J. Appl. Prob. **51,** 921–929 (2014)
Printed in England

IDENTIFIABILITY OF A COALESCENT-BASED POPULATION TREE MODEL

ARINDAM ROYCHOUDHURY,* *Columbia University*

Abstract

Identifiability of evolutionary tree models has been a recent topic of discussion and some models have been shown to be nonidentifiable. A coalescent-based rooted population tree model, originally proposed by Nielsen *et al.* (1998), has been used by many authors in the last few years and is a simple tool to accurately model the changes in allele frequencies in the tree. However, the identifiability of this model has never been proven. Here we prove this model to be identifiable by showing that the model parameters can be expressed as functions of the probability distributions of subsamples, assuming that there are at least two (haploid) individuals sampled from each population. This a step toward proving the consistency of the maximum likelihood estimator of the population tree based on this model.

Keywords: Population tree; phylogenetic tree; identifiability; coalescent

2010 Mathematics Subject Classification: Primary 92B10
Secondary 60G99; 62P10; 92D15

1. Introduction

A rooted evolutionary tree is a directed weighted tree graph; it represents the evolutionary relationship between groups (also called taxa) of organisms (Figure 1(a)). A leaf or a tip is a node with degree 1; each tip represents a modern day taxon. The root (node 0) represents the most recent common ancestor (MRCA) of all the taxa. The direction (of evolution) is from the root to the tips.

A rooted population tree is a rooted evolutionary tree where the taxa are populations from the same species. Two types of parameters are common in any model of the rooted population tree: the tree-topology parameter (a categorical parameter) for the whole tree, and a branch parameter for each branch (also called edge).

The tree topology is the order in which the path from the root separates for the given set of populations; it is represented as a directed tree graph without the edge weights. (In Figures 1(a) and 1(b), the two trees have different tree topologies for the populations 1–4.) A branch parameter is usually a branch length (an edge-weight) or a transition probability matrix that influences the change in allele frequency between the two nodes of a branch. Here we will prove the identifiability of a population tree model by [9] and [12] that uses Kingman's coalescent process [5]. The model was later modified and expanded by various authors, see, for instance ([2], [8], [10], [11], [12]). Coalescent-based models are of significant importance as they model the underlying allele frequency changes with accuracy and relative ease (see [6]).

Due to the underlying structure in evolutionary tree-based models, its identifiability is never obvious. The identifiability of certain evolutionary tree models have been a recent topic of

Received 1 March 2013; revision received 10 January 2014.

* Postal address: Department of Biostatistics, Columbia University, 6th Floor, 722 W. 168th Street, New York, NY 10032, USA. Email address: ar2946@columbia.edu

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

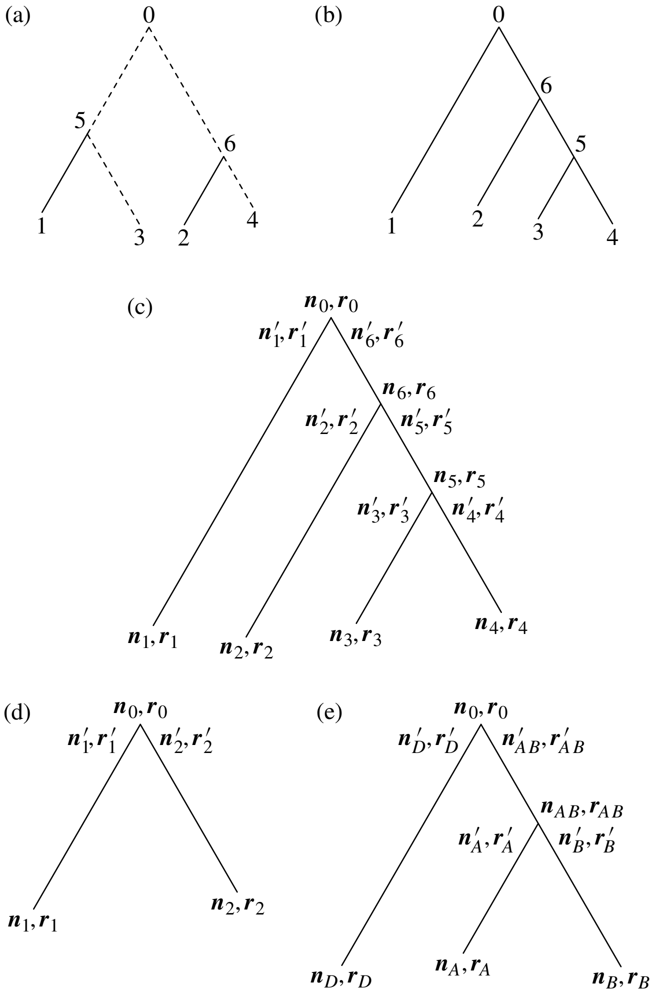

FIGURE 1: (a), (b): Population trees with different tree topologies; (c) Notations. (d) Type-I subtree. (e) Type-II subtree.

discussion. Allman *et al.* [1] proved the identifiability of a general time reversible (GTR) transition probability matrix-based model. Nonidentifiability of another time reversible model was established in [13]. The nonidentifiability of mixture models have been discussed in [7].

