Bull Math Biol (2019) 81:431–451
https://doi.org/10.1007/s11538-018-0399-1

SPECIAL ISSUE: ALGEBRAIC METHODS IN PHYLOGENETICS

# Identifiability of Phylogenetic Parameters from *k*-mer Data Under the Coalescent

**Chris Durden**[1] · **Seth Sullivant**[1]

Received: 19 May 2017 / Accepted: 19 January 2018 / Published online: 1 February 2018


**Abstract** Distances between sequences based on their $k$-mer frequency counts can be used to reconstruct phylogenies without first computing a sequence alignment. Past work has shown that effective use of $k$-mer methods depends on (1) model-based corrections to distances based on $k$-mers and (2) breaking long sequences into blocks to obtain repeated trials from the sequence-generating process. Good performance of such methods is based on having many high-quality blocks with many homologous sites, which can be problematic to guarantee a priori. Nature provides natural blocks of sequences into homologous regions—namely, the genes. However, directly using past work in this setting is problematic because of possible discordance between different gene trees and the underlying species tree. Using the multispecies coalescent model as a basis, we derive model-based moment formulas that involve the species divergence times and the coalescent parameters. From this setting, we prove identifiability results for the tree and branch length parameters under the Jukes–Cantor model of sequence mutations.

**Keywords** $k$-mer method · Coalescent · Algebraic geometry

## 1 Introduction

Phylogenetic tree reconstruction methods that compare character states at homologous sites of molecular sequences require sequence alignment methods that identify the homologous sites. Many sequence alignment methods are progressive—that is, they first compute alignments for sequences from a subset of closely related taxa and then

✉ Seth Sullivant
smsulli2@ncsu.edu

[1] Department of Mathematics, North Carolina State University, Raleigh, NC, USA

assemble these results to generate an alignment for all of the taxa. In this context, a guide tree provides the information about relatedness which is used to control the order in which alignments are assembled. Phylogenetic reconstruction methods which do not rely on aligned sequences are needed to construct such guide trees. Even methods that simultaneously produce an alignment and phylogeny benefit from having a good starting tree, which can make alignment-free tree constructing methods useful in that context as well.

Although multiple sequence alignment algorithms have been designed to reflect an insertion and deletion process that occurs along a phylogenetic tree, most current methods for constructing the guide tree itself are not based on any explicit models of evolution. We therefore aim to develop methods based on widely used evolutionary modeling assumptions that can be used in the context of multiple sequence alignment to reconstruct a phylogenetic tree. A natural approach is to record the $k$-mer frequencies of sequences and to assess relatedness among taxa by computing distances based on these frequencies.

For example, the guide trees computed in MUSCLE (Edgar 2004a, b) and Clustal Omega (Blackshields et al. 2010; Sievers et al. 2011) are based on using $k$-mer frequencies. However, those $k$-mer methods are ad hoc, in the sense that they are not derived based on any evolutionary modeling assumptions. We take the point of view that a desirable property of any phylogenetic method is that it should be *statistically consistent* under widely used phylogenetic modeling assumptions. That is, if the method is applied to data from a standard model, the method should reconstruct the correct tree with probability tending to 1 as the amount of data increases.

In past work, Allman, Rhodes and the second author (Allman et al. 2017) devised a statistically consistent $k$-mer method for phylogenetic tree reconstruction based on models for point substitutions only, and without an insertion and deletion (indel) process. This method utilizes a model-based correction to the $k$-mer distances between pairs of sequences. A key idea in Allman et al. (2017), originating in the work of Daskalakis and Roch (2013), is to break sequences into blocks to get a distribution of $k$-mer distances. From this distribution, various features of the underlying substitution process can be extracted. More specifically, to compute a $k$-mer distance between two sequences $S_1$ and $S_2$, we subdivide each sequence into $r$ subsequences $S_{11}, \ldots, S_{1r}$, and $S_{21}, \ldots, S_{2r}$, respectively. The subdivision is chosen so that, for each $i$, $S_{1i}$ and $S_{2i}$ consist of mostly orthologous sites. (Why this is possible is explained in the next paragraph.) Then, to each pair of subsequences $S_{1i}$ and $S_{2i}$, we compute $k$-mer vectors $X_{1i}$ and $X_{2i}$, compute the appropriate $k$-mer distance between $X_{1i}$ and $X_{2i}$ and average the results from $i = 1, \ldots, r$.

This procedure greatly increases the accuracy of the $k$-mer method if we assume all sequences are drawn from the same underlying phylogenetic process, because the average of independent draws from the same underlying process converges to the true underlying parameter being estimated, by the law of large numbers. Since we do not know a priori exactly when parts of the sequence are orthologous, a subdivision into a small number of parts $r$ of roughly equal lengths guarantees that most sites in the two sequences $S_{1i}$ and $S_{2i}$ are orthologous. Heuristic modification of this procedure is used in the case when the sequences are generated from a model with an indel process. Even with a moderate indel process at work, the number of blocks used cannot be too

big without running into trouble, since the number of orthologous sites within each block becomes insufficient.

Fortunately, nature provides blocks that automatically correspond one to another—namely, the genes. However, comparing sequences from many different genes requires the analysis of more complex probabilistic models, describing how the evolutionary histories of genes are related to the history of the species from which they are derived. The phylogenetic history of a set of species and the history of any given set of orthologous genes from those species are represented by a species tree and a gene tree, respectively. In particular, it is well known that gene trees need not be the same and that gene trees need not be equal to the underlying species tree (Pamilo and Nei 1988). The simplest model to handle this discrepancy is the *multispecies coalescent model*. Given a species tree with branch lengths, the multispecies coalescent gives a probability distribution on gene trees (Takahata et al. 1995). With such a gene tree, we can then use a standard model of sequence evolution to describe the probability distribution of gene sequences.

Our goal in this paper is to extend the results of Allman et al. (2017) to the more general setting from the previous paragraph where blocks correspond to gene sequences, generated by a mutation process on gene trees that come from the multispecies coalescent. While this generalization is not likely to be useful for guide tree construction, it does provide a potential alternate method for species tree construction that is alignment-free. As in Allman et al. (2017), our derivations are based on models without an insertion/deletion process. This is a natural starting point in the derivation of many phylogenetic methods. Simulations in Allman et al. (2017) showed that the formulas based on $k$-mers derived without an insertion/deletion process can be employed on data that is generated with a moderate indel process and still successfully recover the underlying phylogenetic tree.

In Sect. 2, we derive a generalization of the main formula from Allman et al. (2017), which is a calculation of the expected squared Euclidean distance between $k$-mer vectors of sequences, when the underlying gene trees are generated randomly from the coalescent process. Using this formula, in Sects. 3 and 4 we prove identifiability results on the underlying model parameters (unrooted species tree and numerical parameters) which is the first step toward developing a statistically consistent method based on $k$-mers. Section 5 contains some further identifiability results where combinations of $k$-mer vectors of different sizes are used. We conclude in Sect. 6 with a discussion of further directions and ideas about how our identifiability results might lead to new algorithms for constructing trees from $k$-mer data.

## 2 Expected $k$-mer Distances

In this section, we give a derivation of the expected Euclidean distance between $k$-mer vectors of sequences when their gene tree is generated by the coalescent model and mutations arise via any stationary Markov model. We also present the special form of this distance in the case of the Jukes–Cantor substitution model, from which we derive our later results.

We first define what we mean by the $k$-mer vector of a sequence. As mentioned above, the $k$-mers of a given sequence are the subsequences of length $k$. We form a $k$-

mer vector by recording the number of times each $k$-mer occurs in the given sequence. More precisely, let $S = s_1 s_2 \dots s_m \in [L]^m$ be a sequence of length $m$ in the alphabet $[L] = \{1, \dots, L\}$. For $k \leq m$, a $k$-mer is a subsequence $s_p s_{p+1} \dots s_{p+k-1}$ for some starting position $p \in \{1, \dots, m-k+1\}$. The $k$-mer count vector $X$, associated with the sequence $S$, is the vector of length $L^k$ whose coordinates are indexed by the words $W \in [L]^k$, and where the component $X^W$, corresponding to word $W$, records the number of times $W$ appears as a subsequence in $S$.

