Phylogeny of Mixture Models: Robustness of Maximum Likelihood and Non-Identifiable Distributions

DANIEL ŠTEFANKOVIČ[1] and ERIC VIGODA[2]

ABSTRACT

We address phylogenetic reconstruction when the data is generated from a mixture distribution. Such topics have gained considerable attention in the biological community with the clear evidence of heterogeneity of mutation rates. In our work we consider data coming from a mixture of trees which share a common topology, but differ in their edge weights (i.e., branch lengths). We first show the pitfalls of popular methods, including maximum likelihood and Markov chain Monte Carlo algorithms. We then determine in which evolutionary models, reconstructing the tree topology, under a mixture distribution, is (im)possible. We prove that every model whose transition matrices can be parameterized by an open set of multilinear polynomials, either has non-identifiable mixture distributions, in which case reconstruction is impossible in general, or there exist linear tests which identify the topology. This duality theorem, relies on our notion of linear tests and uses ideas from convex programming duality. Linear tests are closely related to linear invariants, which were first introduced by Lake, and are natural from an algebraic geometry perspective.

Key words: phylogeny, maximum likelihood, linear invariants, linear tests, Markov chain Monte Carlo.

1. INTRODUCTION

A MAJOR OBSTACLE to phylogenetic inference is the heterogeneity of genomic data. For example, mutation rates vary widely between genes, resulting in different branch lengths in the phylogenetic tree for each gene. In many cases, even the topology of the tree differs between genes. Within a single long gene, we are also likely to see variations in the mutation rate, see Hellmann et al. (2005) for a current study on regional mutation rate variation.

Our focus is on phylogenetic inference based on single nucleotide substitutions. In this paper, we study the effect of mutation rate variation on phylogenetic inference. The exact mechanisms of single nucleotide substitutions are still being studied; hence, the causes of variations in the rate of these mutations are unresolved, see Hellmann et al. (2003), Meunier and Duret (2004), and Webster et al. (2004). In this paper, we study phylogenetic inference in the presence of heterogeneous data.

For homogenous data, i.e., data generated from a single phylogenetic tree, there is considerable work on consistency of various methods, such as likelihood (Chang, 1996) and distance methods, and inconsistency

[1] Department of Computer Science, University of Rochester, Rochester, New York.
[2] College of Computing, Georgia Institute of Technology, Atlanta, Georgia.

of other methods, such as parsimony (Felsenstein, 1978) Consistency means that the methods converge to the correct tree with sufficiently large amounts of data. We refer the interested reader to Felsenstein (2004) for an introduction to these phylogenetic approaches.

There are several works showing the pitfalls of popular phylogenetic methods when data is generated from a mixture of trees, as opposed to a single tree. We review these works in detail shortly. The effect of mixture distributions has been of marked interest recently in the biological community, for instance, see the recent publications of Kolaczkowski and Thornton (2004), and Mossel and Vigoda (2005).

In our setting, the data is generated from a mixture of trees which have a common tree topology, but can vary arbitrarily in their mutation rates. We address whether it is possible to infer the tree topology. We introduce the notion of a *linear test*. For any mutational model whose transition probabilities can be parameterized by an open set (see the following subsection for a precise definition), we prove that the topology can be reconstructed by linear tests, or it is impossible in general due to a non-identifiable mixture distribution. For several of the popular mutational models we determine which of the two scenarios (reconstruction or non-identifiability) hold.

The notion of a linear test is closely related to the notion of linear invariants. In fact, Lake's invariants are a linear test. There are simple examples where linear tests exist and linear invariants do not (in these examples, the mutation rates are restricted to some range). However, for the popular mutation models, such as Jukes-Cantor and Kimura's 2 parameter model (both of which are closed under multiplication) we have no such examples. For the Jukes-Cantor and Kimura's 2 parameter model, we prove the linear tests are essentially unique (up to certain symmetries). In contrast to the study of invariants, which is natural from an algebraic geometry perspective, our work is based on convex programming duality.

We present the background material before formally stating our new results. We then give a detailed comparison of our results with related previous work.

An announcement of the main results of this paper, along with some applications of the technical tools presented here, are presented in Stefankovic and Vigoda (2007).

1.1. Background

A phylogenetic tree is an unrooted tree T on n leaves (called taxa, corresponding to n species) where internal vertices have degree three. Let $E(T)$ denote the edges of T and $V(T)$ denote the vertices. The mutations along edges of T occur according to a continuous-time Markov chain. Let Ω denote the states of the model. The case $|\Omega| = 4$ is biologically important, whereas $|\Omega| = 2$ is mathematically convenient.

The model is defined by a phylogenetic tree T and a distribution π on Ω. Every edge e has an associated $|\Omega| \times |\Omega|$ rate matrix R_e, which is reversible with respect to π, and a time t_e. Note, since R_e is reversible with respect to π, then π is the stationary vector for R_e (i.e., $\pi R_e = 0$). The rate matrix defines a continuous time Markov chain. Then, R_e and t_e define a transition matrix $P_e = \exp(t_e R_e)$. The matrix is a stochastic matrix of size $|\Omega| \times |\Omega|$, and thus defines a discrete-time Markov chain, which is time-reversible, with stationary distribution π (i.e., $\pi P_e = \pi$).

Given $\overrightarrow{P} = (P_e)_{e \in E(T)}$ we then define the following distribution on labellings of the vertices of T. We first orient the edges of T away from an arbitrarily chosen root r of the tree. (We can choose the root arbitrarily since each P_e is reversible with respect to π.) Then, the probability of a labeling $\ell : V(T) \to \Omega$ is

$$\mu'_{T,\overrightarrow{P}}(\ell) = \pi(\ell(r)) \prod_{\overrightarrow{uv} \in E(T)} \overrightarrow{P}_{uv}(\ell(u), \ell(v)). \tag{1}$$

Let $\mu_{T,\overrightarrow{P}}$ be the marginal distribution of $\mu'_{T,\overrightarrow{P}}$ on the labelings of the leaves of T ($\mu_{T,\overrightarrow{P}}$ is a distribution on Ω^n where n is the number of leaves of T). The goal of phylogeny reconstruction is to reconstruct T (and possibly $\overrightarrow{P}$) from $\mu_{T,\overrightarrow{P}}$ (more precisely, from independent samples from $\mu_{T,\overrightarrow{P}}$).

The simplest four-state model has a single parameter α for the off-diagonal entries of the rate matrix. This model is known as the Jukes-Cantor model, which we denote as JC. Allowing two parameters in the rate matrix is Kimura's two-parameter model, which we denote as K2 (see Section 5.2. for a formal definition). The K2 model accounts for the higher mutation rate of transitions (mutations within purines or pyrimidines) compared to transversions (mutations between a purine and a pyrimidine). Kimura's three-parameter model, which we refer to as K3, accounts for the number of hydrogen bonds altered by the mutation (see Section 7

for a formal definition of the K3 model). For $|\Omega| = 2$, the model is binary and the rate matrix has a single parameter α. This model is known as the CFN (Cavender-Farris-Neyman) model. For any examples in this paper involving the CFN, JC, K2, or K3 models, we restrict the model to rate matrices where all the entries are positive, and times t_e which are positive and finite.

We will use $\mathcal{M}$ to denote the set of transition matrices obtainable by the model under consideration, i.e.,

$$\mathcal{M} = \{P_e = \exp(t_e R_e : t_e \text{ and } R_e \text{ are allowed in the model } \}.$$

The above setup allows additional restrictions in the model, such as requiring $t_e > 0$ which is commonly required.

In our framework, a model is specified by a set $\mathcal{M}$, and then each edge is allowed any transition matrix $P_e \in \mathcal{M}$. We refer to this framework as the *unrestricted framework*, since we are not imposing any dependencies on the choice of transition matrices between edges. This set-up is convenient since it gives a natural algebraic framework for the model as we will see in some later proofs. A similar set-up was required in the work of Allman and Rhodes (2007), also to utilize the algebraic framework.

An alternative framework (which is typical in practical works) requires a common rate matrix for all edges, specifically $R = R_e$ for all e. Note that we cannot impose such a restriction in our unrestricted framework, since each edge is allowed any matrix in $\mathcal{M}$. We will refer to this framework as the *common rate framework*. Note, for the Jukes-Cantor and CFN models, the unrestricted and common rate frameworks are identical, since there is only a single parameter for each edge in these models. We will discuss how our results apply to the common rate framework when relevant, but the default setting of our results is the unrestricted model.

Returning to our setting of the unrestricted framework, recall under the condition $t_e > 0$ the set $\mathcal{M}$ is not a compact set (and is parameterized by an open set as described shortly). This will be important for our work since our main result will only apply to models where $\mathcal{M}$ is an open set. Moreover we will require that $\mathcal{M}$ consists of multilinear polynomials. More precisely, a polynomial $p \in \mathbb{R}[x_1, \ldots, x_m]$ is *multi-linear* if for each variable x_i the degree of p in x_i is at most 1. Our general results will apply when the model is a set of multi-linear polynomials which are parameterized by an open set which we now define precisely.

Definition 1. *We say that a set $\mathcal{M}$ of transition matrices is* parameterized by an open set *if there exists a finite set Ω, a distribution π over Ω, an integer m, an open set $O \subseteq \mathbb{R}^m$, and multi-linear polynomials $p_{ij} \in \mathbb{R}[x_1, \ldots, x_m]$ such that*

$$\mathcal{M} = \{(p_{ij})_{i,j=1}^{\Omega} \mid (x_1, \ldots, x_m) \in O\},$$

where $\mathcal{M}$ is a set of stochastic matrices which are reversible with respect to π (thus π is their stationary distribution).

Typically the polynomials p_{ij} are defined by an appropriate change of variables from the variables defining the rate matrices. Some examples of models that are paraemeterized by an open set are the general Markov model considered by Allman and Rhodes (2007); Jukes-Cantor, Kimura's two-parameter and three-parameter, and Tamura-Nei models. For the Tamura-Nei model (which is a generalization of Jukes-Cantor and Kimura's models), we show in Stefankovic and Vigoda (2007) how the model can be re-parameterized in a straightforward manner so that it consists of multi-linear polynomials, and thus fits the parameterized by an open set condition (assuming the additional restriction $t_e > 0$).

1.2. Mixture models

In our setting, we will generate assignments from a mixture distribution. We will have a single tree topology T, a collection of k sets of transition matrices $\overrightarrow{P_1}, \overrightarrow{P_2}, \ldots, \overrightarrow{P_k}$ where $\overrightarrow{P_i} \in \mathcal{M}^{E(T)}$ and a set of non-negative reals $q_1, q_2, \ldots, q_k$ where $\sum_i q_i = 1$. We then consider the mixture distribution:

$$\mu = \sum_i q_i \mu_{T, \overrightarrow{P_i}}$$

Thus, with probability q_i we generate a sample according to $\mu_{T, \overrightarrow{P_i}}$. Note the tree topology is the same for all the distributions in the mixture (thus there is a notion of a generating topology). In several of our simple examples we will set $k = 2$ and $q_1 = 1/2$, thus we will be looking at a uniform mixture of two trees.

1.3. Maximum likelihood and MCMC results

We begin by showing a simple class of mixture distributions where popular phylogenetic algorithms fail. In particular, we consider maximum likelihood methods, and Markov chain Monte Carlo (MCMC) algorithms for sampling from the posterior distribution.

In the following, for a mixture distribution μ, we consider the likelihood of a tree T as, the maximum over assignments of transition matrices $\overrightarrow{P}$ to the edges of T, of the probability that the tree $(T, \overrightarrow{P})$ generated μ. Thus, we are considering the likelihood of a pure (non-mixture) distribution having generated the mixture distribution μ. More formally, the maximum expected log-likelihood of tree T for distribution μ is defined by

$$\mathcal{L}_T(\mu) = \max_{\overrightarrow{P} \in \mathcal{M}^E} \mathcal{L}_{T,\overrightarrow{P}}(\mu),$$

where

$$\mathcal{L}_{T,\overrightarrow{P}}(\mu) = \sum_{y \in \Omega^n} \mu(y) \ln(\mu_{T,\overrightarrow{P}}(y))$$

Recall for the CFN, JC, K2 and K3 models, $\mathcal{M}$ is restricted to transition matrices obtainable from positive rate matrices R_e and positive times t_e.

Chang (1996b) constructed a mixture example where likelihood (maximized over the best single tree) was maximized on the wrong topology (i.e., different from the generating topology). In Chang's examples one tree had all edge weights sufficiently small (corresponding to invariant sites). We consider examples with less variation within the mixture and fewer parameters required to be sufficiently small. We consider (arguably more natural) examples of the same flavor as those studied by Kolaczkowski and Thornton (2004), who showed experimentally that in the JC model, likelihood appears to perform poorly on these examples.

Figure 1 shows the form of our examples where C and x are parameters of the example. We consider a uniform mixture of the two trees in the figure. For each edge, the figure shows the mutation probability, i.e., it is the off-diagonal entry for the transition matrix. We consider the CFN, JC, K2, and K3 models.

We prove that in this mixture model, maximum likelihood is not robust in the following sense: when likelihood is maximized over the best single tree, the maximum likelihood topology is different from the generating topology.

In our example all of the off-diagonal entries of the transition matrices are identical. Hence for each edge we specify a single parameter and thus we define a set $\overrightarrow{P}$ of transition matrices for a 4-leaf tree by a 5-dimensional vector where the i-th coordinate is the parameter for the edge incident leaf i, and the last coordinate is the parameter for the internal edge.

Here is the statement of our result on the robustness of likelihood.

Theorem 2. *Let $C \in (0, 1/|\Omega|)$. Let $\overrightarrow{P_1} = (C+x, C-x, C-x, C+x, x^2)$ and $\overrightarrow{P_2} = (C-x, C+x, C+x, C-x, x^2)$. Consider the following mixture distribution on T_3:*

$$\mu_x = \left(\mu_{T_3,\overrightarrow{P_1}} + \mu_{T_3,\overrightarrow{P_2}}\right)/2.$$

1. *In the CFN model, for all $C \in (0, 1/2)$, there exists $x_0 > 0$ such that for all $x \in (0, x_0)$ the maximum-likelihood tree for μ_x is T_1.*
2. *In the JC, K2 and K3 models, for $C = 1/8$, there exists $x_0 > 0$ such that for all $x \in (0, x_0)$ the maximum-likelihood tree for μ_x is T_1.*

Recall, likelihood is maximized over the best pure (i.e., non-mixture) distribution.

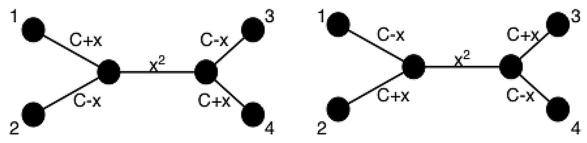


FIG. 1. In the binary CFN model, the parameters C and x define the mutation probabilities. For all choices of C and for x sufficiently small, maximum likelihood tree is different than the generating tree.

Note, for the above theorem, we are maximizing the likelihood over assignments of valid transition matrices for the model. For the above models, valid transition matrices are those obtainable with finite and positive times t_e, and rate matrices R_e where all the entries are positive.

A key observation for our proof approach of Theorem 2 is that the two trees in the mixture example are the same in the limit $x \to 0$. The $x = 0$ case is used in the proof for the $x > 0$ case. We expect the above theorem holds for more a general class of examples (such as arbitrary x, and any sufficiently small function on the internal edge), but our proof approach requires x sufficiently small. Our proof approach builds upon the work of Mossel and Vigoda (2005).

Our results also extend to show, for the CFN and JC models, MCMC methods using NNI transitions converge exponentially slowly to the posterior distribution. This result requires the 5-leaf version of mixture example from Figure 1. We state our MCMC result formally in Theorem 32 in Section 9 after presenting the background material. Previously, Mossel and Vigoda (2005) showed a mixture distribution where MCMC methods converge exponentially slowly to the posterior distribution. However, in their example, the tree topology varies between the two trees in the mixture.

1.4. Duality theorem: non-identifiablity or linear tests

Based on the above results on the robustness of likelihood, we consider whether there are any methods which are guaranteed to determine the common topology for mixture examples. We first found that in the CFN model there is a simple mixture example of size 2, where the mixture distribution is non-identifiable. In particular, there is a mixture on topology T_1 and also a mixture on T_3 which generate identical distributions. Hence, it is impossible to determine the correct topology in the worst case. It turns out that this example does not extend to models such as JC and K2. In fact, all mixtures in JC and K2 models are identifiable. This follows from our following duality theorem which distinguishes which models have non-identifiable mixture distributions, or have an easy method to determine the common topology in the mixture.