The identifiability for the model by Felsenstein [4] has been proven in [3]. To the best of the author's knowledge the identifiability of the coalescent-based model of [9] and [12] has never been proven.

For estimating evolutionary trees each independent genetic locus is viewed as a single data point, as opposed to viewing each individual as a data point; see, for example, [10]. Thus, identifiability would mean that the model parameters can be identified from the distribution of allele-types for a set of individuals at a single genetic locus.

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

2. The model

In this section we will describe the underlying model of [9] and [12]. First we provide an intuitive description of the model. In this model, a coalescent theoretic framework is assumed, where the lineages of observed haploid samples coalesce as they go back in time and, consequently, reduce in number. The lineages may coalesce at a given time point only if they belong to the same population (i.e. same branch or edge) at that time point. When two populations have merged (going back in time from the present) their lineages are able to coalesce. The allele-types are assigned (according to a 'root distribution') at the root among the lineages that are still left to coalesce at the root, and then the allele-types are tracked to the present day through the lineages. By tracking the number of coalescent events and allele-types along these lineages, the probabilities of observed allele-types are determined for each individual at each population.

Next we provide a mathematical description of the model. We start by defining our notations (see also Figure 1(c)). We define a P-tip population tree as $T = (\Lambda^{(P)}, \boldsymbol{\Psi}, \boldsymbol{\theta})$. The parameter $\Lambda^{(P)}$ is the tree topology, an unweighted directed tree graph. It takes finitely many discrete categorical values; the (P) in the superscript denotes the number of tips. The parameter $\boldsymbol{\Psi} = (\tau_1, \tau_2, \dots, \tau_{2P-2})$ is a vector of length $2P - 2$ consisting of the branch lengths τ_i for each branch i in $\Lambda^{(P)}$. A strictly bifurcating tree topology has exactly $2P-2$ branches. If $\Lambda^{(P)}$ is non-bifurcating then it has fewer branches and the remaining elements of $\boldsymbol{\Psi}$ are populated by zeros. The parameter $\boldsymbol{\theta}$ is a vector containing the parameters of root distribution which we will define later in this section. We also define $S(\Lambda^{(P)})$ as the set of tips at $\Lambda^{(P)}$.

At each tip z there are $n_z (\geq 1)$ lineages, each having allele-type '0' or '1'. The allele types among these lineages at each tip are the observable random variables. Similarly, at each non-tip node x, the random variable $n_x (\geq 1)$ is the (random) number of lineages that are ancestral to the tips below x along the tree. We also define the random variable r_x at each node x (tip or non-tip), as the count of allele '1' among the n_x lineages. In this paper we will use the term 'allele-count' to refer to the count of allele '1'. For each tip z, the allele-count r_z is observable.

Consider a branch with lower (towards the tips) node x and upper (towards the root) node y. Let n'_x be the number of lineages in y that are ancestral to the n_x lineages at x ($n'_x \leq n_x$). Also, let r'_x be the allele-count among these n'_x lineages ($r'_x \leq r_x$). If y is the upper node of ν branches with lower nodes $x_1, x_2, \dots, x_\nu$ then $n_{x_1}, n_{x_2}, \dots, n_{x_\nu}$ are independent, and

$$n_y = \sum_{k=1}^{\nu} n'_{x_k} \tag{2.1}$$

and also $r_y = \sum_{k=1}^{\nu} r'_{x_k}$. (For a strictly bifurcating tree $\nu = 2$.)

From the model parameters $T = (\Lambda^{(P)}, \boldsymbol{\Psi}, \boldsymbol{\theta})$ we compute the probability of observed vector of allele-counts $\boldsymbol{r} = (r_1, r_2, \dots, r_P)$ from samples of sizes $\boldsymbol{n} = (n_1, n_2, \dots, n_P)$ at P tips $(1, 2, \dots, P)$ as follows. Consider a branch with length τ_{x_1}, with upper node y and lower node x_1. Given the probability mass function (PMF) of n_{x_1} (the number of lineages at x_1), the PMF of n'_{x_1} is computed as

$$\mathbb{P}_{\boldsymbol{n}}(n'_{x_1} = i' \mid n_{x_1} = i; \tau_{x_1}) = \biggl(\prod_{j=i'+1}^{i} \lambda_j\biggr) \sum_{j=i'}^{i} \frac{\mathrm{e}^{-\lambda_j \tau_{x_1}}}{\prod_{j'=i', j' \neq j}^{i} (\lambda_{j'} - \lambda_j)}, \tag{2.2}$$

where $\lambda_j = j(j-1)/2$ and $\mathbb{P}_{\boldsymbol{n}}$ is the probability under the samples of size $\boldsymbol{n} = (n_1, n_2, \dots, n_P)$ [14]. Then, the PMF of n_y is determined from (2.1).

