,

\[
\begin{pmatrix}
& p_{xxyy}^* & p_{xyxz}^* & p_{xyzx}^* & \cdots & p_{xzyw}^* \\
p_{xyxz}^* & p_{xxyy}^* & p_{xyxz}^* & p_{xyzx}^* & \cdots & p_{xyzw}^* \\
p_{xyzx}^* & p_{xyxz}^* & p_{xxyy}^* & p_{xyxz}^* & \cdots & p_{xzyw}^* \\
& \vdots & \vdots & \vdots & \ddots & \vdots \\
p_{xzyw}^* & p_{xyxz}^* & p_{xyzx}^* & p_{yxxx}^* & \cdots & p_{yxzw}^*
\end{pmatrix}.
\]

The off-diagonal \( (\kappa - 1) \times (\kappa - 1) \) blocks of \( F^* \) can be obtained from the block described below by appropriately permuting rows and columns.

\[
\begin{align*}
12 & \quad 13 & \quad 14 & \quad 15 & \quad \ldots & \quad 1\kappa \\
21 & \quad | & \quad p_{xxyy}^* & \quad p_{xyxy}^* & \quad p_{yxxx}^* & \quad \cdots & \quad p_{xyzw}^* \\
23 & \quad | & \quad p_{xyxy}^* & \quad p_{xxyy}^* & \quad p_{xyxy}^* & \quad \cdots & \quad p_{xyzw}^* \\
24 & \quad | & \quad p_{yxxx}^* & \quad p_{xxyy}^* & \quad p_{xyxy}^* & \quad \cdots & \quad p_{xyzw}^* \\
25 & \quad | & \quad p_{xyzw}^* & \quad p_{xyxy}^* & \quad p_{xyxy}^* & \quad \cdots & \quad p_{xyzw}^* \\
26 & \quad | & \quad p_{xyxy}^* & \quad p_{xyxy}^* & \quad p_{xyxy}^* & \quad \cdots & \quad p_{xyzw}^* \\
& \quad \vdots & \quad \vdots & \quad \vdots & \quad \ddots & \quad \vdots \\
2\kappa & \quad | & \quad p_{xzyw}^* & \quad p_{xyxy}^* & \quad p_{xyxy}^* & \quad \cdots & \quad p_{yxzw}^* \\
\end{align*}
\]

The submatrix \( F^* \) is symmetric. In addition, it is straightforward to show that row (column) sums are all the same and they equal to

\[
p_{xxyy}^* + p_{xyxy}^* + (\kappa - 2)(p_{xxyy}^* + 2p_{xyxy}^* + p_{yxxx}^* + (\kappa - 3)p_{xyzw}^*).
\]

With the choice of parameters made above and using precise formulas for the site pattern probabilities for the generalized Jukes-Cantor model as described in Section 6 and Supplement A, we find that \( F^* \) is strictly diagonally dominant. We say that a matrix \( A = (a_{ij}) \) is diagonally dominant if \( |a_{ii}| \geq \sum_{i \neq j} |a_{ij}| \) for all \( i \). It is well-known that strictly diagonally dominant matrices are nonsingular and if \( A \) is real symmetric with non-negative diagonal entries then it is positive definite. Thus, generically \( \text{rank}(F^*) = \kappa(\kappa - 1) \) (see Mathematica supplementary files for computations).
Since there exists at least one parameter choice in $U_S$ that does not lie in $W$, then $W$ is a proper analytic subvariety of $U_S$ for a 4-leaf species tree $S$ and hence of dimension strictly less than the dimension of $U_S$. We conclude that for a non-valid split $L_1 | L_2$

$$\text{rank}(\text{Flat}_{L_1 | L_2} (P^*_{(S, \tau)})) > \left( \frac{\kappa + 1}{2} \right),$$

for generic distributions $P^*_{(S, \tau)}$ arising from the model, establishing (2).

**Remark 5.2.** Recall that the probability distribution $P^*_{(S, \tau)}$ arises from a fixed 4-leaf species tree topology $S$. Thus, we can use the vanishing of the the $((\kappa+1) \times (\kappa+1))$-minors of the $\text{Flat}_{L_1 | L_2} (P^*_{(S, \tau)})$ to identify $S$ for generic parameters. In particular, using the notation of the Theorem 5.1 we conclude that the unrooted 4-leaf species tree $S$ is identifiable from $P^*_{(S, \tau)} = \psi_S(u)$ for any parameters $u \in U_S \setminus W$.

In the single 4-leaf gene tree case (for the general $\kappa$-state Markov model) under the molecular clock assumption, even though the columns of the flattening labeled by the $cd$-indices $kl$ and $lk$ for distinct $k, l \in \kappa$ are identical, the rank of the flattening for a valid split is less than or equal to $\kappa$. Thus, one might wonder if the rank of the flattening for the species tree under the coalescent model might be less than or equal to $\kappa$ for all distributions arising from the model $C_{GTR}(\kappa)$. For a valid split and using our computations for JC69 $\kappa$-state model under the coalescent (Section 6) we have found several $(\kappa+1) \times (\kappa+1)$ minors for $\kappa = 4$ that do not vanish identically. We believe that this is true in general for all $\kappa$, leading to the following conjecture.

**Conjecture 5.3.** Adopt the notation and assumptions of Theorem 5.1. If $L_1 | L_2$ is a valid split for $S$, then for all distributions $P^*_{(S, \tau)}$ arising from the model

$$\kappa < \text{rank}(\text{Flat}_{L_1 | L_2} (P^*_{(S, \tau)})) \leq \left( \frac{\kappa + 1}{2} \right).$$

Now we turn our attention to the identifiability of the unrooted $n$-taxon species tree from the induced quartets.

**Remark 5.4.** Let $S_n$ be an $n$-taxon species tree. We are going to show that distribution for $S_n$ can be marginalized to $S_{n-1}$. We demonstrate the idea on the 5-taxon species tree $S_5$; the $n$-taxon case follows easily from the one described below.

