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
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).

1.1. Background

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).

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 $\Omega$ 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 $\pi$ on $\Omega$. Every edge $\ell$ has an associated $|\Omega| \times |\Omega|$ rate matrix $R_{\ell}$, which is reversible with respect to $\pi$, and a rate $t_{\ell}$. Note, since $R_{\ell}$ is reversible with respect to $\pi$, then $\pi$ is the stationary vector for $R_{\ell}$ (i.e., $\pi R_{\ell} = 0$). The rate matrix defines a continuous time Markov chain. Then, $R_{\ell}$ and $t_{\ell}$ define a transition matrix $P_{\ell} = \exp(t_{\ell} R_{\ell})$. 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 $\pi$ (i.e., $\pi P_{\ell} = \pi$).

Given $\tilde{P} = (P_{\ell})_{\ell \in E(T)}$ 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 $P_{\ell}$ is reversible with respect to $\pi$.) Then, the probability of a labeling $\ell : V(T) \to \Omega$ is

$$
\mu_{\pi,\tilde{P}}(\ell) = \pi(\ell(r)) \prod_{u \sim r} \tilde{P}_{u \to v}(\ell(u), \ell(v)). \quad (1)
$$

Let $\mu_{\pi,\tilde{P}}$ be the marginal distribution of $\mu_{\pi,\tilde{P}}$ on the labellings of the leaves of $T$ ($\mu_{\pi,\tilde{P}}$ is a distribution on $\Omega^n$ where $n$ is the number of leaves of $T$). The goal of phylogeny reconstruction is to reconstruct $T$ (and possibly $\tilde{P}$) from $\mu_{\pi,\tilde{P}}$ (more precisely, from independent samples from $\mu_{\pi,\tilde{P}}$).

The simplest four-state model has a single parameter $\alpha$ 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 refer to 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 $|O| = 2$, the model is binary and the rate matrix has a single parameter $\alpha$. 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 $t$, which are positive and finite.

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

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

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 $M$, and then each edge is allowed any transition matrix $P_e \in M$. 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 $M$. 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 $M$ 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 $M$ is an open set. Moreover we will require that $M$ 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 $M$ of transition matrices is parameterized by an open set if there exists a finite set $\Omega$, a distribution $\pi$ over $\Omega$, an integer $n$, an open set $O \subseteq \mathbb{R}^n$, and multi-linear polynomials $\mu_{ij} \in \mathbb{R}[x_1, \ldots, x_m]$ such that

$$M = \{ (\mu_{ij})_{i,j \in [1]} \big| (x_1, \ldots, x_m) \in O \},$$

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

Typically the polynomials $\mu_{ij}$ are defined by an appropriate change of variables from the variables defining the rate matrices. Some examples of models that are parameterized 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 $P_1, P_2, \ldots, P_k$ where $P_i \in M^{T}(T_i)$ and a set of non-negative reals $q_1, q_2, \ldots, q_k$ where $\sum q_i = 1$. We then consider the mixture distribution:

$$\mu = \sum_i q_i \mu_{ij}^{T_i},$$

Thus, with probability $q_i$ we generate a sample according to $\mu_{ij}^{T_i}$. 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 $q_i = 1/2$, 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 $\mu$, we consider the likelihood of a tree $T$ as the maximum over assignments of transition matrices $\rightarrow P$ to the edges of $T$, of the probability that the tree $(T, \rightarrow P)$ generated $\mu$.

Thus, we are considering the likelihood of a pure (non-mixture) distribution having generated the mixture distribution $\mu$. More formally, the maximum expected log-likelihood of tree $T$ for distribution $\mu$ is defined by

$$L_T(\mu) = \max_{\rightarrow P \in M} \mathbb{E}_{\mu} \left[ L_T, \rightarrow P(\mu) \right],$$

where

$$L_T, \rightarrow P(\mu) = \sum_{y \in \Omega^n} \mu(y) \ln(\mu_T, \rightarrow P(y))$$

Recall for the CFN, JC, K2 and K3 models, $M$ is restricted to transition matrices obtainable from positive rate matrices $R_i$ and positive times $t_i$.

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 $\rightarrow P$ 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 $C \in (0, 1/2)$. Let $P_1 = (C + x, C - x, C - x, C + x, x^2)$ and $P_2 = (C - x, C + x, C + x, C - x, x^2)$. Consider the following mixture distribution on $T_3$:

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

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 $\mu_x$ 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 $\mu_x$ 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.
Thus, in this case, it is impossible to distinguish these two distributions. Hence, we cannot infer which particular, there is a mixture on topology.

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 convergence is exponentially slow to the posterior distribution. This result requires the 5-leaf version of mixture examples. We first found that in the work of Mossel and Vigoda (2005).

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, 2006). Every mixture distribution $\mu$ 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 $z = (z_1, \ldots, z_4)$ and $z_1 = \mu(0000), z_2 = \mu(0001), z_3 = \mu(0010), \ldots, z_4 = \mu(1111)$. Let $C_i$ denote the set of points corresponding to distributions $\mu(T_i, \mathcal{P})$ 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 $T_1$ and $T_2$. A linear test is a vector $t \in \mathbb{R}^{30}$ such that $t^T \mu_{\mathcal{P}} > 0$ for any mixture distribution $\mu_{\mathcal{P}}$ arising from $T_1$ and $t^T \mu_{\mathcal{P}} < 0$ for any mixture distribution $\mu_{\mathcal{P}}$ arising from $T_2$.

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 $t^{(5)}$ distinguishes the mixtures from $T_1$ and the mixtures from $T_2$, where $(5)$ swaps the labels for leaves 1 and 3. More precisely, for all $(a_1, a_2, a_3, a_4) \in |\Omega|^4$,

$$t^{(5)} = \frac{t_{(a_1, a_2, a_3, a_4)} - t_{a_1, a_2, a_3, a_4}}{2}$$

(2)
Theorem 4. For any model whose set $M$ 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-identifiability 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 $\pi \in S_4$, we let $T_\pi$ denote tree $T$ permuted by $\pi$. It is easily checked that the following group $K$ (Klein group) fixes every $T_\pi$:

$$K = \{(1), (12)(34), (13)(24), (14)(23)\} \leq S_4.$$

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


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

$$\mu_{T', w}(a_1, \ldots, a_n) = \mu_{T, w}(a_1, \ldots, a_n).$$  \hfill (4)

The actions on weighted trees and on distributions are compatible:

$$\mu_{T', w} = (\mu_{T, w})^T.$$

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.

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 $C_2 = \{(0, y) \mid y > 0\} \cup \{(x, y) \mid x > 0\}$). The sets of mixtures are special - they are convex hulls of images of open sets under a multi-linear polynomial map.

Lemma 9. Let $p_i(x_1, \ldots, x_n)$ be multi-linear polynomials in $x_1, \ldots, x_n$. Let $p = (p_1, \ldots, p_n)$. Let $D \subseteq \mathbb{R}^n$ be an open set. Let $C$ be the convex hull of $\{p(x) \mid x \in D\}$. Assume that $0 \notin C$. There exists $s \in \mathbb{R}^n$ 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 $p' = (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n)$. Let $C'$ be the convex hull of $\{p'(x) \mid x \in D\}$. Then

$$(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \mapsto (y_1, \ldots, y_{i-1}, \sum_{j \neq i} c_j y_{i+1}, \ldots, y_n)$$

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 $s \in \mathbb{R}^n$ 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 $s' \in \mathbb{R}^n$ such that $s'^T p(x) > 0$ for all $x \in D$). Hence, without loss of generality, we can assume that the polynomials $p_1, \ldots, p_n$ 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, $p_i(x) = p_i(x + a)$), 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 $p_1, \ldots, p_n$ 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 $y_i \neq 0$ for $x_j$ which occur in $w$ and $y_i = 0$ 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 $p_1, \ldots, p_n$.

We now prove our duality theorem.

Proof of Theorem 4. Clearly there cannot exist both a non-identifiable mixture and a linear test. Let $C$ 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 $M$. 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 $s^T p(x) > 0$ 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 $G \leq S_n$ (where $G$ is the intersection of the automorphism groups of the weighted graphs corresponding to the matrices in $M$). 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 $T^G$ which are in the same orbit of the action of $G$ on $T^O$ have the same probability in any distribution arising from the model.