We next consider two sequences $S_1$ and $S_2$ descended from a common ancestor, and we define a $k$-mer distance between them as follows. First, we assume that each site in each sequence is generated independently by a Markov mutation process and that the distribution at each site is stationary. Let $Q$ be the rate matrix of the mutation process, and let $\pi$ be the associated stationary distribution. By stationarity, the probability of a $k$-mer in either sequence is $(\pi^W)_{W \in [L]^k}$, where $\pi^W = \prod_{i \in [k]} \pi^{w_i}$. The $k$-mer distance between sequences $S_1$ and $S_2$ is then $\sum_{W \in [L]^k} \frac{1}{\pi^W}(X_1^W - X_2^W)^2$. Allman, Rhodes, and the second author previously derived the expected value of this $k$-mer distance for a pair of orthologous sequences with divergence time $\tau$ under such a mutation process.

**Proposition 2.1** (Allman et al. 2017) *Let $S_1$ and $S_2$ be two sequences of length $m$ generated from an indel-free Markov model with transition matrix $M = \exp(Q\tau)$, where $Q$ is the rate matrix, and stationary initial distribution $\pi$. Let $X_1$ and $X_2$ be the resulting $k$-mer count vectors. Then*

$$\mathbf{E}\left[\sum_{W \in [L]^k} \frac{1}{\pi^W}\left(X_1^W - X_2^W\right)^2\right] = 2(m-k+1)\left(L^k - (\operatorname{tr} M)^k\right).$$

If $\lambda_1, \dots, \lambda_L$ are the (not necessarily distinct) eigenvalues of $Q$, then the trace of the transition matrix $M = \exp(Q\tau)$ can be computed from these eigenvalues:

$$\operatorname{tr} M = \sum_{l=1}^{L} e^{\lambda_l \tau}.$$

Note that, to use Proposition 2.1 in practice, we estimate the expected $k$-mer distance by dividing the sequence into blocks and computing an average over the blocks, as discussed in the introduction. However, to discuss the value of the expectation, we only work with a single sequence (or, equivalently, a single block) here and throughout. Thus, we have suppressed the index $i$ that we used in the introduction to refer to individual blocks.

We next generalize the expected $k$-mer distance formula of Proposition 2.1 by allowing the divergence times of the sequences to vary according to the multispecies coalescent. We start with a brief overview of the coalescent and multispecies coalescent models. For our purposes, it is sufficient to consider a special case of the multispecies coalescent model, consisting of only two species. We thus describe the two-species coalescent model in detail, and we use this to derive our coalescent-based $k$-mer distance formula.

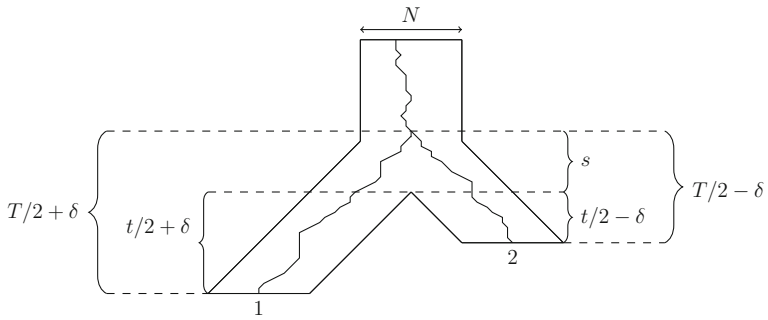


**Fig. 1** The two-species coalescent model. The quantity $\delta$ is introduced solely for the purpose of illustrating that the species tree is not ultrametric. The species divergence time $t$ represents the total evolutionary time separating species 1 and 2. The sequence divergence time $T$ represents the total evolutionary time separating a pair of orthologous sequences from species $i$ and $j$. The time between speciation and coalescence is denoted by $s$. In the coalescent model, the distribution of $s$ depends on the ancestral population size $N$

The original *n-coalescent*, as formulated by Kingman (1982), is a stochastic (Markov) process that represents the hierarchical history of family relationships of a set of $n$ objects. These objects are said to coalesce when two blocks of the partition merge to form a single larger block. In the context of this paper, the coalescent process provides a stochastic model of the genealogical history of $n$ molecular sequences. A realization of the coalescent process gives a gene tree, which represents the pattern of the coalescence events among the lineages as they extend back in time to their most recent common ancestor. The $n$-coalescent is commonly used to model gene trees of sequences, based on the assumption that they are sampled from a single, large, randomly mating population. In this model, all possible coalescence events (associated with all pairs of distinct lineages) occur at the same rate, which depends on the population size.

The *multispecies coalescent model* with $n$ species extends Kingman's $n$-coalescent model. In the multispecies coalescent process, coalescence events are constrained according to a species tree parameter, so that the process only generates gene trees that are consistent with the phylogenetic relationships given by the species tree. Specifically, genes derived from two taxa cannot coalesce more recently than the taxa themselves diverged, as delineated by the species tree. To represent this pictorially, the branches of the species tree are drawn with thick branches, and gene trees are drawn embedded within the species tree, as in Fig. 1. We will assume, for simplicity, that there is a one-to-one correspondence between the leaves of the species tree and the leaves of any embedded gene tree.

Once a population size is assigned to each ancestral taxon in the species tree, the multispecies coalescent model determines a probability distribution on the gene trees. The full probability distribution of gene trees under the $n$-species coalescent model has been described by Rannala and Yang (2003), but the 2-species case is sufficient for our purposes, since we only calculate expected values of $k$-mer distances for one pair of species at a time. The basic structure of the two-species coalescent model, showing the gene tree embedded within a species tree with two taxa, is illustrated in Fig. 1.

The expected $k$-mer distance between two orthologous sequences from two species only depends on the species divergence time for the corresponding two-leaf gene trees.

We thus use the distribution of the species divergence time, based on the two-species coalescent model, along with Proposition 2.1 to derive our generalized expected $k$-mer distance. After computing the expected $k$-mer distance for each species pair, we will assemble these pairwise distances together to obtain a collection of expected $k$-mer distances, which we would like to use to estimate phylogenetic parameters.

We now formulate precisely the model that we use to derive the expected $k$-mer distance between sequences when the gene tree varies according to the multispecies coalescent. We consider two species, and we let $t$ denote the speciation time, in true time units (e.g., proportional to years). This time represents the total evolutionary time separating the species. Thus, it is given by the sum of the branch lengths along the species tree leading from the point where the species diverge to each of the two leaves (see Fig. 1). The times along each of these branches are not assumed to be equal. In other words, we do not assume that our species tree is ultrametric. We consider pairs of orthologous sequences, where each pair consists of a sequence from one species along with an ortholog from the other species. The divergence time of any pair of orthologs will exceed the speciation time $t$ by some amount $2s$, representing twice the time between speciation and coalescence (Takahata 1986). Thus, the sequence divergence time is given by $\tau = t + 2s$. The coalescence time $s$ depends on the size of the population ancestral to the two species, which we denote by $N$. Note that in many implementations of the multispecies coalescent, variability of the population size across multiple populations is allowed. In this paper, we assume that the population size $N$ is fixed across all branches of the tree. A graphical representation of the quantities in this model is shown in Fig. 1.