Let $\mathcal{Q}(S_5)$ be a collection of 4-leaf species trees that are induced by $S_5$. Without loss of generality, consider $S_5 = ((a, b), (c, (d, e)))$ as in Figure 3 and let $S \in \mathcal{Q}(S_5)$ be $((a, b), (d, e))$. We would like to show that we can express each site pattern probability for an induced quartet $S$ as,

$$p^{\ast}_{i_1 i_2 i_4 i_5} | (S_{(S, \tau)}) = \sum_{x \in [\kappa]} p^{\ast}_{i_1 i_2 x i_4 i_5} | (S_{5, \tau_x}).$$ (19)
by \( \{G_h\} \) the set of all histories for a gene tree \( G \) conditional on the species tree \( S \), and let \( \mathcal{H}_4 = \{\{G_h\}\} \) be the set of all histories for the 4-taxon species tree. \( \mathcal{H}_5 \) will denote the set of all histories for the 5-taxon tree \( S_5 \), and \( G_{h_5} \) a history for a gene tree \( G \) conditional on \( S_5 \). Let \( t^* \) be the coalescent time for taxon \( C \) and define \( t := t_\bullet \setminus t^* \), where \( t_\bullet \) is a vector of coalescent times for \( G_{h_5} \). Thus, \( t \) is a vector of coalescent times for \( G_{h_5} \). Now, using Equation 11 as applied to a 5-taxon species tree \( S_5 \), the right-hand side of Equation 19 is equal to,

\[
\sum_{x \in \kappa} p^*_x(x_{i_1i_2i_4i_5}|(S_5, \tau_\bullet)) = \sum_{x \in \kappa} \sum_{G_{h_5} \in \mathcal{H}_5} \int_{t_\bullet} p_{\sigma(i_1i_2i_4i_5)|(G_{h_5}, t_\bullet)} f(G_{h_5}, t_\bullet)|(S_5, \tau_\bullet) dt_\bullet
\]

\[
= \sum_{G_{h_5} \in \mathcal{H}_5} \int_{t_\bullet} \sum_{x \in \kappa} p_{\sigma(i_1i_2i_4i_5)|(G_{h_5}, t_\bullet)} f(G_{h_5}, t_\bullet)|(S_5, \tau_\bullet) dt_\bullet
\]

\[
= \sum_{G_{h_5} \in \mathcal{H}_5} \int_{t_\bullet} p_{\sigma(i_1i_2i_4i_5)|(G_{h_5}, t_\bullet)} f(G_{h_5}, t_\bullet)|(S_5, \tau_\bullet) dt_\bullet
\]

\[= \sum_{G_h \in \mathcal{H}_4} \sum_{G_{h_5} \in \mathcal{H}_5} \int_{t_\bullet} p_{\sigma(i_1i_2i_4i_5)|(G_h, t_\bullet)} f(G_{h_5}, t_\bullet)|(S_5, \tau_\bullet) dt_\bullet \tag{20}\]

The second sum in the last expression, \( \sum_{G_h \in \mathcal{H}_4} G_{h_5} \in \mathcal{H}_5 \), means that for each 4-taxon history we sum over all placements of the 5\textsuperscript{th} taxon. Define \( B^\ast \) to be the branch on which \( t^* \) occurs, and let \( B \) be the set of all branches regardless of where \( t^* \) occurs. Then according to Equation 7 we can write the gene tree density given \( S_5 \) as

\[
f(G_{h_5}, t_\bullet)|(S_5, \tau_\bullet) = \prod_{b \in B} f_{P_b}(t_{u_b}, t_{u_{b-1}}, \ldots, t_{v_{b+1}})
\]

\[
= \left( \prod_{b \in B \setminus B^\ast} f_{P_b}(t) \right) \cdot f_{P_{B^\ast}}(t, t^*) \tag{21}\]
Substituting $21$ into the last expression of $20$ we get

$$\sum_{x \in \kappa} p_{i_1i_2i_3i_4i_5}^x (S, \tau)$$

$$= \sum_{G_1 \in H_1} \sum_{G_2 \in H_2} \int_{t_*}^{t} p_{\sigma(i_1i_2i_3i_4i_5)} (G_1, t) f(G_2, t) \left( \prod_{b \in B \setminus B^*} f_{P_b}(t) \right) f_{P_{B^*}}(t, t^*) dt^* dt$$

$$= \sum_{G_1 \in H_1} \int_{t_*}^{t} p_{\sigma(i_1i_2i_3i_4i_5)} (G_1, t) \left( \prod_{b \in B \setminus B^*} f_{P_b}(t) \int_{t_*}^{t} f_{P_{B^*}}(t, t^*) dt^* \right) dt$$

$$= \sum_{G_1 \in H_1} \int_{t_*}^{t} p_{\sigma(i_1i_2i_3i_4i_5)} (G_1, t) f(G_2, t) \prod_{b \in B \setminus B^*} f_{P_b}(t) \int_{t_*}^{t} f_{P_{B^*}}(t, t^*) dt^* d\tau$$

Let $U_S$ denote the continuous parameter space for the species tree $S$ and $U_S$ the continuous parameter space for the induced 4-leaf tree $S$. From the argument above it is easy to see that we have a commutative diagram of analytic maps with $\alpha_S$ surjective and $\beta_S$ a marginalization map, e.g. we sum over indices for taxon C (Figure 3 and Equation 19).

$$U_S \xrightarrow{\psi_S} \Delta^{\kappa-1}$$

$$\alpha_S \downarrow \quad \downarrow \beta_S$$

$$U_S \xrightarrow{\psi_S} \Delta^{\kappa-1}$$

It is straightforward to extend all of these ideas to the $n$-taxon case. By applying an argument similar to that of Corollary 3 on page 1109 in Allman and Rhodes (2006) [4], we get the following result.

**Corollary 5.5.** Let $C_{S_n}$ denote the coalescent phylogenetic model for the $n$-taxon species tree $S_n$. Then the unrooted species tree $S_n$ is identifiable for generic parameters.

## 6 Generalized Jukes-Cantor Coalescent $\kappa$-state Model

To show generic identifiability in Theorem 5.1 we have used precise formulas for the generalized Jukes-Cantor $\kappa$-state model under the coalescent for a 4-taxon species tree. In this section we describe computations of the JC69 for a symmetric and asymmetric 4-leaf species trees.