Let $O'$ be the orbits of $T^G$ 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

$$t_{o_1,o_2}(\mu) = \sum_{x \in o_1} \mu(x) - \sum_{x \in o_2} \mu(x).$$  \hspace{1cm} (5)

Lemma 10. Suppose that $M$ has a linear test $s$. Then $M$ has a linear test $t$ which is a linear combination of the $t_{o_1,o_2}$.
Proof. Let $s$ be a linear test. Let 

$$e' = \sum_{a \in K} s^a$$

and

$$t = e' - e^{(1,4)}.$$ 

Let $\mu_2$ be a mixture arising from $T_2$. For any $g \in K$ the mixture $\mu^g_2$ arises from $T_2$ and hence

$$(t')^T \mu_2 = \sum_{a \in K} (s^a)^T \mu_2 = \sum_{a \in K} s^T (\mu^g_2) > 0.$$ 

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

Now we show that $t$ is a linear test as well. Let $\mu_2$ arise from $T_2$. Note that $\mu_3 = \mu_{2}^{(1,4)}$ arises from $T_3$ and hence

$$(t')^T \mu_2 = (t' - e^{(1,4)})^T \mu_2 = (t')^T \mu_2 - (t')^T \mu_3 > 0.$$ 

Similarly $t(\mu_3) < 0$ for $\mu_3$ 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 $t_{o_1, o_2}$. 


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., $I \not\in M$, 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 $\text{cl}(M)$ the generated distribution is arbitrarily close to a distribution generated from the model.

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

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

Suppose that the identity matrix $I \in \text{cl}(M)$. Let $\mu$ arise from $T_2$ with weights such that the internal edge has weight $I$. Then $\mu_3$ arises also from $T_3$ with the same weights. A linear test has to be positive for mixtures from $T_2$ and negative for mixtures from $T_3$. Hence we have:

Observation 12. Let $\mu$ 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 $\Omega_{JC} = \{0, 1, 2, 3\}$ and its semigroup $M_{JC}$ consists of matrices

$$M_{JC}(E) = \frac{1}{4}E + \begin{pmatrix}
\frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\
-\frac{1}{2} & \frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} \\
-\frac{1}{2} & -\frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\
-\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{3}{2}
\end{pmatrix}. $$

Let $\tau_{o_1, o_2}$ be a linear test. Let

$$\tau = \sum_{x \in \{0, 1, 2, 3\}} \tau(x) e^{(x)}$$

and

$$\tilde{\tau} = \sum_{x \in \{0, 1, 2, 3\}} \tau^{(x)} e^{(x)}.$$ 

Then

$$\tau = \sum_{x \in \{0, 1, 2, 3\}} \tau(x) e^{(x)}$$

and

$$\tilde{\tau} = \sum_{x \in \{0, 1, 2, 3\}} \tau^{(x)} e^{(x)}.$$ 

Similarly

$$\tau^{(x)} = \tau(x) e^{(x)}$$

and

$$\tilde{\tau}^{(x)} = \tau^{(x)} e^{(x)}.$$ 

Now we show that $\tau$ is a linear test as well. Let $\mu_2$ arise from $T_2$. Note that $\mu_3 = \mu_{2}^{(1,4)}$ arises from $T_3$ and hence

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

Similarly $\tilde{\tau}(\mu_3) < 0$ for $\mu_3$ arising from $T_3$ and hence $\tilde{\tau}$ is a linear test.
Thus, $\hat{x}$ means both endpoints have the same assignment. Whereas independent.

$$\Phi$$

where $\Omega_{\text{IC}} = \Omega_4$. There are 15 orbits in $\Omega_{\text{IC}}^2$ under the action of $G_{\text{IC}}$ (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 $(1 4)$:

$$
\begin{array}{cccc}
0000 & 0110 & 0121 & 0111 \\
0112 & 0100 & 0011 & 0102 \\
0212 & 0211 & 0122 & 0112 \\
0001 & & & \\
\end{array}
$$

(6)

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

$$t_1 = \mu(0011) - \mu(0101), \quad \text{and} \quad t_2 = \mu(0122) - \mu(0121) - \mu(0102) + \mu(0012).$$

(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 $\mu$ be a single-tree mixture arising from a tree $T$ on 4 leaves. Let $t$ be defined by

$$t^T \mu = t_2 - t_1 = \mu(0122) - \mu(0121) + \mu(0101) - \mu(0011) + \mu(0102) - \mu(0012).$$

(8)

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

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

In particular $t$ is a linear test.

**Proof.** Label the 4 leaves as $v_1, \ldots, v_4$, and let $\hat{x}_1, \ldots, \hat{x}_4$ denote the centered edge weight of the edge incident to the respective leaf. Let $\tilde{x}_i$ denote the centered edge weight of the internal edge. Let $\mu_j$ arise from $T_j$ with centered edge weights $\hat{x}_1, \ldots, \hat{x}_4$, $j \in \{1, 2, 3\}$. Let $\Phi_j$ be the multi-linear polynomial $t^T \mu_j$. If $\hat{x}_i = 0$ then $\mu_j$ does not depend on the label of $v_i$ and hence, for all $a \in \Omega_{\text{IC}}$,

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

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

$$\Phi_j = \ell_{12} \ell_{23} \ell_{34} \ell_{41} \ell_{12} (\ell_{12}),$$

where $\ell(x)$ is a linear polynomial in $K_0$.

Let $\mu'_i$ arise from $T_i$ with $\hat{x}_1 = \cdots = \hat{x}_4 = 1/4$. In leaf-labelings with non-zero probability in $\mu'_i$ 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 $\Phi_j = 0$ if $\hat{x}_1 = \cdots = \hat{x}_4 = 1/4$. Hence $\ell(\hat{x}_i)$ is the zero polynomial and $\Phi_i$ is the zero polynomial as well.

Now we consider $T_2$. If $\hat{x}_1 = 1/4$ then, by Observation 12, $\Phi_2 = 0$. Thus $1/4$ is a root of $\ell$ and hence $\Phi_2 = \alpha \cdot \ell_{12} \ell_{23} \ell_{34} (1/4 - \hat{x}_4)$. We plug in $\hat{x}_4 = 0$ and $\hat{x}_1 = 2 \hat{x}_2 = \hat{x}_3 = 1/4$ to determine $\alpha$. Let $\mu''_2$ be the distribution generated by these weights. The leaf-labelings for which $\mu''_2$ is non-zero must have the same label for $v_1, v_2$ and the same label for $v_3, v_4$. Thus $\Phi_2 = \mu''_2(0101) = 1/16$ and hence $\alpha = 64$. We have

$$\Phi_2 = 64 \ell_{12} \ell_{23} \ell_{34} (1 - 4 \hat{x}_1).$$

Note that $\Phi_2$ is always positive. The action of $(1 4)$ switches the signs of the $t_i$ and hence $\Phi_2 = -\Phi_2$. Thus $\Phi_2$ is always negative.
Thus \( \alpha \) the following orbits are swapped by model in terms of the times with 
\[
\begin{align*}
\bar{x} &= \bar{x}_4 = \bar{x}_5 = 1/4 \text{ and } \bar{x}_1 = \bar{x}_3 = 0 \text{ on } T_2. 
\end{align*}
\]
By Observation 12, we must have \( t^T \mu_1 = 0 \). Note that
\[
\mu_1(a_1 a_2 a_3 a_4) = \begin{cases} 1/4 & \text{if } a_2 = a_4, \\
0 & \text{otherwise.} \end{cases}
\]
Hence
\[
t^T \mu_1 = -\alpha_1 (0101) - \alpha_2 (0121) = -1/4(\alpha_1 + \alpha_2) = 0.
\]
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 \( \Omega_{K2} = \{0, 1, 2, 3\} \) and its semigroup \( M_{K2} \) consists of matrices
\[
M_{K2}(\bar{x}, \bar{y}) = \frac{1}{4} \mathcal{K} + \begin{pmatrix} \bar{x} + 2\bar{y} & -\bar{x} & -\bar{y} & -\bar{y} \\
-\bar{x} & \bar{x} + 2\bar{y} & -\bar{y} & -\bar{y} \\
-\bar{y} & -\bar{y} & -\bar{x} & -\bar{x} + 2\bar{y} \\
-\bar{y} & -\bar{y} & -\bar{x} + 2\bar{y} & \bar{x} + 2\bar{y} \end{pmatrix},
\]
with \( \bar{x} \leq \bar{y} < 1/4 \) and \( \bar{x} + \bar{y} > 0 \). See Felsenstein (2004) for closed form of the transition matrices of the model in terms of the times \( t_i \) and rate matrices \( R_i \). One can then derive the equivalence of the conditions there with the conditions \( \bar{x} \leq \bar{y} < 1/4, \bar{x} + \bar{y} > 0 \).