To generalize Proposition 2.1, we allow the gene trees of orthologous sequences to vary according to the multispecies coalescent, so that the sequence divergence times $\tau$ are realizations of a random variable $T$. Under this model, the sequence divergence time is given by $T = t + 2s$, where the coalescence time $s$ is exponentially distributed. Since the size of the ancestral population is $N$, the rate of coalescence is $1/N$, and

$$s \sim \text{Exp}(1/N).$$

We compute the expected $k$-mer distance with respect to the distribution of $T$, by interpreting Proposition 2.1 as the expected $k$-mer distance given that $T = \tau$, and using the law of total expectation:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{W\in[L]^k} \frac{1}{\pi^W}\left(X_1^W - X_2^W\right)^2\right] \\
&\quad = \mathbf{E}\left[\mathbf{E}\left[\sum_{W\in[L]^k} \frac{1}{\pi^W}\left(X_1^W - X_2^W\right)^2 \,\middle|\, T = t + 2s\right]\right] \\
&\quad = \mathbf{E}\left[2(m-k+1)\left(L^k - \left(\sum_{l=1}^{L} \exp(\lambda_l(t+2s))\right)^k\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= 2(m-k+1)\left(L^k - \sum_{q_1+\cdots+q_L=k} \binom{k}{q_1,\ldots,q_L} \mathbf{E}\left[\exp\left((\lambda_1 q_1 + \cdots \right.\right.\right. \\
&\quad \left.\left.\left. + \lambda_L q_L)(t+2s)\right)\right]\right) \\
&= 2(m-k+1)\left(L^k - \sum_{q_1+\cdots+q_L=k} \binom{k}{q_1,\ldots,q_L} \exp\left((\lambda_1 q_1 + \cdots + \lambda_L q_L)t\right)\right. \\
&\quad \left. \cdot \mathbf{E}\left[\exp(2(\lambda_1 q_1 + \cdots + \lambda_L q_L)s)\right]\right) \\
&= 2(m-k+1)\left(L^k - \sum_{q_1+\cdots+q_L=k} \binom{k}{q_1,\ldots,q_L} \exp((\lambda_1 q_1 + \cdots + \lambda_L q_L)t)\right. \\
&\quad \left. \cdot \int_0^\infty \exp(2(\lambda_1 q_1 + \cdots + \lambda_L q_L)s) \frac{1}{N} \exp(-s/N) ds\right) \\
&= 2(m-k+1)\left(L^k - \sum_{q_1+\cdots+q_L=k} \binom{k}{q_1,\ldots,q_L} \frac{\exp((\lambda_1 q_1 + \cdots + \lambda_L q_L)t)}{1 - 2N(\lambda_1 q_1 + \cdots + \lambda_L q_L)}\right).
\end{aligned}
$$

The preceding calculation proves the following proposition.

**Proposition 2.2** *Let $S_1$ and $S_2$ be sequences of length $m$ with divergence time distributed according to the multispecies coalescent process with population size $N$ and species divergence time $t$. Suppose $S_1$ and $S_2$ are generated from an indel-free Markov model with transition rate matrix $Q$ and stationary distribution $\pi$, and let $\lambda_1, \ldots, \lambda_L$ be the eigenvalues of $Q$. Let $X_1$ and $X_2$ be the resulting $k$-mer count vectors. Then,*

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{W\in[L]^k} \frac{1}{\pi^W}(X_1^W - X_2^W)^2\right] \\
&\quad = 2(m-k+1)\left(L^k - \sum_{q_1+\cdots+q_L=k} \binom{k}{q_1,\ldots,q_L} \frac{\exp((\lambda_1 q_1 + \cdots + \lambda_L q_L)t)}{1 - 2N(\lambda_1 q_1 + \cdots + \lambda_L q_L)}\right).
\end{aligned}
$$

In practice, in order to obtain precise estimates of the expected $k$-mer distance of Proposition 2.2, we would like to compute an average $k$-mer distance over many independently-generated sequences. We have in mind the context in which a sequence (e.g., a genome) is divided into blocks corresponding to genes, and the expected $k$-mer distance is estimated by taking the average $k$-mer distance over all blocks. Under our model, the sequence data from distinct loci are independent if the gene trees for those loci are independent, given the species tree. We expect this to be the case for unlinked loci, but it may also hold for some linked loci if the population sizes, population structure, and recombination rates are sufficient to decouple inheritance (Wakeley 2009; McVean 2002).

In order to use Proposition 2.2 to derive statistically consistent model-based corrections to the $k$-mer distances, the map from parameter space to the collection of all $k$-mer distances must be one-to-one. For the remainder of this paper, we study this problem for the special case where $L = 4$ and $Q$ is the Jukes–Cantor rate matrix

$$Q = \begin{pmatrix} -\mu & \mu/3 & \mu/3 & \mu/3 \\ \mu/3 & -\mu & \mu/3 & \mu/3 \\ \mu/3 & \mu/3 & -\mu & \mu/3 \\ \mu/3 & \mu/3 & \mu/3 & -\mu \end{pmatrix}.$$

The eigenvalues of $Q$ are $\lambda_1 = 0$ and $\lambda_2 = -4\mu/3$, with multiplicities 1 and 3, respectively. The stationary distribution of the corresponding Markov process is $\pi^W = \frac{1}{4^k}$ for all $W \in [4]^k$. So, Proposition 2.2 reduces to the following:

**Corollary 2.3** *Let $S_1$ and $S_2$ be sequences of length $m$ with divergence time distributed according to the multispecies coalescent process with population size $N$ and species divergence time $t$. Suppose $S_1$ and $S_2$ are generated from an indel-free Jukes–Cantor mutation model. Let $X_1$ and $X_2$ be the resulting $k$-mer count vectors. Then*

$$\mathbf{E}[\|X_1 - X_2\|_2^2] \;=\; 2(m-k+1)\left(1 - \frac{1}{4^k}\sum_{h=0}^{k}\binom{k}{h}3^h \frac{\exp(-4h\mu t/3)}{1+8h\mu N/3}\right).$$

To analyze the dependence of the expected $k$-mer distance on the parameters $t$, $N$, and $\mu$, we define a function which maps these parameters to the expected $k$-mer distance:

$$f_k(t, N, \mu) = 2(m-k+1)\left(1 - \frac{1}{4^k}\sum_{h=0}^{k}\binom{k}{h}3^h \frac{\exp(-4h\mu t/3)}{1+8h\mu N/3}\right).$$

We note that

$$f_k(\alpha t, \alpha N, \mu/\alpha) = f_k(t, N, \mu) \tag{1}$$

for any $\alpha > 0$. This can be seen as a specific consequence of the fact that both the mutation process and the coalescence process are invariant under similar transformations of the parameters $t$, $\mu$ and $N$: These processes are unaffected by changing evolutionary times from $t$ to $\alpha t$, while simultaneously changing the coalescence rate from $1/N$ to $1/\theta = 1/N\alpha$ and the mutation from $\mu$ to $\mu/\alpha$. Thus, we have a fundamental inability to determine the values of the parameters $\mu$, $N$, and $t$ from data generated by these processes.

We use a standard approach of adjusting time units to suppress the invariance described by Eq. 1: We assume that time is measured in substitution units—that is, time units in which $\mu = 1$. Then, $t$ represents time in units of the expected number of substitutions. In these time units, the rate of coalescence is $1/N\mu$. We let $\theta$ denote the quantity $N\mu$ in the denominator. With respect to the new time units, we obtain the expected $k$-mer distance as a function of $t$ and $\theta$:

$$f_k(t,\theta) = 2(m-k+1)\left(1 - \frac{1}{4^k}\sum_{h=0}^{k}\binom{k}{h}3^h\frac{\exp(-4ht/3)}{1+8h\theta/3}\right).$$

To facilitate an algebraic analysis of model identifiability, we reparameterize $f_k$ by introducing a quantity $x$, which we call the transformed species divergence time, defined by the following equation:

$$x = \exp(-4t/3),$$

where the species divergence time $t$ is given in substitution units, as mentioned previously. With this change of variables, the expected $k$-mer distance of Corollary 2.3 can be written as a rational function of $x$ and $\theta$:

$$f_k(x,\theta) = 2(m-k+1)\left(1 - \frac{1}{4^k}\sum_{h=0}^{k}\binom{k}{h}3^h\frac{x^h}{1+8h\theta/3}\right). \tag{2}$$

The parameterization of the expected $k$-mer distance in Eq. 2 depends continuously on two parameters, so these parameters are not identifiable from a single expected $k$-mer distance between one pair of species. We thus consider a situation in which we have $n$ species ($n > 2$). For any two distinct species, labeled by $i, j \in [n]$, we consider $k$-mer vectors $X_i$ and $X_j$ generated by the coalescent and mutation processes. We let $t_{ij}$ denote the species divergence time, and we let $x_{ij}$ be the corresponding transformed species divergence time. We suppose that the $\binom{n}{2}$ $k$-mer distances between pairs of species are parameterized by a common set of parameters which, taken together, describe the evolutionary history of these species. The simplest assumption which provides a model that is potentially identifiable—having fewer than $\binom{n}{2}$ independent parameters—is that the species divergence times $t_{ij}$ are distances among the leaves of a species tree, and all of the ancestral populations have the same size $N$ (and therefore the same value of $\theta$). Under these assumptions, we can parameterize all $\binom{n}{2}$ pairs of expected $k$-mer distances by $2n-2$ parameters: $2n-3$ parameters giving the branch lengths of the species tree, and one $\theta$ parameter.