We prove, that for any model which is parameterized by an open set, either there exists a linear test (which is a strictly separating hyperplane as defined shortly), or the model has *non-identifiable* mixture distributions in the following sense. Does there exist a tree T, a collection $\overrightarrow{P_1}, \ldots, \overrightarrow{P_k}$, and distribution $p_1, \ldots, p_k$, such that there is another tree $T' \neq T$, a collection $\overrightarrow{P'_1}, \ldots, \overrightarrow{P'_k}$ and a distribution $p'_1, \ldots, p'_k$ where:

$$\sum_{i=1}^{k} p_i \mu_{T, \overrightarrow{P_i}} = \sum_{i=1}^{k} p'_i \mu_{T', \overrightarrow{P'_i}}$$

Thus, in this case, it is impossible to distinguish these two distributions. Hence, we can not even infer which of the topologies T or T' is correct. If the above holds, we say the model has *non-identifiable mixture distributions*.

In contrast, when there is no non-identifiable mixture distribution we can use the following notion of a linear test to reconstruct the topology. A *linear test* is a hyperplane strictly separating distributions arising from two different 4 leaf trees (by symmetry the test can be used to distinguish between the 3 possible 4 leaf trees). It suffices to consider trees with 4 leaves, since the full topology can be inferred from all 4 leaf subtrees (Bandelt and Dress, 1986).

Our duality theorem uses a geometric viewpoint (for a nice introduction to a geometric approach, see Kim, 2000). Every mixture distribution μ on a 4-leaf tree T defines a point $z \in \mathbb{R}^N$ where $N = |\Omega|^4$. For example, for the CFN model, we have $z = (z_1, \ldots, z_{2^4})$ and $z_1 = \mu(0000), z_2 = \mu(0001), z_3 = \mu(0010), \ldots, z_{2^4} = \mu(1111)$. Let C_i denote the set of points corresponding to distributions $\mu(T_i, \overrightarrow{P})$ for the 4-leaf tree T_i, $i = 1, 2, 3$. A linear test is a hyperplane which strictly separates the sets for a pair of trees.

Definition 3. *Consider the 4-leaf trees T_2 and T_3. A* linear test *is a vector $t \in \mathbb{R}^{|\Omega|^4}$ such that $t^T \mu_2 > 0$ for any mixture distribution μ_2 arising from T_2 and $t^T \mu_3 < 0$ for any mixture distribution μ_3 arising from T_3.*

There is nothing special about T_2 and T_3 - we can distinguish between mixtures arising from any two 4 leaf trees, e. g., if t is a test then $t^{(1\,3)}$ distinguishes the mixtures from T_1 and the mixtures from T_2, where $(1\,3)$ swaps the labels for leaves 1 and 3. More precisely, for all $(a_1, a_2, a_3, a_4) \in |\Omega|^4$,

$$t^{(1\,3)}_{a_1, a_2, a_3, a_4} = s_{a_4, a_2, a_3, a_1} \tag{2}$$

Theorem 4. *For any model whose set $\mathcal{M}$ of transition matrices is parameterized by an open set (of multilinear polynomials), exactly one of the following holds:*

- *There exist non-identifiable mixture distributions, or*
- *There exists a linear test.*

For the JC and K2 models, the existence of a linear test follows immediately from Lake's (1987) linear invariants. Hence, our duality theorem implies that there are no non-identifiable mixture distributions in this model. In contrast for the K3 model, we prove there is no linear test, hence there is an non-identifiable mixture distribution. We also prove that in the K3 model in the common rate matrix framework, there is a non-identifiable mixture distribution.

To summarize, we show the following:

Theorem 5.

1. *In the CFN model, there is an ambiguous mixture of size 2.*
2. *In the JC and K2 model, there are no ambiguous mixtures.*
3. *In the K3 model there exists a non-identifiable mixture distribution (even in the common rate matrix framework).*

Steel et al. (1994) previously proved the existence of a non-identifiable mixture distribution in the CFN model, but their proof was non-constructive and gave no bound on the size of the mixture. Their result had the more appealing feature that the trees in the mixture were scalings of each other.

Allman and Rhodes (2007) recently proved identifiability of the topology for certain classes of mixture distributions using invariants (not necessarily linear). Rogers (2001) proved that the topology is identifiable in the general time-reversible model when the rates vary according to what is known as the invariable sites plus gamma distribution model. Much of the current work on invariants uses ideas from algebraic geometry, whereas our notion of a linear test is natural from the perspective of convex programming duality.

Note, that even in models that do not have non-identifability between different topologies, there is non-identifiability within the topology. An interesting example was shown by Evans and Warnow (2004).

1.5. Outline of paper

We prove, in Section 3, Theorem 4 that a phylogenetic model has a non-identifiable mixture distribution or a linear test. We then detail Lake's linear invariants in Section 5, and conclude the existence of a linear test for the JC and K2 models. In Sections 6 and 7, we prove that there are non-identifiable mixtures in the CFN and K3 models, respectively. We also present a linear test for a restricted version of the CFN model in Section 6.3. We prove the maximum likelihood results stated in Theorem 2 in Section 8. The maximum likelihood results require several technical tools which are also proved in Section 8. The MCMC results are then stated formally and proved in Section 9.

2. PRELIMINARIES

Let the permutation group S_4 act on the 4-leaf trees $\{T_1, T_2, T_3\}$ by renaming the leaves. For example $(14) \in S_4$ swaps T_2 and T_3, and fixes T_1. For $\pi \in S_n$, we let T^π denote tree T permuted by π. It is easily checked that the following group K (Klein group) fixes every T_i:

$$K = \{(), (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\} \leq S_4. \tag{3}$$

Note that $K \trianglelefteq S_4$, i. e., K is a normal subgroup of S_4.

For weighted trees we let S_4 act on $(T, \overrightarrow{P})$ by changing the labels of the leaves but leaving the weights of the edges untouched. Let $\pi \in S_n$ and let $T' = T^\pi$. Note that the distribution $\mu_{T',w}$ is just a permutation of the distribution $\mu_{T,w}$:

$$\mu_{T',\overrightarrow{P}}(a_{\pi_1}, \ldots, a_{\pi_n}) = \mu_{T,\overrightarrow{P}}(a_1, \ldots, a_n). \tag{4}$$

The actions on weighted trees and on distributions are compatible:

$$\mu_{(T,\overrightarrow{P})^\pi} = (\mu_{T,\overrightarrow{P}})^\pi.$$

3. DUALITY THEOREM

In this section we prove the duality theorem (i.e., Theorem 4).

Our assumption that the transition matrices of the model are parameterized by an open set implies the following observation.

Observation 6. *For models parameterized by an open set, the coordinates of $\mu'_{T,w}$ are multi-linear polynomials in the parameters.*

We now state a classical result that allows one to reduce the reconstruction problem to trees with 4 leaves. Note there are three distinct leaf-labeled binary trees with 4 leaves. We will call them T_1, T_2, T_3 see Figure 2.

For a tree T and a set S of leaves, let $T|_S$ denote the induced subgraph of T on S where internal vertices of degree 2 are removed.

Theorem 7 [Bandelt and Dress (1986)]. *For distinct leaf-labeled binary trees T and T' there exist a set S of 4 leaves where $T|_S \neq T'|_S$. Hence, the set of induced subgraphs on all 4-tuples of leaves determines a tree.*

The above theorem also simplifies the search for non-identifiable mixture distributions.

Corollary 8. *If there exists a non-identifiable mixture distribution then there exists a non-identifiable mixture distribution on trees with 4 leaves.*

Recall from the Introduction, the mixtures arising from T_i form a convex set in the space of joint distributions on leaf labelings. A test is a hyperplane *strictly* separating the mixtures arising from T_2 and the mixtures arising from T_3. For general disjoint convex sets, a strictly separating hyperplane need not exist (e. g., take $C_1 = \{(0,0)\}$ and $C_2 = \{(0,y) \,|\, y > 0\} \cup \{(x,y) \,|\, x > 0\}$). The sets of mixtures are special - they are convex hulls of images of open sets under a multi-linear polynomial map.

Lemma 9. *Let $p_1(x_1, \ldots, x_n), \ldots, p_m(x_1, \ldots, x_n)$ be multi-linear polynomials in $x_1, \ldots, x_n$. Let $p = (p_1, \ldots, p_m)$. Let $D \in \mathbb{R}^n$ be an open set. Let C be the convex hull of $\{p(x) \,|\, x \in D\}$. Assume that $0 \notin C$. There exists $s \in \mathbb{R}^m$ such that $s^T p(x) > 0$ for all $x \in D$.*

Proof. Suppose the polynomial p_i is a linear combination of the other polynomials, i.e., $p_i(x) = \sum_{j \neq i} c_j p_j(x)$. Let $p' = (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m)$. Let C' be the convex hull of $\{p'(x) \,|\, x \in D\}$. Then

$$(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m) \mapsto \left(y_1, \ldots, y_{i-1}, \sum_{j \neq i} c_j y_j, y_{i+1}, \ldots, y_m \right)$$

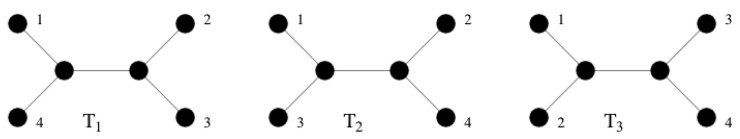


FIG . 2. All leaf-labeled binary trees with $n = 4$ leaves.

is a bijection between points in C' and C. Note, $0 \notin C'$. There exists a strictly separating hyperplane between 0 and C (i.e., there exists $s \in \mathbb{R}^m$ such that $s^T p(x) > 0$ for all $x \in D$) if and only if there exists a strictly separating hyperplane between 0 and C' (i.e., there exists $s' \in \mathbb{R}^m$ such that $s'^T p'(x) > 0$ for all $x \in D$). Hence, without loss of generality, we can assume that the polynomials $p_1, \ldots, p_m$ are linearly independent.

Since C is convex and $0 \notin C$, by the separating hyperplane theorem, there exists $s \neq 0$ such that $s^T p(x) \geq 0$ for all $x \in D$. If $s^T p(x) > 0$ for all $x \in D$ we are done.

Now suppose $s^T p(a) = 0$ for some $a \in D$. If $a \neq 0$, then by translating D by $-a$ and changing the polynomials appropriately (namely, $p_j(x) := p_j(x + a)$), without loss of generality, we can assume $a = 0$.

Let $r(x_1, \ldots, x_n) = s^T p(x_1, \ldots, x_n)$. Note that r is multi-linear and $r \neq 0$ because $p_1, \ldots, p_m$ are linearly independent. Since $a = 0$, we have $r(0) = 0$ and hence r has no constant monomial.

Let w be the monomial of lowest total degree which has a non-zero coefficient in r. Consider $y = (y_1, \ldots, y_m)$ where $y_j \neq 0$ for x_j which occur in w and $y_j = 0$ for all other x_j. Then, $r(y) = w(j)$ since there are no monomials of smaller degree, and any other monomials contain some y_j which is 0. Hence by choosing y sufficiently close to 0, we have $y \in D$ (since D is open) and $r(y) < 0$ (by choosing an appropriate direction for y). This contradicts the assumption that s is a separating hyperplane. Hence $r = 0$ which is a contradiction with the linear independence of the polynomials $p_1, \ldots, p_m$. ■

We now prove our duality theorem.

Proof of Theorem 4. Clearly there cannot exist both a non-identifiable mixture and a linear test. Let C_i be the convex set of mixtures arising from T_i (for $i = 2, 3$). Assume that $C_2 \cap C_3 = \emptyset$, i. e., there is no non-identifiable mixture in $\mathcal{M}$. Let $C = C_2 - C_3$. Note that C is convex, $0 \notin C$, and C is the convex hull of an image of open sets under multi-linear polynomial maps (by Observation 6). By Lemma 3 there exists $s \neq 0$ such that $s^T x > 0$ for all $x \in C$. Let $t = s - s^{(1\,4)}$ (where $s^{(1\,4)}$ is defined as in (2)). Let $\mu_2 \in C_2$ and let $\mu_3 = \mu_2^{(1\,4)}$ where $\mu_2^{(1\,4)}$ is defined analogously to $s^{(1\,4)}$. Then

$$t^T \mu_2 = (s - s^{(1\,4)})^T \mu_2 = s^T \mu_2 - s^T \mu_3 = s^T (\mu_2 - \mu_3) > 0.$$

Similarly for $\mu_3 \in C_3$ we have $t^T \mu_3 < 0$ and hence t is a test.

4. SIMPLIFYING THE SEARCH FOR A LINEAR TEST

The CFN, JC, K2, and K3 models all have a natural group-theoretic structure. We show some key properties of linear tests utilizing this structure. These properties will simplify proofs of the existence of linear tests in JC and K2 (and restricted CFN) models and will also be used in the proof of the non-existence of linear tests in K3 model. Our main objective is to use symmetry inherent in the phylogeny setting to drastically reduce the dimension of the space of linear tests.

Symmetric phylogeny models have a group of symmetries $G \leq S_\Omega$ (G is the intersection of the automorphism groups of the weighted graphs corresponding to the matrices in $\mathcal{M}$). The probability of a vertex labeling of T does not change if the labels of the vertices are permuted by an element of G. Thus the elements of Ω^n which are in the same orbit of the action of G on Ω^n have the same probability in any distribution arising from the model.

Let O' be the orbits of Ω^4 under the action of G. Let O be the orbits of O' under the action of K. Note that the action of $(1\,4)$ on O is well defined (because K is a normal subgroup of S_4). For each pair $o_1, o_2 \in O$ that are swapped by $(1\,4)$ let

$$\ell_{o_1,o_2}(\mu) = \sum_{a \in o_1} \mu(a) - \sum_{a \in o_2} \mu(a). \tag{5}$$

Lemma 10. *Suppose that $\mathcal{M}$ has a linear test s. Then $\mathcal{M}$ has a linear test t which is a linear combination of the ℓ_{o_1,o_2}.*

Proof. Let s be a linear test. Let

$$t' = \sum_{g \in K} s^g \quad \text{and} \quad t = t' - t'^{(1\,4)}.$$

Let μ_2 be a mixture arising from T_2. For any $g \in K$ the mixture μ_2^g arises from T_2 and hence

$$(t')^T \mu_2 = \sum_{g \in K} (s^g)^T \mu_2 = \sum_{g \in K} s^T(\mu_2^g) > 0.$$

Similarly $(t')^T \mu_3 < 0$ for μ_3 arising from T_3 and hence t' is a linear test.

Now we show that t is a linear test as well. Let μ_2 arise from T_2. Note that $\mu_3 = \mu_2^{(1\,4)}$ arises from T_3 and hence

$$t^T \mu_2 = (t' - t'^{(14)})^T \mu_2 = (t')^T \mu_2 - (t')^T \mu_3 > 0.$$

Similarly $t(\mu_3) < 0$ for μ_3 arising from T_3 and hence t is a linear test.

Note that t is zero on orbits fixed by $(1\,4)$. On orbits o_1, o_2 swapped by $(1\,4)$ we have that t has opposite value (i. e., a on o_1, and $-a$ on o_2 for some a). Hence t is a linear combination of the ℓ_{o_1,o_2}. ■

4.1. A simple condition for a linear test

For the later proofs, it will be convenient to label the edges by matrices which are not allowed by the phylogenetic models. For example the identity matrix I (which corresponds to zero length edge) is an invalid transition matrix, i.e., $I \notin \mathcal{M}$, for the models considered in this paper.

The definition (1) is continuous in the entries of the matrices and hence for a weighting by matrices in $\mathrm{cl}(\mathcal{M})$ (the closure of $\mathcal{M}$) the generated distribution is arbitrarily close to a distribution generated from the model.

Observation 11. *A linear test for $\mathcal{M}$ (which is a strictly separating hyperplane for $\mathcal{M}$) is a separating hyperplane for* $\mathrm{cl}(\mathcal{M})$.

The above observation follows from the fact that if a continuous function $f : cl(A) \to \mathbb{R}$ is positive on some set A then it is non-negative on $\mathrm{cl}(A)$.

