Pitfalls of Heterogeneous Processes for Phylogenetic Reconstruction

DANIEL ŠTEFANKOVIČ[1] AND ERIC VIGODA[2]

[1]*Department of Computer Science, University of Rochester, Rochester, New York 14627, USA; and Comenius University, Bratislava; E-mail: stefanko@cs.rochester.edu*

[2]*College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332, USA; E-mail: vigoda@cc.gatech.edu*

Abstract.—Different genes often have different phylogenetic histories. Even within regions having the same phylogenetic history, the mutation rates often vary. We investigate the prospects of phylogenetic reconstruction when all the characters are generated from the same tree topology, but the branch lengths vary (with possibly different tree shapes). Furthering work of Kolaczkowski and Thornton (2004, Nature 431: 980–984) and Chang (1996, Math. Biosci. 134: 189–216), we show examples where maximum likelihood (under a homogeneous model) is an inconsistent estimator of the tree. We then explore the prospects of phylogenetic inference under a heterogeneous model. In some models, there are examples where phylogenetic inference under any method is impossible—despite the fact that there is a common tree topology. In particular, there are nonidentifiable mixture distributions, i.e., multiple topologies generate identical mixture distributions. We address which evolutionary models have nonidentifiable mixture distributions and prove that the following duality theorem holds for most DNA substitution models. The model has either: (i) *nonidentifiability*—two different tree topologies can produce identical mixture distributions, and hence distinguishing between the two topologies is impossible; or (ii) *linear tests*—there exist linear tests which identify the common tree topology for character data generated by a mixture distribution. The theorem holds for models whose transition matrices can be parameterized by open sets, which includes most of the popular models, such as Tamura-Nei and Kimura's 2-parameter model. The duality theorem relies on our notion of linear tests, which are related to Lake's linear invariants. [Inconsistency of likelihood; linear invariants; Markov chain; mixture models; Monte Carlo; non-identifiability; phylogenetic invariants; phyogenetics; rate variation; tree identifiability.]

It is now clear that there is considerable heterogeneity in substitution rates within a genome (see, e.g., Hellmann, 2005; Pond and Muse, 2005). Variation in evolutionary forces is an obvious cause, but even within neutrally evolving regions heterogeneity is relevant. For phylogenetic studies based on multiple genes, heterogeneity is especially pertinent because gene trees are well known to differ. Even for studies relying on a single (ideally long) gene, substitution rates within the gene might vary due to a variety of factors. For example, recombination rates, which are known to affect substitution rates, can vary dramatically over the scale of kilobases (see Hellmann et al., 2003, 2005; McVean et al., 2004; Myers et al., 2005). The latest versions of popular phylogeny programs, such as MrBayes (Rohnquist and Huelsenbeck, 2003), now allow for partitioned models to account for varying phylogenetic histories. However, the effects of partition-heterogeneity on phylogenetic studies are still poorly understood.

Our work considers models where different subsets of sites evolve at different rates (the partitioning of the sites into homogenous subsets is not known a priori). More precisely, we consider a single tree topology generating the data, but the branch lengths can vary between sites. Thus, the character data are produced from (possibly) multiple tree shapes, though they share a common topology. This differs from models, such as the gamma rate-heterogeneity model, which assume a common tree shape across all sites.

We first look at the effects of heterogeneity on phylogenetic inference under homogeneous models. Several works, such as Kolaczkowski and Thornton (2004) and Chang (1996), have presented mixture examples where maximum likelihood (under a homogeneous model) is inconsistent; i.e., the maximum likelihood topology is different from the generating topology. We present several new, simple examples (along with new mathematical tools for their analysis) showing inconsistency. Moreover, there are examples where the maximum likelihood is achieved on multiple topologies.

In some settings, even under heterogeneous models, inference can fail. In particular, for certain models, there are mixture distributions that are nonidentifiable. More precisely, mixtures on different topologies generate identical distributions (of site patterns). Hence, it is impossible to distinguish (using any methods) between the multiple topologies that generate the distribution.

This inspires the study of which evolutionary models have nonidentifiable mixture distributions. We present a new duality theorem that says that a model either has nonidentifiable mixture distributions, or there is a simple method for reconstructing the common topology in a mixture.

Our work builds upon several theoretical works (Chang, 1996; Mossel and Vigoda, 2005) and experimental work (Kolaczkowski and Thornton, 2004) showing examples where likelihood methods fail in the presence of mixtures. Our nonidentifiability results are also related to results of Steel et al. (1994). We discuss these results in more detail when presenting related results. Our duality theorem uses a geometric viewpoint (see Kim, 2000, for a nice introduction to a geometric approach).

DEFINITION OF MIXTURE DISTRIBUTIONS

Consider an evolutionary model on a set of states Ω, such as the Jukes-Cantor model on $\Omega = \{A, C, T, G\}$). Let S denote the number of states in the data. Thus for the Jukes-Cantor and Kimura's 2-parameter model $S = 4$.

For a tree topology T, the probabilities of change across a branch (or edge) e are determined by an instantaneous rate matrix $Q(e)$ and the length of the branch $\ell(e)$. These

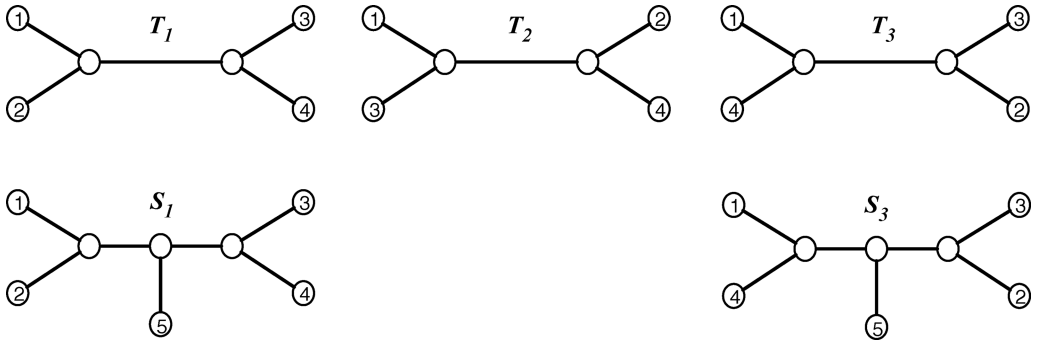

FIGURE 1. The three 4-taxon trees (T_1, T_2, T_3), and the two 5-taxon trees (S_1, S_2) of interest.

then define a transition probability matrix $w(e)$. For a tree T with a set w of transition probability matrices for each branch of T, let $\mu(T, w)$ denote the corresponding distribution on the labelings of leaves (or taxa) of T by states in Ω.