We now describe this parameterization. Suppose $T$ is a species tree with leaves labeled by the taxa $[n]$, and let $\{w_e \in \mathbb{R}_{\geq 0} : e \in E(T)\}$ be an edge weighting of $T$, giving evolutionary times along edges of $T$. If the species divergence times $(t_{ij})_{1\leq i<j\leq n}$ are the distances between leaves $i$ and $j$ of $T$, then they are obtained by summing edge weights along paths in $T$:

$$t_{ij} = \sum_{e\in P(i,j)} w_e, \tag{3}$$

where $P(i, j)$ is the set of edges of the unique path in $T$ connecting leaves $i$ and $j$. The transformed species divergence times $(x_{ij})_{1\leq i<j\leq n}$ can then be expressed analogously: For each edge $e \in E(T)$, we define a transformed edge weight $a_e = \exp(-4w_e/3)$. Then, the transformed species divergence times are obtained by multiplying branch

lengths along corresponding paths in $T$:

$$x_{ij} = \prod_{e \in P(i,j)} a_e.$$

Since $w_e \geq 0$, we have $a_e \in (0, 1]$. In the analyses which follow, we will parameterize the expected $k$-mer distance between species $i$ and $j$ using these transformed edge weights: $\mathbf{E}[\|X_i - X_j\|_2^2] = f_k(\prod_{e \in P(i,j)} a_e, \theta)$.

We note that Proposition 2.2 gives a $k$-mer distance formula that depends on the total evolutionary distance between the species, and not on the individual branch lengths leading to their common ancestor. Thus, the $k$-mer distances of Proposition 2.2 do not depend on the choice of a root for the tree $T$. The following observations clarify why this is true. First, the distribution of the distances between the leaves of the gene tree does not depend on where the species tree is rooted. This can be seen by observing that changing the root is equivalent to varying the quantity $\delta$ in Fig. 1, which does not affect any of the divergence times. Second, the expected $k$-mer distances, for a fixed gene tree, as given by Proposition 2.1 also do not depend on the choice of a root for the gene tree. In fact, by the stationarity of the Markov mutation process, the joint distribution of the sequences also does not depend on the root. Together, these observations imply that the rooted tree is not identifiable from $k$-mer distances under our model. Thus, we will consider only the unrooted tree when we formulate our identifiability results.

For a given $n$-leaf binary tree $T$ and for each $k \in \mathbb{N}$, we construct a rational map whose inputs are the parameter $\theta$ and the transformed branch lengths $a_e = \exp(-4\mu w_e/3)$, and which maps to the $\binom{n}{2}$ expected $k$-mer distances between pairs of species. This map, denoted by $\phi_{k,T}$, has the following form:

$$\begin{aligned} \phi_{k,T} &: \mathbb{R}^{2n-2} \to \mathbb{R}^{n(n-1)/2} \\ (a_e : e \in E(T), \theta) &\mapsto \left( f_k \left( \prod_{e \in P(i,j)} a_e, \theta \right) \right)_{1 \leq i < j \leq n}. \end{aligned}$$

The parameters $a_e = \exp(-4w_e/3)$ are restricted to lie in $(0, 1]$, and the parameter $\theta = \mu N$ is positive. The parameter space of $\phi_{k,T}$ is thus $\Theta_T = (0, 1]^{2n-3} \times \mathbb{R}_{>0}$. We note that the smallest $n$ for which the dimension of the image is at least the number of parameters is $n = 4$. For $n = 4$, $\phi_{k,T} : (0, 1]^5 \times \mathbb{R}_{>0} \to \mathbb{R}^6$, so that it is in principle possible for the parameters to be identifiable. We will see in the next section that the parameters are in fact locally identifiable in this case if $k > 1$.

*Example 2.4* Figure 2 shows the set of edge length parameters $(a_e)_{e \in P(i,j)}$ arising from a particular labeling of the 5 edges of a 4-taxon tree. Suppose that $k = 2$. Then, $\phi_{k,T}$ takes the following form:

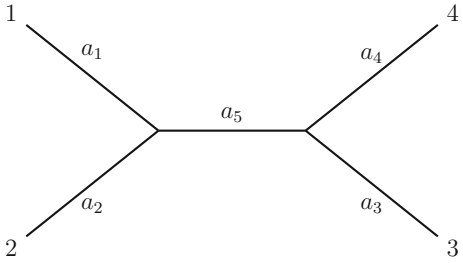


**Fig. 2** Example labeling of a 4-taxon tree with transformed edge weight parameters $(a_e)_{e\in P(i,j)}$. The parameters are defined by $a_e = \exp(-4w_e/3)$ where $w_e$ represents the evolutionary time along edge $e$ in substitution units. In the parameterization of our model, the transformed species divergence time for taxa 1 and 3, for example, is $x_{13} = a_1a_5a_3$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ \theta \end{pmatrix} \mapsto \begin{pmatrix} f_{2,T}(a_1a_2,\theta) \\ f_{2,T}(a_1a_5a_3,\theta) \\ f_{2,T}(a_1a_5a_4,\theta) \\ f_{2,T}(a_2a_5a_3,\theta) \\ f_{2,T}(a_2a_5a_4,\theta) \\ f_{2,T}(a_3a_4,\theta) \end{pmatrix} = \begin{pmatrix} 2(m-1)(1-\frac{1}{16}(1+\frac{18a_1a_2}{3+8\theta}+\frac{27a_1^2a_2^2}{3+16\theta}) \\ 2(m-1)(1-\frac{1}{16}(1+\frac{18a_1a_5a_3}{3+8\theta}+\frac{27a_1^2a_5^2a_3^2}{3+16\theta}) \\ 2(m-1)(1-\frac{1}{16}(1+\frac{18a_1a_5a_4}{3+8\theta}+\frac{27a_1^2a_5^2a_4^2}{3+16\theta}) \\ 2(m-1)(1-\frac{1}{16}(1+\frac{18a_2a_5a_3}{3+8\theta}+\frac{27a_2^2a_5^2a_3^2}{3+16\theta}) \\ 2(m-1)(1-\frac{1}{16}(1+\frac{18a_2a_5a_4}{3+8\theta}+\frac{27a_2^2a_5^2a_4^2}{3+16\theta}) \\ 2(m-1)(1-\frac{1}{16}(1+\frac{18a_3a_4}{3+8\theta}+\frac{27a_3^2a_4^2}{3+16\theta}) \end{pmatrix}$$

## 3 Local Identifiability from $k$-mer Distances

In this section, we prove our first main identifiability result. We show that the map $\phi_{k,T}$ is generically finite-to-one if $n \geq 4$ and $k > 1$. In other words, both $\theta$ and the branch length parameters of the phylogenetic tree are locally identifiable from $k$-mer distances provided $k > 1$ and there are at least 4 species. This is proven in Theorem 3.3 and Corollary 3.4. We will use this result in the next section to show that the (unrooted) tree parameter is identifiable.

**Definition 3.1** A map $\phi : S \to \mathbb{R}^d$ defined on an open set $S \subset \mathbb{R}^n$ is called *generically finite-to-one* if there exists a proper algebraic subset $\tilde{S} \subset \mathbb{R}^n$ such that the fiber

$$\mathcal{F}_\phi(s') = \{s \in S \mid \phi(s) = \phi(s')\}$$

is finite for all $s' \in S\backslash\tilde{S}$.

The following folklore result is the standard tool to prove local identifiability. A recent proof can be found in Leung et al. (2016):

**Lemma 3.2** *If $\phi : S \to \mathbb{R}^d$ is a polynomial or rational map defined on an open set $S \subset \mathbb{R}^n$ then $\phi$ is generically finite-to-one on $S$ if and only if the Jacobian matrix of $\phi$ generically has full column rank.*

**Theorem 3.3** *For $k > 1$ and $n = 4$, the map $\phi_{k,T} : \mathbb{R}^6 \to \mathbb{R}^6$ is generically finite-to-one. In particular, it is generically finite-to-one on $\Theta_T = (0, 1]^5 \times \mathbb{R}_{>0}$.*

*Proof* We show that the Jacobian matrix of $\phi_{k,T}$ has full column rank. We first write $\phi_{k,T}$ as a composite map and use the chain rule to write the Jacobian matrix as a product of matrices: $\phi_{k,T}$ can be expressed as the composition $\phi_{k,T} = F_{k,T} \circ \xi_T$ where

$$F_{k,T} = (f_k(x_{12}, \theta), f_k(x_{13}, \theta), f_k(x_{14}, \theta), f_k(x_{23}, \theta), f_k(x_{24}, \theta), f_k(x_{34}, \theta)),$$
$$\xi_T(a, \theta) = (a_1a_2, a_1a_3a_5, a_1a_4a_5, a_2a_3a_5, a_2a_4a_5, a_3a_4, \theta).$$

(here we have identified the input vector $(x_{12}, x_{13}, x_{14}, x_{23}, x_{24}, x_{34}, \theta)$ of $F_{k,T}$ with the output vector of $\xi_T$).