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

Using (2.1) and (2.2), starting from $\boldsymbol{n} = (n_1, n_2, \dots, n_P)$ and going upward, we compute the PMF of n_z and n'_z for any non-tip non-root node z, and finally n_0 at the root (node 0). Then a 'root distribution' with parameter $\boldsymbol{\theta}$ gives the PMF of (allele-count) r_0 given n_0 at the root

$$\mathrm{G}(0) = (\mathbb{P}_{\boldsymbol{n}}(r_0 = j \mid n_0 = i; \boldsymbol{\theta}),\ j = 0, 1, \dots, n_0; i = 1, 2, \dots, m_0^{(\mathrm{b})}),$$

where

$$m_0^{(\mathrm{b})} = \sum_{z \text{ is a tip}} n_z$$

is the maximum possible value of n_0 (number of lineages at the root). Different authors have used different root distributions. In particular [12] used a symmetric beta-binomial distribution

$$\mathbb{P}_{\boldsymbol{n}}(r_0 = j \mid n_0 = i; \theta) = \binom{i}{j} \frac{\beta(j + \theta)\beta(i - j + \theta)}{\beta(\theta, \theta)}, \tag{2.3}$$

where $\beta(\cdot, \cdot)$ is the beta function; $\theta > 0$ is a parameter to be estimated.

Then, from the distribution of n_0, r_0, and (n_z, n'_z) for all non-root nodes z, we compute the distribution of r_z (allele-counts) at the rest of the nodes as follows. Consider a node y where ν branches merge from the bottom with the bottom nodes $x_1, x_2, \dots, x_\nu$. Recall that we already have the distributions of n_y, n_{x_i}, and n'_{x_i}, $i = 1, 2, \dots, \nu$. The PMF of r'_{x_i} is computed from the PMF of r_y using

$$\begin{aligned}\mathbb{P}_{\boldsymbol{n}}(r'_{x_1} = j'_1,\ r'_{x_2} = j'_2, \dots, r'_{x_\nu} = j'_\nu \mid r_y = j,\ n_y = i,\ n'_{x_1} = i'_1,\ n'_{x_2} = i'_2, \dots, n'_{x_\nu} = i'_\nu)\\ = \frac{\binom{j}{j'_1, j'_2, \dots, j'_\nu}\binom{i-j}{i'_1 - j'_1, i'_2 - j'_2, \dots, i'_\nu - j'_\nu}}{\binom{i}{i'_1, i'_2, \dots, i'_\nu}}.\end{aligned} \tag{2.4}$$

Then the PMF of r_{x_k} is computed from the above PMF using the following (from an expression in [12]):

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{n}}(r_{x_k} = j_k \mid r'_{x_k} = j'_k,\ n'_{x_k} = i'_k,\ n_{x_k} = i_k)\\ &\quad = \begin{cases} \dfrac{\beta(j_k, i_k - j_k)}{\beta(j'_k, i'_k - j'_k)} \dbinom{i_k - i'_k}{j_k - j'_k}, & 0 < j_k < i_k \text{ and } 0 < j'_k < i'_k, \\ 1, & 0 = j_k = j'_k \text{ or } 0 = i_k - j_k = i'_k - j'_k, \\ 0, & \text{otherwise;} \end{cases}\end{aligned} \tag{2.5}$$

$k = 1, 2, \dots, \nu$ [12]. Thus, starting with G(0) at the root, we compute the joint PMF of $(r_1, r_2, \dots, r_P)$ from (2.4) and (2.5). Note that in (2.1), (2.2), (2.3), (2.4), and (2.5), probability 'flows' up along ns and then flows down along rs.

Now that we have completely described the model, we will proceed to prove the identifiability of this model in the next section.