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

The group of symmetries is \( G_{K2} = \langle (01), (02)(13) \rangle \) (it has 8 elements). There are 36 orbits in \( \mathbb{R}^4 \) 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):

\[
\begin{align*}
0111 & \leftrightarrow 0222 \leftrightarrow 0233 \\
0000 & \leftrightarrow 0110 \leftrightarrow 0220 \leftrightarrow 0231 \\
0221 & \leftrightarrow 0100 \leftrightarrow 0209 \leftrightarrow 0210 \\
0230 & \leftrightarrow 0010 \leftrightarrow 0020 \leftrightarrow 0120 \\
0001 & \leftrightarrow 0002 \leftrightarrow 0012 \leftrightarrow 0102 \\
0011 & \leftrightarrow 0101 \leftrightarrow 0022 \leftrightarrow 0202 \leftrightarrow 0213 \leftrightarrow 0233 \\
0122 & \leftrightarrow 0212 \leftrightarrow 0211 \leftrightarrow 0232 \leftrightarrow 0023 \leftrightarrow 0232 \\
0012 & \leftrightarrow 0013 \leftrightarrow 0102 \leftrightarrow 0101 \leftrightarrow 0212 \leftrightarrow 0122 \\
0023 & \leftrightarrow 0123 \leftrightarrow 0202 \leftrightarrow 0203 \leftrightarrow 0032 \leftrightarrow 0102.
\end{align*}
\]

The following orbits are swapped by (1.4):

\[
\begin{align*}
0011 & \leftrightarrow 0101 \leftrightarrow 0022 \leftrightarrow 0202 \leftrightarrow 0123 \leftrightarrow 0233 \\
0122 & \leftrightarrow 0212 \leftrightarrow 0211 \leftrightarrow 0232 \leftrightarrow 0023 \leftrightarrow 0232 \\
0023 & \leftrightarrow 0232 \leftrightarrow 0123 \leftrightarrow 0212 \leftrightarrow 0122 \leftrightarrow 0212 \\
0032 & \leftrightarrow 0102 \leftrightarrow 0202 \leftrightarrow 0203 \leftrightarrow 0021 \leftrightarrow 0032 \\
0012 & \leftrightarrow 0013 \leftrightarrow 0102 \leftrightarrow 0101 \leftrightarrow 0212 \leftrightarrow 0122 \\
0023 & \leftrightarrow 0123 \leftrightarrow 0202 \leftrightarrow 0203 \leftrightarrow 0032 \leftrightarrow 0102.
\end{align*}
\]

By Lemma 10, any linear test for the K2 model is a linear combination of
\[
\begin{align*}
t_1 &= \mu(0011) - \mu(0101), \\
t_2 &= \mu(0233) + \mu(0211) - \mu(0232) - \mu(0201), \\
&= -\mu(0121) + \mu(0021) - \mu(0102) + \mu(0012), \\
t_3 &= \mu(0022) - \mu(0202), \\
t_4 &= \mu(0122) + \mu(0023) - \mu(0122) - \mu(0203), \\
t_5 &= \mu(0123) - \mu(0213).
\end{align*}
\]
Lemma 15. Let $\mu$ be a single-tree mixture arising from a tree $T$ on 4 leaves. Let $t$ be defined by

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

(9)

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

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

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 $M_{E_j}(\hat{x}_i, \hat{y}_i)$, and the internal edge has $M_{E_j}(\hat{x}_5, \hat{y}_5)$. Let $\mu_j$ be the generated distribution, and let $\Phi_j$ be the multi-linear polynomial $t^T \mu_j$.

If $\hat{y}_i = -\hat{x}_i$ then the matrix on the edge incident to leaf $v_i$ 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:

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

and we obtain,

$$
\Phi_j = 0.
$$

(10)

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

$$
\Phi_j = (\hat{x}_1 - \hat{y}_1)...(\hat{x}_4 - \hat{y}_4) f_j(\hat{x}_5, \hat{y}_5),
$$

(11)

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

Now let $\hat{x}_i = \hat{y}_i = 1/4$ 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

$$
\Phi_j = \mu_j(0202) - \mu_j(0022).
$$

(12)

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

$$
\Phi_1 = 0.
$$

(13)

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

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

(14)

Note that $\hat{x}_i - \hat{y}_i < 0$ and $\hat{y}_i < 1/4$ 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).
Thus, $\alpha$ orbits fixed: further reduces the number of orbits to $K$. Edge probabilities are restricted to some interval. Non-identifiable mixture distribution. We then show that there is a linear test for the CFN model when 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

$$t^T \mu_1 = -\alpha_3 \mu_3(0101) - \alpha_2 \mu_1(0121) = -\frac{1}{16} (\alpha_1 \hat{e} + \alpha_2 \hat{g}) = 0.$$ 

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

Let $\mu_2$ be generated from $T_1$ with centered weights $(\hat{x}_2, \hat{y}_2) = (\hat{x}_4, \hat{y}_4) = (1/4, 1/4), (\hat{x}_1, \hat{y}_1) = (1/4, 0)$, and $(\hat{x}_2, \hat{y}_2) = (\hat{x}_1, \hat{y}_1) = (1/4, 0)$. 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 $t_1, t_2, t_5$ are 0202, 0213, 0212, and 0203. Hence

$$t^T \mu_3 = \mu_1(0202) - \mu_2(0212) - \mu_3(0213) + \alpha_3 \mu_3(0213) = \frac{1}{256} (3 - 1 - \alpha_5) = 0.$$ 

Thus, $\alpha_5 = -1$ and all the $\alpha_i$ 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 $\Omega_{\text{CFN}} = \{0, 1\}$, and its semigroup $M_{\text{CFN}}$ consists of matrices

$$M_{\text{CFN}}(\alpha) = \frac{1}{2} E + \left( \begin{array}{cc} \alpha & -\alpha \\ -\alpha & \alpha \end{array} \right)$$

with $0 < \alpha < 1/2$.

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

$$\mu_{\alpha, \beta}(a_1, \ldots, a_n) = \mu_{\beta, \alpha}(1 - a_1, \ldots, 1 - a_n).$$

Hence, the group of symmetries of $\Omega_{\text{CFN}}$ is $G_{\text{CFN}} = \{011\} = \mathbb{Z}/(22)$. There are 8 orbits of the action of $G_{\text{CFN}}$ on $\Omega_{\text{CFN}}$ (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:

$$0111 \quad 0100 \quad 0010 \quad 0010 \quad 0011 \leftrightarrow 0101 \quad 0001$$
By Lemma 10, if there exists a linear test for CFN then a linear test yields 
\[ t_1 = \mu(0011) - \mu(0101) \] is a linear test. Let \( \mu \) arise from \( T_2 \) with the edge incident to leaf \( v_i \) labeled by \( M_{\text{CFN}}(\hat{x}_i) \), for \( i = 1, \ldots, 4 \), and the internal edge labeled by \( M_{\text{CFN}}(\hat{x}_5) \). A short calculation yields
\[ \mu(0011) - \mu(0101) = \frac{\hat{x}_1 \hat{x}_2 + \hat{x}_3 \hat{x}_4}{2} . \] (19)
Note that (19) is negative if \( \hat{x}_5 \) is much smaller than the other \( \hat{x}_i \); and (19) is positive if \( \hat{x}_1, \hat{x}_3 \) are much smaller than the other \( \hat{x}_i \). 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 \( \vec{P} = (w_1, w_2, w_3, w_4, w_5) \) 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
\[ \vec{P}_1 = \frac{1}{2} \begin{pmatrix} 1 & a & b & a & c \\ 1 & b & a & b & d \end{pmatrix} \] and
\[ \vec{P}_2 = \frac{1}{2} \begin{pmatrix} 1 & b & a & b & d \\ 1 & a & b & a & d \end{pmatrix} . \]
where \( 1 = (1, 1, 1, 1, 1) \) and
\[ c = \frac{z}{p}, \]
\[ d = \frac{z}{(1 - p)} , \text{ and} \]
\[ z = \frac{ab}{2(a^2 + b^2)} . \]
Let
\[ \mu = p\mu_{T_2, P_1} + (1 - p)\mu_{T_2, P_2} . \]

The distribution \( \mu \) is invariant under \( \pi = (1 4) \). Hence, \( \mu \) is also generated by a mixture from \( T' \), a leaf-labeled tree different from \( T = T_2 \). In particular, the following holds:
\[ \mu = p\mu_{T_2, P_1} + (1 - p)\mu_{T_2, P_2} . \]

Hence, whenever \( c \) and \( d \) satisfy \( 0 < c, d < 1/2 \) then \( \mu \) 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
\[ \mu(0011) - \mu(0101) = ab - (a^2 + b^2)(p + (1 - p)d) , \]
and
\[ \mu(0001) - \mu(0111) = (a^2 - b^2)(p - (1 - p)d) . \]
which are both zero for our choice of \( c \) and \( d \). This implies that \( \mu \) 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 \( \bar{x} \) for the CFN model is restricted to the interval \((a, \sqrt{a \cdot a^2})\), then there is a linear test.