The Jacobian matrix of the outer function $F_{k,T}$ is

$$d_{(\mathbf{x},\theta)}F_{k,T} = \begin{pmatrix} D_{12} & 0 & 0 & 0 & 0 & 0 & B_{12} \\ 0 & D_{13} & 0 & 0 & 0 & 0 & B_{13} \\ 0 & 0 & D_{14} & 0 & 0 & 0 & B_{14} \\ 0 & 0 & 0 & D_{23} & 0 & 0 & B_{23} \\ 0 & 0 & 0 & 0 & D_{24} & 0 & B_{24} \\ 0 & 0 & 0 & 0 & 0 & D_{34} & B_{34} \end{pmatrix} \tag{4}$$

where

$$D_{ij}(\mathbf{x}, \theta) = \frac{\partial f_k}{\partial x_{ij}}(x_{ij}, \theta) = -2(m - k + 1)\frac{1}{4^k}\sum_{h=1}^{k}\binom{k}{h}\frac{3^h h x_{ij}^{h-1}}{1 + 8\theta h/3}, \text{ and}$$
$$B_{ij}(\mathbf{x}, \theta) = \frac{\partial f_k}{\partial \theta}(x_{ij}, \theta) = 2(m - k + 1)\frac{1}{4^k}\sum_{h=1}^{k}\binom{k}{h}\frac{3^h x_{ij}^h 8h/3}{(1 + 8\theta h/3)^2}.$$

The Jacobian matrix of $\xi_T$ is:

$$d_{(a,\theta)}\xi_T = \begin{pmatrix} a_2 & a_1 & 0 & 0 & 0 & 0 \\ a_3a_5 & 0 & a_1a_5 & 0 & a_1a_3 & 0 \\ a_4a_5 & 0 & 0 & a_1a_5 & a_1a_4 & 0 \\ 0 & a_3a_5 & a_2a_5 & 0 & a_2a_3 & 0 \\ 0 & a_4a_5 & 0 & a_2a_5 & a_2a_4 & 0 \\ 0 & 0 & a_4 & a_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

If we use block notation to express matrices (4) and (5) as $[D\ B]$ and $\begin{bmatrix} E & 0 \\ 0 & 1 \end{bmatrix}$ (where $D$ is $6\times6$, $B$ is $6\times1$, and $E$ is $6\times5$), then the Jacobian determinant of $F_{k,T} \circ \xi_T : \mathbb{R}^6 \to \mathbb{R}^6$ is

$$\begin{aligned}\det(d_{(a,\theta)}(F_{k,T}\circ\xi_T)) &= \det(d_{\xi_T(a,\theta)}F_{k,T}\cdot d_{(a,\theta)}\xi_T)\\ &= \det\left([D\ B]\cdot\begin{bmatrix}E & 0\\ 0 & 1\end{bmatrix}\right)\\ &= \det([DE\ B])\\ &= 4D_{12}\Big(a_1a_2^2a_3a_4^2a_5^2B_{13}D_{14}D_{23}D_{24}\\ &\quad - a_1a_2^2a_3^2a_4a_5^2D_{13}B_{14}D_{23}D_{24}\\ &\quad - a_1^2a_2a_3a_4^2a_5^2D_{13}D_{14}B_{23}D_{24}\\ &\quad + a_1^2a_2a_3^2a_4a_5^2D_{13}D_{14}D_{23}B_{24}\Big)D_{34}.\end{aligned}$$

To prove that $\det(d_{(a,\theta)}(F_{k,T}\circ\xi_T))$ is not the zero function, it suffices to collect terms and show that there is a nonzero term in this expression. Here, we think of $\det(d_{(a,\theta)}(F_{k,T}\circ\xi_T))$ as an element of the ring $\mathbb{R}(\theta)[a_1,a_2,a_3,a_4,a_5]$. Ignoring the common factor of $4D_{12}D_{34}a_1a_2a_3a_4a_5^2$, we need to show that

$$\begin{aligned}&a_2a_4B_{13}D_{14}D_{23}D_{24} - a_2a_3D_{13}B_{14}D_{23}D_{24} - a_1a_4D_{13}D_{14}B_{23}D_{24}\\ &\quad + a_1a_3D_{13}D_{14}D_{23}B_{24}\end{aligned}$$

is not the zero polynomial. Expanding this as an element of $\mathbb{R}(\theta)[a_1,a_2,a_3,a_4,a_5]$, and ignoring the common factor of $\left(2(m-k+1)\frac{1}{4^k}\right)^4$, we see that the lowest order term in the lexicographic order with $a_1>a_2>a_3>a_4>a_5$ when $k\geq 2$ is

$$-a_1a_2^2a_3a_4^2a_5^2\frac{648k^4(k-1)}{(1+8\theta/3)^3(1+16\theta/3)^2}.$$

The coefficient of this monomial is a nonzero function of $\theta$, so for $k>1$ the Jacobian determinant of $F_{k,T}\circ\xi_T$ is nonzero generically, and thus $d_{(a,\theta)}F_{k,T}\circ\xi_T$ has full column rank. By Lemma 3.2, the map $\phi_{k,T}=F_{k,T}\circ\xi_T$ is generically finite-to-one on $\mathbb{R}^6$. Therefore, $\phi_{k,T}$ is generically finite-to-one on $\Theta_T=(0,1]^5\times\mathbb{R}_{>0}$. □

The preceding theorem shows that $\phi_{k,T}$ is generically finite-to-one for $n=4$ leaf trees. This easily extends to arbitrary trees with $n\geq 4$ leaves.

**Corollary 3.4** *For $k>1$ and $n\geq 4$, the map $\phi_{k,T}:\mathbb{R}^{2n-2}\to\mathbb{R}^{n(n-1)/2}$ is generically finite-to-one. In particular, it is generically finite-to-one on $\Theta_T=(0,1]^{2n-3}\times\mathbb{R}_{>0}$.*

*Proof* We argue by induction. Suppose that $\phi_{k,S}$ is generically finite-to-one for every $n-1$-leaf subtree $S$ of $T$. Let $i$ be a leaf from a cherry of $T$, and let $j$ be a leaf which does not form a cherry with $i$. Then, every metric of $T$ in the fiber $\mathcal{F}_{\phi_{k,T}}(a,\theta)=\phi_{k,T}^{-1}(\phi_{k,T}(a,\theta))$ is determined by its induced metrics on the subtrees $T\backslash i$ and $T\backslash j$. (Here, we use $T\backslash i$ to denote the binary phylogenetic tree obtained by deleting leaf vertex $i$, and then suppressing the remaining vertex of the edge incident to $i$.) Since

$\phi_{k,T\setminus i}$ and $\phi_{k,T\setminus j}$ are finite-to-one generically, $\phi_{k,T}$ is finite-to-one generically. The result follows by induction, starting from the 4-leaf subtrees $Q$ of $T$, for which $\phi_{k,Q}$ is finite-to-one generically by Theorem 3.3. □

Note that for a 4 leaf tree with $k = 1$ the Jacobian determinant is just zero, and so the model is not locally identifiable in this case. In fact, for monomers, and any tree, the model $\phi_{1,T}$ is never identifiable. This can be seen from the special form of the function $f_1$ which is

$$f_1(x, \theta) = 2m\left(1 - \frac{1}{4}\left(1 + \frac{9x}{3 + 8\theta}\right)\right).$$

From this, we see that for a given vector $(a_e : e \in E(T), \theta)$ and a given $\theta_0$, if we can replace all leaf edge parameters with $b_e = \frac{\sqrt{3+8\theta_0}}{\sqrt{3+8\theta}} a_e$, and make $b_e = a_e$ for internal edges, then

$$\phi_{1,T}(a, \theta) = \phi_{1,T}(b, \theta_0).$$

In particular, this shows that the dimension of the image of $\phi_{1,T}$ is strictly less than $2n - 2$, so the parameters could not be locally identifiable.