As was mentioned in Section 2.2 if we let the rates of mutation and relative frequencies of the states at equilibrium be $\mu := \mu_1 = \cdots = \mu_{\frac{1}{2}\kappa(\kappa-1)}$ and $\pi = \left( \frac{1}{\kappa}, \frac{1}{\kappa}, \cdots, \frac{1}{\kappa} \right)$.
in equation [8], respectively, then the resulting model is the Mk model as described in [24], which is a generalized Jukes-Cantor $\kappa$-state model. The $\kappa \times \kappa$ instantaneous rate matrix $Q$ for the $\kappa$-state JC69 model is

$$Q = \alpha \begin{pmatrix} 1 - \kappa & 1 & \cdots & 1 \\
1 & 1 - \kappa & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1 - \kappa \end{pmatrix}, \quad (22)$$

where $\alpha = \frac{\mu}{\kappa}$ is the instantaneous rate of any transition between states. The transition probability matrix $P(t)$ takes the form,

$$P_{ij}(t) = \begin{cases} \frac{1}{\kappa} + \frac{\kappa-1}{\kappa}e^{-\mu t} & i = j, \\
\frac{1}{\kappa} - \frac{1}{\kappa}e^{-\mu t} & i \neq j. \end{cases} \quad (23)$$

The symmetry of the Jukes-Cantor model makes it possible to write each site pattern probability for a species tree concisely for any $\kappa \geq 2$. Notice that the only part of the site pattern probabilities $p_{i_1i_2i_3i_4}(S, \tau)$ for a species tree that depends on $\kappa$ is the site pattern probability for each gene tree.

6.1 Site pattern probabilities: gene tree

Consider a 4-leaf unrooted gene tree labeled as in Figure 4. Recall that the site pattern probability for a particular observation $i_1i_2i_3i_4$, $i_j \in [\kappa]$, and the branch length $v_e$, $e \in \{1, 2, 3, 4, 5\}$ is given by

$$p_{i_1i_2i_3i_4} = \sum_{r=1}^{\kappa} \sum_{s=1}^{\kappa} \pi_r P_{r1}(v_1)P_{r2}(v_2)P_{rs}(v_3)P_{si_3}(v_4)P_{si_4}(v_5). \quad (24)$$

To make notation simpler, we let $p_{ii_e}^e := P_{ii}(t_e)$ be the probability of no change in the state over a time interval of length $t_e$ and let $p_{ij}^e := P_{ij}(t_e)$ be the probability of a state change along the edge $e \in \{1, 2, 3, 4, 5\}$ with length $t_e$. Then the probability of the pattern $xxxx$, where $x \in [\kappa]$ is
\[ p_{xxxx} = \frac{1}{\kappa}(p_1^1 p_2^1 p_3^4 p_4^5 + (\kappa - 1)p_1^2 p_3^4 p_4^5 + (\kappa - 1)p_2^3 p_3^4 p_4^5 + (\kappa - 1)p_3^4 p_4^5) \]

\[ + (\kappa - 1)p_1^2 p_3^4 p_4^5 + (\kappa - 2)p_1^3 p_3^4 p_4^5). \]  

Indeed, this is easily computed by observing that when \( r = s = x \) in Equation (24) then we will have exactly one term of the form
\[ p_1^1 p_2^1 p_3^4 p_4^5. \]

For \( r = s \neq x \) we will have \((\kappa - 1)\) terms of the form
\[ p_1^1 p_2^1 p_3^4 p_4^5. \]

The other three cases for \( r \neq s \) are as follows:

- \( r = x \) and \( s \in [\kappa]\{r\} \), then we have \((\kappa - 1)\) terms of the form \( p_1^1 p_2^1 p_3^4 p_4^5 \);
- \( s = x \) and \( r \in [\kappa]\{s\} \), then we have \((\kappa - 1)\) terms of the form \( p_1^1 p_2^1 p_3^4 p_4^5 \);
- \( r, s \neq x \), then we have \((\kappa - 1)(\kappa - 2)\) terms of the form \( p_1^1 p_2^1 p_3^4 p_4^5 \).

The complete list of all site pattern probabilities on the single gene tree can be found in the Supplement A.

### 6.2 Site pattern probabilities: species tree

Now we are ready to describe computations for two 4-leaf species trees, the symmetric tree \(((a, b), (c, d))\) and the asymmetric tree \((a, (b, (c, d)))\), under the generalized JC69 coalescent \(\kappa\)-state model. For these species trees we have 15 rooted gene trees \((G, t)\) with 25 histories for a symmetric tree and 34 histories for an asymmetric species tree. Vectors \( \tau = (\tau_1, \tau_2, \tau_3) \) and \( t = (t_1, t_2, t_3) \) denote speciation and coalescent times, respectively. The gene tree density function and limits of integration depend on where the coalescent events happen. Figure 1 shows two particular histories out of 25 for a symmetric tree \(S\). The gene tree in Figure 1A call it History A, is symmetric and is in agreement with the tree \(S\). It has one coalescent event happening along the branch of length \(\tau_3 - \tau_1\) and two other events above the root of the tree \(S\). In contrast, the gene tree in Figure 1B call it History B, is asymmetric with lineages \(a\) and \(c\) being sister leaves and all coalescent events happening above the root of the tree \(S\). Using Equation (7) we compute the gene tree densities for Histories A and B as follows.

**History A gene tree density**

\[ f^A_{(G, t)| (S, \tau)} = \left(\frac{2}{\theta}\right)^3 e^{-2\tau_1} e^{-2\tau_2} e^{-2(t_3 - t_2)} e^{\omega} (\tau_3 - \tau_1). \]
History B gene tree density

\[ f_{G,t}^B((S,t) | (S,\tau)) = \left(\frac{2}{\theta}\right)^3 e^{\frac{\mu}{\theta} (\tau_3 - \tau_1)} e^{\frac{\mu}{\theta} (\tau_3 - \tau_2)} e^{\frac{\mu}{\theta} (t_2 - t_1)} e^{\frac{\mu}{\theta} (t_3 - t_2)}. \]

The site pattern probability for the pattern \(xxxx\) and Histories A and B for a symmetric species tree is now easily computed by appropriately substituting gene tree branch lengths as listed in Figure 1 into Equations (23), (24), and (25), multiplied by the gene tree density function and integrated with respect to \(t\). Notice the different limits of integration for the histories. As was mentioned previously, the limits depend on the location of coalescent events:

History A

\[ p_{xxxx}^*(S,\tau) = \int_0^\infty \int_0^t \int_0^t p_{xxxx} \cdot f_{G,t}^A((S,\tau)) dt_1 dt_2 dt_3. \]