3. Identifiability

Let $T = (\Lambda^{(P)}, \boldsymbol{\Psi}, \boldsymbol{\theta})$ be a tree with $S(\Lambda^{(P)}) = \{1, 2, \dots, P\}$. We define a subtree T^* of T as a tree formed by a subset S^* (cardinality $P' \le P$) of $S(\Lambda^{(P)})$ by tracking the tips in S^* along the tree to their MRCA node. Thus, $T^* = (\Lambda^{(P')*}, \boldsymbol{\Psi}^*, \boldsymbol{\theta})$, where $\Lambda^{(P')*}$ is the tree topology with P' tips of S^*. For example, in Figure 1(a), $P = 4$, $S(\Lambda^{(4)}) = \{1, 2, 3, 4\}$, $S^* = \{3, 4\}$ and the subtree T^* is drawn with the dotted lines.

Consider two distinct trees $T_1 = (\Lambda_1^{(P)}, \boldsymbol{\Psi}_1, \boldsymbol{\theta}_1)$ and $T_2 = (\Lambda_2^{(P)}, \boldsymbol{\Psi}_2, \boldsymbol{\theta}_2)$ with a common set of tips $S_{T_{1,2}} = S(\Lambda_1^{(P)}) = S(\Lambda_2^{(P)})$.

If $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$ then there must be at least one doubleton subset $\{z_1, z_2\} \subseteq S_{T_{1,2}}$ with the following property: the subtrees $T_1^* = (\Lambda^{(2)}, \boldsymbol{\Psi}_1^*, \boldsymbol{\theta})$ and $T_2^* = (\Lambda^{(2)}, \boldsymbol{\Psi}_2^*, \boldsymbol{\theta})$, formed by tracking z_1 and z_2 to the root in T_1 and T_2 (respectively), are distinct. That is, if $\boldsymbol{\Psi}_l^* = (\tau_{1l}, \tau_{2l})$ and τ_{jl} is the path distance (total branch length) between z_j and the MRCA of z_1 and z_2 along the subtree T_l^* $(j, l = 1, 2)$, then $(\tau_{11}, \tau_{21}) \neq (\tau_{12}, \tau_{22})$. (Note that there is only one possible tree topology for a two-tip tree, denoted as $\Lambda^{(2)}$ above.) Thus, the set of all two tip subtrees, along with $\boldsymbol{\theta}$, uniquely identifies the tree.

We assign the two-tip subtrees into two categories: Type-I subtrees are those with the root as the MRCA of the two tips. For example in Figure 1(a), the subtree formed by tips $\{3, 4\}$ has the root as the MRCA of the two tips 3 and 4. Thus, it is of type-I. All other two-tip subtrees are type-II subtrees. For example, in Figure 1(a), if a subtree is formed by tips 2 and 4, it will be a type-II subtree as their MRCA is node 6, and not the root. We will deal with these two types of subtrees separately.

We note that the parameters of the root distribution of (2.3) from [12] are identifiable, given the lineage and allele count distribution at the root (as it is beta-binomial). Next, we will prove the identifiability of the whole model by assuming a general identifiable root distribution that has parameter vector $\boldsymbol{\theta}$. (In particular, our proof would work with beta-binomial as the root distribution.)

Theorem 3.1. *Suppose that we have a tree T with the underlying model as described in Section 2. Also, suppose that we have $N_k \geq 2$ lineages sampled at each tip k and the root distribution is identifiable. Then the parameters of T are identifiable from the distribution of allele types at the tips.*

To prove the above theorem, we will show that the parameters of each two-tip subtree can be expressed as a function of the joint PMF

$$(\mathbb{P}_{\boldsymbol{n}}((R_1, R_2, \ldots, R_P) = (J_1, J_2, \ldots, J_P); T),\ J_k = 0, 1, 2, \ldots, N_k,\ k = 1, 2, \ldots, P).$$

This will complete the proof as the set of all two-tip subtrees, along with $\boldsymbol{\theta}$, uniquely identifies the tree.

3.1. Identifiability of type-I subtrees

Suppose that $T = (\Lambda^{(2)}, \{\tau_1, \tau_2\}, \boldsymbol{\theta})$ is a type-I subtree with the underlying model as described in Section 2. Let z_1 and z_2 be its two tips. Let the root be denoted as '0' (Figure 1(d)) and let τ_k be the path distance between z_k and the root $(k = 1, 2)$.

Proposition 3.1. *Suppose that we have at least two lineages sampled at each of z_1 and z_2 and the root distribution is identifiable. Then τ_1, τ_2, and $\boldsymbol{\theta}$ can be expressed as functions of the joint PMF of allele types in z_1 and z_2, and, hence, they are identifiable.*

Proof. Suppose that we have samples of N_1 and N_2 lineages from z_1 and z_2, respectively, and the allele-counts among these lineages are R_1 and R_2, respectively. Let the joint PMF of (R_1, R_2) be $f_{\boldsymbol{N},\boldsymbol{R}}$. Consider random subsamples (without replacement) of size n_1 and n_2 from z_1 and z_2, respectively, with $n_k \leq 2$, $k = 1, 2$. Rather than working with the allele-counts R_k at the original samples, we will work with allele-counts r_k at the subsamples.