The following example shows that $\phi_{k,T}$ it is not one-to-one generically on $\Theta_T$ for $k = 2$ and $n = 4$.

*Example 3.5* This example shows that for $k = 2, n = 4$ the numerical parameters are not generically identifiability. Let $T = 12|34$. Then $\phi_{2,T}(a_1, \theta_1) = \phi_{2,T}(a_2, \theta_2)$ for the following values of $(a_1, \theta_1)$ and $(a_2, \theta_2)$:

$$\begin{aligned}(a_1, \theta_1) &= (a_{1,1}, a_{1,2}, a_{1,3}, a_{1,4}, a_{1,5}, \theta_1)\\ &\approx (0.7199, 0.6687, 0.9623, 0.9950, 0.9907, 0.9146)\\ (a_2, \theta_2) &= (a_{2,1}, a_{2,2}, a_{2,3}, a_{2,4}, a_{2,5}, \theta_2)\\ &\approx (0.5624, 0.5202, 0.7642, 0.7916, 0.9902, 0.3400)\end{aligned}$$

By direct computation, the Jacobian has full rank at $(a_1, \theta_1)$ and $(a_2, \theta_2)$. By the Implicit Function Theorem, in a union of small open neighborhoods of $(a_1, \theta_1)$ and $(a_2, \theta_2)$, $\phi_{2,T}$ will be 2-to-1. This shows that the numerical parameters are not generically identifiable in this case.

## 4 Identifiability of the Tree Topology

In this section, we provide a proof of the generic identifiability of the unrooted tree topology from Jukes–Cantor $k$-mers distances for all binary trees with $n \geq 5$ leaves. The proof is based on applying some ideas from algebraic statistics together with tools from combinatorial phylogenetics.

To discuss identifiability of the tree parameter, we need to work explicitly with the image of $\phi_{k,T}$, that is, the set of pairwise $k$-mer distances that are compatible with our model assumption and the underlying tree topology $T$. We denote this set $\mathcal{M}_{k,T}$.

**Definition 4.1** The tree parameter of the multispecies coalescent model is *generically identifiable* from $k$-mer distances on trees with $n$ leaves if for all (unrooted) binary trees $T, T'$ on $n$ leaves with $T \neq T'$, we have

$$\dim(\mathcal{M}_{k,T} \cap \mathcal{M}_{k,T'}) < \min(\dim(\mathcal{M}_{k,T}), \dim(\mathcal{M}_{k,T'})).$$

An interpretation of the definition is that if we sample a point $p \in \cup_T \mathcal{M}_{k,T}$, then with probability one, there is a unique binary tree $T$ such that $p \in \mathcal{M}_{k,T}$. So, generic identifiability of the tree topology means that with probability one, if we have a point $p$ that came from some tree and some choice of continuous parameters, we can tell which tree it came from. One approach to prove the dimension inequality is by using the vanishing ideals of the sets $\mathcal{M}_{k,T}$.

**Definition 4.2** Let $S \subseteq \mathbb{R}^d$ and let $\mathbb{R}[p_1, \ldots, p_d]$ be the polynomial ring in indeterminates $p_1, \ldots, p_d$ with real coefficients. The *vanishing ideal* of $S$ is the set

$$I(S) = \{f \in \mathbb{R}[p_1, \ldots, p_d] : f(a) = 0 \text{ for all } a \in S\}.$$

For example, if $S = \{(t, t^2) \in \mathbb{R}^2 : t \in \mathbb{R}\}$, then $I(S) = \langle p_1^2 - p_2 \rangle \subseteq \mathbb{R}[p_1, p_2]$. The polynomial $p_1^2 - p_2$ is called a generator of the ideal $I(S)$. We refer the reader to Cox et al. (2015) for more background on the necessary algebra, and (Allman et al. 2011) for an example of how this approach is used in studying the identifiability under the coalescent model in other contexts.

In our case, we will look at the vanishing ideals of the sets $\mathcal{M}_{k,T} \subseteq \mathbb{R}^{n(n-1)/2}$ and so the appropriate polynomial ring is $\mathbb{R}[p_{ij} : 1 \leq i < j \leq n]$. The following proposition is the key general result about vanishing ideals that is often used to prove identifiability.

**Proposition 4.3** *Suppose that $S_1$ and $S_2 \subseteq \mathbb{R}^d$ are two parameterized sets such that there are nonzero polynomials $f_1 \in I(S_1) \setminus I(S_2)$ and $f_2 \in I(S_2) \setminus I(S_1)$. Then*

$$\dim(S_1 \cap S_2) < \min(\dim S_1, \dim S_2).$$

When the trees have 4 leaves, the model has 6 parameters and there are $\binom{4}{2}$ pairwise distances. Since the numerical parameters are generically locally identifiable in this case when $k \geq 2$, the vanishing ideals will all be the zero ideals. This suggests that the tree parameters might not be generically identifiable when $n = 4$ and $k \geq 2$, though we have not been able to find specific parameter values realizing this. On the other hand, when $k = 1$ and $n = 4$, the numerical parameters are not identifiable and the tree parameters *are* identifiable.

**Proposition 4.4** *Let $k = 1$ and $n = 4$, and $T$ and 4-leaf binary tree. Then, the vanishing ideal of $I(\mathcal{M}_{1,T})$ is a principal ideal, and the generator can be used to distinguish between the different trees.*

*Proof* Fix the specific tree $T = 12|34$. The parameterization in the $k = 1$ case has the following form:

$$\begin{aligned}
p_{12} &= 2m\left(1 - \frac{1}{4}\left(1 + a_1a_2\frac{9}{3+8\theta}\right)\right)\\
p_{13} &= 2m\left(1 - \frac{1}{4}\left(1 + a_1a_3a_5\frac{9}{3+8\theta}\right)\right)\\
p_{14} &= 2m\left(1 - \frac{1}{4}\left(1 + a_1a_4a_5\frac{9}{3+8\theta}\right)\right)\\
p_{23} &= 2m\left(1 - \frac{1}{4}\left(1 + a_2a_3a_5\frac{9}{3+8\theta}\right)\right)\\
p_{24} &= 2m\left(1 - \frac{1}{4}\left(1 + a_2a_4a_5\frac{9}{3+8\theta}\right)\right)\\
p_{34} &= 2m\left(1 - \frac{1}{4}\left(1 + a_3a_4\frac{9}{3+8\theta}\right)\right).
\end{aligned}$$

These expressions satisfy the single relation:

$$\left(p_{13} - \frac{3m}{2}\right)\left(p_{24} - \frac{3m}{2}\right) - \left(p_{14} - \frac{3m}{2}\right)\left(p_{23} - \frac{3m}{2}\right).$$

As the indeterminates that appear in this equation are different from the ones that occur in the analogous equation for one of the other trees on 4 leaves, this shows that we can apply Proposition 4.3 to deduce that the tree parameters are generically identifiable in this case. □

Now, we show the analogous result for $n = 5$ leaf trees and for $k \geq 2$.

**Proposition 4.5** *Let $k \geq 2$ and consider the five leaf tree $T$ with nontrivial splits $12|345$ and $123|45$. Then, none of the generators of the vanishing ideal $I(\mathcal{M}_{k,T})$ involve the indeterminates $p_{12}$ and $p_{45}$. Furthermore, there are generators of $I(\mathcal{M}_{k,T})$ that involve the other indeterminates of $\mathbb{R}[p]$ in a nontrivial way.*

*Proof* First of all, let us make sense of the statement that "none of the generators of the vanishing ideal $I(\mathcal{M}_{k,T})$ involve the indeterminates $p_{12}$ and $p_{45}$". This is equivalent to saying that if $(p_{12}, p_{13}, \ldots, p_{35}, p_{45})$ is a generic point in $\mathcal{M}_{k,T}$ then so is $(p_{12} + \epsilon_1, p_{13}, \ldots, p_{35}, p_{45} + \epsilon_2)$ for small $\epsilon_1$ and $\epsilon_2$.