History B

\[ p_{xxxx}^*(S,\tau) = \int_0^\infty \int_0^t \int_0^t p_{xxxx} \cdot f_{G,t}^B((S,\tau)) dt_1 dt_2 dt_3. \]

We perform these computations for all 25 histories for the symmetric species tree and for all 34 histories for the asymmetric species tree, sum them up and arrive at the following site pattern probabilities for observation \(xxxx\). The complete list for all patterns can be found in Supplement A. We list site pattern probabilities in the parameterized form for a cleaner output. For each \(i \in \{1, 2, 3\}\) let \(x_i = e^{-\tau_i}\), then

1) Symmetric species 4-leaf tree:

\[ p_{xxxx}^*(S,\tau) = \frac{1}{k^4} + \frac{(k - 1)x_1^{2\mu}}{\kappa^4(1 + \mu \theta)} + \frac{(k - 1)x_2^{2\mu}}{\kappa^4(1 + \mu \theta)} + \frac{(k - 1)^2x_1^{2\mu}x_2^{2\mu}}{\kappa^4(1 + \mu \theta)^2} + \frac{4(k - 1)x_{1,2}^{2\mu}}{\kappa^4(1 + \mu \theta)} + \frac{4(k - 2)(k - 1)x_4^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)(2 + \mu \theta)} + \frac{4(k - 2)^2(k - 1)x_1^{2\mu}x_2^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2(2 + \mu \theta)^2} + \frac{2(k - 1)\mu \theta(k + \mu \theta)(k + \mu \theta)x_1^{2\mu}x_2^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2(2 + \mu \theta)^2(3 + \mu \theta)}. \]

2) Asymmetric species 4-leaf tree:

\[ p_{xxxx}^*(S,\tau) = \frac{1}{k^4} + \frac{(k - 1)x_1^{2\mu}}{\kappa^4(1 + \mu \theta)} + \frac{2(k - 1)x_2^{2\mu}}{\kappa^4(1 + \mu \theta)} + \frac{2(k - 2)(k - 1)x_1^{2\mu}x_2^{2\mu}}{\kappa^4(1 + \mu \theta)(2 + \mu \theta)} + \frac{(k - 1)^2x_1^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2} + \frac{3(k - 1)x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2} + \frac{2(k - 2)(k - 1)x_1^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2} + \frac{(k - 1)^2x_1^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2} + \frac{4(k - 2)(k - 1)x_1^{2\mu}x_2^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)(2 + \mu \theta)^2} + \frac{4(k - 2)^2(k - 1)x_1^{2\mu}x_2^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2(2 + \mu \theta)^2} + \frac{2(k - 1)\mu \theta(k + \mu \theta)(k + \mu \theta)x_1^{2\mu}x_2^{2\mu}x_3^{2\mu}}{\kappa^4(1 + \mu \theta)^2(2 + \mu \theta)^2(3 + \mu \theta)}. \]
7 Species Tree Inference

The results presented here can be used to develop methodology for inferring species trees given DNA sequence data at the tips of the tree in cases where the coalescent model is appropriate. Indeed, we have recently developed and done some preliminary testing with one such method [10], and we briefly describe the basic ideas behind the method here. These ideas are modeled after the work of Eriksson [11] for the single gene case. In particular, consider a data set consisting of $R$ unlinked single nucleotide polymorphisms (SNPs) for a collection of $n$ species. Each SNP site is assumed to have its own gene tree, and by “unlinked” we mean that SNPs on the same chromosome are located far enough apart that their gene trees are independent given the species tree. In other words, we assume a data set of $R$ observations arising under the model in expression (12).

For a given subset of four species from the $n$ species under study, we can consider the three possible splits and their associated flattenings. In the empirical setting, we do not observe the flattening matrices, but these matrices can be estimated using the counts of the observed site patterns in the $R$ data points. The question of interest is then which of the three possible splits gives a flattening that is closest to a rank 10 matrix. The relevant rank is 10, because $\kappa = 4$ for DNA sequence data and $\binom{4+1}{2} = 10$.

To assess this, we compute the distance to the nearest rank 10 matrix in the Frobenius norm by defining the SVD score for a split $L_1 | L_2$, $SVD(L_1|L_2)$, to be

$$SVD(L_1|L_2) = \sqrt{\sum_{i=11}^{16} \hat{\sigma}_i^2}$$

where $\hat{\sigma}_i$ are the singular values of $Flat_{L_1|L_2}(\hat{P})$, for $i \in \{11, \ldots, 16\}$, and $Flat_{L_1|L_2}(\hat{P})$ refers to the flattening matrix for the split $L_1 | L_2$ estimated from the data. The split yielding the smallest score of the three possible splits can be inferred to be the valid split for the collection of four taxa. We show [10] that the SVD score has good ability to infer the correct split under a variety of conditions.

We note, also, that although the results presented in this paper apply to unlinked SNP data, many existing data sets consist of collections of genes for which the entire DNA sequence is available for each species. This setting is different than that considered here, in that each site in the DNA sequence for a specific gene is generally assumed to have arisen from the same gene tree. When the number of genes is large, these multi-locus data sets will approximately satisfy the model, and we expect that the inference method proposed here will still perform well. We tested this using simulation [10] and found that the method was still very effective at distinguishing the valid split.

To estimate the entire species tree, our proposed algorithm works by sampling collections of four taxa at random from those included in the data set. For each sampled collection of four taxa, the valid split is inferred by computation of the SVD score for the three possible splits. The collection of valid four-taxon splits can then be used in a quartet assembly algorithm to construct the species tree. Many possible
algorithms for quartet assembly have been proposed [32, 31, 30]. We find that the
method of Snir and Rao (2010) [30] is a fast and effective technique for constructing
the species tree estimate from the collection of inferred splits.

Based on the good performance of this method in our initial study, we are currently
working on expanding the scope of the data sets to which it can be applied. In partic-
ular, we will consider data set for which (i) multiple individuals are sampled within a
species; (ii) there are ambiguities in the observed nucleotide at a site, and (iii) sequence
data other than DNA, such as codon or amino acid data, have been collected. Overall,
the method shows good promise in addressing this important problem in phylogenetic
inference under the coalescent process, and we believe that it can be effectively used
for practical phylogenetic inference.