Proof. We will show that (19) is positive if the \( \bar{x}_i \) are in the interval \((a, \sqrt{a \cdot a^2})\). Let \( b = \sqrt{a \cdot a^2} \).

Note that \( 0 < a < b < 1/2 \).

Since (19) is multi-linear, its extrema occur when the \( \bar{x}_i \) are from the set \([a, b]\) (we call such a setting of the \( \bar{x}_i \) extremal). Note that the \( \bar{x}_i \) are positive and \( \bar{x}_i \) occurs only in terms with negative sign. Thus a minimum occurs for \( \bar{x}_i = 0 \). The only extremal settings of the \( \bar{x}_i \) which have \( \bar{x}_i \bar{x}_j + \bar{x}_j \bar{x}_k + \bar{x}_k \bar{x}_i \) are \( \bar{x}_1 = \bar{x}_3 = b, \bar{x}_2 = \bar{x}_4 = a \) and \( \bar{x}_1 = \bar{x}_3 = a, \bar{x}_2 = \bar{x}_4 = b \). For the other extremal settings (19) is positive, since \( b < 1/2 \). For \( \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4 = a \) 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. Moreover, 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 \( \Omega_{K3} = \{0, 1, 2, 3\} \) and its semigroup \( \mathcal{M}_{K3} \) consists of matrices of the following form (which we have centered around their stationary distribution):

\[
M_{K3}(\bar{x}, \bar{y}, \bar{z}) = \frac{1}{4} E + \begin{pmatrix}
\bar{x} + \bar{y} + \bar{z} & -\bar{x} & -\bar{y} & -\bar{z} \\
-\bar{x} & \bar{x} + \bar{y} + \bar{z} & -\bar{y} & -\bar{z} \\
-\bar{y} & -\bar{x} & \bar{x} + \bar{y} + \bar{z} & -\bar{z} \\
-\bar{z} & -\bar{y} & -\bar{x} & \bar{x} + \bar{y} + \bar{z}
\end{pmatrix}
\]

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

The group of symmetries is \( G_{K3} = \{(01)(23), (02)(13)\} \) (which is again the Klein group). There are 64 orbits in \( \Omega^4 \) 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 (14):

\[
0000, 0010, 0020, 0300, 0320, 0330, 0220, 0100, 0120, 0200, 0110, 0202, 0302,
0111, 0211, 0311, 0101, 0121, 0212, 0312, 0112, 0132, 0213, 0231, 0313, 0011
0022, 0023, 0033, 0122, 0132, 0222, 0233, 0322, 0333
0220, 0230, 0320, 0330, 0202, 0203, 0302, 0303, 0200, 0100, 0101
0201, 0202, 0203, 0301, 0302, 0303, 0202, 0203, 0201, 0200, 0203
0202, 0203, 0201, 0200, 0100, 0101, 0102, 0103, 0104, 0105
\]

The following orbits switch as indicated under the action of (14):

\[
0011 \leftrightarrow 0100, 0022 \leftrightarrow 0202, 0033 \leftrightarrow 0303, 0123 \leftrightarrow 0213, 0132 \leftrightarrow 0231, 0231 \leftrightarrow 0321.
\]

7.1. Non linear test for K3

By Lemma 10, any test is a linear combination of \( t_1 = \mu(0011) - \mu(0101) \).
PHYLOGENY OF MIXTURE MODELS

Any \((\mu(0322) + \mu(0311) - \mu(0232) - \mu(0201) - \mu(0131) - \mu(0102) + \mu(0021) + \mu(0012))\) \(\Phi\) (we ignore the labels in this expression) with this property are non-identifiable with non-zero probability, leaves \(t_2\) and \(t_3\), but the unique test for \(K_2\) has a non-trivial structure and hence has at most 6 terms. A test for \(K_3\) must be a test for \(K_2\), but the unique test for \(K_2\) has 8 terms. Thus, \(\Phi\) cannot be a test for \(K_3\).

Lemma 20. There does not exist a linear test for the \(K_3\) model.

Corollary 21. There exists a non-identifiable mixture in the \(K_3\) model.

Proof of Lemma 20. Suppose that \(t = \alpha_1 t_1 + \cdots + \alpha_5 t_5\) is a test. Let \(\tilde{t}_i = (\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)\), \(1 \leq i \leq 5\). For \(1 \leq i \leq 4\), \(\tilde{w}_i\) denotes the centered parameters for the edge incident to leaf \(i\), and \(\tilde{w}_5\) are the centered parameters for the internal edge.

In the definitions of \(\mu_1\) and \(\mu_2\) below, we will set \(\tilde{w}_2 = \tilde{w}_1 = \tilde{w}_5 = (0, 0, 0)\). 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, \(\Phi(\mu_1) = 0\). Let \(\mu_2\) be generated from \(T_2\) with \(\tilde{v}_1 = (1/4, \tilde{y}_1, \tilde{z}_1)\), and \(\tilde{v}_3 = (1/4, 0, 0)\). 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 \(t_1, \ldots, t_5\) with this property are \(0101, 0232, 0323\). Thus,

\[
\begin{align*}
0^T \mu_1 &= -\alpha_1 \mu(0101) - \alpha_2 \mu(0232) - \alpha_3 \mu(0323) = \frac{1}{16} (\alpha_1 + \alpha_2 \tilde{y}_1 + \alpha_3 \tilde{z}_1) = 0. \tag{20}
\end{align*}
\]

Any \(y, z \in [1/8 - \varepsilon, 1/8 + \varepsilon]\) with \(y' \geq z\) gives a valid matrix in the \(K_3\) model. For (20) to be always zero we must have \(\alpha_1 = \alpha_2 = \alpha_3 = 0\) (since \(\tilde{y}_1, \tilde{z}_1\) are linearly independent polynomials).

Let \(\mu_2\) be generated from \(T_2\) with \(\tilde{v}_1 = (1/4, \tilde{y}_1, \tilde{z}_1)\), and \(\tilde{v}_3 = (1/4, \tilde{y}_2, \tilde{z}_2)\). The only labels in \(t_4, \ldots, t_5\) 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,

\[
\begin{align*}
0^T \mu_1 &= -\alpha_4 \mu(0202) - \alpha_5 \mu(0313) - \alpha_5 \mu(0212) - \alpha_5 \mu(0303) = 0.
\end{align*}
\]

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

Thus, \(t^2\) is a linear combination of \(t_1, t_2, t_3\) and hence has at most 6 terms. A test for \(K_3\) must be a test for \(K_2\), but the unique test for \(K_2\) has 8 terms. Thus, \(\Phi\) cannot be a test for \(K_3\).
We will now consider rate matrices for the model, as opposed to the transition probability matrices. The rate matrix for the K3 model is

\[
R = R(\alpha, \beta, \gamma) = \begin{pmatrix}
-\alpha & -\alpha - \beta - \gamma & \alpha \\
\alpha & -\alpha - \beta - \gamma & \beta \\
\beta & \gamma & -\alpha - \beta - \gamma \\
\gamma & \beta & \alpha
\end{pmatrix},
\]

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 \(\alpha, \beta, \gamma\) be chosen independently from the uniform distribution on \([0, 1]\). Then the probability (over the choice of \(\alpha, \beta, \gamma\)) that there does not exist a test for the K3 model with the rate matrix \(R(\alpha, \beta, \gamma)\) is

\[
\frac{1}{m}.
\]

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

\[
\sum_{i=1}^{m} a_i(u_1, \ldots, u_n) e^{b_i(u_1, \ldots, u_n)},
\]

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 Szegő, 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 \(u_1\). Since \(a_i\) is a non-zero polynomial there are only finitely many choices of \(c \in [0, 1]\) for which \(a_i(c, u_2, \ldots, u_n)\) is the zero polynomial. Thus with probability 1 over the choice of \(u_1 = c\) we have that \(a_i(c, u_2, \ldots, u_n)\) is a non-zero polynomial. Now, by the induction
We have changed notation from our earlier definition of $\gamma$; have that there exists no test with probability $\sigma$ these as functions of $x$. Let $\xi_1, \ldots, \xi_9$ be defined in Section 7. Suppose that $t = \xi_1 t_1 + \cdots + \xi_9 t_9$ is a test. We will use Observation 12 to show that $\xi_1 = \cdots = \xi_9 = 0$ and hence $t$ cannot be a test. Let $\sigma = (\xi_1, \ldots, \xi_9)$.