First, we consider a simplified version of the parameterization where

$$\begin{aligned}
x_{12} &= a_1a_2, x_{13} = a_1a_3a_6, x_{14} = a_1a_4a_6a_7, x_{15} = a_1a_5a_6a_7,\\
x_{23} &= a_2a_3a_6, x_{24} = a_2a_4a_6a_7, x_{25} = a_2a_5a_6a_7, x_{34} = a_3a_4a_7, x_{35}\\
&= a_3a_5a_7, x_{45} = a_4a_5.
\end{aligned}$$

This is the parameterization of a certain toric ideal associated with initial ideals of the Grassmannian (Speyer and Sturmfels 2004), and it is known that the vanishing ideal is generated by the following polynomials:

$$x_{13}x_{24} - x_{14}x_{23}, x_{13}x_{25} - x_{15}x_{23}, x_{14}x_{25} - x_{15}x_{24}, x_{15}x_{34} - x_{14}x_{25}, x_{25}x_{34} - x_{24}x_{35}.$$

It is directly seen that none of these polynomials involve either of the variables $x_{12}$ or $x_{45}$; in particular, these coordinates can be moved freely while staying in the variety defined by these equations. Let $I$ be the ideal defined by these equations and $\mathcal{M}_T$ the resulting image of the parameterization for all $a_i$ parameters in $(0, 1]$.

Now, our model $\mathcal{M}_{k,T}$ is obtained from $\mathcal{M}_T$ by applying the function

$$f_k(y, \theta) = 2(m - k + 1)\left(1 - \frac{1}{4^k}\sum_{h=0}^{k}\binom{k}{h}y^h\frac{3^{h+1}}{3 + 8\theta h}\right)$$

simultaneously to each coordinate while letting $\theta$ range over $(0, \infty)$. Since $f_k(y, \theta)$ is not the zero function and depends nontrivially on $y$, we can see that if we want to make a small perturbation to the value $p_{12} = f_k(x_{12}, \theta)$, without changing any of the other $p_{ij}$ values, we can do this by perturbing $x_{12}$ and leaving all the other $x_{ij}$'s fixed. This is possible because no ideal generators of $I$ involved the variables $x_{12}$. A similar argument holds with respect to the coordinate $p_{45}$ which proves the first part of the proposition.

To see that there are equations in $I(\mathcal{M}_{k,T})$ that do involve all the other variables, we make two observations. First, the natural symmetry group of the tree $T$, translates into symmetries of the ideal $I(\mathcal{M}_{k,T})$. In particular, the variables fall into 3 orbits under this symmetry group $\{p_{12}, p_{45}\}$, $\{p_{14}, p_{15}, p_{24}, p_{25}\}$, and $\{p_{13}, p_{23}, p_{34}, p_{35}\}$. The set $\mathcal{M}_{k,T}$ has dimension 8, by local identifiability of parameters when $k \geq 2$. There cannot be any equation involving only the variables $\{p_{14}, p_{15}, p_{24}, p_{25}\}$, since such equations would already by implied when looking at 4 leaf trees (since only 4 indices are involved) and we know there are no such relations. There also cannot be any relations involving only the variables $\{p_{13}, p_{23}, p_{34}, p_{35}\}$, because if there were, we could also find such a relation in the ideal $I$ associated with the parameterization in the $x_{ij}$. But there is clearly no such equation. Since $\mathcal{M}_{k,T}$ has dimension 8, there must exist some equation, any such equation must involve some variables from the set $\{p_{14}, p_{15}, p_{24}, p_{25}\}$ and some from the set $\{p_{13}, p_{23}, p_{34}, p_{35}\}$. Then, taking the orbit of such an equation and adding the equations together (with some random coefficients if necessary) will produce an equation in $I(\mathcal{M}_{k,T})$ involving all of the variables except $p_{12}$ and $p_{45}$, as desired. □

**Theorem 4.6** *For $k = 1$, the tree parameter of the 1-mer multispecies Jukes–Cantor coalescent model is identifiable for all trees on $n \geq 4$ leaves. For $k \geq 2$, the tree parameter of the $k$-mer multispecies Jukes–Cantor coalescent model is identifiable for all trees on $n \geq 5$ leaves.*

*Proof* For $k = 1$, Proposition 4.4 shows that we can distinguish between 4 leaf trees using the invariants. Then, we use the basic fact that if we know the subtrees on all 4-leaf subsets, we can recover the underlying tree (that is, the quartets in the tree determine the tree, see, e.g., Semple and Steel 2003).

For $k \geq 2$, Proposition 4.5 shows that we can distinguish between 5 leaf trees using the invariants of the tree. Indeed, for each 5 leaf tree $T$ there will be a distinct set of pair of variables which are the ones that do not appear in any generator of $I(\mathcal{M}_{k,T})$. This guarantees that for any $T, T'$ that are different 5 leaf trees that there are nontrivial polynomials in $I(\mathcal{M}_{k,T}) \setminus I(\mathcal{M}_{k,T'})$, guaranteeing identifiability of the tree parameter from Proposition 4.3. Once 5 leaf trees are identified, this tells us all 4 leaf subtrees in our underlying tree, which identifies the tree for an arbitrary number of leaves $\geq 5$. □

An anonymous referee suggested that an alternate approach to proving a variation on Theorem 4.6 is to connect 1-mer frequency correlations to the $p$-distance between aligned sequences, and then use the results of Dasarathy et al. (2015).

## 5 Identifiability from Combinations of $k$-mer Distances

In previous sections, we studied the identifiability of parameters from the expected $k$-mer distances assembled for multiple pairs of taxa. Here, we consider a single pair of taxa, and we instead assemble expected $k$-mer distances for two distinct values of $k$, which we denote $k$ and $l$. In this section, we show that numerical parameters are identifiable in this context for any set of two or more taxa, when the data consist of both pairwise $k$-mer distances and pairwise $l$-mer distances for $l \neq k$.

**Theorem 5.1** *Let $1 \leq k < l$ and let $\phi : (0, 1] \times \mathbb{R}_{>0} \to \mathbb{R}^2$ be the map given by $\phi(x, \theta) = (f_k(x, \theta), f_l(x, \theta))$, where $f_k$ is defined by Eq. 2. Then, $\phi$ is one-to-one.*

*Proof* To simplify notation, we reparameterize $f_k$ of Eq. 2 as follows: Let $y = 3x$, $\xi = 8\theta/3$, and define

$$g_k(y, \xi) = 1 - \frac{1}{4^k} \sum_{h=0}^{k} \binom{k}{h} \frac{y^h}{1 + \xi h}.$$

Then, $g_k(y, \xi) = f_k(x, \theta)$. We consider the map $\psi : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}^2$ defined by $\psi(y, \xi) = (g_k(y, \xi), g_l(y, \xi))$. It suffices to show that $\psi$ is one-to-one.

By Theorem 7 of Gale and Nikaido (1965), a function $F : \Omega \to \mathbb{R}^2$ is one-to-one on a rectangular region $\Omega$ if its partial derivatives are continuous and no principal minors of its Jacobian vanish. The partial derivatives of $\psi$ are continuous on the positive orthant, by the following formulas:

$$\frac{\partial g_k}{\partial \xi} = \sum_{h=1}^{k} \binom{k}{h} \frac{h y^h}{(1 + \xi h)^2}$$

$$\frac{\partial g_k}{\partial y} = -\sum_{h=1}^{k} \binom{k}{h} \frac{h y^{h-1}}{1 + \xi h}.$$

Thus, to show that $\psi$ is one-to-one, it suffices to show that the principal minors of the Jacobian matrix of $\psi$ do not vanish on the positive orthant. This is clear for the $1 \times 1$ minors, so it suffices to show that the Jacobian determinant of $\psi$ does not vanish.