In our framework, there is a single tree topology T and a set of k different processes. The processes are defined by a collection of k sets of transition probability matrices $w = (w_1, \ldots, w_k)$ where, for each i, w_i assigns each edge of T a transition matrix. In other words, the ith tree is defined by (T, w_i). The proportion of sites contributed by the ith tree (which is also the probability that a randomly sampled site is generated under $\mu(T, w_i)$) is denoted p_i. More precisely, we consider the mixture distribution:

$$\mu(T, w) = \sum_{i=1}^{k} p_i \mu(T, w_i).$$

Thus, with probability p_i we generate a character according to $\mu(T, w_i)$. Note the tree topology is the same for all the distributions in the mixture.

Let M^k denote the above class of mixture distributions of size k. We will often consider a uniform mixture of several trees. Thus, let U^k denote the class of mixture distributions of size k where $p_1 = p_2 = \cdots = p_k = 1/k$. In many cases we will consider examples from U^2 (i.e., a uniform mixture of two trees). For each of our results we will detail the precise setting.

Our focus in this paper is on the three 4-taxon trees T_1, T_2, and T_3 and the two specific 5-taxon trees S_1 and S_3 depicted in Figure 1.

The surprising properties of maximum likelihood that we study in this paper are best identified in the simplest models. Hence we often consider examples in the binary CFN (Cavender-Farris-Neyman) model and the Jukes-Cantor model. The CFN model is the two-state version of the Jukes-Cantor model. Thus in the CFN model $S = 2$.

The Jukes-Cantor and CFN models have a single parameter for the substitution rate for each branch. For these models we can use the probability of a change of state across a branch as the branch length parameter (instead of the usual parameterization in which the branch length is the expected number of changes per site).

For branch e, the probability is denoted $p(e)$. Note that for the CFN model the branch lengths satisfy $0 < p(e) < 1/2$, whereas for the the Jukes-Cantor model the branch lengths satisfy $0 < p(e) < 1/4$.

PHASE TRANSITION FOR INCONSISTENCY OF MAXIMUM LIKELIHOOD

We consider a class of 4-taxon mixture examples where maximum likelihood has intriguing properties. Figure 2 presents the class of mixture examples we study for the CFN and Jukes-Cantor models. We take a uniform mixture of the two trees. We will use μ to denote the mixture distribution. Note, the trees have a common topology T_1 and only differ in their branch probabilities. In our notation, the example is in U^2, which is the class of uniform mixtures of size 2.

The terminal branch probabilities are a function of the parameters x and C. The parameter C is any valid branch probability, thus $0 \leq C \leq 1/S$. The parameter x controls the variation of the branch lengths between the two trees. We need that $C + x$ and $C - x$ are valid branch probabilities, thus we require $0 \leq x \leq \min\{C, \frac{1}{S} - C\}$.

When $x = 0$ the two trees are identical. The internal branch probability is defined by a third parameter α where $0 \leq \alpha \leq 1/S$. We will study the properties of the likelihood function as α varies.

This class of examples was studied by Kolaczkowski and Thornton (2004). Using computational simulations, they showed that when α is sufficiently small, maximum likelihood (over M^1, i.e., under a homogeneous model) is inconsistent. Also, under a violated model, maximum parsimony performed better than maximum likelihood for some range of α. We delve into the properties of maximum likelihood on these examples. Our aim is to

FIGURE 2. Mixture distribution on tree T_1. C is a parameter that can take any values $0 < C < 1/2$ in the CFN model and $0 < C < 1/4$ in the Jukes-Cantor model; x is a parameter controlling the variation between the two trees; and α is the internal branch length.