Suppose that the identity matrix $I \in \mathrm{cl}(\mathcal{M})$. Let μ arise from T_2 with weights such that the internal edge has weight I. Then μ_2 arises also from T_3 with the same weights. A linear test has to be positive for mixtures form T_2 and negative for mixtures from T_3. Hence we have:

Observation 12. *Let μ arise from T_2 with weights such that the internal edge has transition matrix I. Let t be a linear test. Then $t^T \mu = 0$.*

5. LINEAR TESTS FOR JC AND K2

In this section, we show a linear test for JC and K2 models. In fact, we show that the linear invariants introduced by Lake (1987) are linear tests. We expect that this fact is already known, but we include the proof for completeness and since it almost elementary given the preliminaries from the previous section.

5.1. A linear test for the Jukes-Cantor model

To simplify many of the upcoming expressions throughout the following section, we center the transition matrix for the Jukes-Cantor (JC) model around its stationary distribution in the following manner. Recall the JC model has $\Omega_{\mathrm{JC}} = \{0, 1, 2, 3\}$ and its semigroup $\mathcal{M}_{\mathrm{JC}}$ consists of matrices

$$M_{\mathrm{JC}}(\widehat{x}) = \frac{1}{4}E + \begin{pmatrix} 3\widehat{x} & -\widehat{x} & -\widehat{x} & -\widehat{x} \\ -\widehat{x} & 3\widehat{x} & -\widehat{x} & -\widehat{x} \\ -\widehat{x} & -\widehat{x} & 3\widehat{x} & -\widehat{x} \\ -\widehat{x} & -\widehat{x} & -\widehat{x} & 3\widehat{x} \end{pmatrix},$$

where E is the all ones matrix (i.e., $E(i,j) = 1$ for all $0 \le i, j < |\Omega|$) and $0 < \widehat{x} < 1/4$.

We refer to $\widehat{x}$ as the *centered edge weight*. Thus, a centered edge weight of $\widehat{x} = 1/4$ (which is not valid) means both endpoints have the same assignment. Whereas $\widehat{x} = 0$ (also not valid) means the endpoints are independent.

The group of symmetries of Ω_{JC} is $G_{\mathrm{JC}} = S_4$. There are 15 orbits in Ω^4_{JC} under the action of G_{JC} (each orbit has a representative in which i appears before j for any $i < j$). The action of K further decreases the number of orbits to 9. Here we list the 9 orbits and indicate which orbits are swapped by $(1\,4)$:

$$0000,\ 0110,\ 0123,\ \begin{matrix}0112\\0120\end{matrix},\ \begin{matrix}0111\\0100\\0010\end{matrix},\ 0011 \leftrightarrow 0101,\ \begin{matrix}0122\\0012\end{matrix} \leftrightarrow \begin{matrix}0121\\0102\end{matrix}. \tag{6}$$

$$0001$$

By Lemma 10, every linear test in the JC model is a linear combination of

$$\begin{aligned} t_1 &= \mu(0011) - \mu(0101), \text{ and} \\ t_2 &= \mu(0122) - \mu(0121) - \mu(0102) + \mu(0012). \end{aligned} \tag{7}$$

We will show that $t_2 - t_1$ is a linear test and that there exist no other linear tests (i. e., all linear tests are multiples of $t_2 - t_1$).

Lemma 13. *Let μ be a single-tree mixture arising from a tree T on 4 leaves. Let t be defined by*

$$\begin{aligned} t^T\mu = t_2 - t_1 = \mu(0122) - \mu(0121) + \mu(0101) \\ -\mu(0011) + \mu(0012) - \mu(0102). \end{aligned} \tag{8}$$

Let μ_i arise from T_i, for $i = 1, 2, 3$. We have

$$t^T\mu_1 = 0, \quad t^T\mu_2 > 0, \text{ and } t^T\mu_3 < 0.$$

In particular t is a linear test.

Proof. Label the 4 leaves as $v_1, \ldots, v_4$, and let $\widehat{x}_1, \ldots, \widehat{x}_4$ denote the centered edge weight of the edge incident to the respective leaf. Let $\widehat{x}_5$ denote the centered edge weight of the internal edge.

Let μ_j arise from T_j with centered edge weights $\widehat{x}_1, \ldots, \widehat{x}_5$, $j \in \{1, 2, 3\}$. Let Φ_j be the multi-linear polynomial $t^T\mu_j$. If $\widehat{x}_4 = 0$ then μ_j does not depend on the label of v_4 and hence, for all $a \in \Omega_{JC}$,

$$\Phi_j = \mu_j(012a) - \mu_j(012a) + \mu_j(010a) - \mu_j(001a) + \mu_j(001a) - \mu_j(010a) = 0.$$

Thus, $\widehat{x}_4$ divides Φ_j. The t_i are invariant under the action of K (which is transitive on $1, 2, 3, 4$) and hence Φ_j is invariant under the action of K. Hence $\widehat{x}_i$ divides Φ_j for $i = 1, \ldots, 4$. We have

$$\Phi_j = \widehat{x}_1\widehat{x}_2\widehat{x}_3\widehat{x}_4\ell(\widehat{x}_5),$$

where $\ell(\widehat{x}_5)$ is a linear polynomial in $\widehat{x}_5$.

Let μ_1' arise from T_1 with $\widehat{x}_1 = \cdots = \widehat{x}_4 = 1/4$. In leaf-labelings with non-zero probability in μ_1' the labels of v_1, v_4 agree and the labels of v_2, v_3 agree. None of the leaf-labelings in (8) satisfy this requirement and hence $\Phi_1 = 0$ if $\widehat{x}_1 = \cdots = \widehat{x}_4 = 1/4$. Hence $\ell(\widehat{x}_5)$ is the zero polynomial and Φ_1 is the zero polynomial as well.

Now we consider T_2. If $\widehat{x}_5 = 1/4$ then, by Observation 12, $\Phi_2 = 0$. Thus $1/4$ is a root of ℓ and hence $\Phi_2 = \alpha \cdot \widehat{x}_1\widehat{x}_2\widehat{x}_3\widehat{x}_4(1/4 - \widehat{x}_5)$. We plug in $\widehat{x}_5 = 0$ and $\widehat{x}_1 = \widehat{x}_2 = \widehat{x}_3 = \widehat{x}_4 = 1/4$ to determine α. Let μ_2' be the distribution generated by these weights. The leaf-labelings for which μ_2' is non-zero must have the same label for v_1, v_3 and the same label for v_2, v_4. Thus $\Phi_2 = \mu_2'(0101) = 1/16$ and hence $\alpha = 64$. We have

$$\Phi_2 = 64\widehat{x}_1\widehat{x}_2\widehat{x}_3\widehat{x}_4(1 - 4\widehat{x}_5).$$

Note that Φ_2 is always positive. The action of $(1\,4)$ switches the signs of the t_i and hence $\Phi_3 = -\Phi_2$. Thus Φ_3 is always negative. ■

We now show uniqueness of the above linear test, i.e., any other linear test is a multiple of $t_2 - t_1$.

Lemma 14. *Any linear test in the JC model is a multiple of (8).*

Proof. Let $t = \alpha_1 t_1 + \alpha_2 t_2$ be a linear test. Let μ_1 be the distribution generated by centered weights $\widehat{x}_2 = \widehat{x}_4 = \widehat{x}_5 = 1/4$ and $\widehat{x}_1 = \widehat{x}_3 = 0$ on T_2. By Observation 12, we must have $t^T \mu_1 = 0$. Note that

$$\mu_1(a_1a_2a_3a_4) = \begin{cases} 1/64 & \text{if } a_2 = a_4, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$t^T \mu_1 = -\alpha_1 \mu_1(0101) - \alpha_2 \mu(0121) = -1/64(\alpha_1 + \alpha_2) = 0.$$

Thus $\alpha_1 = -\alpha_2$ and hence t is a multiple of (8). ■

5.2. A linear test for Kimura's two-parameter model

Mutations between two purines (A and G) or between two pyrimidines (C or T) are more likely than mutations between a purine and a pyrimidine. Kimura's two-parameter model (K2) tries to model this fact.

We once again center the transition matrices to simplify the calculations. The K2 model has $\Omega_{\mathrm{K2}} = \{0, 1, 2, 3\}$ and its semigroup $\mathcal{M}_{\mathrm{K2}}$ consists of matrices

$$M_{\mathrm{K2}}(\widehat{x}, \widehat{y}) = \frac{1}{4}E + \begin{pmatrix} \widehat{x}+2\widehat{y} & -\widehat{x} & -\widehat{y} & -\widehat{y} \\ -\widehat{x} & \widehat{x}+2\widehat{y} & -\widehat{y} & -\widehat{y} \\ -\widehat{y} & -\widehat{y} & \widehat{x}+2\widehat{y} & -\widehat{x} \\ -\widehat{y} & -\widehat{y} & -\widehat{x} & \widehat{x}+2\widehat{y} \end{pmatrix},$$

with $\widehat{x} \le \widehat{y} < 1/4$ and $\widehat{x} + \widehat{y} > 0$. See Felsenstein (2004) for closed form of the transition matrices of the model in terms of the times t_e and rate matrices R_e. One can then derive the equivalence of the conditions there with the conditions $\widehat{x} \le \widehat{y} < 1/4, \widehat{x} + \widehat{y} > 0$.

Note, $\widehat{x}$ can be negative, and hence certain transitions can have probability $> 1/4$ but are always $< 1/2$. Observe that $M_{\mathrm{K2}}(\widehat{x}, \widehat{x}) = M_{\mathrm{JC}}(\widehat{x})$, i.e., the JC model is a special case of the K2 model.

The group of symmetries is $G_{\mathrm{K2}} = \langle (01), (02)(13) \rangle$ (it has 8 elements). There are 36 orbits in Ω^4 under the action of G_{K2} (each orbit has a representative in which 0 appears first and 2 appears before 3). The action of K further decreases the number of orbits to 18. The following orbits are fixed by $(1\,4)$:

$$0000, 0110, 0220, 0231, \begin{matrix} 0221 \\ 0230 \end{matrix}, \begin{matrix} 0111 & 0222 & 0223 \\ 0100 & 0200 & 0210 \\ 0010 & 0020 & 0120 \\ 0001 & 0002 & 0112 \end{matrix}.$$

The following orbits are swapped by $(1\,4)$:

$$0011 \leftrightarrow 0101, 0022 \leftrightarrow 0202, 0123 \leftrightarrow 0213, \begin{matrix} 0122 \\ 0023 \end{matrix} \leftrightarrow \begin{matrix} 0212 \\ 0203 \end{matrix}, \begin{matrix} 0233 & 0232 \\ 0211 & 0121 \\ 0021 & 0201 \\ 0012 & 0102 \end{matrix} \leftrightarrow \begin{matrix} 0232 \\ 0121 \\ 0201 \\ 0102 \end{matrix}.$$

By Lemma 10, any linear test for the K2 model is a linear combination of

$$\begin{aligned} t_1 &= \mu(0011) - \mu(0101), \\ t_2 &= \mu(0233) + \mu(0211) - \mu(0232) - \mu(0201) \\ &\quad -\mu(0121) + \mu(0021) - \mu(0102) + \mu(0012), \\ t_3 &= \mu(0022) - \mu(0202), \\ t_4 &= \mu(0122) + \mu(0023) - \mu(0212) - \mu(0203), \\ t_5 &= \mu(0123) - \mu(0213) \end{aligned}$$

Lemma 15. *Let μ be a single-tree mixture arising from a tree T on 4 leaves. Let t be defined by*

$$t^T\mu = \mu(0122) - \mu(0212) + \mu(0202) - \mu(0022) + \mu(0023) - \mu(0203) + \mu(0213) - \mu(0123). \tag{9}$$

Let μ_i arise from T_i, for $i = 1, 2, 3$. We have

$$t^T\mu_1 = 0, \quad t^T\mu_2 > 0, \text{ and } t^T\mu_3 < 0.$$

In particular t is a linear test.

Proof. Let $T = T_j$ for some $j \in \{1, 2, 3\}$. Let the transition matrix of the edge incident to leaf v_i be $M_{\mathrm{K2}}(\widehat{x}_i, \widehat{y}_i)$, and the internal edge has $M_{\mathrm{K2}}(\widehat{x}_5, \widehat{y}_5)$. Let μ_j be the generated distribution, and let Φ_j be the multi-linear polynomial $t^T\mu_j$.

If $\widehat{y}_4 = -\widehat{x}_4$ then the matrix on the edge incident to leaf v_4 has the last two columns the same. Hence roughly speaking this edge forgets the distinction between labels 2 and 3, and therefore in (9), we can do the following replacements:

$$0122 \to 0123$$
$$0202 \to 0203$$
$$0022 \to 0023$$
$$0213 \to 0212,$$

and we obtain,

$$\Phi_j = 0. \tag{10}$$

Thus $\widehat{x}_4 - \widehat{y}_4$ divides Φ_j. Since Φ_j is invariant under the action of K we have that $\widehat{x}_i - \widehat{y}_i$ divides Φ_j for $i = 1, \ldots, 4$ and hence

$$\Phi_j = (\widehat{x}_1 - \widehat{y}_1)\ldots(\widehat{x}_4 - \widehat{y}_4)\ell_j(\widehat{x}_5, \widehat{y}_5), \tag{11}$$

where $\ell_j(\widehat{x}_5, \widehat{y}_5)$ is linear in $\widehat{x}_5$ and $\widehat{y}_5$.

Now let $\widehat{x}_i = \widehat{y}_i = 1/4$ for $i = 1, \ldots, 4$. The label of the internal vertices v_j for $j = 5, 6$ must agree with the labels of neighboring leaves and hence

$$\Phi_j = \mu_j(0202) - \mu_j(0022). \tag{12}$$

Now plugging $j = 1$ into (11) and (12), for this setting of $\widehat{x}_i, \widehat{y}_i$, we have

$$\Phi_1 = 0 \tag{13}$$

By plugging $j = 2, 3$ into (11) and (12) we have

$$\Phi_2 = -\Phi_3 = (\widehat{x}_1 - \widehat{y}_1)(\widehat{x}_2 - \widehat{y}_2)(\widehat{x}_3 - \widehat{y}_3)(\widehat{x}_4 - \widehat{y}_4)(1 - 4\widehat{y}_5). \tag{14}$$

Note that $\widehat{x}_i - \widehat{y}_i < 0$ and $\widehat{y}_i < 1/4$ and hence (14) is always positive. Linearity of the test $\mu \mapsto t^T\mu$ implies that $t^T\mu$ is positive for any mixture generated from T_2 and negative for any mixture generated from T_3. ■

Lemma 16. *Any linear test in the K2 model is a multiple of (9).*

Proof. Let $t = \alpha_1 t_1 + \cdots + \alpha_5 t_5$ be a linear test. A linear test in the K2 model must work for JC model as well. Applying symmetries G_{JC} we obtain

$$\begin{aligned} t = \alpha_1(\mu(0011) - \mu(0101)) + 2\alpha_2(\mu(0122) - \mu(0121) - \mu(0102) + \mu(0012)) \\ +\alpha_3(\mu(0011) - \mu(0101)) + \alpha_4(\mu(0122) + \mu(0012) - \mu(0121) - \mu(0102)). \end{aligned} \tag{15}$$

Comparing (15) with (8) we obtain

$$\alpha_1 + \alpha_3 = -1 \text{ and } \alpha_4 + 2\alpha_2 = 1. \tag{16}$$

Let μ_1 arise from T_2 with centered weights $(\widehat{x}_2, \widehat{y}_2) = (\widehat{x}_4, \widehat{y}_4) = (\widehat{x}_5, \widehat{y}_5) = (1/4, 1/4)$, $(\widehat{x}_1, \widehat{y}_1) = (1/4, 0)$, and $(\widehat{x}_3, \widehat{y}_3) = (\widehat{x}, \widehat{y})$. From observation 12, it follows that $\Phi_{\mu_1} = 0$. The leaf-labelings with non-zero probability must give the same label to v_2 and v_4, and the labels of v_1 and v_2 must either be both in $\{0, 1\}$ or both in $\{2, 3\}$. The only such leaf-labelings involved in $t_1, \ldots, t_5$ are $0101, 0121$. Thus

$$t^T \mu_1 = -\alpha_1 \mu_1(0101) - \alpha_2 \mu_1(0121) = -\frac{1}{16}(\alpha_1 \widehat{x} + \alpha_2 \widehat{y}) = 0. \tag{17}$$

Thus, $\alpha_1 = \alpha_2 = 0$ and from (16) we get $\alpha_3 = -1$ and $\alpha_4 = 1$.