8 Conclusions

Within the last 15 years, numerous methods for inferring species-level phylogenies from
genome-scale data have been proposed, and several are commonly used in empirical
practice (e.g., [25, 17], [22], [20], [8]). However, identifiability of the species tree from
sequence data has never been formally established. We have shown here that the un-
rooted \(n\)-taxon species tree topology is identifiable from the distribution of site pattern
probabilities under the coalescent model with a general \(\kappa\)-state time-reversible substi-
tution model for the mutational process. This is an important step in understanding
phylogenetic estimation under more complicated models of DNA sequence evolution,
and it has already led to the development of a promising new method for inference [10].

Several important problems in this area remain to be addressed. For example, open
questions include whether the root of the species tree is identifiable, and whether other
parameters in the species tree (e.g., branch lengths and effective population sizes) and
in the substitution model (e.g., parameters in the transition probability matrix) are
identifiable under the model. We consider these questions in future work.

9 Acknowledgements

We thank Elizabeth Allman for helpful comments on an earlier draft of this manuscript.
We acknowledge support from the National Science Foundation under award DMS-
1106706 (to J. C. and L. K.); from the Mathematical Biosciences Institute at The
Ohio State University (J. C. and L. K.); and from NIH Cancer Biology Training Grant
T32-CA079448 at Wake Forest School of Medicine (J.C.).

References

[1] E. S. Allman, C. Ané, and J. A. Rhodes. Identifiability of a Markovian model of
molecular evolution with gamma-distributed rates. Advances in Applied Probabil-
ity, 40:228–249, 2008.
[2] 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.

[3] E. S. Allman, S. Petrovic, J. A. Rhodes, and S. Sullivant. Identifiability of 2-tree mixtures for group-based models. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, page to appear, 2010.

[4] E. S. Allman and J. A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13(5):1101–1113, 2006.

[5] E. S. Allman and J. A. Rhodes. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. *Mathematical Biosciences*, 211:18–33, 2008.

[6] E. S. Allman and J. A. Rhodes. The identifiability of covarion models in phylogenetics. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 6:76–88, 2009.

[7] Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. Determining species tree topologies from clade probabilities under the coalescent. *J. Theoret. Biol.*, 289:96–106, 2011.

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

[9] J. A. Cavender. Mechanized derivation of linear invariants. *Molecular Biology and Evolution*, 6(3):301–316, 1989.

[10] J. Chifman and L. Kubatko. Quartet inference from SNP data under the coalescent model. *submitted to Bioinformatics*, 2014.

[11] N Eriksson. Tree construction using Singular Value Decomposition. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, chapter 19, pages 347–358. Cambridge University Press, 2005.

[12] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–76, 1981.

[13] Y. Fu and W. Li. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201 – 228, 1992.

[14] YX. Fu. Linear invariants under Jukes’ and Cantor’s one-parameter model. *Journal of Theoretical Biology*, 173(4):339–352, 1995.

[15] R.C. Gunning and H. Rossi. *Analytic Functions of Several Complex Variables*. Ams Chelsea Publishing. AMS Chelsea Pub., 2009.
[16] M. Hasegawa, H. Kishino, and T. Yano. Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160174, 1985.

[17] J. Heled and A. J. Drummond. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, in press, 2010.

[18] T.H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–123. Academic Press, New York, 1969.

[19] M. Kimura. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111120, 1980.

[20] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[21] J. F. C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13:235–248, 1982.

[22] L. S. Kubatko, B. C. Carstens, and L. L. Knolwes. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973, 2009.

[23] J. A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution*, 4(2):167–191, 1987.

[24] Paul O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925, 2001.

[25] 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.

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

[27] Wayne P. Maddison. Gene trees in species trees. *Syst. Biol.*, 46:523–536, 1997.

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

[29] Bruce Rannala and Ziheng Yang. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 164:1645–1656, 2003.

[30] S. Snir and S. Rao. Quartet maxcut: A fast algorithm for amalgamating quartet trees. *Mol. Phylogen. Evol.*, 62:1–8, 2012.
[31] K. Strimmer, N. Goldman, and A. von Haeseler. Bayesian probabilities and quartet puzzling. *Mol. Biol. Evol.*, 14:210–213, 1997.

[32] K. Strimmer and A. von Haeseler. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969, 1996.

[33] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3):512526, 1993.

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

[35] S. Tavare. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences (American Mathematical Society*, 17:5786, 1986.)
Supplement A

Main article title: Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes

Julia Chifman  
Wake Forest School of Medicine

Laura Kubatko  
The Ohio State University

1 List of Site Pattern Probabilities for Generalized JC69 coalescent $\kappa$-state model

In this supplement we provide a list of site pattern probabilities for gene and species trees as described in Section 6 of the main article.

1.1 Four-leaf gene tree site pattern probabilities

\[ p_{xxxx} = \frac{1}{\kappa} \left(p_{i_1i_1}^{1,2,3}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}\right) \]

\[ p_{xxyy} = \frac{1}{\kappa} \left(p_{i_1i_1}^{1,2,3}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}\right) \]

\[ p_{xyyx} = \frac{1}{\kappa} \left(p_{i_1i_1}^{1,2,3}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}\right) \]

\[ p_{yxxy} = \frac{1}{\kappa} \left(p_{i_1i_1}^{1,2,3}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}\right) \]

\[ p_{yyxx} = \frac{1}{\kappa} \left(p_{i_1i_1}^{1,2,3}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1} + \left(\kappa - 1\right)p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}p_{i_1i_1}^{1}\right) \]
\[ p_\text{xxy} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{xzy} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{xzz} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{yyx} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{yyz} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{yyw} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]

\[ p_\text{zw} = \frac{1}{k} (p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + p_1^2 p_2^3 p_3^4 p_4^5 + p_1^2 p_2^1 p_3^1 p_4^1 p_5^1 + \cdots) \]
1.2 Symmetric four-leaf species tree