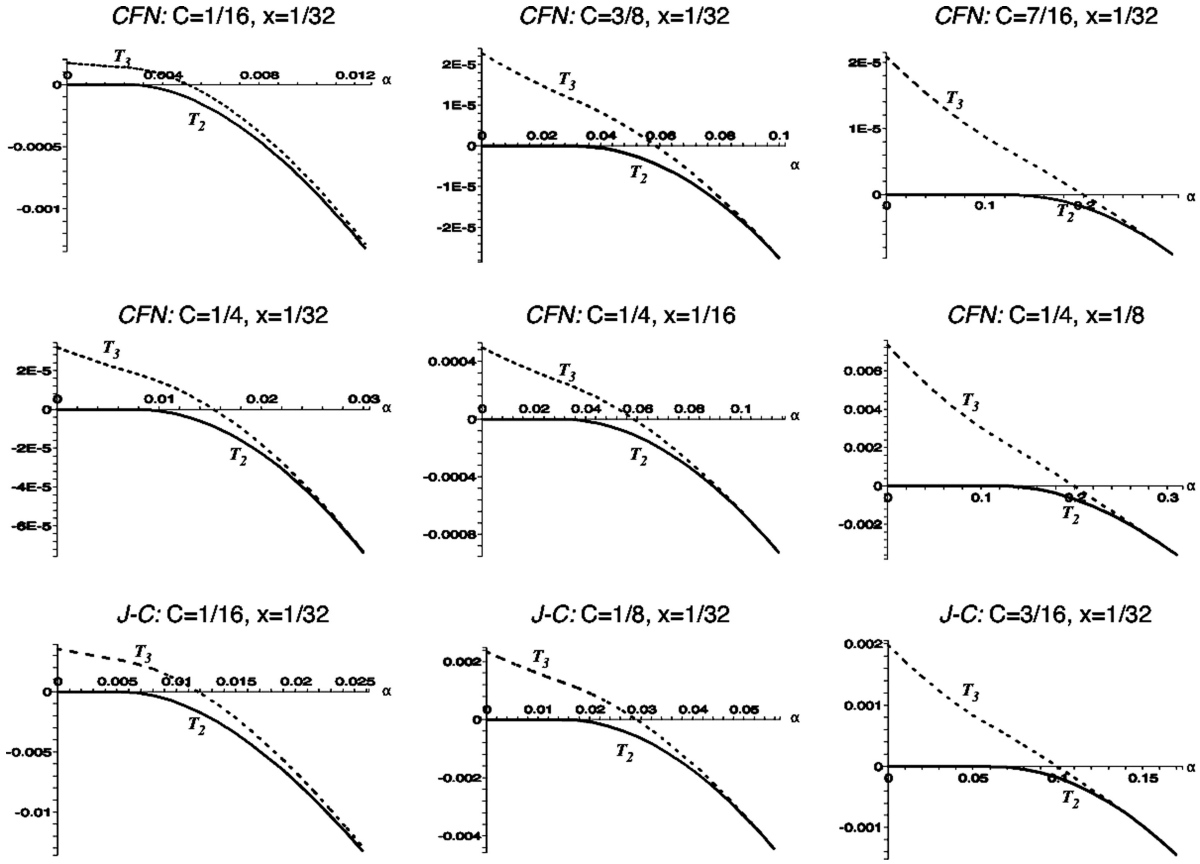

FIGURE 3. For the CFN and Jukes-Cantor model, we consider the mixture distribution μ on tree T_1, which is defined in Figure 2. C, x, and α are parameters of this mixture example from Figure 2. Recall $\mathcal{L}_S(\mu)$ is the maximum expected log-likelihood of μ on tree S. For various choices of C and x, we plot $\mathcal{L}_{T_3}(\mu) - \mathcal{L}_{T_1}(\mu)$ in the curve marked T_3, and $\mathcal{L}_{T_2}(\mu) - \mathcal{L}_{T_1}(\mu)$ in the curve marked T_2. The x-axis is α, which is the choice of internal branch probability. Note, when the internal branch probability α is small, the likelihood of T_3 is larger than that of the generating tree T_1. When the curve for T_3 hits the x-axis (at $\alpha = \alpha_c$), maximum likelihood is ambiguous; i.e., $\mathcal{L}_{T_3}(\mu) = \mathcal{L}_{T_1}(\mu)$.

formally establish a more precise picture of the behavior of maximum likelihood, and to devise mathematical tools for the analysis of likelihood methods. We will also discover interesting new properties, and further questions will arise that we will explore later in the paper.

In Figure 3 we study maximum likelihood of the three 4-taxon trees on μ for the CFN and Jukes-Cantor models. For each tree we look at the maximum expected log-likelihood of a homogeneous model. Thus likelihood is maximized over M^1, whereas the character data is generated over U^2. More precisely, the maximum expected log-likelihood is defined as:

$$\mathcal{L}_T(\mu) = \max_w \mathcal{L}_{T,w}(\mu),$$

where

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

w is an assignment of a branch length to each branch of the tree topology T, and $\mathcal{P}$ is the set of all possible data patterns (i.e., $\mathcal{P} = \Omega^n$). Thus, for tree T, we are finding the single set w of branch lengths that maximizes the sum over data patterns of the frequency of the pattern multiplied by natural log of the probability the pattern is produced by the proposed tree (T, w).

The character data are generated from a mixture distribution on tree T_1, and thus one would presume that tree T_1 has the maximum likelihood. However, the behavior of the likelihood function as α varies has a phase transition at the critical point $\alpha = \alpha_c$ (which is a function of C and x) as depicted in Figure 3.

For $\alpha > \alpha_c$, tree T_1 is the maximum likelihood tree. However, this changes at $\alpha = \alpha_c$. When $\alpha < \alpha_c$ the maximum likelihood tree is T_3; thus, maximum likelihood (under a homogeneous model) is an inconsistent estimator of the phylogeny. We prove the inconsistency holds for all choices of C, all $\alpha < \alpha_c$, and for all x sufficiently small. (We expect the result holds for all x.) Our proof uses a new approach which we outline in our methodology section. A detailed proof is included in the supplemental material available at http://systematicbiology.org (for a detailed statement, see Supplemental Material, Theorem 1 for the Jukes-Cantor model and Theorem 2 for the CFN model).

At the critical point $\alpha = \alpha_c$, there are multiple topologies achieving the maximum likelihood. In particular, we have $\mathcal{L}_{T_3} = \mathcal{L}_{T_1}$. Hence, likelihood cannot distinguish between these two tree topologies. We prove this ambiguity of maximum likelihood holds for all choices of C and x in the CFN model. In the Jukes-Cantor model, we prove that maximum likelihood is ambiguous at α_c for all choices of C and all x sufficiently small.

In the CFN model we prove that

$$\alpha_c = \alpha_c(C, x) := \frac{x^2}{\frac{1}{4} - C + C^2 + x^2}.$$

The simplest case is $C = 1/4$, in which case $\alpha_c = \frac{16x^2}{1+16x^2} \approx 16x^2$ for small x. Note, $\alpha_c \to 0$ as $x \to 0$. In other words, as the heterogeneity decreases (i.e., $x \to 0$), the inconsistency of likelihood arises from a shorter internal edge (i.e., $\alpha_c \to 0$). This is necessary because at $x = 0$ the two trees are the same, and then ambiguity of likelihood only occurs for $\alpha_c = 0$. In the Jukes-Cantor model, we prove there exists some $\alpha_c > 0$, but we do not know its exact value.

In contrast to our results, we note that Chang (1996) proved inconsistency of maximum likelihood on a different class of mixture examples. In particular, his examples included invariable sites.