Let μ_2 be generated from T_2 with centered weights $(\widehat{x}_4, \widehat{y}_4) = (\widehat{x}_5, \widehat{y}_5) = (1/4, 1/4)$, $(\widehat{x}_1, \widehat{y}_2) = (\widehat{x}_3, \widehat{y}_3) = (0, 0)$, and $(\widehat{x}_2, \widehat{y}_2) = (1/4, 0)$. In leaf-labelings with non-zero probability the labels of v_2 and v_4 are either both in $\{0, 1\}$ or both in $\{2, 3\}$. The only such leaf labelings in t_3, t_4, t_5 are 0202,0213,0212, and 0203. Hence

$$t^T \mu_3 = \mu_3(0202) - \mu_3(0212) - \mu_3(0203) - \alpha_5 \mu_3(0213) = \frac{1}{256}(3 - 3 - 1 - \alpha_5) = 0.$$

Thus, $\alpha_5 = -1$ and all the α_i are determined. Hence, the linear test is unique (up to scalar multiplication). ■

6. NON-IDENTIFIABILITY AND LINEAR TESTS IN CFN

In this section, we consider the CFN model. We first prove there is no linear test, and then we present a non-identifiable mixture distribution. We then show that there is a linear test for the CFN model when the edge probabilities are restricted to some interval.

6.1. No linear test for CFN

Again, when considering linear tests we look at the model with its transition matrix centered around its stationary distribution. The CFN model has $\Omega_{\mathrm{CFN}} = \{0, 1\}$, and its semigroup $\mathcal{M}_{\mathrm{CFN}}$ consists of matrices

$$M_{\mathrm{CFN}}(\widehat{x}) = \frac{1}{2}E + \begin{pmatrix} \widehat{x} & -\widehat{x} \\ -\widehat{x} & \widehat{x} \end{pmatrix}$$

with $0 < \widehat{x} < 1/2$.

In the CFN model, note that the roles of 0 and 1 are symmetric, i. e.,

$$\mu_{T,\overrightarrow{P}}(a_1, \ldots, a_n) = \mu_{T,\overrightarrow{P}}(1 - a_1, \ldots, 1 - a_n). \tag{18}$$

Hence, the group of symmetries of Ω_{CFN} is $G_{\mathrm{CFN}} = \langle (01) \rangle = \mathbb{Z}/(2\mathbb{Z})$. There are 8 orbits of the action of G_{CFN} on Ω^4_{CFN} (one can choose a representative for each orbit to have the first coordinate 0). The action of K further reduces the number of orbits to 5. The action of $(1\,4)$ swaps two of the orbits and keeps 3 of the orbits fixed:

$$0000, 0110, \begin{matrix} 0111 \\ 0100 \\ 0010 \\ 0001 \end{matrix}, 0011 \leftrightarrow 0101.$$

By Lemma 10, if there exists a linear test for CFN then (a multiple of) $t_1 = \mu(0011) - \mu(0101)$ is a linear test. Let μ arise from T_2 with the edge incident to leaf v_i labeled by $M_{\mathrm{CFN}}(\widehat{x}_i)$, for $i = 1, \ldots, 4$, and the internal edge labeled by $M_{\mathrm{CFN}}(\widehat{x}_5)$. A short calculation yields

$$\mu(0011) - \mu(0101) = \widehat{x}_5(\widehat{x}_1\widehat{x}_3 + \widehat{x}_2\widehat{x}_4) - \frac{\widehat{x}_1\widehat{x}_2 + \widehat{x}_3\widehat{x}_4}{2}. \tag{19}$$

Note that (19) is negative if $\widehat{x}_5$ is much smaller than the other $\widehat{x}_i$; and (19) is positive if $\widehat{x}_1, \widehat{x}_3$ are much smaller than the other $\widehat{x}_i$. Thus, t_1 is not a linear test and hence there does not exist a linear test in the CFN model. By Theorem 4, there exists a non-identifiable mixture. The next result gives an explicit family of non-identifiable mixtures.

6.2. Non-identifiable mixture for CFN

For each edge e, we will give the edge probability $0 < w_e < 1/2$, which is the probability the endpoints receive different assignments (i.e., it is the off-diagonal entry in the transition matrix. For a 4-leaf tree T, we specify a set of transition matrices for the edges by a 5-dimensional vector $\overrightarrow{P} = (w_1, w_2, w_3, w_4, w_5)$ where, for $1 \leq i \leq 4$, w_i is the edge probability for the edge incident to leaf labeled i, and w_5 is the edge probability for the internal edge.

Proposition 17. *For $0 < a, b < 1/2$ and $0 < p \leq 1/2$, set*

$$\overrightarrow{P_1} = \frac{1}{2}\mathbf{1} - (a, b, a, b, c) \text{ and}$$
$$\overrightarrow{P_2} = \frac{1}{2}\mathbf{1} - (b, a, b, a, d),$$

where $\mathbf{1} = (1, 1, 1, 1, 1)$ and

$$c = z/p,$$
$$d = z/(1-p), \text{ and}$$
$$z = \frac{ab}{2(a^2 + b^2)}.$$

Let

$$\mu = p\mu_{T_3, \overrightarrow{P_1}} + (1-p)\mu_{T_3, \overrightarrow{P_2}}.$$

The distribution μ is invariant under $\pi = (14)$. Hence, μ is also generated by a mixture from T^π, a leaf-labeled tree different from $T = T_3$. In particular, the following holds:

$$\mu = p\mu_{T_2, \overrightarrow{P_1}} + (1-p)\mu_{T_2, \overrightarrow{P_2}}.$$

Hence, whenever c and d satisfy $0 < c, d < 1/2$ then μ is in fact a distribution and there is non-identifiability. Note, for every $0 < p \leq 1/2$, there exist a and b which define a non-identifiable mixture distribution.

Proof. Note that $\pi = (1\,4)$ fixes leaf labels $0000, 0010, 0100, 0110$ and swaps 0011 with 0101 and 0001 with 0111.
A short calculation yields

$$\mu(0011) - \mu(0101) = ab - (a^2 + b^2)(pc + (1-p)d), \text{ and}$$
$$\mu(0001) - \mu(0111) = (a^2 - b^2)(pc - (1-p)d).$$

which are both zero for our choice of c and d. This implies that μ is invariant under the action of $(1\,4)$, and hence non-identifiable. ■

6.3. *Linear test for CFN with restricted weights*

Lemma 18. *Let $a \in (0, 1/2)$. If the centered edge weight $\widehat{x}$ for the CFN model is restricted to the interval $(a, \sqrt{a-a^2})$ then there is a linear test.*

Proof. We will show that (19) is positive if the $\widehat{x}_i$ are in the interval $(a, \sqrt{a-a^2})$. Let $b = \sqrt{a-a^2})$. Note that $0 < a < b < 1/2$.

Since (19) is multi-linear, its extrema occur when the $\widehat{x}_i$ are from the set $\{a, b\}$ (we call such a setting of the $\widehat{x}_i$ extremal). Note that the $\widehat{x}_i$ are positive and $\widehat{x}_5$ occurs only in terms with negative sign. Thus a minimum occurs for $\widehat{x}_5 = b$. The only extremal settings of the $\widehat{x}_i$ which have $\widehat{x}_1\widehat{x}_3 + \widehat{x}_2\widehat{x}_4 > \widehat{x}_1\widehat{x}_2 + \widehat{x}_3\widehat{x}_4$ are $\widehat{x}_1 = \widehat{x}_3 = b, \widehat{x}_2 = \widehat{x}_4 = a$ and $\widehat{x}_1 = \widehat{x}_3 = a, \widehat{x}_2 = \widehat{x}_4 = b$. For the other extremal settings (19) is positive, since $b < 1/2$. For $\widehat{x}_1 = \widehat{x}_3 = b, \widehat{x}_2 = \widehat{x}_4 = a$ the value of (19) is $b(a - (a^2 + b^2))$. ■

Remark 19. *In contrast to the above lemma, it is known that there is no linear invariant for the CFN model. This implies that there is also no linear invariant for the restricted CFN model considered above, since such an invariant would then extend to the general model. This shows that the notion of linear test is more useful in some settings than linear invariants.*

7. NON-IDENTIFIABILITY IN K3

In this section, we prove there exists a non-identifiable mixture distribution in the K3 model. Our result holds even when the rate matrix is the same for all edges in the tree (the edges differ only by their associated time), i.e., the common rate matrix framework. Morevoer, we will show that for most rate matrices R in the K3 model there exists a non-identifiable mixture in which all transition matrices are generated from R.

The Kimura's three-parameter model (K3) has $\Omega_{\mathrm{K2}} = \{0, 1, 2, 3\}$ and its semigroup $\mathcal{M}_{\mathrm{K3}}$ consists of matrices of the following form (which we have centered around their stationary distribution):

$$M_{\mathrm{K3}}(\widehat{x}, \widehat{y}, \widehat{z}) = \frac{1}{4}E + \begin{pmatrix} \widehat{x}+\widehat{y}+\widehat{z} & -\widehat{x} & -\widehat{y} & -\widehat{z} \\ -\widehat{x} & \widehat{x}+\widehat{y}+\widehat{z} & -\widehat{z} & -\widehat{y} \\ -\widehat{y} & -\widehat{z} & \widehat{x}+\widehat{y}+\widehat{z} & -\widehat{x} \\ -\widehat{z} & -\widehat{y} & -\widehat{x} & \widehat{x}+\widehat{y}+\widehat{z} \end{pmatrix},$$

with $\widehat{x} \le \widehat{y} \le \widehat{z} < 1/4$, $\widehat{x} + \widehat{y} > 0$, and $(\widehat{x} + \widehat{y}) > 2(\widehat{x} + \widehat{z})(\widehat{y} + \widehat{z})$. Note that $M_{\mathrm{K3}}(\widehat{x}, \widehat{y}, \widehat{y}) = M_{\mathrm{K2}}(\widehat{x}, \widehat{y})$, i. e., the K2 model is a special case of the K3 model.

The group of symmetries is $G_{\mathrm{K3}} = \langle (01)(23), (02)(13) \rangle$ (which is again the Klein group). There are 64 orbits in Ω^4 under the action of G_{K3} (each orbit has a representative in which 0 appears first). The action of K further decreases the number of orbits to 28. The following orbits are fixed by $(1\,4)$:

$$0000,\ 0110,\ 0220,\ 0330,\ \begin{matrix} 0331 & 0332 & 0223 & 0111 & 0222 & 0333 \\ 0320 & 0310 & 0210 & 0100 & 0200 & 0300 \\ 0230\text{'} & 0130\text{'} & 0120\text{'} & 0010\text{'} & 0020\text{'} & 0030 \\ 0221 & 0112 & 0113 & 0001 & 0002 & 0003 \end{matrix}$$

The following orbits switch as indicated under the action of $(1\,4)$:

$$\begin{matrix} 0322 \\ 0311 \\ 0021 \\ 0012 \end{matrix} \leftrightarrow \begin{matrix} 0232 \\ 0201 \\ 0131\text{'} \\ 0102 \end{matrix} \begin{matrix} 0233 \\ 0211 \\ 0031 \\ 0013 \end{matrix} \leftrightarrow \begin{matrix} 0323 \\ 0301 \\ 0121\text{'} \\ 0103 \end{matrix} \begin{matrix} 0133 \\ 0122 \\ 0032 \\ 0023 \end{matrix} \leftrightarrow \begin{matrix} 0313 \\ 0302 \\ 0212 \\ 0203 \end{matrix}$$

and $0011 \leftrightarrow 0101, 0022 \leftrightarrow 0202, 0033 \leftrightarrow 0303, 0123 \leftrightarrow 0213, 0132 \leftrightarrow 0312, 0231 \leftrightarrow 0321$.

7.1. *No linear test for K3*

By Lemma 10, any test is a linear combination of

$$t_1 = \mu(0011) - \mu(0101),$$

$$
\begin{aligned}
t_2 &= \mu(0322) + \mu(0311) - \mu(0232) - \mu(0201) - \\
&\qquad \mu(0131) - \mu(0102) + \mu(0021) + \mu(0012), \\
t_3 &= \mu(0233) + \mu(0211) - \mu(0323) - \mu(0301) \\
&\qquad + \mu(0031) + \mu(0013) - \mu(0121) - \mu(0103), \\
t_4 &= \mu(0022) - \mu(0202), \\
t_5 &= \mu(0133) + \mu(0122) + \mu(0032) + \mu(0023) \\
&\qquad - \mu(0313) - \mu(0302) - \mu(0212) - \mu(0203), \\
t_6 &= \mu(0033) - \mu(0303), \\
t_7 &= \mu(0123) - \mu(0213), \\
t_8 &= \mu(0132) - \mu(0312), \\
t_9 &= \mu(0231) - \mu(0321).
\end{aligned}
$$

We first present a non-constructive proof of non-identifiability by proving that there does not exist a linear test, and then Theorem 4, implies there exists a non-identifiable mixture. We then prove the stronger result where the rate matrix is fixed.

Lemma 20. *There does not exist a linear test for the K3 model.*

Corollary 21. *There exists a non-identifiable mixture in the K3 model.*

Proof of Lemma 20. Suppose that $t = \alpha_1 t_1 + \cdots + \alpha_9 t_9$ is a test. Let $\widehat{w}_i = (\widehat{x}_i, \widehat{y}_i, \widehat{z}_i)$, $1 \leq i \leq 5$. For $1 \leq i \leq 4$, $\widehat{w}_i$ denotes the centered parameters for the edge incident to leaf i, and $\widehat{w}_5$ are the centered parameters for the internal edge.

In the definitions of μ_1 and μ_2 below, we will set $\widehat{w}_2 = \widehat{w}_4 = \widehat{w}_5 = (0, 0, 0)$. This ensures that in labelings with non-zero probability, leaves v_2, v_4 and both internal vertices all have the same label. Moreover, by observation 12, $\Phi(\mu_i) = 0$.

Let μ_1 be generated from T_2 with $\widehat{w}_1 = (1/4, \widehat{y}_1, \widehat{z}_1)$, and $\widehat{w}_3 = (1/4, 0, 0)$. In labelings with non-zero probability, the labels of v_2 and v_3 have to both be in $\{0, 1\}$ or both in $\{2, 3\}$. The only labels in $t_1, \ldots, t_9$ with this property are $0101, 0232, 0323$. Thus,

$$
t^T \mu_1 = -\alpha_1 \mu(0101) - \alpha_2 \mu_1(0232) - \alpha_3 \mu_1(0323) = -\frac{1}{16}(\alpha_1/4 + \alpha_2 \widehat{y}_1 + \alpha_3 \widehat{z}_1) = 0. \tag{20}
$$

Any $\widehat{y}, \widehat{z} \in [1/8 - \varepsilon, 1/8 + \varepsilon]$ with $y' \geq z'$ gives a valid matrix in the K3 model. For (20) to be always zero we must have $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (since $1, \widehat{y}_1, \widehat{z}_1$ are linearly independent polynomials).

Let μ_2 be generated from T_2 with $\widehat{w}_1 = (1/4, \widehat{y}_1, \widehat{z}_1)$, and $\widehat{w}_3 = (1/4, \widehat{y}, \widehat{z})$. The only labels in $t_4, \ldots, t_9$ with v_2 and v_4 having the same label are

$$
0202, 0313, 0212, 0303
$$

(we ignore the labels in t_1, t_2, t_3 because $\alpha_1 = \alpha_2 = \alpha_3 = 0$). Thus

$$
\begin{aligned}
t^T \mu_1 &= -\alpha_4 \mu_2(0202) - \alpha_5 \mu_2(0313) - \alpha_5 \mu_2(0212) - \alpha_6 \mu_2(0303) \\
&= -\frac{1}{4}(\alpha_4 (\widehat{y}_1)^2 + 2\alpha_5 \widehat{y}_1 \widehat{z}_1 + \alpha_6 (\widehat{z}_1)^2) = 0.
\end{aligned} \tag{21}
$$

Polynomials $(\widehat{y}_1)^2, \widehat{y}_1 \widehat{z}_1, (\widehat{z}_1)^2$ are linearly independent and hence $\alpha_4 = \alpha_5 = \alpha_6 = 0$.