Each site pattern probability for a symmetric species 4-leaf tree will be of the following form:

\[
p^*_{s_t, s_i, s_3} = c_0 + c_1 x_1^{2\mu} + c_2 x_2^{2\mu} + c_3 x_1 x_2^{2\mu} + c_4 x_3^{2\mu} + c_5 x_1 x_3^{2\mu} + c_6 x_2 x_3^{2\mu} + c_7 x_1 x_2 x_3^{2\mu} + c_8 x_1 x_2 x_3^{2\mu},
\]

where \(x_i = e^{-\tau_i}, i \in \{1, 2, 3\}\), \(\kappa\) is the number of states, \(\theta > 0\) is the effective population size and \(\mu > 0\) is the instantaneous rate of any transition between states. Table 1 lists all the coefficients \(c_i\) for each pattern. Under the molecular clock assumption, rooted balanced 4-leaf species tree \((a, b), (c, d)\) will have 9 distinct site patterns probabilities

\[
\begin{align*}
p^*_{xxxx} &= p^*_{xxyx}, p^*_{xxyx} = p^*_{yxxx}, p^*_{yyxy} = p^*_{yyyy},
p^*_{xyxx} &= p^*_{yxxz}, p^*_{yzyx}, p^*_{zyyx}, p^*_{yzyz}.
\end{align*}
\]

Table 1: Coefficients for site pattern probabilities: symmetric species 4-leaf tree

|   | xxxx   | xxyy   | xyxx   |
|---|--------|--------|--------|
| \(c_0\) | \(\frac{1}{\kappa^4}\) | \(\frac{1}{\kappa^4}\) | \(\frac{1}{\kappa^4}\) |
| \(c_1\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) | \(-\frac{1}{\kappa^4(1+\mu\theta)}\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) |
| \(c_2\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) | \(-\frac{1}{\kappa^4(1+\mu\theta)}\) |
| \(c_3\) | \(-\frac{(2\kappa-1)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{(2\kappa-1)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{(2\kappa-1)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_4\) | \(\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(\frac{2(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(\frac{2(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_5\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_6\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_7\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_8\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |

|   | xxyy   | xxyz   |
|---|--------|--------|
| \(c_0\) | \(\frac{1}{\kappa^4}\) | \(\frac{1}{\kappa^4}\) | \(\frac{1}{\kappa^4}\) |
| \(c_1\) | \(-\frac{1}{\kappa^4(1+\mu\theta)}\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) | \(-\frac{1}{\kappa^4(1+\mu\theta)}\) |
| \(c_2\) | \(-\frac{1}{\kappa^4(1+\mu\theta)}\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) | \(\frac{(\kappa-1)}{\kappa^4(1+\mu\theta)}\) |
| \(c_3\) | \(-\frac{1}{\kappa^4(1+\mu\theta)^2}\) | \(\frac{(2\kappa-1)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{(2\kappa-1)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_4\) | \(\frac{2(\kappa-2)}{\kappa^4(1+\mu\theta)}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_5\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_6\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_7\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
| \(c_8\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) | \(-\frac{8(\kappa-2)}{\kappa^4(1+\mu\theta)^2}\) |
1.3 Asymmetric four-leaf species tree

Each site pattern probability for an asymmetric species 4-leaf tree will be of the following form:

\[
p_{i1i2i3i4}(S, \tau) = c_0 + c_1 x_1^{2\mu} + c_2 x_2^{2\mu} + c_3 x_1 x_2^{2\mu} + c_4 x_2 x_3^{2\mu} + c_5 x_1^{2\mu} x_3 + c_6 x_1^{2\mu} x_2 + c_7 x_2^{2\mu},
\]

where \( x_i = e^{-\tau_i}, i \in \{1, 2, 3\} \), \( \kappa \) is the number of states, \( \theta > 0 \) is the effective population size and \( \mu > 0 \) is the instantaneous rate of any transition between states. Table 2 lists all the coefficients \( c_i \) for each pattern. Under the molecular clock assumption, rooted asymmetric 4-leaf species tree \((a, (b, (c,d)))\) will have 11 distinct site patterns probabilities

\[
P^*_x_{xxx}, P^*_x_{xxxy} = P^*_y_{xxy}, P^*_x_{yxzx}, P^*_x_{xxyz}, P^*_x_{yxxz}, P^*_x_{yxzx}, P^*_x_{xyzx}, P^*_y_{xyxx}, P^*_x_{yxyx}, P^*_x_{yxxz}, P^*_x_{xyzw}, P^*_y_{xzyx}, P^*_x_{yzyx}, P^*_x_{xyzw}.
\]
| Coefficient | xxxx | xxyy | xyxx | yxxx |
|-------------|------|------|------|------|
| $c_0$       | $\frac{(\kappa-1)}{\kappa(1+\mu\theta)}$ | $\frac{1}{\kappa(1+\mu\theta)}$ | $\frac{(\kappa-1)}{\kappa(1+\mu\theta)}$ | $\frac{(\kappa-1)}{\kappa(1+\mu\theta)}$ |
| $c_1$       | $\frac{2(\kappa-1)}{\kappa(1+\mu\theta)(2+\mu\theta)}$ | $(\kappa-2)$ | $\frac{2}{\kappa(1+\mu\theta)}$ | $\frac{2(\kappa-1)}{\kappa(1+\mu\theta)}$ |
| $c_2$       | $\frac{2(\kappa-1)(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{2(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{2(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{2(\kappa-1)(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ |
| $c_3$       | $\frac{2(\kappa-2)(\kappa-1)}{\kappa(1+\mu\theta)(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa(1+\mu\theta)(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa(1+\mu\theta)(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa(1+\mu\theta)(2+\mu\theta)}$ |
| $c_4$       | $\frac{(\kappa-1)}{\kappa^2(1+\mu\theta)^2}$ | $\frac{(\kappa-1)}{\kappa^2(1+\mu\theta)^2}$ | $\frac{(\kappa-1)}{\kappa^2(1+\mu\theta)^2}$ | $\frac{(\kappa-1)}{\kappa^2(1+\mu\theta)^2}$ |
| $c_5$       | $\frac{4(\kappa-2)(\kappa-1)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{4(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{4(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{4(\kappa-2)}{\kappa^2(1+\mu\theta)(2+\mu\theta)^2}$ |
| $c_6$       | $\frac{4(\kappa-1)(\kappa-2)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4}$ |
| $c_7$       | $\frac{2(\kappa-1)(\kappa-2)(\kappa+\mu)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^4(3+\mu\theta)}$ |