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

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

where

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

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

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

We have changed notation from our earlier definition of $M_{K3}$ by exchanging $A, B, C$ for $\tilde{A}, \tilde{B}, \tilde{C}$ to indicate these as functions of $\alpha, \beta$ and $\gamma$.

Let $\mu_2$ be generated from $T_2$ 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 $t^T \mu_2 = 0$ for all $x \geq 0$.

Let $f(x) = t^T \mu_2$ 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 $\alpha, \beta, \gamma$, and $x$. Thus, the Wronskian (i.e., det $W(x)$) is a generalized polynomial in variables $\alpha, \beta, \gamma$, and $x$.

Now we show that for a particular choice of $\alpha, \beta, \gamma$, and $x$ we have det $W(x) \neq 0$. Let $\alpha = \pi, \beta = 2\pi, \gamma = -2\pi$, 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. Tedium computation (best performed using a computer algebra system) yields that the Wronskian matrix is the following. The first seven columns are as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\alpha + \beta$</th>
<th>$\alpha + \gamma$</th>
<th>$\beta + \gamma$</th>
<th>$\alpha + \beta + \gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>2</td>
<td>2.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>2.0000</td>
<td>2.0000</td>
<td>0.0000</td>
<td>4.0000</td>
</tr>
<tr>
<td>3</td>
<td>3.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>3.0000</td>
<td>3.0000</td>
<td>0.0000</td>
<td>6.0000</td>
</tr>
<tr>
<td>4</td>
<td>4.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>4.0000</td>
<td>4.0000</td>
<td>0.0000</td>
<td>8.0000</td>
</tr>
<tr>
<td>5</td>
<td>5.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>5.0000</td>
<td>5.0000</td>
<td>0.0000</td>
<td>10.0000</td>
</tr>
<tr>
<td>6</td>
<td>6.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>6.0000</td>
<td>6.0000</td>
<td>0.0000</td>
<td>12.0000</td>
</tr>
<tr>
<td>7</td>
<td>7.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>7.0000</td>
<td>7.0000</td>
<td>0.0000</td>
<td>14.0000</td>
</tr>
</tbody>
</table>

The last two columns are as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\alpha + \beta + \gamma$</th>
<th>$\alpha \beta \gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>4.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>6.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>8.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>10.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>12.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>14.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The determinant of $W(x)$ is $33920150890618370273798016000000\pi^{36}$, which is non-zero. Thus, det $W(x)$ is a non-zero generalized polynomial and hence for random $\alpha, \beta, \gamma$, 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 $w_1 = \sigma^T \mu_2$. Since $t$ is a test we have $w_1 = 0$. The second entry $w_2$ of $w$ is given by $w_2 = (\partial/\partial x)^T \mu_2$. 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 $\sigma$ 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 \( 0 < x_e < 1/\delta \). In K2, there are 2 parameters and in K3 there are 3 parameters. We use the term weights \( \vec{w} \) to denote the setting of the parameters defining the transition matrix for the edge \( e \).

And then we let \( \vec{w} = (\vec{w}_e)_{e \in T} \) denote the set of weights \( \vec{w} \) 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, \( L_T(\mu) \) is maximized over the set of weights \( \vec{w} \). Hence our consideration of weights \( \vec{w} \). There are typically constraints on \( \vec{w} \) in order to define valid transition matrices. For example, in the CFN model we have \( e_{\vec{w}} = x_e \), where \( 0 < x_e < 1/2 \), and similarly \( 0 < x_e < 1/4 \) in the JC model. In the K2 model we have \( e_{\vec{w}} = (x, y) \) 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 \( P_1 = P_2 \) and hence \( \mu_0 = \mu_F, \vec{w} \) where \( \vec{w} = (1/2, 1/2, 1/2, 1/2) \) are the weights for the CFN model, i.e., \( \mu_0 \) is generated by a pure distribution from \( T_2 \). 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 \( \vec{w} \) where all internal edges have zero weight. For all trees \( S \neq T \) on \( n \) leaves, there is a unique weight \( \vec{w}' \) such that \( \mu_{S, \vec{w}} = \mu_{T, \vec{w}'}. \)

**Proof.** Let \( \vec{w}' \) have the same weight as \( \vec{w} \) for all terminal edges, and zero weight for all internal edges. Note, then we have \( \mu_{S, \vec{w}} = \mu_{T, \vec{w}'} \) and it remains to prove uniqueness of \( \vec{w}' \).

Let \( \Omega \) be a set of weights where \( \mu_{S, \vec{w}} = \mu_{T, \vec{w}'} \). Let \( S' \) be obtained from \( S \) by contracting all the edges of zero weight in \( \vec{w} \), and let \( \vec{w} \) be the resulting set of weights for the remaining edges. The tree \( S' \) has all internal vertices of degree \( \geq 3 \) and internal edges have non-zero weight.

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

For \( \vec{w} \) and \( \vec{w}' \) defined for the CFN model in Theorem 2, for any 4-leaf tree \( S \), the maximum of \( L_S(\mu_0) \) is achieved on \( v = (1/4, 1/4, 1/4, 1/4, 0) \). For any \( \vec{v}' \neq \vec{v} \) the distribution \( \mu_{S, \vec{v}'} \) is different from \( \mu_0 \) and hence \( L_{S, \vec{v}'}(\mu_0) < L_{S, \vec{v}}(\mu_0) = \mu_{S, \vec{v}}^0 \) in \( \mu_0 \). Intuitively, for small \( \epsilon \) the maximum of \( L_S(\mu_0) \) should be realized on a \( \vec{w}' \) which is near \( \vec{v} \). 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 \( L_S(\mu_0) \) in terms of \( L_S(\mu_0) \).

**Lemma 26.** Let \( \mu \) be a probability distribution on \( V_n \) 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 \( \mu_{S, \vec{v}} = \mu \). Then for every \( \delta > 0 \) there exists \( \epsilon > 0 \) such that for any \( \vec{w} \) with \( ||\vec{w}' - \vec{w}|| \leq \epsilon \) the global optima of \( L_S(\mu_0) \) 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 \( \mu_{S, \vec{v}} = \mu_0 \). Note, the requirement that every element in \( \{0, 1\}^n \) has non-zero probability, is
We assumed thus there are no branches of length $\mu$. In particular, for the CFN model, the branch lengths are in the interval $[\mu - \delta, \mu + \delta]$ where $\delta$ is a weight for which the optimum of $L_{S,v}(\mu')$ is attained. By the optimality of the $\mu'$ we have

$$L_{S,v}(\mu') \geq L_{S,v}(\mu).$$

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

$$\lim_{i \to \infty} L_{S,v}(\mu') = \lim_{i \to \infty} (\mu'^T \ln \mu_{S,v,i}) = \mu^T \ln \mu_{S,v} = L_{S,v}(\mu).$$

Take a subsequence of the $\mu'$ which converges to some $\nu'$. Note, $\nu'$ is in the closure of the model.