$$\begin{aligned}
\det(d\psi)(y,\xi) &= \frac{\partial g_k}{\partial y}\frac{\partial g_l}{\partial \xi} - \frac{\partial g_l}{\partial y}\frac{\partial g_k}{\partial \xi} = -\frac{\partial g_l}{\partial y}\frac{\partial g_k}{\partial \xi} + \frac{\partial g_k}{\partial y}\frac{\partial g_l}{\partial \xi} \\
&= \sum_{h=1}^{k}\binom{k}{h}\frac{hy^h}{(1+\xi h)^2}\sum_{i=1}^{l}\binom{l}{i}\frac{iy^{i-1}}{1+\xi i} \\
&\quad - \sum_{h=1}^{k}\binom{k}{h}\frac{hy^{h-1}}{1+\xi h}\sum_{i=1}^{l}\binom{l}{i}\frac{iy^{i}}{(1+\xi i)^2} \\
&= \sum_{h=1}^{k}\sum_{i=1}^{l}\binom{k}{h}\binom{l}{i}\frac{hy^h}{(1+\xi h)^2}\frac{iy^{i-1}}{1+\xi i} - \frac{hy^{h-1}}{1+\xi h}\frac{iy^i}{(1+\xi i)^2} \\
&= \sum_{h=1}^{k}\sum_{i=1}^{l}\binom{k}{h}\binom{l}{i}\frac{(1+\xi i)-(1+\xi h)}{(1+\xi h)^2(1+\xi i)^2}hiy^{i+h-1} \\
&= \sum_{h=1}^{k}\sum_{i=1}^{l}\binom{k}{h}\binom{l}{i}\frac{\xi hi(i-h)}{(1+\xi h)^2(1+\xi i)^2}y^{i+h-1}.
\end{aligned}$$

Let

$$a_{hi} = \binom{k}{h}\binom{l}{i}\frac{\xi hi(i-h)}{(1+\xi h)^2(1+\xi i)^2}.$$

Then, the coefficient of $y^m$ in $\det(d\psi)$ is:

$$[y^m]\det(d\psi) = \sum_{i+h-1=m} a_{hi}. \tag{6}$$

To show that $\det(d\psi)$ does not vanish on the positive orthant, it suffices to show that the above sum is positive for all $m$ and for all $\xi > 0$. To see this, we view the $a_{hi}$ as the entries of a matrix $(a_{hi})_{(h,i)\in[k]\times[l]}$. For fixed $m$, any term $a_{hi}$ which appears in (6), lies along the "cross-diagonal" $h + i - 1 = m$ of $A$. The signs of the entries of this matrix are

$$(\operatorname{sign}(a_{hi}))_{(h,i)\in[k]\times[l]} = \begin{pmatrix} 0 & + & + & + & + & + & \cdots & + \\ - & 0 & + & + & + & + & \cdots & + \\ - & - & 0 & + & + & + & \cdots & + \\ \vdots & & & \ddots & & & & \vdots \\ - & - & \cdots & - & 0 & + & \cdots & + \end{pmatrix}.$$

From the pattern of signs, we see that if $a_{hi}$ is negative, then $h > i$. Since $l > k$, the term $a_{ih}$ is also a term in (6), and it is positive. Thus, it suffices to show that $a_{ih} > -a_{hi}$ for $h > i$. By simple algebra, this is equivalent to the following inequality:

$$\binom{k}{i}\binom{l}{h} - \binom{k}{h}\binom{l}{i} > 0 \text{ for } h > i.$$

This can be verified by expanding the binomial coefficients:

$$\begin{aligned}
\binom{k}{i}\binom{l}{h} - \binom{k}{h}\binom{l}{i} &= \frac{k!}{i!(k-i)!}\frac{l!}{h!(l-h)!} - \frac{k!}{h!(k-h)!}\frac{l!}{i!(l-i)!} \\
&= \frac{1}{i!h!}\left(\frac{k!}{(k-i)!}\frac{l!}{(l-h)!} - \frac{k!}{(k-h)!}\frac{l!}{(l-i)!}\right) \\
&= \frac{1}{i!h!}\left((k)_i(l)_h - (k)_h(l)_i\right) \\
&= \frac{(l)_i(k)_i}{i!h!}\left((l-i)_{h-i} - (k-i)_{h-i}\right) > 0
\end{aligned}$$

Thus, all principal minors of the Jacobian $d\psi$ are non-vanishing on the positive orthant. Therefore, $\psi$ is injective on the positive orthant by Theorem 7 of Gale and Nikaido (1965). Thus, $\phi$ is injective. □

**Corollary 5.2** *Let $k \neq l$ be positive integers. Let $T$ be an unrooted tree with $n \geq 2$ leaves and $m$ edges. The map $\phi_{k,l,T} : (0,1]^m \times \mathbb{R}_{>0} \to \mathbb{R}^{n(n-1)}$ with*

$$\phi_{k,l,T}(a,\theta) = (\phi_{k,T}(a,\theta), \phi_{l,T}(a,\theta))$$

*is one-to-one. In particular, the numerical parameters of the model $a, \theta$ are identifiable given $k$-mer and $l$-mer distances between all pairs of taxa.*

*Proof* From Theorem 5.1, we know that we can recover all the products $x_{ij} = \prod_{e \in P(i,j)} a_e$ for each pair of taxa $i, j$ from the $k$-mer and $l$-mer vectors. Since none of the $a_e$ are zero, none of the $x_{ij}$ are zero either. Taking logarithms, we have:

$$-\log x_{ij} = -\sum_{e \in P(i,j)} \log a_e = \frac{4}{3}\sum_{e \in P(i,j)} w_e,$$

so the matrix of $-\log x_{ij}$ is an additive tree distance. The branch lengths in an additive tree can be recovered from the pairwise distances. □

## 6 Discussion

The expected $k$-mer distance derived here provides the basis for a statistically consistent $k$-mer-based method which generalizes the method devised by Allman, Rhodes, and the second author. This generalization extends the $k$-mer method to the case in which sequence blocks correspond to genes, whose gene trees are modeling by the coalescent. We have derived our results under a relatively simple setting where there is a global unknown effective population size $N$ that we use over the entire tree, and we work with the Jukes–Cantor substitution model. It would be natural to try to extend these results to more general settings with a more general substitution model and allowing the effective population size to vary over branches of the species tree.

Our identifiability results make a first suggestion for an algorithm for reconstructing the species tree based on $k$-mer distances from multiple genes. Namely, using the results in Sect. 5, for distinct positive integers $k$ and $l$, $k$-mer and $l$-mer frequency distributions can be used to estimate the species divergence time between each pair of taxa. These pairwise distances can be used in a distance-based method like Neighbor-Joining to reconstruct an evolutionary tree. It remains to be seen how this methodology will perform against other methods for reconstructing species trees. Our identifiability results are the first step in showing that methods based on $k$-mers can be derived for these problems.

**Acknowledgements** Chris Durden was partially supported by the US National Science Foundation (DMS 1615660). Seth Sullivant was partially supported by the US National Science Foundation (DMS 1615660) and by the David and Lucille Packard Foundation.

## References

Allman ES, Degnan JH, Rhodes JA (2011) Determining species tree topologies from clade probabilities under the coalescent. J Theor Biol 289:96–106

Allman ES, Rhodes JA, Sullivant S (2017) Statistically consistent k-mer methods for phylogenetic tree reconstruction. J Comput Biol 24(2):153–171

Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG (2010) Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol Biol 5:21

Cox DA, Little J, O'Shea D (2015) Ideals, varieties, and algorithms. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, An introduction to computational algebraic geometry and commutative algebra

Dasarathy G, Nowak R, Roch S (2015) Data requirement for phylogenetic inference from multiple loci: A new distance method. IEEE/ACM Trans Comput Biol Bioinf 12:422–432

Daskalakis C, Roch S (2013) Alignment-free phylogenetic reconstruction: sample complexity via a branching process analysis. Ann Appl Probab 23:693–721

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf 5:113

Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Gale D, Nikaido H (1965) The Jacobian matrix and global univalence of mappings. Math Ann 159(2):81–93

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

Leung D, Drton M, Hara H et al (2016) Identifiability of directed Gaussian graphical models with one latent source. Electron J Stat 10(1):394–422

McVean GAT (2002) A genealogical interpretation of linkage disequilibrium. Genetics 162(2):987–991

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

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. Genetics 164(4):1645–1656

Semple C, Steel M (2003) Phylogenetics, volume 24 of Oxford lecture series in mathematics and its applications. Oxford University Press, Oxford

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Speyer D, Sturmfels B (2004) The tropical Grassmannian. Adv Geom 4(3):389–411

Takahata N (1986) An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet Res 48(03):187–190

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48(2):198–221

Wakeley J (2009) Coalescent theory: an introduction, vol 1. Roberts & Company Publishers Greenwood Village, Colorado