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

We compute the joint PMF of (r_1, r_2) from $f_{\boldsymbol{N},\boldsymbol{R}}$ as

$$\mathbb{P}_{\boldsymbol{n}}(r_k = j_k,\ k = 1,2 \mid R_k = J_k,\ N_k = I_k,\ n_k = i_k,\ k = 1,\ 2;\ \tau_1, \tau_2, \boldsymbol{\theta})$$
$$= \sum_{J_1=j_1}^{I_1-(i_1-j_1)} \sum_{J_2=j_2}^{I_2-(i_2-j_2)} \left(\prod_{k=1}^{2} \frac{\binom{J_k}{j_k}\binom{I_k-J_k}{i_k-j_k}}{\binom{I_k}{i_k}} \right) f_{\boldsymbol{N},\boldsymbol{R}}(J_1, J_2).$$

We will argue that in order to identify the parameters τ_1, τ_2, and $\boldsymbol{\theta}$ it is enough to use the joint PMFs (r_1, r_2) for $(n_1, n_2) = (1,1), (1,2)$ and $(2,1)$. As before, let n'_k be the number of lineages ancestral to subsamples at z_k that are present at the top node (the root) (see Figure 1(d)) and r'_k be the allele-count out of these n'_k; $(k = 1, 2)$. Also, let $n_0 = n'_1 + n'_2$ be the number of lineages at the root ancestral to the subsampled lineages at z_1 and z_2, and $r_0 = r'_1 + r'_2$ be the allele-count out of these n_0 lineages. First, consider the case $n_1 = n_2 = 1$. Then $r_k = 0$ or 1 for $k = 1, 2$. From (2.2) it follows that $n'_1 = n'_2 = 1$; thus, $\mathbb{P}_{\boldsymbol{n}}(n'_k = i' \mid n_k = i; \tau_k)$ and, hence, $\mathbb{P}(r_1 = j_1, r_2 = j_2 \mid n_1 = n_2 = 1; \tau_1, \tau_2, \boldsymbol{\theta})$ does not involve τ_1 and τ_2. From (2.5) it also follows that $r_k = r'_k$, $k = 1, 2$. Also, $n_0 = n'_1 + n'_2 = 1 + 1 = 2$. Note that $r_0 = r'_1 + r'_2$ and $r_k = r'_k$ $(k = 1, 2)$ are counts. Thus,

$$(r_1, r_2) = (0, 0) \iff (r'_1, r'_2) = (0, 0) \iff r_0 = 0.$$

Using a symmetric argument

$$(r_1, r_2) = (1, 1) \iff (r'_1, r'_2) = (1, 1) \iff r_0 = 2.$$

Thus,

$$\mathbb{P}((r_1, r_2) = (j, j) \mid n_1 = n_2 = 1; \boldsymbol{\theta}) = \mathbb{P}(r_0 = 2j \mid n_0 = 2; \boldsymbol{\theta}), \qquad j = 0, 1. \tag{3.1}$$

It follows that

$$\begin{aligned}
&\mathbb{P}((r_1, r_2) = (0, 1) \mid n_1 = n_2 = 1; \boldsymbol{\theta}) + \mathbb{P}((r_1, r_2) = (1, 0) \mid n_1 = n_2 = 1; \boldsymbol{\theta}) \\
&\quad = 1 - \mathbb{P}((r_1, r_2) = (0, 0) \mid n_1 = n_2 = 1; \boldsymbol{\theta}) - \mathbb{P}((r_1, r_2) = (1, 1) \mid n_1 = n_2 = 1; \boldsymbol{\theta}) \\
&\quad = 1 - \mathbb{P}(r_0 = 0 \mid n_0 = 2; \boldsymbol{\theta}) - \mathbb{P}(r_0 = 2 \mid n_0 = 2; \boldsymbol{\theta}) \\
&\quad = \mathbb{P}(r_0 = 1 \mid n_0 = 2; \boldsymbol{\theta}).
\end{aligned} \tag{3.2}$$