Thus, t is a linear combination of t_7, t_8, t_9 and hence has at most 6 terms. A test for K3 must be a test for K2, but the unique test for K2 has 8 terms. Thus, Φ cannot be a test for K3. ■

7.2. *Non-identifiable mixtures in the K3 model with a fixed rate matrix*

We will now consider rate matrices for the model, as opposed to the transition probability matrices. The rate matrix for the K3 model is

$$R = R(\alpha, \beta, \gamma) = \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & -\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & -\alpha-\beta-\gamma \end{pmatrix}, \tag{22}$$

where the rates usually satisfy $\alpha \geq \beta \geq \gamma > 0$. For our examples we will also assume that $\alpha, \beta, \gamma \in [0, 1]$. Since the examples are negative they work immediately for the above weaker constraint.

Recall that the K2 model is a submodel of the K3 model: the rate matrices of the K2 model are precisely the rate matrices of K3 model with $\beta = \gamma$. By Lemma 15, there exists a test in the K2 model and hence there are no non-identifiable mixtures. We will show that the existence of a test in the K2 model is a rather singular event: for almost all rate matrices in the K3 model, there exist non-identifiable mixtures and hence no test.

We show the following result.

Lemma 22. *Let α, β, γ be chosen independently from the uniform distribution on $[0, 1]$. With probability 1 (over the choice of α, β, γ) there does not exist a test for the K3 model with the rate matrix $R(\alpha, \beta, \gamma)$.*

To prove Lemma 7.2, we need the following technical concept. A *generalized polynomial* is a function of the form

$$\sum_{i=1}^{m} a_i(u_1, \ldots, u_n) e^{b_i(u_1, \ldots, u_n)}, \tag{23}$$

where the a_i are non-zero polynomials and the b_i are distinct linear polynomials. Note that the set of generalized polynomials is closed under addition, multiplication, and taking derivatives. Thus, for example, the Wronskian of a set of generalized polynomials (with respect to one of the variables, say u_1) is a generalized polynomial.

For $n = 1$, we have the following bound on the number of roots of a generalized polynomial (see Pólya and Szegȍ, 1976, part V, problem 75):

Lemma 23. *Let $G = \sum_{i=1}^{m} a_i(u) e^{b_i(u)}$ be a univariate generalized polynomial. Assume $m > 0$. Then G has at most*

$$(m-1) + \sum_{i=1}^{m} \deg a_i(u)$$

real roots, where $\deg a_i(u)$ is the degree of the polynomial $a_i(u)$.

Corollary 24. *Let $G(u_1, \ldots, u_n) = \sum_{i=1}^{m} a_i(u_1, \ldots, u_n) e^{b_i(u_1, \ldots, u_n)}$ be a generalized polynomial. Assume $m > 0$ (i. e., that G is not the zero polynomial). Let $r_1, \ldots, r_n$ be picked in independently and randomly from uniform distribution on $[0, 1]$. Then with probability 1 (over the choice of $r_1, \ldots, r_n$) we have $G(r_1, \ldots, r_n) \neq 0$.*

Proof of Corollary 24. We will proceed by induction on n. The base case $n = 1$ follows from Lemma 23 since the probability of a finite set is zero.

For the induction step, consider the polynomial a_1. Since a_1 is a non-zero polynomial there are only finitely many choices of $c \in [0, 1]$ for which $a_1(c, u_2, \ldots, u_n)$ is the zero polynomial. Thus with probability 1 over the choice of $u_1 = c$ we have that $a_1(c, u_2, \ldots, u_n)$ is a non-zero polynomial. Now, by the induction

hypothesis, with probability 1 over the choice of $u_2, \ldots, u_n$ we have that $G(c, u_2, \ldots, u_n) \neq 0$. Hence, with probability 1 over the choice of $u_1, \ldots, u_n$, we have that $G(u_1, \ldots, u_n) \neq 0$. ■

Proof of Lemma 22. Any test has to be a linear combination of $t_1, \ldots, t_9$ defined in Section 7. Suppose that $t = \sigma_1 t_1 + \ldots \sigma_9 t_9$ is a test. We will use Observation 12 to show that $\sigma_1 = \cdots = \sigma_9 = 0$ and hence t cannot be a test. Let $\sigma = (\sigma_1, \ldots, \sigma_9)$.

The transition matrix of the process with rate matrix R, at time s is $T(s) = \exp(sR)$. We have (Felsenstein, 2004)

$$T(s) = M_{K3} = \frac{1}{4}E + \begin{pmatrix} A+B+C & -A & -B & -C \\ -A & A+B+C & -C & -B \\ -B & -C & A+B+C & -A \\ -C & -B & -A & A+B+C \end{pmatrix}, \tag{24}$$

where

$$A = (1 + e^{-2s(\alpha+\beta)} + e^{-2s(\alpha+\gamma)} - e^{-2s(\beta+\gamma)})/4,$$
$$B = (1 + e^{-2s(\alpha+\beta)} - e^{-2s(\alpha+\gamma)} + e^{-2s(\beta+\gamma)})/4,$$
$$C = (1 - e^{-2s(\alpha+\beta)} + e^{-2s(\alpha+\gamma)} + e^{-2s(\beta+\gamma)})/4.$$

We have changed notation from our earlier definition of M_{K3} by exchanging A, B, C for $\widehat{x}, \widehat{y}, \widehat{z}$ to indicate these as functions of α, β and γ.

Let μ_x be generated from T_3 with edge weights $T(x), T(2x), T(3x), T(4x), T(0)$. The internal edge has length 0 (i. e., it is labeled by the identity matrix), and hence, by Observation 12, we must have $t^T \mu_x = 0$ for all $x \geq 0$.

Let $f_i(x) = t_i^T \mu_x$ for $i = 1, \ldots, 9$. Let $W(x)$ be the Wronskian matrix of $f_1(x), \ldots, f_9(x)$ with respect to x. The entries of $W(x)$ are generalized polynomials in variables α, β, γ, and x. Thus, the Wronskian (i. e., $\det W(x)$) is a generalized polynomial in variables α, β, γ, and x.

Now we show that for a particular choice of α, β, γ and x we have $\det W(x) \neq 0$. Let $\alpha = \pi i$, $\beta = 2\pi i$, $\gamma = -2\pi i$, and $x = 1$. Of course complex rates are not valid in the K3 model, we use them only to establish that $W(x)$ is a non-zero generalized polynomial. Tedious computation (best performed using a computer algebra system) yields that the Wronskian matrix is the following. The first seven columns are as follows:

$$\begin{pmatrix} 8 & -16 & -16 & 8 & 16 & 8 & 0 \\ 40i\pi & -240i\pi & 80i\pi & 56i\pi & 80i\pi & 24i\pi & 48i\pi \\ -1052\pi^2 & 3736\pi^2 & 856\pi^2 & -780\pi^2 & -824\pi^2 & -300\pi^2 & 112\pi^2 \\ -12244i\pi^3 & 52920i\pi^3 & -4520i\pi^3 & -13988i\pi^3 & -2600i\pi^3 & 732i\pi^3 & -4224i\pi^3 \\ 206468\pi^4 & -923656\pi^4 & -134536\pi^4 & 187188\pi^4 & 46856\pi^4 & 9108\pi^4 & 1520\pi^4 \\ 3199060i\pi^5 & -12454200i\pi^5 & -280600i\pi^5 & 3574316i\pi^5 & 103400i\pi^5 & -354276i\pi^5 & 576168i\pi^5 \\ -46789172\pi^6 & 230396776\pi^6 & 25447336\pi^6 & -47807340\pi^6 & -2740904\pi^6 & 1992900\pi^6 & -2781848\pi^6 \\ -826613044i\pi^7 & 3116686680i\pi^7 & 184720120i\pi^7 & -908923988i\pi^7 & -4488200i\pi^7 & 82589532i\pi^7 & -98067024i\pi^7 \\ 11742908228\pi^8 & -58371817096\pi^8 & -5047384456\pi^8 & 12445876068\pi^8 & 163759496\pi^8 & -776675292\pi^8 & 840993440\pi^8 \end{pmatrix}$$

The last two columns are as follows:

$$\begin{pmatrix} 0 & 0 \\ -48i\pi & -96i\pi \\ 912\pi^2 & 800\pi^2 \\ 14880i\pi^3 & 17568i\pi^3 \\ -212400\pi^4 & -198560\pi^4 \\ -3692808i\pi^5 & -4100016i\pi^5 \\ 49984152\pi^6 & 51191600\pi^6 \\ 919913280i\pi^7 & 1003352208i\pi^7 \\ -12613821600\pi^8 & -13323533120\pi^8 \end{pmatrix}.$$

The determinant of $W(x)$ is $339201508906183707450958527237980160000000\pi^{36}$, which is non-zero. Thus, $\det W(x)$ is a non-zero generalized polynomial and hence for random α, β, γ and x we have that $\det W(x)$ is non-zero with probability 1.

Assume now that $\det W(x) \neq 0$. Let $w = W(x)\sigma$. The first entry w_1 of w is given by $w_1 = t^T \mu_x$. Since t is a test we have $w_1 = 0$. The second entry w_2 of w is given by $w_2 = (\partial/\partial x) t^T \mu_x$. Again we must have $w_2 = 0$, since t is a test. Similarly we show $w_3 = \cdots = w_9 = 0$. Note that $W(x)$ is a regular matrix and hence σ must be the zero vector. Thus t is not a test. Since $\det W(x) \neq 0$ happens with probability 1 we have that there exists no test with probability 1. ■

8. MAXIMUM LIKELIHOOD RESULTS

Here we prove Theorem 2.

8.1. Terminology

In the CFN and JC model there is a single parameter defining the transition matrix for each edge, namely a parameter x_e where $0 < x_e < 1/|\Omega|$. In K2, there are 2 parameters and in K3 there are 3 parameters. We use the term weights $\overrightarrow{w}_e$ to denote the setting of the parameters defining the transition matrix for the edge e. And then we let $\overrightarrow{w} = (\overrightarrow{w}_e)_{e \in T}$ denote the set of vectors defining the transition matrices for edges of tree T.

We use the term zero weight edge to denote the setting of the weights so that the transition matrix is the identity matrix I. Thus, in the CFN and JC models, this corresponds to setting $x_e = 0$. We refer to a non-zero weight edge as a setting of the parameters so that all entries of the transition matrix are positive. Hence, in the CFN and JC models this corresponds to $x_e > 0$.

Note, $\mathcal{L}_{\mathcal{T}}(\mu)$ is maximized over the set of weights $\overrightarrow{w}$. Hence our consideration of weights $\overrightarrow{w}$. There are typically constraints on $\overrightarrow{w}$ in order to define valid transition matrices. For example, in the CFN model we have $\overrightarrow{w}_e = x_e$ where $0 < x_e < 1/2$, and similarly $0 < x_e < 1/4$ in the JC model. In the K2 model we have $\overrightarrow{w}_e = (x, y)$ where $x \geq y > 0, x + y < 1/2$. Finally, in the K3 model, we have further constraints as detailed in Section 7.

8.2. Technical tools

Our proof begins starts from the following observation. Consider the CFN model. Note that for $x = 0$ we have $\overrightarrow{P_1} = \overrightarrow{P_2}'$ and hence $\mu_0 = \mu_{T_3, \overrightarrow{v}}$ where $\overrightarrow{v} = (1/2, 1/2, 1/2, 1/2, 0)$ are the weights for the CFN model, i.e., μ_0 is generated by a pure distribution from T_3. In fact it can be generated by a pure distribution from any leaf-labeled tree on 4 leaves since the internal edges have zero length.

Observation 25. *Consider a tree T on n leaves and weights w where all internal edges have zero weight. For all trees $S \neq T$ on n leaves, there is a unique weight $\overrightarrow{v}$ such that $\mu_{S,\overrightarrow{v}} = \mu_{T,\overrightarrow{w}}$.*

Proof. Let $\overrightarrow{v}$ have the same weight as $\overrightarrow{w}$ for all terminal edges, and zero weight for all internal edges. Note, we then have $\mu_{S,\overrightarrow{v}} = \mu_{T,\overrightarrow{w}}$ and it remains to prove uniqueness of $\overrightarrow{v}$.

Let $\overrightarrow{u}$ be a set of weights where $\mu_{S,\overrightarrow{u}} = \mu_{S,\overrightarrow{w}}$. Let S' be obtained from S by contracting all the edges of zero weight in $\overrightarrow{u}$, and let $\overrightarrow{u'}$ be the resulting set of weights for the remaining edges. The tree S' has all internal vertices of degree ≥ 3 and internal edges have non-zero weight.

It follows from the work of Buneman (1971) that $S', \overrightarrow{u'}$ is unique among trees with non-zero internal edge weights and without vertices of degree two. If $\overrightarrow{u} = \overrightarrow{v}$, then S' is a star, and hence every $\overrightarrow{u}$ must contract to a star. This is only possible if $\overrightarrow{u}$ assigns zero weight to all internal edges. Therefore, $\overrightarrow{v}$ is the unique weight. ■

For w and w' defined for the CFN model in Theorem 2, for any 4-leaf tree S, the maximum of $\mathcal{L}_S(\mu_0)$ is achieved on $v = (1/4, 1/4, 1/4, 1/4, 0)$. For any $v' \neq v$ the distribution $\mu_{S,v'}$ is different from μ_0 and hence $\mathcal{L}_{S,v'}(\mu_0) < \mathcal{L}_{S,v}(\mu_0) = \mu_0^T \ln \mu_0$. Intuitively, for small x the maximum of $\mathcal{L}_S(\mu_x)$ should be realized on a v'' which is near v. Now we formalize this argument in the following lemma. Then in Lemma 28 we will use the Hessian and Jacobian of the expected log-likelihood functions to bound $\mathcal{L}_S(\mu_x)$ in terms of $\mathcal{L}_S(\mu_0)$.

Lemma 26. *Let μ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n leaves. Suppose that there exists a unique v in the closure of the model such that $\mu_{S,v} = \mu$. Then for every $\delta > 0$ there exists $\epsilon > 0$ such that for any μ' with $||\mu' - \mu|| \leq \epsilon$ the global optima of $\mathcal{L}_S(\mu')$ are attained on v' for which $||v' - v|| \leq \delta$.*

Remark 27. *In our application of the above lemma we have $\mu = \mu_0$. Consider a tree S and its unique weight v where $\mu_{S,v} = \mu_0$. Note, the requirement that every element in $\{0, 1\}^n$ has non-zero probability, is*

satisfied if the terminal edges have non-zero weights. In contrast, the internal edges have zero weight so that Observation 25 applies, and in some sense the tree is achievable on every topology.

Proof. We will prove the lemma by contradiction. Roughly speaking, we now suppose that there exists μ' close to μ where the maximum of μ' is achieved far from the maximum of μ. Formally, suppose that there exists $\delta > 0$ and sequences μ'_i and v'_i such that $\lim_{i\to\infty} ||\mu - \mu'_i|| = 0$ and $||v'_i - v|| > \delta$ where v'_i is a weight for which the optimum of $\mathcal{L}_S(\mu'_i)$ is attained. By the optimality of the v'_i we have

$$\mathcal{L}_{S,v'_i}(\mu'_i) \geq \mathcal{L}_{S,v}(\mu'_i). \tag{25}$$

We assumed $\mu_{S,v} = \mu$ and hence the entries of $\ln \mu_{S,v}$ are finite since we assumed μ has positive entries. Thus

$$\lim_{i\to\infty} \mathcal{L}_{S,v}(\mu'_i) = \lim_{i\to\infty} ({\mu'}_i^T \ln \mu_{S,v}) = \mu^T \ln \mu_{S,v} = \mathcal{L}_{S,v}(\mu). \tag{26}$$

Take a subsequence of the v'_i which converges to some v'. Note, v' is in the closure of the model.

Let ϵ be the smallest entry of μ. For all sufficiently large i, μ'_i has all entries $\geq \epsilon/2$.