Even more intriguing ambiguity properties occur at the critical point $\alpha = \alpha_c$, which we explore now. Later in the paper (section 5-Taxon Mixtures and Slow Mixing of MCMC Methods) we discuss the implications of the above likelihood results to Markov chain Monte Carlo algorithms for sampling from the posterior distribution.

NONIDENTIFIABILITY AT THE CRITICAL POINT $\alpha = \alpha_c$

We now explore the above examples at the critical point α_c. We address whether any method (even using a heterogeneous model) can infer the common generating topology.

In the CFN model, at $\alpha = \alpha_c$, not only is maximum likelihood ambiguous, but the distribution itself is nonidentifiable. In particular, there is a mixture distribution μ' on tree T_3 that is identical to the distribution μ on T_1. Consequently, no phylogenetic reconstruction method can distinguish between the two tree topologies. Figure 4 presents the mixture on tree T_3 where the resulting distribution μ' satistifies $\mu' = \mu$. (Note, μ' and μ are both in the class U^2.) Given characters sampled from μ (or equivalently μ'), it is impossible to determine if the character data are generated from topology T_1 or T_3. No phylogenetic methods can distinguish between the two topologies. This nonidentifiability holds for any choice of the parameters C and x. See Theorem 5 in Supplemental Material for a precise statement of the nonidentifiability result and an extension of the result to a nonuniform mixture of two trees.

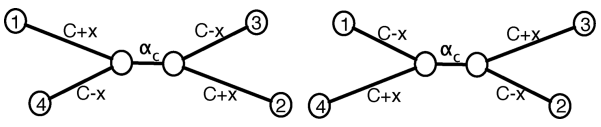

FIGURE 4. Mixture distribution on tree T_3, which is identical to the distribution on tree T_1 in Figure 2 for the CFN model. The parameters C and x are the same parameters as defined in Figure 2. The parameter α_c is a setting of the parameter α from Figure 2 where nonidentifiability holds in the CFN model.

A nonidentifiable mixture distribution was previously shown for the CFN model by Steel et al. (1994); however, their result was nonconstructive (i.e., the existence of such a mixture was proven without constructing a specific example or determining the number of trees in the mixtures). However, their result had the more appealing feature that the set of trees in the mixture were scalings of each other (i.e., the tree shape was preserved).

For the Jukes-Cantor model, at $\alpha = \alpha_c$, maximum likelihood is ambiguous. However, unlike the case for the CFN model, the mixture distribution μ at the critical point α_c is identifiable; i.e., there is no mixture distribution on another tree topology which is identical to μ. In fact, we prove there are no nonidentifiable mixture distributions in the Jukes-Cantor model. This raises the general question: which models have nonidentifiable mixture distributions? Our duality theorem addresses this question.

GEOMETRIC INTUITION

Before presenting our duality theorem, it is useful to look at nonidentifiable mixture distributions from a geometric perspective. The geometric viewpoint presented here is closely related to the work of Kim (2000) (and we encourage the interested reader to refer to that work for useful illustrations of some concepts presented in this section). This geometric approach is especially useful for the proof of our duality theorem.

Consider the 2-state CFN model with a tree topology T on 4 taxa and a set w of transition matrices for the branches. This defines a distribution $\mu(T, w)$ on assignments of $\{0, 1\}$ to the 4 taxa. The distribution $\mu(T, w)$ defines a point $z \in \mathbb{R}^{2^4}$ where $z = (z_1, \ldots, z_{2^4})$ and $z_1 = \mu(0000), z_2 = \mu(0001), z_3 = \mu(0010), \ldots, z_{2^4} = \mu(1111)$. (In other words, the first coordinate of z is defined by the probability of all taxa getting assigned 0, and so on for the 2^4 possible assignments to the 4 taxa.) Similarly, for a 4-state model, a distribution $\mu(T, w)$ defines a point in 4^4 dimensional space.

Let D_1 denote the set of points corresponding to distributions $\mu(T_1, w)$ for the 4-taxon tree T_1. Similarly, define D_2 for T_2, and D_3 for T_3. Figures 5a and b are different illustrations of what these sets of points might look like for the three 4-taxon trees. (These are 2-dimensional representations of a high dimensional set, hence they are only for illustration purposes.) These curves are referred to as the "model manifold" by Kim (2000).

The set of mixture distributions obtainable from topology T_1 is the set of convex combinations of points in D_1, which we denote as the set H_1. Note that these are the set of all mixture distributions without any constraint on the size of the mixture (parameter k) and for any choice of the distribution on the trees (parameters $p_1, \ldots, p_k$).

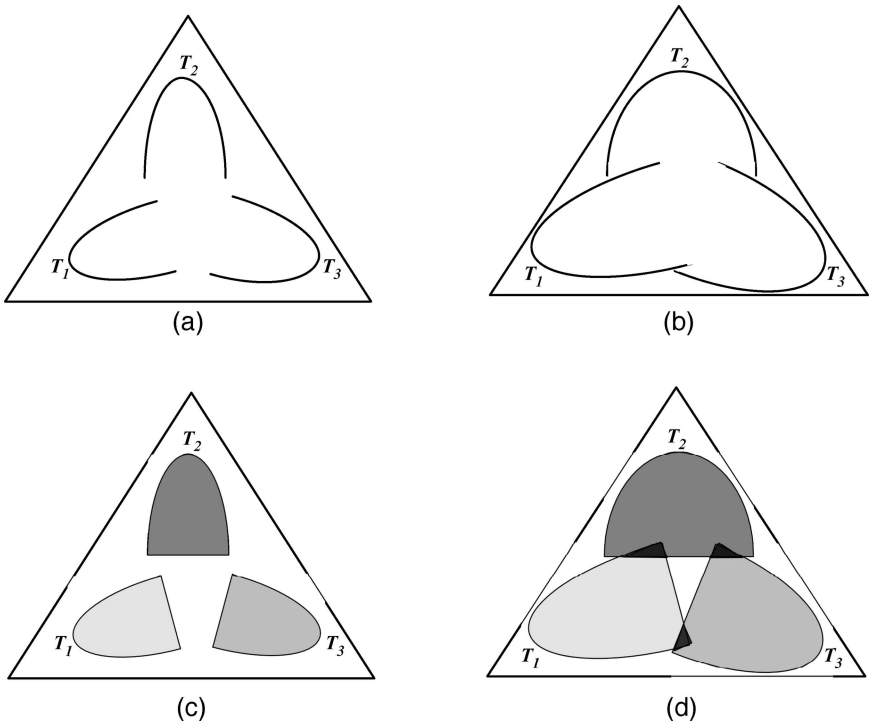

FIGURE 5. (a and b) Illustrations of the distributions obtainable from a homogenous model. The corresponding set of mixture distributions are shown in (c) and (d). The points in the intersection of two sets in (d) correspond to nonidentifiable mixture distribution.