| Coefficient | xxyy | xyxx | xxyz | yxxx |
|-------------|------|------|------|------|
| $c_0$       | $\frac{1}{\kappa^2}$ | $\frac{1}{\kappa^2}$ | $\frac{1}{\kappa^2}$ | $\frac{1}{\kappa^2}$ |
| $c_1$       | $\frac{(\kappa-1)}{\kappa^3(1+\mu\theta)}$ | $\frac{1}{\kappa^3(1+\mu\theta)}$ | $\frac{1}{\kappa^3(1+\mu\theta)}$ | $\frac{1}{\kappa^3(1+\mu\theta)}$ |
| $c_2$       | $\frac{2(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ | $(\kappa-2)$ | $\frac{2}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ | $\frac{2(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ |
| $c_3$       | $\frac{2(\kappa-2)(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ | $\frac{2(\kappa-2)(\kappa-1)}{\kappa^3(1+\mu\theta)^2(2+\mu\theta)}$ |
| $c_4$       | $\frac{(\kappa-1)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{(\kappa-1)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{(\kappa-1)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{(\kappa-1)}{\kappa^3(1+\mu\theta)^3}$ |
| $c_5$       | $\frac{4(\kappa-2)(\kappa-1)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{4(\kappa-2)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{4(\kappa-2)}{\kappa^3(1+\mu\theta)^3}$ | $\frac{4(\kappa-2)}{\kappa^3(1+\mu\theta)^3}$ |
| $c_6$       | $\frac{4(\kappa-1)(\kappa-2)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4}$ | $\frac{4(\kappa-2)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4}$ |
| $c_7$       | $\frac{2(\kappa-1)(\kappa-2)(\kappa+\mu)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4(3+\mu\theta)}$ | $\frac{2\mu(\kappa-1)(\kappa-2)(\kappa+\kappa-1)(\kappa-\theta)}{\kappa^5(1+\mu\theta)^4(2+\mu\theta)^4(3+\mu\theta)}$ |
One checks that for all observations $\sigma = i_1 i_2 i_3 i_4$, $i_j \in \kappa$

$$
\sum_i P^*_\sigma(S, \kappa) = \kappa p_{xxxx}^* + 2\kappa(\kappa - 1)p_{xxyy}^* + \kappa(\kappa - 1)p_{xyxx}^* + \kappa(\kappa - 1)p_{yxxx}^* + \kappa(\kappa - 1)p_{xxyy}^*
$$
$$
+ 2\kappa(\kappa - 1)p_{xyxy}^* + \kappa(\kappa - 1)(\kappa - 2)p_{xyxz}^* + 2\kappa(\kappa - 1)(\kappa - 2)p_{yxxz}^* + 2\kappa(\kappa - 1)(\kappa - 2)p_{yxyz}^* + \kappa(\kappa - 1)(\kappa - 2)p_{yzxx}^* + \kappa(\kappa - 1)(\kappa - 2)(\kappa - 3)p_{xyzw}^* = 1.
$$

| \multicolumn{1}{c}{c_0} | \multicolumn{1}{c}{\frac{1}{\kappa^4}} | \multicolumn{1}{c}{\frac{1}{\kappa^4}} | \multicolumn{1}{c}{\frac{1}{\kappa^4}} |
|-------------------|-------------------|-------------------|
| \multicolumn{1}{c}{c_1} | $-\frac{1}{\kappa^4(1+\mu\theta)}$ | $-\frac{1}{\kappa^4(1+\mu\theta)}$ | $-\frac{1}{\kappa^4(1+\mu\theta)}$ |
| \multicolumn{1}{c}{c_2} | $-\frac{2}{\kappa^4(1+\mu\theta)}$ | $\frac{(\kappa - 2)}{\kappa^4(1+\mu\theta)}$ | $-\frac{2}{\kappa^4(1+\mu\theta)}$ |
| \multicolumn{1}{c}{c_3} | $\frac{4}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ | $-\frac{2(\kappa - 2)}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ | $\frac{4}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ |
| \multicolumn{1}{c}{c_4} | $-\frac{2(\kappa - 2)}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ | $\frac{4}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ | $\frac{4}{\kappa^4(1+\mu\theta)(2+\mu\theta)}$ |
| \multicolumn{1}{c}{c_5} | $\frac{4(\kappa - 4)}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{4(\kappa - 4)}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ | $-\frac{8}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ |
| \multicolumn{1}{c}{c_6} | $\frac{4(\kappa - 4)}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ | $\frac{4(\kappa - 4)}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ | $-\frac{16}{\kappa^4(1+\mu\theta)(2+\mu\theta)^2}$ |
| \multicolumn{1}{c}{c_7} | $-\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $-\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $\frac{2\mu^3 \theta^3}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ |
| \multicolumn{1}{c}{c_8} | $\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $\frac{2\mu^3 \theta^3}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ |
| \multicolumn{1}{c}{c_9} | $\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $\frac{\mu^2 \theta^2 \kappa(\kappa - 2\mu\theta)}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ | $\frac{2\mu^3 \theta^3}{\kappa^4(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$ |
Main article title: Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes

Julia Chifman, Wake Forest School of Medicine
Laura Kubatko, The Ohio State University

Site Pattern probabilities for generalized JC69 under the coalescent model for the symmetric 4-leaf species tree.
In the proof we set $\tau_3 = \tau_2$ which implies that asymmetric species tree site pattern probabilities will coincide with symmetric species tree. For cleaner description we set $x_i = \exp[-\tau_i]$, $n = \Theta$, $m = \mu$. 

\[
\begin{align*}
\text{P}_{x,x,x} &= \frac{1}{k^4} \left[ \frac{(-1+k)x_1^{2n}}{k^4(1+mn)} + \frac{(-1+k)x_2^{2n}}{k^4(1+mn)} + \frac{(-1+k)x_1^{2n}x_2^{2n}}{k^4(1+mn)^2} \right] \\
&+ \frac{4(-1+k)x_3^{2n}}{k^4(1+mn)} + \frac{4(-2+k)(-1+k)x_1^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)} + \frac{4(-2+k)(-1+k)x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)} \\
&+ \frac{4(-2+k)^2(-1+k)x_1^{2n}x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)}
\end{align*}
\]

\[
\begin{align*}
\text{P}_{x,y,x} &= \frac{1}{k^4} \left[ \frac{x_1^{2n}}{k^4(1+mn)} + \frac{(-1+k)x_2^{2n}}{k^4(1+mn)} \right] \\
&+ \frac{(-1+k)x_1^{2n}x_2^{2n}}{k^4(1+mn)} + \frac{2(-2+k)x_3^{2n}}{k^4(1+mn)(2+mn)} + \frac{4(-2+k)(-1+k)x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)} \\
&+ \frac{4(-2+k)^2x_1^{2n}x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)}
\end{align*}
\]