Hence,

$$\mathcal{L}_{S,v'_i}(\mu'_i) = \sum_{z\in\Omega^n} \mu'_i(z) \ln(\mu_{S,v'_i}(z)) \leq \frac{\epsilon}{2} \min_{z\in\Omega^n} \ln(\mu_{S,v'_i}(z)) \tag{27}$$

Because of (26), for all sufficiently large i, $\mathcal{L}_{S,v}(\mu'_i) \geq 2\mathcal{L}_{S,v}(\mu)$ (recall that the log-likelihoods are negative). Combining with (27), we have that the entries of $\ln \mu_{S,v'_i}$ are bounded from below by $\mathcal{L}_{S,v}(\mu)/4\epsilon$. Thus, both μ'_i and $\ln \mu_{S,v'_i}$ are bounded. For bounded and convergent sequences,

$$\lim_{n\to\infty} a_n b_n = \lim_{n\to\infty} a_n \lim_{n\to\infty} b_n$$

Therefore,

$$\lim_{i\to\infty} \mathcal{L}_{S,v'_i}(\mu'_i) = (\lim_{i\to\infty} {\mu'}_i^T)(\lim_{i\to\infty} \ln \mu_{S,v'_i}) = \mu^T \ln \mu_{S,v'} = \mathcal{L}_{S,v'}(\mu). \tag{28}$$

From (25), (26), and (28), we have $\mathcal{L}_{S,v'}(\mu) \geq \mathcal{L}_{S,v}(\mu)$. Since $v' \neq v$ we get a contradiction with the uniqueness of v. ■

We now bound the difference of $\mathcal{L}_S(\mu')$ and $\mathcal{L}_S(\mu)$ when the previous lemma applies. This will then imply that for x sufficiently small, $\mathcal{L}_S(\mu_x)$ is close to $\mathcal{L}_S(\mu_0)$.

Here is the formal statement of the lemma.

Lemma 28. *Let $\mu \in \mathbb{R}^{|\Omega|^n}$ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n vertices. Suppose that there exists $\overrightarrow{v}$ in the closure of the model such that $\mu_{S,\overrightarrow{v}} = \mu$ and that $\overrightarrow{v}$ is the unique such weight. Let $\Delta\mu_x \in \mathbb{R}^{|\Omega|^n}$ be such that $\Delta\mu_x^T 1 = 0$, and $x \mapsto \mu_x$ is continuous around 0 in the following sense: $\Delta\mu_x \to \Delta\mu_0$ as $x \to 0$.*

Let $g(\overrightarrow{w}) = \mathcal{L}_{S,\overrightarrow{w}}(\mu)$, and $h_x(\overrightarrow{w}) = (\Delta\mu_x)^T \ln \mu_{S,\overrightarrow{w}}$. Let H be the Hessian of g at $\overrightarrow{v}$ and J_x be the Jacobian of h_x at $\overrightarrow{v}$. Assume that H has full rank. Then

$$\mathcal{L}_S(\mu + x\Delta\mu_x) \leq \mu^T \ln \mu + x h_x(\overrightarrow{v}) - \frac{x^2}{2} J_0 H^{-1} J_0^T + o(x^2). \tag{29}$$

Moreover, if $(H^{-1}J^T)_i \leq 0$ for all i such that $\overrightarrow{v}_i = 0$ then the inequality in (29) can be replaced by equality.

Remark 29. *When $(H^{-1}J^T)_i < 0$ for all i such that $v_i = 0$ then the likelihood is maximized at non-trivial branch lengths. In particular, for the CFN model, the branch lengths are in the interval $(0, 1/2)$, that is, there are no branches of length 0 or $1/2$. Similarly for the Jukes-Cantor model the lengths are in $(0, 1/4)$.*

Proof. For notational convenience, let $f(\overrightarrow{w}) = \ln \mu_{S,\overrightarrow{w}}$. Thus, $g(\overrightarrow{w}) = \mu^T f(\overrightarrow{w})$. Note that

$$f(\overrightarrow{v}) = \ln \mu. \tag{30}$$

The function f maps assignments of weights for the $2n-3$ edges of S to the logarithm of the distribution induced on the leaves. Hence, the domain of f is the closure of the model, which is a subspace of $\mathbb{R}^{d(2n-3)}$, where d is the dimension of the parameterization of the model, e.g., in the CFN and JC models $d = 1$ and in the K2 model $d = 2$. We denote the closure of the model as Λ. Note that the range of f is $[-\infty, 0]^{|\Omega|^n}$.

Let $J_f = (\partial f_i/\partial w_j)$ be the Jacobian of f and $H_f = (\partial f_i/\partial w_j \partial w_k)$ be the Hessian of f (H_f is a rank 3 tensor).

If $\overrightarrow{v}$ is in the interior of Λ, then since $\overrightarrow{w}$ optimizes $\mathcal{L}_{S,\overrightarrow{v}}(\mu)$ we have

$$\mu^T J_f(\overrightarrow{v}) = 0. \tag{31}$$

We will now argue that equality (31) remains true even when $\overrightarrow{v}$ is on the boundary of Λ. Function $\mu_{S,\overrightarrow{v}}$ is a polynomial function of the coordinates of $\overrightarrow{v}$. The coordinates of $\mu_{S,\overrightarrow{v}}$ sum to the constant 1 function. We assumed that $\mu = \mu_{S,\overrightarrow{v}}$ is strictly positive and hence for all $\overrightarrow{w}$ in a small neighborhood of $\overrightarrow{v}$ we have that $\mu_{S,\overrightarrow{w}}$ is still a distribution (note that $\overrightarrow{w}$ can be outside of Λ). Suppose that $\mu^T J_f(\overrightarrow{v}) \neq 0$. Then there exists $\overrightarrow{\Delta v}$ such that $\mu^T \ln \mu_{S,\overrightarrow{v}+\overrightarrow{\Delta v}} > \mu^T \ln \mu_{S,\overrightarrow{v}} = \mu^T \ln \mu$. This contradicts the fact that the maximum (over all distributions ν) of $\mu^T \ln(\nu)$ is achieved at $\nu = \mu$. Thus, (31) holds.
If we perturb μ by $x\Delta\mu_x$ and $\overrightarrow{v}$ by $\overrightarrow{\Delta v}$ the likelihood changes as follows:

$$\begin{aligned}
&\mathcal{L}_{T,\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\Delta\mu_x) \\
&= (\mu + x\Delta\mu_x)^T f(v + \Delta v) \\
&= (\mu + x\Delta\mu_x)^T \left(f(\overrightarrow{v}) + (J_f(\overrightarrow{v}))\,(\overrightarrow{\Delta v}) + \frac{1}{2}(\overrightarrow{\Delta v})^T\,(H_f(\overrightarrow{v}))\,(\overrightarrow{\Delta v}) + O(||\overrightarrow{\Delta v}||^3) \right) \\
&= \mu^T \ln \mu + \mu^T \left((J_f(\overrightarrow{v}))\,(\overrightarrow{\Delta v}) + \frac{1}{2}(\overrightarrow{\Delta v})^T H_f(\overrightarrow{v})(\overrightarrow{\Delta v}) \right) \\
&\quad + x(\Delta\mu_x)^T \left(f(\overrightarrow{v}) + J_f(\overrightarrow{v})(\overrightarrow{\Delta v}) \right) + O\left(||\overrightarrow{\Delta v}||^3 + x||\overrightarrow{\Delta v}||^2 \right) \\
&= \mu^T \ln \mu + \mu^T \left(\frac{1}{2}(\Delta v)^T H_f(\overrightarrow{v})(\overrightarrow{\Delta v}) \right) + x(\Delta\mu_x)^T \left(f(\overrightarrow{v}) + J_f(\overrightarrow{v})(\overrightarrow{\Delta v}) \right) \\
&\quad + O\left(||\overrightarrow{\Delta v}||^3 + x||\overrightarrow{\Delta v}||^2 \right),
\end{aligned} \tag{32}$$

where in the third step we used (30) and in the last step we used (31). In terms of H, J defined in the statement of the theorem, we have

$$\begin{aligned}
\mathcal{L}_{T,\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + x h_x(\overrightarrow{v}) + \frac{1}{2}(\overrightarrow{\Delta v})^T H(\overrightarrow{\Delta v}) + x J_x(\overrightarrow{\Delta v}) \\
+ O\left(||\overrightarrow{\Delta v}||^3 + x||\overrightarrow{\Delta v}||^2 \right).
\end{aligned} \tag{33}$$

We will now prove that H is negative definite. First note that we assumed that H is full-rank, thus all of its eigenvalues are non-zero. Moreover, $g(\overrightarrow{w}) = \mathcal{L}_{S,\overrightarrow{w}}(\mu)$ is maximized for $\overrightarrow{w} = \overrightarrow{v}$. Plugging $x = 0$ into (33), we obtain

$$g(\overrightarrow{v} + \overrightarrow{\Delta v}) = \mu^T \ln \mu + \frac{1}{2}(\overrightarrow{\Delta v})^T H(\overrightarrow{\Delta v}) + O\left(||\overrightarrow{\Delta v}||^3 \right) \tag{34}$$

Hence, if H was not negative definite, then it would have at least one positive eigenvalue. Let $\overrightarrow{z}$ denote the corresponding eigenvector. Let $\overrightarrow{\Delta v}$ be a sufficiently small multiple of $\overrightarrow{z}$. By (34) we have $g(\overrightarrow{v} + \overrightarrow{\Delta v}) > g(\overrightarrow{v})$ which contradicts the earlier claim that $\overrightarrow{w} = \overrightarrow{v}$ maximizes $g(\overrightarrow{w})$. Hence, all of the eigenvalues of H are

negative, i.e., it is negative definite. Let $\lambda_{\max}$ be the largest eigenvalue of H, i.e., closest to zero. We will use that $\lambda_{\max} < 0$. Note,

$$(\overrightarrow{\Delta v})^T H(\overrightarrow{\Delta v}) \leq \lambda_{\max}||\overrightarrow{\Delta v}||^2 \tag{35}$$

Set $\overrightarrow{\Delta v}$ so that $\mathcal{L}_{\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\overrightarrow{\Delta}\mu_x)$ is maximized. ($\overrightarrow{\Delta v}$ is a function of x.) Now we will prove that $||\overrightarrow{\Delta v}|| = O(x)$. For any $\delta > 0$, by Lemma 26, for all sufficiently small x, we have $||\overrightarrow{\Delta v}|| \leq \delta$. By (33) and (35), we have

$$\mathcal{L}_{S,\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\Delta\mu_x) \leq \mu^T \ln \mu + xh_x(\overrightarrow{v}) + \lambda_{\max}||\overrightarrow{\Delta v}||^2 + xJ_x(\overrightarrow{\Delta v}) + O(||\overrightarrow{\Delta v}||^3 + x||\overrightarrow{\Delta v}||^2).$$

Assume $x = o(||\overrightarrow{\Delta v}||)$ (i.e., assume that $||\overrightarrow{\Delta v}||$ goes to zero more slowly than x). Then, we have

$$\mathcal{L}_{S,\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\Delta\mu) \leq \mu^T \ln \mu + xh_x(\overrightarrow{v}) + \lambda_{\max}||\overrightarrow{\Delta v}||^2 + o(||\overrightarrow{\Delta v}||^2).$$

On the other hand, we have

$$\mathcal{L}_{S,\overrightarrow{v}}(\mu + x\Delta\mu) = (\mu + x\Delta\mu)^T \ln \mu_{S,\overrightarrow{v}} = \mu^T \ln \mu + xh_x(\overrightarrow{v}).$$

Since $\lambda_{\max}||\overrightarrow{\Delta v}||^2$ is negative then for sufficiently small $||\overrightarrow{\Delta v}||$ (recall we can choose any $\delta > 0$ where $||\overrightarrow{\Delta v}|| \leq \delta$), we have

$$\mathcal{L}_{\overrightarrow{v}+\overrightarrow{\Delta v}}(\mu + x\Delta\mu) \leq \mathcal{L}_{\overrightarrow{v}}(\mu + x\Delta\mu).$$

Thus, we may restrict ourselves to $\overrightarrow{\Delta v}$ such that $||\overrightarrow{\Delta v}|| = O(x)$. Hence

$$\mathcal{L}_S(\mu + x\Delta\mu_x) \leq \mu^T \ln \mu + xh_x(\overrightarrow{v}) + \max_{\overrightarrow{\Delta w}} \left(\frac{1}{2}(\overrightarrow{\Delta w})^T H(\overrightarrow{\Delta w}) + xJ_x(\overrightarrow{\Delta w}) \right) + O(x^3). \tag{36}$$

The maximum of

$$\frac{1}{2}(\overrightarrow{\Delta w})^T H(\overrightarrow{\Delta w}) + xJ_x(\overrightarrow{\Delta w}) \tag{37}$$

occurs at $\overrightarrow{\Delta} z := -xH^{-1}J_x^T$; for this $\overrightarrow{\Delta z}$ the value of (37) is $-\frac{x^2}{2}J_xH^{-1}J_x^T$. Therefore,

$$\mathcal{L}_S(\mu + x\Delta\mu) \leq \mu^T \ln \mu + xh_x(\overrightarrow{v}) - \frac{x^2}{2}J_xH^{-1}J_x^T + O(x^3).$$

From $\mu_x \to \mu_0$, we have $J_xH^{-1}J_x^T = (1 + o(1))J_0H^{-1}J_0^T$, and hence

$$\mathcal{L}_S(\mu + x\Delta\mu) \leq \mu^T \ln \mu + xh_x(\overrightarrow{v}) - \frac{x^2}{2}J_0H^{-1}J_0^T + o(x^2).$$

This completes the proof of the first part of the lemma. It remains to prove the case when the inequality can be replaced by equality.

Note that in general the inequality cannot be replaced by equality, since $\overrightarrow{v} + x\overrightarrow{\Delta z}$ can be an invalid weight (i.e., outside of Λ) for all x. If $(\overrightarrow{\Delta v})_i \geq 0$ whenever $\overrightarrow{v}_i = 0$ then $\overrightarrow{v} + x\overrightarrow{\Delta z}$ is a valid weight vector for sufficiently small x and hence plugging directly into (33) we have

$$\begin{aligned} \mathcal{L}_S(\mu + x\Delta\mu_x) \geq \mu^T \ln \mu + xh_x(\overrightarrow{v}) - \frac{x^2}{2}J_xH^{-1}J_x^T + O(x^3) = \\ \mu^T \ln \mu + xh_x(\overrightarrow{v}) - \frac{x^2}{2}J_0H^{-1}J_0^T + o(x^2). \end{aligned} \tag{38}$$

■

8.3. Proof of theorem 2 in CFN: $C = 1/4$

In this section, we deal with the CFN model. We prove Part 1 of Theorem 2. For simplicity, we first present the proof for the case $C = 1/4$.

Let μ_1 be generated from T_3 with weights $\overrightarrow{P_1} = (1/4{+}x, 1/4{-}x, 1/4{-}x, 1/4{+}x, x^2)$ and μ_2 be generated from T_3 with weights $\overrightarrow{P_2} = (1/4{-}x, 1/4{+}x, 1/4{+}x, 1/4{-}x, x^2)$. Let $\mu = (\mu_1{+}\mu_2)/2$. Note that $(1\,4)(2\,3)$ fixes μ_1, and μ_2; and $(1\,2)(3\,4)$ swaps μ_1 and μ_2. Hence μ is invariant under $K = \langle (1\,2)(3\,4), (1\,4)(2\,3)\rangle$. This simplifies many of the following calculations.

One can verify that the Hessian is the same for all the trees:

$$H = \begin{pmatrix} \frac{-1552}{615} & \frac{-16}{41} & \frac{-16}{41} & \frac{-16}{41} & \frac{-80}{123} \\ \frac{-16}{41} & \frac{-1552}{615} & \frac{-16}{41} & \frac{-16}{41} & \frac{-80}{123} \\ \frac{-16}{41} & \frac{-16}{41} & \frac{-1552}{615} & \frac{-16}{41} & \frac{-80}{123} \\ \frac{-16}{41} & \frac{-16}{41} & \frac{-16}{41} & \frac{-1552}{615} & \frac{-80}{123} \\ \frac{-80}{123} & \frac{-80}{123} & \frac{-80}{123} & \frac{-80}{123} & \frac{-400}{369} \end{pmatrix}.$$

The above is straightforward to verify in any symbolic algebra system, such as Maple.