Similarly define the sets H_2 and H_3 for trees T_2 and T_3, respectively. By definition, the sets H_1, H_2, and H_3 are convex sets. Figures 5c and d show the sets H_1, H_2, and H_3 for the examples in Figures 5a and b.

In Figure 5d the sets intersect. Thus there are points z that lie in H_1 and H_3. That means the distribution defined by z is obtainable from a mixture on topology T_1 and also from a mixture on topology T_3. This corresponds to a *nonidentifiable mixture distribution*, which we studied in the section Nonidentifiability at the Critical Point, for the CFN model. If H_1 and H_3 intersect, it is impossible (using any methods) to separate the sets H_1 and H_3.

Do these sets H_1 and H_3 overlap in the commonly used evolutionary models? This is the focus of our duality theorem presented in the next section. We prove that either these sets H_1 and H_3 overlap (and there is nonidentifiability), or there is a simple way to separate the sets. In particular, in the latter case there is a hyperplane that strictly separates the sets. By strictly separating, we mean that no point in H_1 or H_3 is on the hyperplane, and H_1 lies on one side while H_3 lies on the other side. A strictly separating hyperplane implies a method, which we refer to as a linear test, for determining whether the mixture distribution is in H_1 or H_3.

Consequently, we can address for many models whether there is nonidentifiability by proving whether or not there is a strictly separating hyperplane. For the symmetric models (CFN, Jukes-Cantor, and Kimura's 2- and 3-parameter models) we address the existence of nonidentifiable mixture distributions in the section Implications of the Duality Theorem.

The duality theorem relies on intuition from convex programming, which is a central topic in operations research and theoretical computer science. Convex programming refers to the optimization of a linear function over a convex set. The development of polynomial-time algorithms for convex programming (e.g., ellipsoid methods) relied on convex programming duality. One view of convex programming duality says that for any 2-convex sets, either the sets have nonempty intersection or there is a separating hyperplane. In contrast to the above perspective of a strictly separating hyperplane, in this setting the sets might both intersect the hyperplane and then the hyperplane is not useful for our purposes. Using properties of the sets that can arise from evolutionary models, we prove that a separating hyperplane is in fact a strictly separating hyperplane if the sets do not intersect.

NEW DUALITY THEOREM: NON-IDENTIFIABLE MIXTURES OR LINEAR TESTS

We begin by formally defining nonidentifiable mixture distributions, and then present our duality theorem. Recall the formal definition of a mixture distribution in the section Definition of Mixture Distributions, defined by a topology T, a collection of assignments of transition matrices $w = (w_1, \ldots, w_k)$, and a distribution $p_1, \ldots, p_k$ on the k trees.

We say a model has a *nonidentifiable mixture distribution* if there exists a collection of transition matrices $w = (w_1, \ldots, w_k)$ and distribution $p_1, \ldots, p_k$, such that there is another tree topology $T' \neq T$, a collection

$w' = (w'_1, \ldots, w'_k)$ and a distribution $p'_1, \ldots, p'_k$ such that:

$$\mu(T, w) = \mu(T', w')$$

In other words,

$$\sum_{i=1}^{k} p_i \mu(T, w_i) = \sum_{i=1}^{k} p'_i \mu(T', w'_i).$$

The mixture distributions (which are on different topologies) are identical. Hence, even with unlimited characters, it is impossible to distinguish these two distributions and we cannot infer which of the topologies T or T' is correct. If the above holds, we say the model has nonidentifiable mixture distributions.

Determining which evolutionary models have nonidentifiable mixture distributions relies on a new *duality theorem*, which relates to the evolutionary parsimony method of Lake (1987). (Lake's method is now classified as a linear invariant; see Pachter and Sturmfels, 2005, for an introduction to invariants.)

Recall from section Geometric Inuition the definition of the set H_1 as the points (corresponding to the set of mixture distributions) obtainable from topology T_1. Similarly we have H_3 for topology T_3. If the convex sets H_1 and H_3 intersect, then there is a mixture distribution obtainable by both topologies, and hence the model has nonidentifiable mixture distributions. We prove (under certain assumptions on the model) that if the sets do not intersect, then there is a hyperplane that strictly separates the sets.

The existence of a strictly separating hyperplane immediately yields what we refer to as a linear test. A hyperplane is defined by a vector y. If y is a strictly separating hyperplane for the sets H_1 and H_3, then $y^T z < 0$ for all $z \in H_1$ and $y^T u > 0$ for all $u \in H_3$. We define a linear test as a vector y which defines a strictly separating hyperplane. The existence of such a linear test for T_1 and T_3 immediately yields a linear test for any pair of 4-taxon trees. It suffices to consider trees with 4-taxon, because the full topology can be inferred from all 4-taxon subtrees (Bandelt and Dress, 1986).

Our duality theorem holds for models whose transition matrices can be *parameterized by an open set*. This means that there is an open set W of vectors in $\mathbb{R}^d$ (for some d) and the transition probabilities for the model (i.e., the entries of the transition probability matrices w_e) can be expressed as a set of multilinear polynomials with domain W. This holds for most of the popular models, including Tamura-Nei, HKY, Felsenstein, Kimura's 2- and 3-parameter, Jukes-Cantor, and CFN models. (See Felsenstein, 2004, for an introduction to these models.) For all of the reversible models whose transition rates can be solved analytically, it turns out that they can be parameterized by an open set. We demonstrate this in the supplemental material for the Tamura-Nei model.