\[
\begin{align*}
\text{P}_{y,y,x} &= \frac{1}{k^4} \left[ \frac{(-1+k)x_1^{2n}}{k^4(1+mn)} - \frac{x_2^{2n}}{k^4(1+mn)} \right] \\
&+ \frac{(-1+k)x_1^{2n}x_2^{2n}}{k^4(1+mn)} + \frac{2(-2+k)x_3^{2n}}{k^4(1+mn)(2+mn)} + \frac{4(-2+k)(-1+k)x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)} \\
&+ \frac{4(-2+k)^2x_1^{2n}x_2^{2n}x_3^{2n}}{k^4(1+mn)(2+mn)}
\end{align*}
\]
Computing Determinant for \( k = 2 \) and \( k = 3 \).

\[
\text{FullSimplify} @ \text{Det}[\text{Flat2}] = \frac{390625 \left(-2 + e^{9/50} \left(2 + 301 \left(-1 + e^{9/50}\right)^2 e^{891/50}\right)\right)}{31322180701 e^{891/50}}
\]

One checks that the numerator in the above output is not zero:

\[
\text{N} \left[-2 + e^{9/50} \left(2 + 301 \left(-1 + e^{9/50}\right)^2 e^{891/50}\right)\right] = 7.68701 \times 10^8
\]

\text{PositiveDefiniteMatrixQ}[\text{Flat2}] = \text{True}
$k := 3$
\[n := 1 \text{ to } 10 \]
\[m := 1 \text{ to } 10 \]
\[x_3 := \exp(-1) \]
\[x_2 := \exp(-1) \]
\[x_1 := \exp(-1 \text{ to } 10) \]

Flat3 := \{\]
\{\text{Pxxxx}\{k, n, m, x_3, x_1, x_2\}, \text{Pxxx}\{k, n, m, x_3, x_1, x_2\}, \text{Pxxxy}\{k, n, m, x_3, x_1, x_2\}, \text{Pxyxz}\{k, n, m, x_3, x_1, x_2\}, \text{Pxyyz}\{k, n, m, x_3, x_1, x_2\}, \text{Pxyx}\{k, n, m, x_3, x_1, x_2\}, \text{Pxyy}\{k, n, m, x_3, x_1, x_2\}, \text{Pyxz}\{k, n, m, x_3, x_1, x_2\}, \text{Pyx}\{k, n, m, x_3, x_1, x_2\}, \text{Pyx}\{k, n, m, x_3, x_1, x_2\}, \text{Pyxy}\{k, n, m, x_3, x_1, x_2\}, \text{Pyzz}\{k, n, m, x_3, x_1, x_2\}, \text{Pyz}\{k, n, m, x_3, x_1, x_2\}, \text{Pyy}\{k, n, m, x_3, x_1, x_2\}, \text{Pyy}\{k, n, m, x_3, x_1, x_2\}, \text{Pzz}\{k, n, m, x_3, x_1, x_2\}, \text{Pz}\{k, n, m, x_3, x_1, x_2\}, \text{P}\{k, n, m, x_3, x_1, x_2\} \}
\}

Simplify[Det[Flat3]]
\[
\{-16281603 + 32563206 \exp(-5/9) - 16281603 \exp(-25/9) + 3264481602 \exp(-18) - 646318876800 \exp(-100/100) + 62999654952 \exp(-90/9) + 1305727838400 \exp(-25/100) - 1282876628790 \exp(-15/9) - 65940896100 \exp(-25/9) + 649615597596 \exp(-25/50) + 129260478477600 \exp(-25/50) + 333628323124802 \exp(-25/50) + 1576026656320000 \exp(-25/100) + 17171110664245600 \exp(-25/50) + 339552035523202 \exp(-25/50) + 2560824025632000 \exp(-25/50) + 32906309963202 \exp(-25/50) + 105750443385600000 \exp(-25/50) + 209961062497121600 \exp(-100/100) + 294350967971276798 \exp(-95/9) - 981470180993449600 \exp(-100/100) + 528859084108976808 \exp(-25/50) + 371913222316809600 \exp(-80/9) + 344487045710499210 \exp(-25/100) - 26640781457601600 \exp(-100/100) + 5301910512365604 \exp(-100/100) + 213120976389120000 \exp(-100/100) - 9590523067238656000 \exp(-50/9) + 143859032964279044000 \exp(-25/50) - 399611722157819264020 \exp(-25/50) - 95907604564234886400 \exp(-100/100) + 71931296195724812805 \exp(-25/50) + 1598486452973122400 \exp(-25/100) + 23977494628710728004 \exp(-25/100) + 2664210034494121601 \exp(-100/100) \}
\]
\}

One checks that the numerator in the above output is not zero:
Diagonal Dominance calculation for $k >= 4$.
In particular, we show that for our choice of parameters $P_{xyxy} - (P_{xxxy} + (k-2)(P_{xyxy} + (k-3)P_{xyyw})) > 0$, which establishes that $F^{*} \hat{\text{ }}$ is strictly diagonally dominant and hence generically invertible.

```
k := k
n := 1/10
m := 1/10
x3 := Exp[-1]
x2 := Exp[-1]
x1 := Exp[-1/10]
```

```
N[Collect[P_{xyxy}[k, n, m, x3, x1, x2] - (P_{xxxy}[k, n, m, x3, x1, x2] + (k - 2)(P_{xyxy}[k, n, m, x3, x1, x2] + (k - 3)P_{xyyw}[k, n, m, x3, x1, x2]), k]]
```

\[
0.00156097 + 0.0836308 \frac{k^4}{k^3} + 0.037729 \frac{k^2}{k^2}
\]

A different look at the same calculation.
Collect[Simplify[Pxxxy{k, n, m, x3, x1, x2} - (Pxxxy{k, n, m, x3, x1, x2} + (k - 2) (Pxxxy{k, n, m, x3, x1, x2} + 2 * Pxyz[k, n, m, x3, x1, x2] + Pyzxx[k, n, m, x3, x1, x2] + (k - 3) Pxyzw[k, n, m, x3, x1, x2])]], k]