Thus, from (3.1) and (3.2) $\mathbb{P}(r_0 = j_0 \mid n_0 = 2; \boldsymbol{\theta})$, $j_0 = 0, 1, 2$ can be expressed as functions of $\mathbb{P}((r_1, r_2) = (j_1, j_2) \mid n_1 = n_2 = 1; \boldsymbol{\theta})$, $j_1, j_2 = 0, 1$. The former is the root distribution for $n_0 = 2$, which is identifiable by the condition of Proposition 3.1. Thus, $\boldsymbol{\theta}$ can the expressed as a function of the PMF of r_0 (given $n_0 = 2$), and, thus, as a function of the joint PMF of (r_1, r_2). Hence, it can also be expressed as a function of $f_{\boldsymbol{N},\boldsymbol{R}}$. Next, we consider $n_1 = 2$, $n_2 = 1$. Then $r_1 = 0, 1,$ or 2 and $r_2 = 0$ or 1. From (2.2) it follows that $n'_2 = 1$; thus, $\mathbb{P}_{\boldsymbol{n}}(n'_2 = i'_2 \mid n_2 = i_2; \tau_2)$ and, hence, $\mathbb{P}((r_1, r_2) = (0, 1) \mid n_1 = n_2 = 1; \tau_1, \tau_2, \boldsymbol{\theta})$ does not involve τ_2. Moreover, $n_0 = n'_1 + n'_2 = n'_1 + 1$. Also, from (2.5) it follows that

$$(r_1, r_2) = (0, 1) \iff (r'_1, r'_2) = (0, 1).$$

Thus,

$$
\begin{aligned}
&\mathbb{P}((r_1, r_2) = (0, 1) \mid (n_1, n_2) = (2, 1); \tau_1, \boldsymbol{\theta}) \\
&\quad = \sum_{i'=1}^{2} \mathbb{P}((r_1', r_2') = (0, 1) \mid (n_1', n_2') = (i', 1); \boldsymbol{\theta}) \mathbb{P}(n_1' = i' \mid n_1 = 2; \tau_1) \\
&\quad = \sum_{i'=1}^{2} \mathbb{P}((r_1', r_2') = (0, 1) \mid n_0 = n_1' + 1 = i' + 1; \boldsymbol{\theta}) \mathbb{P}(n_1' = i' \mid n_1 = 2; \tau_1) \\
&\quad = \sum_{i'=1}^{2} \sum_{j_0=0}^{i'+1} \mathbb{P}((r_1', r_2') = (0, 1) \mid r_0 = j_0, n_0 = i' + 1) \\
&\qquad \times \mathbb{P}(r_0 = j_0 \mid n_0 = i' + 1; \boldsymbol{\theta}) \mathbb{P}(n_1' = i' \mid n_1 = 2; \tau_1)
\end{aligned}
$$

Note that $r_0 \neq 1 \implies (r_1', r_2') \neq (0, 1)$. Also, note that

$$\mathbb{P}(r_0 = 1 \mid n_0 = i' + 1; \boldsymbol{\theta})$$

is a function of $\boldsymbol{\theta}$ only (and no other parameters); hence, we call it $c_{i'+1}(\boldsymbol{\theta})$, $i' = 1, 2$. Thus,

$$
\begin{aligned}
&\mathbb{P}((r_1, r_2) = (0, 1) \mid (n_1, n_2) = (2, 1); \tau_1, \boldsymbol{\theta}) \\
&\quad = \sum_{i'=1}^{2} \mathbb{P}((r_1', r_2') = (0, 1) \mid r_0 = 1, n_0 = i' + 1) c_{i'+1}(\boldsymbol{\theta}) \mathbb{P}(n_1' = i' \mid n_1 = 2; \tau_1) \\
&\quad = \frac{c_2(\boldsymbol{\theta})}{2}(1 - \mathrm{e}^{-\tau_1}) + \frac{c_3(\boldsymbol{\theta})}{3}\mathrm{e}^{-\tau_1} \\
&\quad = \mathrm{e}^{-\tau_1}\left(\frac{c_3(\boldsymbol{\theta})}{3} - \frac{c_2(\boldsymbol{\theta})}{2}\right) + \frac{c_2(\boldsymbol{\theta})}{2}
\end{aligned}
$$

from (2.2) and (2.4). From the above equation it follows that

$$\tau_1 = b(\mathbb{P}((r_1, r_2) = (0, 1) \mid (n_1, n_2) = (2, 1); \tau_1, \boldsymbol{\theta}), \boldsymbol{\theta}) \tag{3.3}$$

for some function $b(\cdot, \cdot)$. We have already established that $\boldsymbol{\theta}$ can be expressed as a function of $f_{\boldsymbol{N},\boldsymbol{R}}$. Thus, τ_1 can be expressed as a function of $f_{\boldsymbol{N},\boldsymbol{R}}$ and, hence, τ_1 is identifiable. Using a symmetric argument, we can establish that τ_2 can be expressed as a function of $f_{\boldsymbol{N},\boldsymbol{R}}$ and, hence, it is identifiable. Thus, this proposition is proven.