This assumption on the model implies that the transition probabilities can be expressed as multilinear polynomials in the parameters of the model. Using this form of the model we can prove the following duality theorem.

For every phylogenetic model whose transition matrices can be parameterized by an open set, we prove that exactly one of the following holds:

Nonidentifiable: There is a nonidentifiable mixture distribution on 4-taxon trees. Thus, in the worst case, it is impossible to infer the common tree topology from a mixture distribution, because there are multiple 4-taxon tree topologies that generate identical mixture distributions.

Linear test: There is a linear test that separates any pair of 4-taxon trees. This implies an easy method for reconstructing the common topology from a mixture distribution. Note, the test is determining the generating tree topology, but it has no connections to (or implications for) likelihood methods.

The duality theorem uses a classical result of Bandelt and Dress (1986), which implies that if a model has nonidentifiable mixture distributions, then there is an example with just 4-taxon. This simplifies the search for nonidentifiable mixture distributions, or for proving they do not exist.

Implications of the Duality Theorem

Linear tests are closely related to linear invariants, such as Lake's method. A linear invariant is a hyperplane that contains the entire set H_1 and does not intersect H_3. A linear invariant can be transformed into a linear test. Consequently, Lake's linear invariants for the Jukes-Cantor and Kimura's 2-parameter model give a linear test for these models. Therefore, there are no nonidentifiable mixture distributions in the Jukes-Cantor and Kimura's 2-parameter models. This is in contrast to the example from the section Nonidentifiability at the Critical Point of a nonidentifiable mixture distribution for the CFN model.

Whereas every linear invariant can be transformed into a linear test, the reverse implication is not necessarily true. Thus linear tests are potentially more powerful than linear invariants.

For Kimura's 3-parameter model we show that there are nonidentifiable mixture distributions. We prove this result by showing there is no linear test for this model, and then the duality theorem implies there is a nonidentifiable mixture distribution. This proof is nonconstructive, thus we prove there exist nonidentifiable examples without providing an explicit example. In many cases nonconstructive proofs are substantially simpler than constructive proofs. On the other hand, because of the nonconstructive nature of our proof, we do not have bounds on the size of the mixture (i.e., parameter k). Because Kimura's 3-parameter model is a special case of any super model such as the general time-reversible model (GTR), some examples of GTR models will be nonidentifiable in this setting.

It is, however, unknown at this point how large the set of nonidentifiable mixture distributions is in the CFN or

K3 models. Earlier work of Allman and Rhodes (2006) proved that, when restricting attention to mixtures of size 3 or smaller, the set of nonidentifiable mixture distributions form an insignificant portion of all mixtures (more precisely, the set of nonidentifiable mixtures has measure zero). For larger mixtures, it is an interesting question whether the nonidentifiable mixtures are a significant portion of all mixtures.

In future work we hope to address the existence of nonidentifiable mixture distributions in models such as Tamura-Nei and HKY models. The models considered in this paper are symmetric, which is utilized in the proofs on the existence of linear tests.

5-Taxon Mixtures and Slow Mixing of MCMC Methods

In this section we address the implications of some of our earlier examples for Markov chain Monte Carlo (MCMC) algorithms. In particular, we consider extensions of the 4-taxon mixture examples, which showed inconsistency of maximum likelihood in the CFN and Jukes-Cantor models.

Figure 6 presents a 5-taxon mixture example for the Jukes-Cantor model. (This example is related to the earlier 4-taxon examples with $C = 1/8$ and the internal edge approximately α_c.)

We study maximum likelihood of the 15 5-taxon trees, and analyze the likelihood landscape with respect to nearest-neighbor interchanges (NNI). For sufficiently small x, we prove that the two trees S_1 and S_3 (depicted in Fig. 1) are local maximum with respect to NNI transitions. In particular, we prove that each of these trees has larger expected log-likelihood than any of the 8 5-taxon trees that are connected to S_1 or S_3 by an NNI transition. Thus, in the tree space defined by NNI transitions, S_1 and S_3 are local maxima separated by a "valley" (trees with lower expected likelihood). As the number of characters increases, the valley becomes deeper. Hence, Markov chain Monte Carlo algorithms using NNI transitions take longer to escape from a local maxima as the number of characters is increased. Consequently, MCMC algorithms with NNI transitions converge exponentially slowly (in the number of characters) to the posterior distribution.

These results improve recent work of Mossel and Vigoda (2005), who proved similar results for examples on a mixture of two different tree topologies. Note, in contrast to the work of Mossel and Vigoda, in our setting there is a correct topology. We expect our MCMC

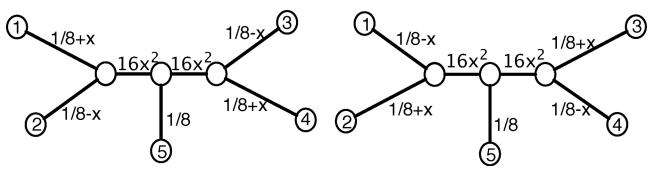

FIGURE 6. 5-Taxon example for Jukes-Cantor model where Markov chains with NNI transitions are exponentially slow. The quantity x is a parameter measuring the variation between the trees, which needs to be sufficiently small for the slow-mixing result to hold.

results to extend (as in Mossel and Vigoda, 2005) to other transitions such as subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR). Note, our results do not imply anything about the convergence rate of Metropolis-coupled Markov chain Monte Carlo (MC^3), which is used in MrBayes (Rohnquist and Huelsenbeck, 2003), but we hope that the mathematical tools we present will be useful in future theoretical work on the convergence properties of MrBayes.

Nonuniform Mixtures

Our examples for the inconsistency of maximum likelihood (section Phase Transition for Inconsistency of Maximum Likelihood), the nonidentifiability for the CFN model (section Nonidentifiability at the Critical Point), and the slow convergence of MCMC algorithms (section 5-Taxon Mixtures and Slow Mixing of MCMC Methods) use a uniform mixture of two trees. A uniform mixture simplifies the mathematical computations but is not an essential feature. There exist nonuniform mixtures with the same phenomenon.