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

Hence,

$$L_{S,v}(\mu') = \sum_{i \in \Omega} \mu'_i(z) \ln(\mu_{S,v,z}(z)) \leq \epsilon \min_{z \in \Omega} \ln(\mu_{S,v,z}(z))$$

Because of (26), for all sufficiently large $i$, $L_{S,v}(\mu') \geq 2L_{S,v}(\mu)$ (recall that the log-likelihoods are negative). Combining with (27), we have that the entries of $\ln \mu_{S,v'}$ are bounded from below by $L_{S,v}(\mu)/4\epsilon$.

Thus, both $\mu'$ and $\ln \mu_{S,v'}$ 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,

$$\lim_{n \to \infty} L_{S,v}(\mu') = \lim_{n \to \infty} (\mu'^T (\ln \mu_{S,v})) = \mu^T \ln \mu_{S,v} = L_{S,v}(\mu).$$

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

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

Here is the formal statement of the lemma.

**Lemma 28.** Let $\mu \in \mathbb{R}^{2n}$ be a probability distribution on $\Omega^n$ such that every element has non-zero probability. Let $S$ be a leaf-labeled binary tree on $n$ vertices. Suppose that there exists $\nu'$ in the closure of the model such that $\mu_{S,v'} = \mu$ and that $\nu'$ is the unique such weight. Let $\Delta_{\nu} \in \mathbb{R}^{2n}$ be such that $\Delta_{\nu}^T 1 = 0$, and $\nu' \to \nu$ is continuous around 0 in the following sense: $\Delta_{\nu} \to \Delta_{\nu_0}$ as $\nu \to 0$.

Let $g(\nu) = L_{S,v}(\mu)$, and $h_1(\nu) = (\Delta_{\nu}^T)^T \ln \mu_{S,v}$. Let $H$ be the Hessian of $g$ at $\nu'$ and $J_{\nu}$ be the Jacobian of $h_1$ at $\nu'$. Assume that $H$ has full rank. Then

$$L_{S,v}(\mu + x \Delta_{\nu}) \leq \mu^T \ln \mu + x h_1(\nu') - \frac{x^2}{2} H^{-1} x^T + o(x^2).$$

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

**Remark 29.** When $(H^{-1} x^T)_{i} < 0$ 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 $(0, 1/4)$. 
We will now argue that equality (31) remains true even when the function $w$ which contradicts the earlier claim that $\mu T f(w) = 0$. Suppose that $\mu T f(w) = 0$. Then there exists $\Delta v$ such that $\mu^T \ln \mu_T + \sum_{v} > 0$. Then, the maximum (over all distributions $v$) of $\mu^T \ln(v)$ is achieved at $v = \mu$. Thus, (31) holds.

If we perturb $\mu$ by $x\Delta \mu$, and $\nu$ by $\Delta \nu$, the likelihood changes as follows:

$$L_{\nu+\Delta \nu} = \mu^T \ln \mu + x\Delta \mu$$
$$= \mu^T \ln \mu + x\Delta \mu.$$  

(30)

We will now argue that equality (31) remains true even when $\nu$ is on the boundary of $\Lambda$. Function $\mu_T$ is still a distribution (note that $\nu$ can be outside of $\Delta$). Suppose that $\mu^T f(\nu) \neq 0$. Then there exists $\Delta \nu$ such that $\mu^T \ln \mu_T + \sum_{v} = 0$. This contradicts the fact that the maximum (over all distributions $v$) of $\mu^T \ln(v)$ is achieved at $v = \mu$. Thus, (31) holds.

If we perturb $\mu$ by $x\Delta \mu$, and $\nu$ by $\Delta \nu$, the likelihood changes as follows:

$$L_{\nu+\Delta \nu} = \mu^T \ln \mu$$
$$= \mu^T \ln \mu + x\Delta \mu$$

(30)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)

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, $g(\nu) = L_{\nu}(\mu)$ is maximized for $\nu = \nu$. Plugging $x = 0$ into (30), we obtain

$$g(\nu + \Delta \nu) = \mu^T \ln \mu + x\Delta \mu = 0.$$  

(31)
negative, i.e., it is negative definite. Let $\lambda_{\text{max}}$ be the largest eigenvalue of $H$, i.e., closest to zero. We will use that $\lambda_{\text{max}} < 0$. Note,

$$\langle \Delta \varphi \rangle^T H(\Delta \varphi) \leq \lambda_{\text{max}} ||\Delta \varphi||^2 \quad (35)$$

Set $\overline{\Delta \varphi}$ so that $L_{\varphi, \overline{\Delta \varphi}}(\mu + x \Delta \mu_0)$ is maximized. ($\Delta \varphi$ is a function of $x$.) Now we will prove that $||\Delta \varphi|| = O(x)$. For any $\delta > 0$, by Lemma 26, for all sufficiently small $x$, we have $||\Delta \varphi|| \leq \delta$. By (33) and (35), we have

$L_{\varphi, \overline{\Delta \varphi}}(\mu + x \Delta \mu_0) \leq \mu^T \ln \mu + x h_s(\overline{\varphi}) + \lambda_{\text{max}} ||\Delta \varphi||^2 + x J_s(\overline{\varphi}) + O(||\Delta \varphi||^3) + x ||\Delta \varphi||^2$.

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

$L_{\varphi, \overline{\Delta \varphi}}(\mu + x \Delta \mu_0) \leq \mu^T \ln \mu + x h_s(\overline{\varphi}) + \lambda_{\text{max}} ||\Delta \varphi||^2 + O(||\Delta \varphi||^2)$.

On the other hand, we have

$L_{\varphi, \overline{\Delta \varphi}}(\mu + x \Delta \mu) = (\mu + x \Delta \mu)^T \ln \mu + x h_s(\overline{\varphi})$.

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

$L_{\varphi, \overline{\Delta \varphi}}(\mu + x \Delta \mu) \leq L_{\varphi}(\mu + x \Delta \mu)$.

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

$L_{\varphi}(\mu + x \Delta \mu) \leq \mu^T \ln \mu + x h_s(\overline{\varphi}) + \max_{\Delta \varphi} \left( \frac{1}{2} (\Delta \varphi)^T H(\Delta \varphi) + x J_s(\overline{\varphi}) \right) + O(x^2)$. \quad (36)

The maximum of

$$\frac{1}{2} (\Delta \varphi)^T H(\Delta \varphi) + x J_s(\overline{\varphi}) \quad (37)$$

occurs at $\overline{\Delta \varphi} := -x H^{-1} J_s^*$; for this $\overline{\Delta \varphi}$ the value of (37) is $- \frac{x^2}{2} J_s H^{-1} J_s^*$. Therefore,

$L_{\varphi}(\mu + x \Delta \mu) \leq \mu^T \ln \mu + x h_s(\overline{\varphi}) - \frac{x^2}{2} J_s H^{-1} J_s^* + O(x^2)$.

From $\mu_0 \rightarrow \mu_0$, we have $J_s H^{-1} J_s^* = (1 + o(1)) J_s H^{-1} J_s^*$, and hence

$L_{\varphi}(\mu + x \Delta \mu) \leq \mu^T \ln \mu + x h_s(\overline{\varphi}) - x^2 J_s H^{-1} J_s^* + o(x^2)$.

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 $\overline{\varphi} + x \overline{\Delta \varphi}$ can be an invalid weight (i.e., outside of $\Lambda$) for all $x$. If $\overline{\Delta \varphi}$, $\overline{\varphi} = 0$ whenever $\overline{\varphi}$, $\overline{\Delta \varphi}$ is a valid weight vector for sufficiently small $x$ and hence plugging directly into (33) we have

$L_{\varphi}(\mu + x \Delta \mu_0) \geq \mu^T \ln \mu + x h_s(\overline{\varphi}) - \frac{x^2}{2} J_s H^{-1} J_s^* + O(x^2) = \mu^T \ln \mu + x h_s(\overline{\varphi}) - \frac{x^2}{2} J_s H^{-1} J_s^* + o(x^2)$. \quad (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 $\mu_1$ be generated from $T_3$ with weights $(1/4 + x, 1/4 - x, 1/4 - x, 1/4 + x, x^2)$ and $\mu_2$ be generated from $T_3$ with weights $(1/4 - x, 1/4 + x, 1/4 + x, 1/4 - x, x^2)$. Let $\mu = (\mu_1 + \mu_2)/2$. Note that $(1/4)(2/3)$ fixes $\mu_1$ and $\mu_2$, and $(1/2)(1/4)$ swaps $\mu_1$ and $\mu_2$. Hence $\mu$ is invariant under $K = \langle (1/2)(1/4), (1/4)(2/3) \rangle$. This simplifies many of the following calculations.