The Jacobians differ only in their last coordinate. For T_3 we have

$$J_0 = \left(\frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{-880}{369}\right),$$

Finally,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{36328}{1845} \approx 19.68997.$$

Hence, for $v = (1/4, 1/4, 1/4, 1/4, 0)$, by Lemma 28, we have

$$\begin{aligned} \mathcal{L}_S(\mu_x) &\le \mu^T \ln \mu + x h(v) - \frac{x^2}{2} J_0 H^{-1} J_0^T + O(x^3) \\ &= \mu^T \ln \mu + x h(v) + x^2 \frac{36328}{1845} + O(x^3) \end{aligned}$$

For T_2 we have

$$J = \left(\frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{-1208}{369}\right),$$

Then,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{244862}{9225} \approx 26.54331.$$

By Lemma 28, we have

$$\mathcal{L}_S(\mu_x) \le \mu^T \ln \mu + x h(v) + x^2 \frac{244862}{9225} + O(x^3)$$

For T_1 we have

$$J_0 = \left(\frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{1744}{615}, \frac{4040}{369}\right),$$

$$-H^{-1} J_0 = \left(\frac{-7}{4}, \frac{-7}{4}, \frac{-7}{4}, \frac{-7}{4}, \frac{143}{10}\right). \tag{39}$$

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{126118}{1845} \approx 68.35664.$$

Note that the last coordinate of $-H^{-1}J_0$ (in (39)) is positive and hence we have equality in Lemma 28. Thus,

$$\mathcal{L}_S(\mu_x) = \mu^T \ln \mu + xh(v) + x^2\frac{126118}{1845} + O(x^3)$$

The largest increase in likelihood is attained on T_1. Thus T_1 is the tree with highest likelihood for sufficiently small x.

8.4. Proof of theorem 2 in CFN: arbitrary C

The Hessian is the same for all four-leaf trees. In this case, we state the inverse of the Hessian which is simpler to state than the Hessian itself. We have

$$H^{-1} = \begin{pmatrix} \frac{16C^4-1}{32C^2} & 0 & 0 & 0 & \frac{(4C^2-1)^2}{128C^3} \\ 0 & \frac{16C^4-1}{32C^2} & 0 & 0 & \frac{(4C^2-1)^2}{128C^3} \\ 0 & 0 & \frac{16C^4-1}{32C^2} & 0 & \frac{(4C^2-1)^2}{128C^3} \\ 0 & 0 & 0 & \frac{16C^4-1}{32C^2} & \frac{(4C^2-1)^2}{128C^3} \\ \frac{(4C^2-1)^2}{128C^3} & \frac{(4C^2-1)^2}{128C^3} & \frac{(4C^2-1)^2}{128C^3} & \frac{(4C^2-1)^2}{128C^3} & \frac{(16C^4-24C^2-3)(4C^2-1)^2}{256C^4(1+4C^2)^2} \end{pmatrix}$$

The first four coordinates of the Jacobians are equal and have the same value for all trees:

$$\frac{16C(64C^6 - 16C^4 + 12C^2 + 1)}{(16C^4 - 1)(16C^4 + 24C^2 + 1)}.$$

Thus, for each tree, we will only list the last coordinate of the Jacobian J_0 which we denote as $J_0[5]$.

Let

$$\beta = (4C^2 - 1)^2(16C^4 + 24C^2 + 1),$$
$$\gamma = \beta \cdot \frac{(1 + 4C^2)^2}{16}$$

Note that $\beta > 0$ and $\gamma > 0$ for $C \in (0, 1/2)$.

For T_1 we have

$$J_0[5] = \frac{1}{\beta} \cdot 128C^2(16C^6 - 24C^4 + 17C^2 + 1),$$

$$\Delta_1 := \frac{1}{2}J_0H^{-1}J_0 = \frac{1}{\gamma} \cdot (-512C^{12} - 2048C^{10} + 3520C^8 - 1856C^6 + 390C^4 + 88C^2 + 3),$$

and the last coordinate of $-H^{-1}J_0$ is

$$L := \frac{48C^6 - 40C^4 + 15C^2 + 2}{2C^2(1 + 4C^2)^2}.$$

It is easily checked that L is positive for $C \in (0, 1/2)$ and hence we have equality in Lemma 28.

For T_2 we have

$$J_0[5] = \frac{1}{\beta} \cdot 128C^4(16C^4 - 40C^2 - 7),$$

$$\Delta_2 := \frac{1}{2} J_0 H^{-1} J_0 = \frac{1}{\gamma} \cdot (-512C^{12} - 5120C^{10} + 960C^8 + 832C^6 + 198C^4 + 28C^2 + 1).$$

For T_3 we have

$$J_0[5] = \frac{1}{\beta} \cdot 256C^4(16C^4 - 8C^2 - 3),$$

$$\Delta_3 := \frac{1}{2} J_0 H^{-1} J_0 = \frac{1}{\gamma} \cdot (2048C^{12} - 2048C^{10} - 512C^8 + 72C^4 + 24C^2 + 1).$$

It remains to show that $\Delta_1 > \Delta_2$ and $\Delta_1 > \Delta_3$. We know that for $C = 1/4$ the inequalities hold. Thus we only need to check that $\Delta_1 - \Delta_2$ and $\Delta_1 - \Delta_3$ do not have roots for $C \in (0, 1/2)$. This is easily done using Sturm sequences, which is a standard approach for counting the number of roots of a polynomial in an interval.

8.5. Proof of theorem 2 in JC, K2, and K3

Our technique requires little additional work to extend the result to JC, K2, and K3 models. Let JC-likelihood of tree T on distribution μ be the maximal likelihood of $\mu_{T,w}$ over all labelings of w, in the JC model. Similarly we define K2-likelihood and K3-likelihood. Note that K3-likelihood of a tree is greater or equal to its K2-likelihood which is greater or equal to its JC-likelihood. In the following we will consider a mixture distribution generated from the JC model, and look at the likelihood (under a non-mixture) for JC, K2, and K3 models.

For the K3 model, the transition matrices are of the form

$$P_{\mathrm{K3}}(\alpha, \beta, \gamma) = \begin{pmatrix} 1-\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & 1-\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & 1-\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & 1-\alpha-\beta-\gamma \end{pmatrix},$$

with $\alpha \geq \beta \geq \gamma > 0$, $\alpha + \beta < 1/2$, and $\gamma > (\alpha+\gamma)(\beta+\gamma)$. The K2 model is the case $\beta = \gamma$, the JC model is the case $\alpha = \beta = \gamma$.

Theorem 30. *Let $\overrightarrow{P_1} = (1/8 + x, 1/8 - x, 1/8 - x, 1/8 + x, x^2)$ and $\overrightarrow{P_2} = (1/8 - x, 1/8 + x, 1/8 + x, 1/8 - x, x^2)$. Let μ_x denote the following mixture distribution on T_3 generated from the JC model:*

$$\mu_x = \left(\mu_{T_3,\overrightarrow{P_1}} + \mu_{T_3,\overrightarrow{P_2}}\right)/2.$$

There exists $x_0 > 0$ such that for all $x \in (0, x_0)$ the JC-likelihood of T_1 on μ_x is higher than the K3-likelihood of T_2 and T_3 on μ_x.

Note that Part 2 of Theorem 2 for the JC, K2, and K3 models is immediately implied by Theorem 30.

First we argue that in the JC model T_1 is the most likely tree. As in the case for the CFN model, because of symmetry, we have the same Hessian for all trees.

$$H = \begin{pmatrix} \frac{-2373504}{112255} & \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-587856}{112255} \\ \frac{-915872}{336765} & \frac{-2373504}{112255} & \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-587856}{112255} \\ \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-2373504}{112255} & \frac{-915872}{336765} & \frac{-587856}{112255} \\ \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-915872}{336765} & \frac{-2373504}{112255} & \frac{-587856}{112255} \\ \frac{-587856}{112255} & \frac{-587856}{112255} & \frac{-587856}{112255} & \frac{-587856}{112255} & \frac{-1130124}{112255} \end{pmatrix},$$

Again the Jacobians differ only in the last coordinate.

For T_3:

$$J_0 = \left(\frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{-7085428}{112255}\right), \tag{40}$$

Then,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{174062924259638}{237159005655} \approx 733.9503.$$

For T_2:

$$J_0 = \left(\frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{-8069818}{112255}\right), \tag{41}$$

Then,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{410113105846051}{474318011310} \approx 864.6374.$$

For T_1:

$$J_0 = \left(\frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{4199248}{112255}, \frac{22878022}{112255}\right), \tag{42}$$

$$-H^{-1} J_0 = \left(\frac{-10499073}{2816908}, \frac{-10499073}{2816908}, \frac{-10499073}{2816908}, \frac{-10499073}{2816908}, \frac{118305233}{4225362}\right). \tag{43}$$

Note, again the last coordinate is positive as required for the application of Lemma 28. Finally,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{1221030227753251}{474318011310} \approx 2574.286. \tag{44}$$

Now we bound the K3-likelihood of T_2 and T_3. The Hessian matrix is now 15×15. It is the same for all the 4-leaf trees and has a lot of symmetry. There are only 8 different entries in H. For distinct $i, j \in [4]$ we have

$$\frac{\partial^2}{\partial p_i \partial p_i} f(\mu_0) = -538996/112255,$$
$$\frac{\partial^2}{\partial p_i \partial p_j} f(\mu_0) = -605684/1010295,$$
$$\frac{\partial^2}{\partial p_i \partial p_5} f(\mu_0) = -132304/112255,$$

$$\frac{\partial^2}{\partial p_i \partial r_i} f(\mu_0) = -126086/112255,$$
$$\frac{\partial^2}{\partial p_i \partial r_j} f(\mu_0) = -51698/336765,$$
$$\frac{\partial^2}{\partial p_i \partial r_5} f(\mu_0) = -2448/8635,$$

$$\frac{\partial^2}{\partial r_i \partial r_i} f(\mu_0) = -268544/112255,$$
$$\frac{\partial^2}{\partial r_i \partial r_5} f(\mu_0) = -54082/112255.$$

For T_3, its Jacobian J_0 is a vector of 15 coordinates. It turns out that $3J_0$ is the concatenation of 3 copies of the Jacobian for the JC model which is stated in (40). Finally, we obtain

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{174062924259638}{237159005655} \approx 733.9503. \tag{45}$$

For T_2 we again obtain that for its Jacobian J_0, $3J_0$ is the concatenation of 3 copies of (41). Then,

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{4101131058460 51}{474318011310} \approx 864.6374. \tag{46}$$

Finally, for T_1, for its Jacobian J_0, $3J_0$ is the concatenation of 3 copies of (42) and $-H^{-1}J_0$ is the concatenation of 3 copies of (43). Then,

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{1221030227753251}{474318011310} \approx 2574.286.$$

Note the quantities $-\frac{1}{2}J_0H^{-1}J_0^T$ are the same in the K3 model are the same as the corresponding quantities in the JC model. It appears that even though the optimization is over the K3 parameters, the optimum assignment is a valid setting for the JC model.

Observe that $-\frac{1}{2}J_0H^{-1}J_0^T$ for T_1 in the JC model (see (44)) is larger than for T_2 and T_3 in the $K3$ model (see (46) and (45)). Applying Lemma 28, this completes the proof of Theorem 30.

9. MCMC RESULTS

The following section has a distinct perspective from the earlier sections. Here we are generating N samples from the distribution and looking at the complexity of reconstructing the phylogeny. The earlier sections analyzed properties of the generating distribution, as opposed to samples from the distribution. In addition, instead of finding the maximum likelihood tree, we are looking at sampling from the posterior distribution over trees. To do this, we consider a Markov chain whose stationary distribution is the posterior distribution, and analyze the chain's mixing time.

For a set of data $\overrightarrow{D} = (D_1, \ldots, D_N)$ where $D_i \in \{0,1\}^n$, its likelihood on tree T with transition matrices $\overrightarrow{P}$ is

$$\begin{aligned}
\mu_{T,\overrightarrow{P}}(\overrightarrow{D}) &= \mathbf{Pr}\left(\overrightarrow{D} \mid T, \overrightarrow{P}\right) \\
&= \prod_i \mathbf{Pr}\left(\overrightarrow{D}_i \mid T, \overrightarrow{P}\right) \\
&= \prod_i \mu_{T,\overrightarrow{P}}(\overrightarrow{D}_i) \\
&= \exp(\sum_i \ln(\mu_{T,\overrightarrow{P}}(\overrightarrow{D}_i))
\end{aligned}$$

Let $\Phi(T, \overrightarrow{P})$ denote a prior density on the space of trees where

$$\sum_T \int_{\overrightarrow{P}} \Phi(T, \overrightarrow{P}) d\overrightarrow{P} = 1.$$

Our results extend to priors that are lower bounded by some ϵ as in Mossel and Vigoda (2006). In particular, for all T, $\overrightarrow{P}$, we require $\Phi(T, \overrightarrow{P}) \geq \epsilon$. We refer to these priors as ϵ-regular priors.

Applying Bayes law we get the posterior distribution:

$$\mathbf{Pr}\left(T, \overrightarrow{P} \mid \overrightarrow{D}\right) = \frac{\mathbf{Pr}\left(\overrightarrow{D} \mid T, \overrightarrow{P}\right)\Phi(T, \overrightarrow{P})}{\mathbf{Pr}\left(\overrightarrow{D}\right)}$$

$$= \frac{\mathbf{Pr}\left(\overrightarrow{D} \mid T, \overrightarrow{P}\right)\Phi(T, \overrightarrow{P})}{\sum_{T'} \int_{\overrightarrow{P}'} \mathbf{Pr}\left(\overrightarrow{D} \mid T', \overrightarrow{P}'\right)\Phi(T', \overrightarrow{P}')d\overrightarrow{P}'}$$

Note that for uniform priors the posterior probability of a tree given $\overrightarrow{D}$ is proportional to $\mathbf{Pr}\left(\overrightarrow{D} \mid T\right)$. Each tree T then has a posterior weight

$$w(T) = \int_{\overrightarrow{P}} \mathbf{Pr}\left(\overrightarrow{D} \mid T, \overrightarrow{P}\right)\Phi(T, \overrightarrow{P})d\overrightarrow{P}.$$

We look at Markov chains on the space of trees where the stationary distribution of a tree is its posterior probability. We consider Markov chains using nearest-neighbor interchanges (NNI). The transitions modify the topology in the following manner (Fig. 3).

Let S_t denote the tree at time t. The transition $S_t \to S_{t+1}$ of the Markov chain is defined as follows. Choose a random internal edge $e = (u, v)$ in S. Internal vertices have degree three, thus let a, b denote the other neighbors of u and y, z denote the other neighbors of v. There are three possible assignments for these 4 subtrees to the edge e (namely, we need to define a pairing between a, b, y, z). Choose one of these assignments at random, denote the new tree as S'. We then set $S_{t+1} = S'$ with probability

$$\min\{1, w(S')/w(S_t)\}.\} \tag{47}$$

With the remaining probability, we set $S_{t+1} = S_t$.

The acceptance probability in (47) is known as the Metropolis filter and implies that the unique stationary distribution π of the Markov chain statisfies, for all trees T:

$$\pi(T) = \frac{w(T)}{\sum_{T'} w(T')}.$$

We refer readers to Felsenstein (2004) and Mossel and Vigoda (2006) for a more detailed introduction to this Markov chain. We are interested in the mixing time T_{mix}, defined as the number of steps until the chain is within variation distance $\leq 1/4$ of the stationary distribution. The constant $1/4$ is somewhat arbitrary, and can be reduced to any δ with $T_{\text{mix}} \log(1/\delta)$ steps.