For example, we can achieve nonidentifiablity in the CFN model with a nonuniform mixture by taking an appropriate modification of our earlier example. In particular, by allowing the branch length of the internal edge to be different between the two trees and choosing these lengths appropriately (as a function of the mixing parameter p), we obtain a nonidentifiable mixture distribution. This construction is detailed in Supplementary Material (see Theorem 5).

Methodology

Proving results on maximum likelihood methods are difficult. Hence, our proof method for inconsistency of maximum likelihood on the mixture examples of Figure 2 may be a useful tool for certain analyses. Note our results are for the region $x > 0$. The proof uses properties of the $x = 0$ case. When $x = 0$ the two trees in the mixture are identical, hence the character data are generated from a pure (i.e., nonmixture) distribution. Moreover, for $x = 0$ we have $\alpha_c = 0$ (i.e., the internal branch length is zero), hence the distribution can be generated from any topology. Therefore, for $x = 0$ we can easily determine the assignments of branch probabilities that maximizes the likelihood. When x is small and non-zero, we consider the Taylor expansion of the likelihood function. Consequently, we obtain the likelihood as a function of the Jacobian and Hessian of the likelihood function.

In Supplementary Material (Lemma 3), we state the main technical lemma, which is proved in Štefankovič and Vigoda (2006), and present an extension of this result (Lemma 4) tailored to the purposes of this paper. We then prove the maximum likelihood results for the 4-taxon mixture examples in the CFN and Jukes-Cantor models. The stated results for MCMC methods on the 5-taxon examples are proved in Štefankovič and Vigoda (2006).

The proof of our duality theorem is related to convex programming duality as noted earlier. For a pair of convex sets, such as H_1 and H_3 defined in the section

Geometric Intuition, convex programming duality implies that if the sets do not intersect, then there is a hyperplane separating the sets. However, the hyperplane may intersect both sets. In particular, for the hyperplane defined by the vector y, we may have $y^T z \leq 0$ for all $z \in H_1$ and $y^T u \geq 0$ for all $u \in H_3$.

We prove that there is in fact a strictly separating hyperplane y. Recall, if it is strictly separating, it implies that $y^T z < 0$ for all $z \in H_1$ and $y^T u > 0$ for all $u \in H_3$. To obtain this we consider a separating hyperplane y and suppose that there is a point $z \in H_1$ where $y^T z = 0$. Using the fact that the set H_1 is the convex hull of a set of multilinear polynomials (with an open set as its domain), we can then argue that for some z' very close to z we have $z' \in H_1$ and $y^T z' > 0$, which contradicts the assumption that y is a separating hyperplane. The details of the proof are contained in Štefankovič and Vigoda (2006).

Concluding Remarks

A nice aspect of our duality theorem is that if a model has no linear test distinguishing 4-taxon trees, then there are ambiguous mixture distributions. For many models, such as Kimura's 3-parameter model, this simplifies the proof that the model has ambiguous mixture distributions. In particular, certain symmetries of the model can be used to narrow the space of possible tests.

We expect our proof approach for analyzing maximum likelihood will be useful for related problems. The proofs rely on the internal edge probabilities approaching zero in the limit. This is a consequence of the proof methodology that uses the first few terms of the Taylor expansion. It appears that all known proof techniques for analyzing maximum likelihood require at least some subset (or all) of the edge probabilities go to zero (e.g., see Chang, 1996; Mossel and Vigoda, 2005). Avoiding these asymptotics seems to be a difficult open question. We expect these results to hold for a much larger class of examples, with larger internal branch lengths. This is supported to some extent by the results of computational experiments reported in Kolaczkowski and Thornton (2004). New mathematical tools will be needed for such extensions.

Acknowledgments

The authors thank the referees, especially Mark Holder, and the associate editor Jack Sullivan for many helpful comments. We also thank Elchanan Mossel for useful discussions. EV's research was supported in part by NSF grant CCF-0455666.

References

Allman, E. S. and J. A. Rhodes. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J. Comp. Biol. 13:1101–1113.

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

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

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

Hellmann, I., K. Prüfer, H. Ji, M. C. Zody, S. Pääbo, and S. E. Ptak. 2005. Why do human diversity levels vary at a megabase scale? Genome Research. 15:1222–1231.

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

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

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

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

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

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Pachter, L., and B. Sturmfels. 2005. Algebraic statistics for computational biology. Cambridge University Press, Cambridge, UK.

Pond, S. L., and S. V. Muse. 2005. Site-to-site variation in synonymous substitution rates. Mol. Biol. Evol. 22:2375–2385.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

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

Štefankovič, D., and E. Vigoda. 2006. Phylogeny of mixture models: Robustness of maximum likelihood and nonidentifiable distributions. To appear in J. Comp. Biol.

Supplementary Material

1 Tamura-Nei can be parameterized by an open set

For the Tamura-Nei model, we will show how the model can be parameterized by an open set. The model has rates α_0, α_1 and β and time t. For $i, j \in \{0, 1, 2, 3\}$, the transition probabilities are the following (see (13.11) in Felsenstein, 2004):

$$\begin{aligned}\mathbf{Pr}(j \mid i, t) &= \exp[-(\alpha_\ell + \beta)t]\delta(i = j) \\ &+ \exp(-\beta t)[1 - \exp(-\alpha_\ell t)]\left[\frac{\pi(j)\epsilon(i, j)}{\sum_k \epsilon(j, k)\pi(k)}\right] \\ &+ [1 - \exp(-\beta t)]\pi(j),\end{aligned}$$

where $\ell = \lfloor i/2 \rfloor$, δ is the standard Kronecker delta function, and $\epsilon(a, b)$ is an indicator function that is 1 if $a, b \in \{0, 1\}$ or $a, b \in \{2, 3\}$, and 0 otherwise. Setting

$$x_0 = \exp(-\alpha_0 t), \quad x_1 = \exp(-\alpha_1 t), \quad y = \exp(-\beta t),$$

we have

$$\begin{aligned}\mathbf{Pr}(j \mid i, t) &= x_\ell y \delta(i = j) + y(1 - x_\ell)\left[\frac{\pi(j)\epsilon(i, j)}{\sum_k \epsilon(j, k)\pi(k)}\right] \\ &+ (1 - y)\pi(j).\end{aligned}$$

Hence, the transition probabilities can be expressed as multi-linear polynomials in x_0, x_1, and y.