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

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

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

$$J_0 = \begin{pmatrix}
1744 & 1744 & 1744 & 1744 & -880 \\
615 & 615 & 615 & 615 & 309
\end{pmatrix},$$

Finally,

$$-\frac{1}{2} H^{-1} J_0^T = \begin{pmatrix}
36328/1845 & \frac{1968997}{284}
\end{pmatrix} \approx 19.68997.$$

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

$$\mathbb{L}_x(\mu_x) \leq \mu^T \ln \mu + x h(v) - \frac{x^2}{2} J_0 H^{-1} J_0^T + O(x^3)$$

For $T_2$ we have

$$J = \begin{pmatrix}
1744 & 1744 & 1744 & 1744 & -1208 \\
615 & 615 & 615 & 615 & 309
\end{pmatrix},$$

Then,

$$-\frac{1}{2} H^{-1} J_0^T = \begin{pmatrix}
244862/9225 & \frac{2654331}{9225}
\end{pmatrix} \approx 26.54331.$$

By Lemma 28, we have

$$\mathbb{L}_x(\mu_x) \leq \mu^T \ln \mu + x h(v) + x^2 \frac{244862}{9225} + O(x^3)$$

For $T_1$ we have

$$J_0 = \begin{pmatrix}
1744 & 1744 & 1744 & 1744 & 4040 \\
615 & 615 & 615 & 615 & 309
\end{pmatrix},$$

$$-H^{-1} J_0 = \begin{pmatrix}
\frac{-7}{4} & \frac{-7}{4} & \frac{-7}{4} & \frac{143}{10}
\end{pmatrix}. $$

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

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

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

$$H^{-1} = \begin{pmatrix}
\frac{(4C^2 - 1)^2}{16C^4} & 0 & 0 & 0 & \frac{(4C^2 - 1)^2}{2(4C^2 - 1)} \\
0 & \frac{16C^2 - 1}{16C^4} & 0 & 0 & \frac{(4C^2 - 1)^2}{4(4C^2 - 1)} \\
0 & 0 & \frac{16C^2 - 1}{16C^4} & 0 & \frac{(4C^2 - 1)^2}{8(4C^2 - 1)} \\
0 & 0 & 0 & \frac{16C^2 - 1}{16C^4} & \frac{(4C^2 - 1)^2}{16(4C^2 - 1)} \\
(4C^2 - 1)(4C^2 - 1)(4C^2 - 1)^2 & (4C^2 - 1)^2 & (4C^2 - 1)^2 & (4C^2 - 1)^2 & (4C^2 - 1)^2
\end{pmatrix}$$

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

$$\frac{16C(64C^6 - 16C^3 + 12C^2 + 1)}{(16C^4 - 1)(16C^4 + 24C^2 + 1)}$$

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

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

$$\gamma = \frac{(1 + 4C^2)^2}{16}.$$  

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

For $T_1$ we have

$$J_0[5] = \frac{1}{17} 128C^2 (16C^6 - 24C^3 + 17C^2 + 1).$$

$$\Delta_1 := \frac{1}{2} J_0 H^{-1} J_0 = \frac{1}{17} (-512C^{12} - 2048C^{10} + 3520C^8 - 1856C^6 + 390C^4 + 88C^2 + 3).$$

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

$$J_0[5] = \frac{1}{17} 128C^2 (16C^6 - 40C^2 - 7).$$
\[ \Delta_2 = \frac{1}{2} J_0 H^{-1} J_0 = \frac{1}{7} \cdot (-512C^{12} - 5120C^{10} + 960C^8 + 512C^6 + 198C^4 + 28C^2 + 1). \]

For \( T_3 \) we have
\[ J_0[5] = \frac{1}{\beta} \cdot 256C^8(16C^4 - 8C^2 - 3). \]
\[ \Delta_3 = \frac{1}{2} J_0 H^{-1} J_0 = \frac{1}{7} \cdot (2048C^{12} - 2048C^{10} - 512C^8 + 72C^4 + 24C^2 + 1). \]

It remains to show that \( \Delta_1 > \Delta_2 \) and \( \Delta_4 > \Delta_3 \). 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 \( \mu \) 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
\[ P_{K3}(\alpha, \beta, \gamma) = \begin{pmatrix}
1 - \alpha - \beta - \gamma & -\alpha & -\beta & -\gamma \\
1 - \alpha - \beta - \gamma & \alpha & \beta & -\gamma \\
1 - \alpha - \beta - \gamma & -\gamma & \alpha & \beta \\
1 - \alpha - \beta - \gamma & \gamma & -\beta & \alpha \\
\end{pmatrix}, \]

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

**Theorem 30.** Let \( P_T = (1/8 + x, 1/8 - x, 1/8 - x, 1/8 - x, x^2) \) and \( P_T = (1/8 - x, 1/8 + x, 1/8 + x, 1/8 - x, x^2) \). Let \( \mu_\alpha \) denote the following mixture distribution on \( T_3 \) generated from the JC model:
\[ \mu_\alpha = \left( \mu_{T_1, T_2} + \mu_{T_2, T_3} \right) / 2. \]

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

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 \( T_3 \) is the most likely tree. As in the case for the CFN model, because of symmetry, we have the same Hessian for all trees:
\[ H = \begin{pmatrix}
-23573954 & -915672 & -915672 & -915672 & 257856 \\
-915672 & 2177304 & -915672 & -915672 & 257856 \\
-915672 & -915672 & 2177304 & -915672 & 257856 \\
-915672 & -915672 & -915672 & 2177304 & 257856 \\
257856 & 257856 & 257856 & 257856 & 257856 \\
\end{pmatrix}. \]

Again the Jacobians differ only in the last coordinate.
For $T_3$:

$$J_0 = \begin{pmatrix} 4199248 & 4199248 & 4199248 & 4199248 & -7085428 \\ 112255 & 112255 & 112255 & 112255 & 112255 \end{pmatrix},$$

Then,

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

For $T_2$:

$$J_0 = \begin{pmatrix} 4199248 & 4199248 & 4199248 & 4199248 & -8069818 \\ 112255 & 112255 & 112255 & 112255 & 112255 \end{pmatrix},$$

Then,

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

For $T_1$:

$$J_0 = \begin{pmatrix} 4199248 & 4199248 & 4199248 & 4199248 & 22878022 \\ 112255 & 112255 & 112255 & 112255 & 112255 \end{pmatrix},$$

$$-H^{-1} J_0 = \begin{pmatrix} -10499073 & -10499073 & -10499073 & -10499073 & 118305233 \\ \frac{2816908}{2816908} & \frac{2816908}{2816908} & \frac{2816908}{2816908} & \frac{2816908}{2816908} & \frac{4225362}{4225362} \end{pmatrix}.$$

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

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

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

(45)

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

$$
\frac{1}{2}J_0H^{-1}J_0^T = \frac{410113105846051}{474118011310} \approx 864.6374.
$$

(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,

$$
\frac{1}{2}J_0H^{-1}J_0^T = \frac{1221030227753251}{474118011310} \approx 2574.286.
$$

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

$$
\mu_{T, \vec{P}}(\vec{D}) = \prod_i \Pr(\vec{D}_i | T, \vec{P})
$$

$$
= \prod_i \Pr(\vec{D}_i | T, \vec{P})
$$

$$
= \prod_i \mu_{T, \vec{P}}(\vec{D}_i)
$$

$$
= \exp \left( \sum_i \ln(\mu_{T, \vec{P}}(\vec{D}_i)) \right)
$$

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

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

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

Applying Bayes law we get the posterior distribution:

$$
\Pr(T, \vec{P} | \vec{D}) = \frac{\Pr(\vec{D} | T, \vec{P}) \Phi(T, \vec{P})}{\Pr(\vec{D})}
$$
Note that for uniform priors the posterior probability of a tree given $\vec{D}$ is proportional to $\Pr(\vec{D} | T)$. Each tree $T$ then has a posterior weight $w(T) = \int_{\vec{P}} \Pr(\vec{D} | T, \vec{P}) \Phi(T, \vec{P}) d\vec{P}$.