For the MCMC result, we consider trees on 5 taxa. Thus trees in this section will have leaves numbered $1, 2, 3, 4, 5$ and internal vertices numbered $6, 7, 8$. Let S_3 denote the tree $(((12), 5), (34))$. Thus, S_3 has edges $e_1 = \{1, 6\}$, $e_2 = \{2, 6\}$, $e_3 = \{3, 8\}$, $e_4 = \{4, 8\}$, $e_5 = \{5, 7\}$, $e_6 = \{6, 7\}$, $e_7 = \{7, 8\}$. We will list

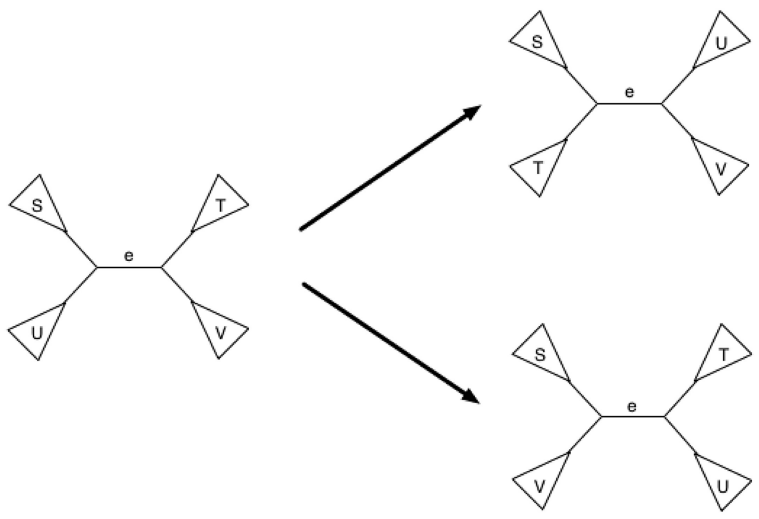


FIG . 3. Illustration of NNI transitions where the subtrees adjacent to internal edge e are re-arranged.

the transition probabilities for the edges of S_3 in this order. For the CFN model, we consider the following vector of transition probabilities. Let

$$\overrightarrow{P_1} = (1/4 + x, 1/4 - x, 1/4 - x, 1/4 + x, 1/4, c, c), \text{ and}$$
$$\overrightarrow{P_2} = (1/4 - x, 1/4 + x, 1/4 + x, 1/4 - x, 1/4, c, c),$$

where

$$c = \frac{1}{2}\left(1 - \sqrt{\frac{1 - 16x^2}{1 + 16x^2}}\right).$$

For the JC model, let

$$\overrightarrow{P_1} = (1/8 + x, 1/8 - x, 1/8 - x, 1/8 + x, 1/8, c', c'), \text{ and}$$
$$\overrightarrow{P_2} = (1/8 - x, 1/8 + x, 1/8 + x, 1/8 - x, 1/8, c', c'),$$

where

$$c' = 16x^2.$$

Let $\mu_1 = \mu_{S_1, \overrightarrow{P_1}}$ and $\mu_2 = \mu_{S_1, \overrightarrow{P_2}}$. We are interested in the mixture distribution:

$$\mu = \frac{1}{2}(\mu_1 + \mu_2)$$

Let S_2 denote the tree $(((14), 5), (23))$.

The key lemma for our Markov chain result states that under μ, the likelihood has local maximum, with respect to NNI connectivity, on S_1 and S_2.

Lemma 31. *For the CFN and JC models, there exists $x_0 > 0$ such that for all $x \in (0, x_0)$ then for all trees S that are one NNI transition from S_1 or S_2, we have*

$$\mathcal{L}_S(\mu) < \mathcal{L}_{S_1}(\mu), \quad \mathcal{L}_S(\mu) < \mathcal{L}_{S_2}(\mu)$$

This then implies the following corollary.

Theorem 32. *There exists a constant $C > 0$ such that for all $\epsilon > 0$ the following holds. Consider a data set with N characters, i.e., $\overrightarrow{D} = (D_1, \ldots, D_N)$, chosen independently from the distribution μ. Consider the Markov chains on tree topologies defined by nearest-neighbor interchanges (NNI). Then with probability $1 - \exp(-CN)$ over the data generated, the mixing time of the Markov chains, with priors which are ϵ-regular, satisfies*

$$T_{mix} \geq \epsilon \exp(CN).$$

The novel aspect of this section is Lemma 31. The proof of Theorem 32 using Lemma 31 is straightforward.

Proof of Lemma 31. The proof follows the same lines as the proof of Theorems 2 and 30. Thus our main task is to compute the Hessian and Jacobians, for which we utilize Maple.

We begin with the CFN model.

The Hessian is the same for all 15 trees on 5-leaves:

$$H = \begin{pmatrix}
\frac{-3880}{1281} & \frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-1241}{1281} & \frac{-190}{427} \\
\frac{-114}{427} & \frac{-3880}{1281} & \frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-1241}{1281} & \frac{-190}{427} \\
\frac{-114}{427} & \frac{-114}{427} & \frac{-3880}{1281} & \frac{-114}{427} & \frac{-114}{427} & \frac{-190}{427} & \frac{-1241}{1281} \\
\frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-3880}{1281} & \frac{-114}{427} & \frac{-190}{427} & \frac{-1241}{1281} \\
\frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-114}{427} & \frac{-3880}{1281} & \frac{-190}{427} & \frac{-190}{427} \\
\frac{-1241}{1281} & \frac{-1241}{1281} & \frac{-190}{427} & \frac{-190}{427} & \frac{-190}{427} & \frac{-6205}{3843} & \frac{-950}{1281} \\
\frac{-190}{427} & \frac{-190}{427} & \frac{-1241}{1281} & \frac{-1241}{1281} & \frac{-190}{427} & \frac{-950}{1281} & \frac{-6205}{3843}
\end{pmatrix},$$

We begin with the two trees of interest:

and

Their Jacobian is

$$J_0 = \left(\frac{17056}{1281}, \frac{17056}{1281}, \frac{17056}{1281}, \frac{17056}{1281}, \frac{2432}{427}, \frac{57952}{3843}, \frac{57952}{3843}\right),$$

Thus,

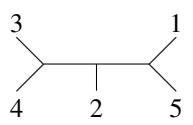
$$-H^{-1}J_0 = (2, 2, 2, 2, 0, 4, 4)\,.$$

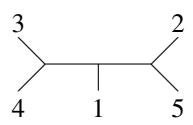
Note the last two coordinates are positive, hence we get equality in the conclusion of Lemma 28. Finally,

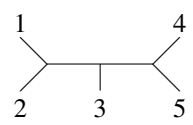
$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{436480}{3843} \approx 113.5779.$$

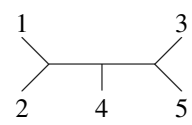
We now consider those trees connected to S_1 and S_2 by one NNI transition. Since there are 2 internal edges each tree has 4 NNI neighbors.

The neighbors of S_1 are

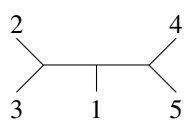


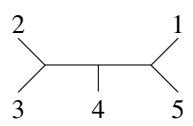


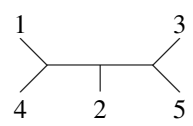


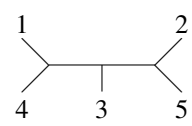


The neighbors of S_2 are









The Jacobian for all 8 of these trees is

$$J_0 = \left(\frac{17056}{1281}, \frac{17056}{1281}, \frac{17056}{1281}, \frac{2432}{427}, \frac{17056}{1281}, \frac{57952}{3843}, \frac{3728}{427}\right),$$

Finally,

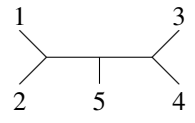
$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{2242633984}{20840589} \approx 107.6090,$$

Note the quantities $-\frac{1}{2}J_0H^{-1}J_0^T$ are larger for the two trees S_1 and S_2. This completes the proof for the CFN model.

We now consider the JC model. Again the Hessian is the same for all the trees:

$$H = \begin{pmatrix}
\frac{-512325018}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-328636353}{41082370} & \frac{-28145979}{8216474} \\
\frac{-36668964}{20541185} & \frac{-512325018}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-328636353}{41082370} & \frac{-28145979}{8216474} \\
\frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-512325018}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-28145979}{8216474} & \frac{-328636353}{41082370} \\
\frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-512325018}{20541185} & \frac{-36668964}{20541185} & \frac{-28145979}{8216474} & \frac{-328636353}{41082370} \\
\frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-36668964}{20541185} & \frac{-512325018}{20541185} & \frac{-28145979}{8216474} & \frac{-28145979}{8216474} \\
\frac{-328636353}{41082370} & \frac{-328636353}{41082370} & \frac{-28145979}{8216474} & \frac{-28145979}{8216474} & \frac{-28145979}{8216474} & \frac{-1273864167}{82164740} & \frac{-134747901}{20541185} \\
\frac{-28145979}{8216474} & \frac{-28145979}{8216474} & \frac{-328636353}{41082370} & \frac{-328636353}{41082370} & \frac{-28145979}{8216474} & \frac{-134747901}{20541185} & \frac{-1273864167}{82164740}
\end{pmatrix}$$

Beginning with tree S_1



We have:

$$J_0 = \left(\frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{1733695536}{20541185}, \frac{5655197244}{20541185}, \frac{5655197244}{20541185}\right),$$

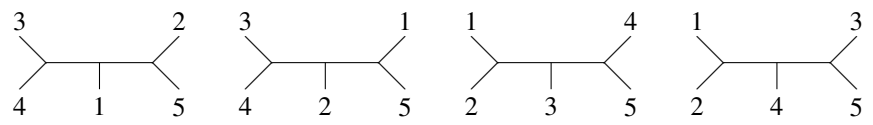
The last two coordinates of $-H^{-1}J_0$ are

$$\frac{5114490004637540016}{593018923302763639}, \frac{5114490004637540016}{593018923302763639}.$$

Since they are positive we get equality in the conclusion of Lemma 28. Finally,

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{4810147291155537042880499 1552}{12181311412062878919972215} \approx 3948.793$$

For the neighbors of S_1:



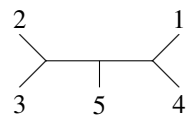
We have:

$$J_0 = \left(\frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{1733695536}{20541185}, \frac{4342624176}{20541185}, \frac{5655197244}{20541185}, \frac{2955839412}{20541185}\right).$$

Hence,

$$-\frac{1}{2}J_0H_1^{-1}J_0^T = \frac{5672580483610108356983726382127006164356508 0096}{15568481282665727860752794372508821870798435} \approx 3643.631$$

Considering S_2:



$$J_0 = \left(\frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{1733695536}{20541185}, \frac{1074039432}{4108237}, \frac{1074039432}{4108237}\right),$$

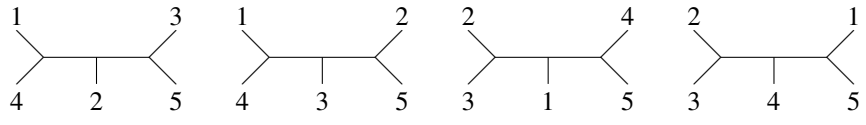
The last two coordinates of $-H^{-1}J_0$ are

$$\frac{13458396942990580792}{1779056769908290917}, \frac{13458396942990580792}{1779056769908290917},$$

which are positive. Finally,

$$-\frac{1}{2}J_0H^{-1}J_0^T = \frac{4536529474419729155571536 8032}{12181311412062878919972215} \approx 3724.172$$

Considering the neighbors of S_2:



We have:

$$J_0 = \left(\frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{4342624176}{20541185}, \frac{1733695536}{20541185}, \frac{4342624176}{20541185}, \frac{1074039432}{4108237}, \frac{2955839412}{20541185}\right)$$

Hence,

$$-\frac{1}{2} J_0 H^{-1} J_0^T = \frac{77561493674724211429726298715535057559621128 08}{22240687546665325515361134817869745529712 05} \approx 3487.369$$

■

Remark 33. *In fact S_1 and S_2 have larger likelihood than any of the 13 other 5-leaf trees. However, analyzing the likelihood for the 5 trees not considered in the proof of Lemma 31 requires more technical work since $-\frac{1}{2} J_0 H^{-1} J_0^T$ is maximized at invalid weights for these 5 trees.*

We now show how the main theorem of this section easily follows from the above lemma. The proof follows the same basic line of argument as in Mossel and Vigoda (2005), we point out the few minor differences.

Proof of Theorem 32. For a set of characters $\overrightarrow{D} = (D_1, \ldots, D_N)$, define the maximum log-likelihood of tree T as

$$\mathcal{L}_T(\overrightarrow{D}) = \max_{\overrightarrow{P} \in \mathcal{M}^E} \mathcal{L}_{T,\overrightarrow{P}}(\overrightarrow{D}),$$

where

$$\mathcal{L}_{T,\overrightarrow{P}}(\overrightarrow{D}) = \sum_i \ln\left(\mathbf{Pr}\left(D_i \mid T, \overrightarrow{P}\right)\right).$$

Consider $\overrightarrow{D} = (D_1, \ldots, D_N)$ where each D_i is independently sampled from the mixture distribution μ. Let S^* be S_1 or S_2, and let S be a tree that is one NNI transition from S^*. Our main task is to show that $\mathcal{L}_{S^*}(\overrightarrow{D}) > \mathcal{L}_S(\overrightarrow{D})$.

Let $\overrightarrow{P^*}$ denote the assignment which attains the maximum for $\mathcal{L}_{S^*}(\mu)$, and let

$$\alpha = \min_{\sigma \in \Omega^5} \mathbf{Pr}\left(\sigma \mid S^*, \overrightarrow{P^*}\right).$$

For $\sigma \in \Omega^5$, let $D(\sigma) = |\{i : D_i = \sigma\}|$. By Chernoff's inequality (Janson et al., 2000) and a union bound over $\sigma \in \Omega^5$, we have for all $\delta > 0$,

$$\mathbf{Pr}\left(\text{ for all } \sigma \in \Omega^5, |D(\sigma) - \mu(\sigma)N| \leq \delta N\right) \geq 1 - 2 \cdot 4^5 \exp(-2\delta^2 N) = 1 - \exp(-\Omega(N)).$$

Assuming the above holds, we then have

$$\mathcal{L}_S(\overrightarrow{D}) \leq N(1-\delta)\mathcal{L}_S(\mu).$$

And,

$$\mathcal{L}_{S^*}(\overrightarrow{D}) \geq \mathcal{L}_{S^*,\overrightarrow{P^*}}(\overrightarrow{D}) \geq N(1-\delta)\mathcal{L}_{S^*}(\mu) + 4^5 \delta N \log(\alpha)$$

Let $\gamma := \mathcal{L}_{S^*,\overrightarrow{P^*}}(\mu) - \mathcal{L}_{S,\overrightarrow{P}}(\mu)$. Note, Lemma 31 states that $\gamma > 0$.

Set $\delta := \frac{\min\{1,\gamma/10\}}{4^5 \log(1/\alpha)}$. Note, $4^{-5} > \delta > 0$. Hence,

$$\mathcal{L}_{S^*}(\overrightarrow{D}) - \mathcal{L}_S(\overrightarrow{D}) \geq N(1-\delta)\gamma - N\gamma/10 > N\gamma/5 = \Omega(N).$$

This then implies that $w(S)/w(S^*) \leq \exp(\Omega(N))$ with probability $\geq 1 - \Omega(N)$ by the same argument as the proof of Lemma 21 in Mossel and Vigoda (2005). Then, the theorem follows from a conductance argument as in Lemma 22 in Mossel and Vigoda (2005).

■

ACKNOWLEDGMENTS

We thank Elchanan Mossel for useful discussions and John Rhodes for helpful comments on an early version of this paper. E.V. was supported in part by the NSF (grant CCF-0455666).

REFERENCES

Allman, E.S., and J.A. Rhodes. 2007. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.* (in press).

Bandelt, H.-J., and A. Dress. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.* 7:309–343.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. in: F.R. Hodson, D.G. Kendall, P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, Edinburgh, 387–395.

Chang, J.T. 1996a. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137:51–73.

Chang, J.T. 1996b. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.* 134:189–216.

Evans, S., and T. Warnow. 2004. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 130–134.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *System. Zool.* 27:401–410.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Hellmann, I.I., I. Ebersberger, S. E. Ptak, et al. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72:1527–1535.

Hellmann, I., K. Prüfer, H. Ji, et al. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* 15:1222–1231.

Janson, S., T. Ĺuczak, and A. Ruciński. 2000. *Random Graphs.* Wiley-Interscience, New York.

Kim, J. 2000. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol. Phylo. Evol.* 17:58–75.

Kolaczkowski, B., and J.W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.

Lake, J.A. 1987. A Rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.* 4:167–191.

Lercher, M.J., and L.D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18:337–340.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc. Natl. Acad. Sci. USA* 98:5688–5692.

Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21:984–990.

Mossel, E. and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309:2207–2209.

Mossel, E., and E. Vigoda. 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Prob.* (in press). Preprint available at *http://arxiv.org/abs/q-bio.PE/0505002*.

Pólya, G., and G. Szegö. 1976. *Problems and Theorems in Analysis. Vol. II.* Springer-Verlag, New York.

Rogers, J.S. 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *System. Biol.* 50:713–722.

Steel, M.A., L. Székely, and M.D. Hendy. 1994. Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.* 1:153–163.

Stefankovic, D., and E. Vigoda. 2007. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *System. Biol.* (in press). Preprint available at *http://www.cc.gatech.edu/∼vigoda*.

Webster, M.T., N.G.C. Smith, M.J. Lercher, et al. 2004. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol. Biol. Evol.* 21:1820–1830.

Address reprint request to:
Dr. Daniel Štefankovič
Department of Computer Science
University of Rochester
Rochester, NY 14627

E-mail: stefanko@cs.rochester.edu