3.2. Identifiability of type-II subtrees

Consider a type-II subtree with tips z_A and z_B. Let the MRCA node of z_A and z_B be denoted as z_{AB}. (By definition z_{AB} is not the root.) Also, consider the path from z_{AB} to the root (node 0) and call it branch AB of the subtree. (Note that, the path from z_{AB} to the root may or may not be a branch in the original tree. But, it is a branch in the subtree, and we will refer to such branches as 'subtree-branches'.) There must be at least another subtree-branch H attached to the root other than the subtree-branch AB (Figure 1(e)). Consider a tip z_D, such that the path between z_D and the root goes through H. Let τ_A be the path distance between z_{AB} and z_A and let τ_B be the path distance between z_{AB} and z_B. Also, let τ_{AB} be the path distance between the root and z_{AB} and let τ_D be the path distance between the root and z_D.

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

Proposition 3.2. *Suppose that we have at least two haploids sampled at each of z_A, z_B, and z_D and the root distribution is identifiable. Then τ_A, τ_B, τ_{AB}, τ_D, and $\boldsymbol{\theta}$ can be expressed as functions of the joint PMF of the allele types at z_A, z_B, and z_D, and, hence, they are identifiable.*

Proof. Suppose that we have samples of N_A, N_B, and N_D lineages from z_A, z_B, and z_D, respectively, and the allele-counts among these lineages are R_A, R_B, and R_D, respectively. Let the joint PMF of (R_A, R_B, R_D) be $f^*_{\boldsymbol{N},\boldsymbol{R}}$.

First we consider the type-I subtree formed by z_A and z_D. From Proposition 3.1 we establish that $\boldsymbol{\theta}$, τ_D and $\tau_A + \tau_{AB}$ can be expressed as a function of the joint PMF of (R_A, R_D) and, hence, of $f^*_{\boldsymbol{N},\boldsymbol{R}}$. A symmetric argument also establishes that $\tau_B + \tau_{AB}$ can be expressed as functions of $f^*_{\boldsymbol{N},\boldsymbol{R}}$. Next we will show that each of z_A, z_B, and z_{AB} can be expressed as a function of $f^*_{\boldsymbol{N},\boldsymbol{R}}$.

Consider a random subsample of size one from each of z_A, z_B, and z_D. Let n_A, n_B, and n_D be the numbers of subsampled haploids at z_A, z_B, and z_D, respectively. (Thus, $n_A = n_B = n_D = 1$.) Let r_A, r_B, and r_D respectively be the observed allele-counts at these subsamples. (Where $r_k = 0$ or 1 for $k = A, B, D$.) As before, let n'_k be the number of lineages ancestral to subsamples at z_k that are present at the top node of the subtree-branch attached to z_k (see Figure 1(e)), and r'_k be the allele-count out of these n'_k $(k = A, B, D)$.

From (2.2) it follows that $n_A = n_B = n_D = n'_A = n'_B = n'_D = 1$ and, thus, $\mathbb{P}(n'_k = i'_k \mid n_k = i_k; \tau_k)$ does not involve τ_k $(k = A, B, D)$. Hence,

$$\mathbb{P}((r_A, r_B, r_D) = (0, 0, 1) \mid n_A = n_B = n_D = 1; \tau_A, \tau_B, \tau_{AB}, \tau_D, \boldsymbol{\theta})$$

does not involve τ_A, τ_B, and τ_D. Also,

$$\begin{aligned}\mathbb{P}&((r_A, r_B, r_D) = (0, 0, 1) \mid n_A = n_B = n_D = 1; \tau_{AB}, \boldsymbol{\theta})\\ &= \sum_{J_A=j_A}^{I_A-(i_A-j_A)} \sum_{J_B=j_B}^{I_B-(i_B-j_B)} \sum_{J_C=j_C}^{I_C-(i_C-j_C)} \Bigg(\prod_{k\in\{A,B,D\}} \frac{\binom{J_k}{j_k}\binom{I_k-J_k}{i_k-j_k}}{\binom{I_k}{i_k}} \Bigg) f^*_{\boldsymbol{N},\boldsymbol{R}}(J_A, J_B, J_D). \qquad (3.4)\end{aligned}$$

Thus, the left-hand side of (3.4) can be expressed as a function of $f^*_{\boldsymbol{N},\boldsymbol{R}}$. It also follows from (2.5) that $r_k = r'_k$, $k = A, B, D$.