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 $S_t \rightarrow S_{t+1}$ 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

$$\min\{1, w(S')/w(S_t)\}.$$  (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 $\pi$ of the Markov chain satisfies, 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 $\delta$ with $T_{mix} \log(1/\delta)$ 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 $S_0$ denote the tree (((12), 5), (34)). Thus, $S_0$ 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
the transition probabilities for the edges of $S_j$ in this order. For the CFN model, we consider the following vector of transition probabilities. Let

$$\mathbf{P}_1 = (1/4 + x, 1/4 - x, 1/4 - x, 1/4 + x, 1/4, c),$$

and

$$\mathbf{P}_2 = (1/4 - x, 1/4 + x, 1/4 + x, 1/4 - x, 1/4, c),$$

where

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

For the JC model, let

$$\mathbf{P}_1 = (1/8 + x, 1/8 - x, 1/8 - x, 1/8 + x, 1/8, c', c''),$$

and

$$\mathbf{P}_2 = (1/8 - x, 1/8 + x, 1/8 + x, 1/8 - x, 1/8, c', c''),$$

where

$$c' = 16\epsilon^2.$$

Let $\mu_1 = \mu_{S_1, \mathbf{P}_1}$ and $\mu_2 = \mu_{S_2, \mathbf{P}_2}$. We are interested in the mixture distribution:

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

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

The key lemma for our Markov chain result states that under $\mu$, the likelihood has local maximum, with respect to NNI connectivity, on $S_j$ and $S_{2j}$.

**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 $S_j$ or $S_{2j}$, we have

$$L_S(\mu) < L_{S_j}(\mu), \quad L_S(\mu) < L_{S_{2j}}(\mu).$$

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., $\mathbf{D} = (D_1, \ldots, D_N)$, chosen independently from the distribution $\mu$. 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 $\epsilon$-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:

$$H = \begin{bmatrix}
1241 & -1241 & 1241 & -1241 & 1241 \\
-1241 & 1241 & -1241 & 1241 & -1241 \\
1241 & -1241 & 1241 & -1241 & 1241 \\
-1241 & 1241 & -1241 & 1241 & -1241 \\
1241 & -1241 & 1241 & -1241 & 1241 \\
\end{bmatrix}.$$
Thus, note the last two coordinates are positive, hence we get equality in the conclusion of Lemma 28. Finally, we now consider the JC model. Again the Hessian is the same for all the trees:

Their Jacobian is

\[ J_0 = \begin{pmatrix} 17056 & 17056 & 17056 & 2432 & 57952 & 57952 \\ 1281 & 1281 & 1281 & 127 & 3843 & 3843 \end{pmatrix} \]

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,

\[ \frac{1}{2} J_0 H^{-1} J_0^T = \frac{436480}{3844} \approx 113.5779.\]

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

\[ J_0 = \begin{pmatrix} 17056 & 17056 & 17056 & 2432 & 17056 & 57952 & 3728 \\ 1281 & 1281 & 1281 & 127 & 3843 & 427 & 427 \end{pmatrix} \]

Finally,

\[ -\frac{1}{2} J_0 H^{-1} J_0^T = \frac{2242633984}{20840589} \approx 107.6090.\]

Note the quantities \(-\frac{1}{2} J_0 H^{-1} J_0^T\) 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:

\[
H = \begin{pmatrix}
11237313 & -36559934 & -36559934 & -36559934 & -36559934 & -36559934 & -29343776 \\
-36559934 & 11237313 & -36559934 & -36559934 & -36559934 & -36559934 & -29343776 \\
-36559934 & -36559934 & 11237313 & -36559934 & -36559934 & -36559934 & -29343776 \\
-36559934 & -36559934 & -36559934 & 11237313 & -36559934 & -36559934 & -29343776 \\
-36559934 & -36559934 & -36559934 & -36559934 & 11237313 & -36559934 & -29343776 \\
-36559934 & -36559934 & -36559934 & -36559934 & -36559934 & 11237313 & -29343776 \\
-36559934 & -36559934 & -36559934 & -36559934 & -36559934 & -36559934 & 11237313
\end{pmatrix}
\]
Beginning with tree $S_1$

We have:

\[
J_0 = \left(\begin{array}{cccccccc}
4342624176 & 4342624176 & 4342624176 & 4342624176 & 1733695536 & 5655197244 & 5655197244 \\
20541185 & 20541185 & 20541185 & 20541185 & 20541185 & 20541185 & 20541185 \\
\end{array}\right)
\]

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

\[
\begin{align*}
5114490004637540016 &
593018923302763639 \\
5114490004637540016 &
593018923302763639 \\
\end{align*}
\]

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

\[
-\frac{1}{2}J_0J_0^T = \frac{48101472911555370428804991552}{12181311412062878919972215} \approx 3948.793
\]

For the neighbors of $S_1$:

We have:

\[
J_0 = \left(\begin{array}{cccccccc}
4342624176 & 4342624176 & 4342624176 & 4342624176 & 1733695536 & 4342624176 & 2955839412 \\
20541185 & 20541185 & 20541185 & 20541185 & 20541185 & 20541185 & 20541185 \\
\end{array}\right)
\]

Hence,

\[
-\frac{1}{2}J_0H^{-1}J_0^T = \frac{56725804836101835698372638212706164356508896}{1556848125865727860752294372598321870798435} \approx 3643.631
\]

Considering $S_2$:

We have:

\[
J_0 = \left(\begin{array}{cccccccc}
4342624176 & 4342624176 & 4342624176 & 4342624176 & 1733695536 & 1074039432 & 1074039432 \\
20541185 & 20541185 & 20541185 & 20541185 & 20541185 & 4108237 & 4108237 \\
\end{array}\right)
\]

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

\[
\begin{align*}
134583963942990580792 &
17793056769908290917 \\
134583963942990580792 &
17793056769908290917 \\
\end{align*}
\]

which are positive. Finally,

\[
-\frac{1}{2}J_0H^{-1}J_0^T = \frac{4536529474419729155715368032}{12181311412062878919972215} \approx 3724.172
\]

Considering the neighbors of $S_2$:  

We have:
\[ J_0 = \left( 4342624176, 4342624176, 4342624176, 1733695536, 4342624176, 4342624176, 4342624176, 1074039432, 2955839412 \right) \]
Hence,
\[ \frac{1}{2} J_0^* H^{-1} J_0^* = \frac{7756149367472421142972629871553505755962112808}{2224068754666532551536113481786974552971205} \approx 3487.369. \]

**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 \( \frac{1}{2} J_0^* H^{-1} J_0^* \) 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 \( \vec{D} = (D_1, \ldots, D_N) \), define the maximum log-likelihood of tree \( T \) as
\[ L_T(\vec{D}) = \max_{\vec{P} \in M^{E(T)}} L_T(\vec{D}, \vec{P}). \]
where
\[ L_T(\vec{D}, \vec{P}) = \sum_i \ln \left( \Pr(D_i | T, \vec{P}) \right). \]
Consider \( \vec{D} = (D_1, \ldots, D_N) \) where each \( D_i \) is independently sampled from the mixture distribution \( \mu \).
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
\[ L_{S^*}(\vec{D}) > L_S(\vec{D}). \]
Let \( \vec{P}^* \) denote the assignment which attains the maximum for \( L_{S^*}(\mu) \), and let
\[ \alpha = \min_{\sigma \in \Omega^5} \Pr(\sigma | S^*, \vec{P}^*). \]

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 \( \delta > 0 \),
\[ \Pr \{ \text{for all } \sigma \in \Omega^5, |D(\sigma) - \mu(\sigma) N| \leq \delta N \} \geq 1 - 2 \cdot 4^\delta \exp(-2\delta^2 N) - 1 - \exp(-\Omega(N)). \]
Assuming the above holds, we then have
\[ L_S(\vec{D}) \leq N(1 - \delta) L_S(\mu). \]
And,
\[ L_{S^*}(\vec{D}) \geq L_{S^*}(\vec{P}^*)(\vec{D}) \geq N(1 - \delta) L_{S^*}(\mu) + 4^\delta N \log(\alpha) \]
Let $\gamma := L_{S^*}(\mu) - L_{S}(\mu)$. Note, Lemma 31 states that $\gamma > 0$.

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

$$L_{S^*}(\bar{D}) - L_{S}(\bar{D}) \geq N(1 - \delta)\gamma - N\gamma/10 > N\gamma/5 = \Omega(N).$$

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