Let $n_{AB} = n'_A + n'_B$ be the total number of lineages from subsamples of z_A and z_B that are present at node AB, and let $r_{AB} = r'_A + r'_B$ be the allele-counts out of these n_{AB} lineages. Also, let n'_{AB} be the number of lineages ancestral to those n_{AB} lineages that are present at the top node (root) of the subtree-branch AB, and let r'_{AB} be the allele-count out of these n'_{AB} lineages. As before, let $n_0 = n'_{AB} + n'_D$ be the total number of lineages at the root ancestral to the subsamples at z_A, z_B, and z_D; let $r_0 = r'_{AB} + r'_D$ be the allele-count out of these n_0 lineages. Note that $n_{AB} = n'_A + n'_B = 2$, $n'_{AB} \le n_{AB}$. From (2.5) and the fact that $r_{AB} = r'_A + r'_B$ it follows that

$$(r_A, r_B) = (0, 0) \iff (r'_A, r'_B) = (0, 0) \iff r_{AB} = 0.$$

Thus,

$$\begin{aligned}\mathbb{P}&((r_A, r_B, r_D) = (0, 0, 1) \mid n_A = n_B = n_D = 1; \tau_{AB}, \boldsymbol{\theta})\\ &= \mathbb{P}((r_{AB}, r_D) = (0, 1) \mid (n_{AB}, n_D) = (2, 1); \tau_{AB}, \boldsymbol{\theta}). \qquad (3.5)\end{aligned}$$

Consider the part of the subtree consisting of the path from z_{AB} and z_D to the root; it is a type-I subtree with z_{AB} and z_D as the tips, and τ_{AB} and τ_D, respectively, as the lengths of the attached

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press

subtree-branches; it has $(n_{AB}, n_D) = (2, 0)$, respectively, as the numbers of observed lineages at z_{AB} and z_D and (r_{AB}, r_D), respectively, as the allele-counts in these lineages. From (3.3) and (3.5)

$$\begin{aligned}\tau_{AB} &= b(\mathbb{P}((r_{AB}, r_D) = (0, 1) \mid (n_{AB}, n_D) = (2, 1); \tau_{AB}, \boldsymbol{\theta}), \boldsymbol{\theta}) \\ &= b(\mathbb{P}((r_A, r_B, r_D) = (0, 0, 1) \mid n_A = n_B = n_D = 1; \tau_{AB}, \boldsymbol{\theta}), \boldsymbol{\theta}).\end{aligned}$$

As we have already established that $\tau_A + \tau_{AB}$, $\tau_B + \tau_{AB}$, τ_D, $\boldsymbol{\theta}$, and the left-hand side of (3.4) can be expressed as functions of $f^*_{N,\boldsymbol{R}}$, it follows that τ_A, τ_B, τ_{AB}, τ_D, and $\boldsymbol{\theta}$ can be expressed as functions of $f^*_{N,\boldsymbol{R}}$. Thus, they are identifiable and this proposition is proven.

Thus, the parameters of the tree are identifiable, as each two-tip subtree along with the root distribution parameter $\boldsymbol{\theta}$ is identifiable.

4. Discussion

We have proven that the model parameters are identifiable under the coalescent-based population tree model of [9] and [12]. Thus, the problem of estimating the population tree from this model is indeed meaningfully stated. Moreover, if the standard statistical theories are applicable to the parameter space, then this also implicitly proves the consistency of the maximum likelihood estimator (MLE) for this model. We have proven the identifiability of the tree parameters for any identifiable root distribution. As a result our proof is valid for different versions of this model (that vary at the root distribution) such as [9], [10], and [12].

References

[1] Allman, E. S., Ané, C. and Rhodes, J. A. (2008). Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv. Appl. Prob.* **40,** 228–249.

[2] Bryant, D. *et al.* (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29,** 1917–1932.

[3] Chai, J. and Housworth, E. A. (2011). On Rogers' proof of identifiability for the GTR + Γ + I model. *Syst. Biol.* **60,** 713–718.

[4] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376.

[5] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

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

[7] Matsen, F. A. and Steel, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* **56,** 767-775.

[8] Nielsen, R. and Slatkin, M. (2000). Likelihood analysis of ongoing gene flow and historical association. *Evolution* **54,** 44–50.

[9] Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. and Slatkin, M. (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52,** 669–677.

[10] RoyChoudhury, A. (2011). Composite likelihood-based inferences on genetic data from dependent loci. *J. Math. Biol.* **62,** 65–80.

[11] RoyChoudhury, A. and Thompson, E. A. (2012). Ascertainment correction for a population tree via a pruning algorithm for likelihood computation. *Theoret. Pop. Biol.* **82,** 59–65.

[12] RoyChoudhury, A., Felsenstein, J. and Thompson, E. A. (2008). A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* **180,** 1095–1105.

[13] Steel, M. A., Székely, L. and Hendy, M. D. (1994). Reconstructing trees when sequence sites evolve at variable rates. *J. Comp. Biol.* **1,** 153–163.

[14] Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110,** 325–344.

https://doi.org/10.1239/jap/1421763318 Published online by Cambridge University Press