2 Main Theorems about Maximum Likelihood

Here we prove the following theorem about the Jukes-Cantor model.

Theorem 1. *Consider the Jukes-Cantor model. For the two trees in Figure 2, let μ denote the mixture distribution defined by a uniform mixture (i.e., $p_1 = p_2$) of these two trees. Note μ is from the class U^2 (the set of uniform mixtures of size 2). Recall the maximum likelihood $\mathcal{L}_\mu(T)$ is over a homogenous model (i.e., the class M^1).*

For all $C \in (0, 1/4)$, there exists x_0, for all $x < x_0$, there exists $\alpha_c = \alpha_c(C, x) \in (0, 1/4)$ such that:

1. *For $\alpha = \alpha_c$, the maximum likelihood of T_1 and T_3 are the same; i.e.,*

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu).$$

2. *For all $\alpha < \alpha_c$, T_3 is the maximum-likelihood tree; i.e.,*

$$\mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_1}(\mu) \text{ and } \mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_2}(\mu).$$

For the CFN model we can prove ambiguity of likelihood at the critical point α_c for all x, and the maximum likelihood is on the "wrong tree" when $\alpha < \alpha_c$ and x is sufficiently small. Here is the formal statement of the theorem.

Theorem 2. *Consider the CFN model. For the two trees in Figure 2, let μ denote the mixture distribution defined by a uniform mixture (i.e., $p_1 = p_2$) of these two trees. Note μ is from the class U^2 (the set of uniform mixtures of size 2). Recall the maximum likelihood $\mathcal{L}_\mu(T)$ is over a homogenous model (i.e., the class M^1).*

For all $C \in (0, 1/2)$, there exists x_0, for all $x < x_0$, there exists $\alpha_c = \alpha_c(C, x) \in (0, 1/2)$ such that:

1. *For $\alpha = \alpha_c$, the maximum likelihood of T_1 and T_3 are the same; i.e.,*

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu).$$

2. *For all $\alpha < \alpha_c$, T_3 is the maximum-likelihood tree; i.e.,*

$$\mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_1}(\mu) \text{ and } \mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_2}(\mu).$$

Remark 3. *For the CFN model, we in fact prove part (1) of Theorem 2 holds for all $x \in (0, \min\{C, 1/2 - C\})$. This follows from Theorem 5.*

Remark 4. *We prove there exists at least one critical point α_c where part 1 holds (i.e., $\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu)$), but there may be many such points α_c. In Section 2.1 we prove that*

$$\alpha_c = \frac{x^2}{\frac{1}{4} - C - C^2 + x^2} \tag{1}$$

is such a critical point. (In fact, at that particular critical point the distribution is also nonidentifiable.) Since there may be multiple critical points, Part 2 holds with respect to the smallest critical α_c, which we cannot determine the exact value of.

We expect that the α_c in (1) is the unique critical point. However, our proof methodology only uses the highest order terms of the likelihood function. Hence, it is not detailed enough to prove the uniqueness of α_c.

Our proofs of the above theorems rely on technical tools developed in Štefankovič and Vigoda (2006). We first describe the relevant technical lemmas in Section 2.2, and then prove Theorems 1 and 2 in Sections 2.4 and 2.3, respectively. Before going into the proofs we formally present the nonidentifiability result at the critical point.

2.1 Nonidentifiabile Mixture for CFN at α_c

We next present the proof of a nonidentifiable mixture distribution in the CFN model. The result is a generalization of the example, and also shows there exists nonuniform mixtures that are nonidentifiable. The theorem implies Part 1 of Theorem 2 as a special case.

In the following we describe the branch lengths on a 4-taxon tree T as a 5-dimensional vector w. For $1 \le i \le 4$, the ith coordinate of w is the branch length of the edge incident to the leaf labeled i. The final coordinate of w is the branch length of the internal edge of T.

In the following theorem, p is the mixing parameter. When $p \neq 1/2$ (i.e., it is not a uniform mixture), then the branch length of the internal edge will differ between the two trees.

Now we can formally state the theorem.

Theorem 5. *Consider the CFN model. For any $0 < a, b < 1/2$ and $0 < p \le 1/2$, let*

$$w = \frac{1}{2} - (a, b, b, a, \gamma) \text{ and}$$
$$w' = \frac{1}{2} - (b, a, a, b, \delta),$$

where

$$\gamma = \eta/p,$$
$$\delta = \eta/(1-p), \text{ and}$$
$$\eta = \frac{ab}{2(a^2+b^2)}.$$

Consider the following mixture distribution, which is in the class M^2:

$$\mu = p\mu(T_1, w) + (1-p)\mu(T_1, w')$$

The distribution μ is invariant under the swapping of leaves 1 and 3. In particular, for the mixture distribution (which is also in M^2)

$$\hat{\mu} = p\mu(T_3, w) + (1-p)\mu(T_3, w')$$

we have

$$\mu = \hat{\mu}.$$

Hence, whenever γ and δ satisfy $0 < \gamma, \delta < 1/2$ then μ and $\hat{\mu}$ are valid distributions and the topology is nonidentifiable (since there is a mixture μ on T_1 that is identical to a distribution $\hat{\mu}$ on T_3). Note for every $0 < p \le 1/2$, there exists a and b where γ and δ are valid, and hence the above construction defines a nonidentifiable mixture distribution.

Since the distribution μ is invariant under the relabeling of leaves 1 and 3, likelihood maximized over M^1 is the same for topology T_1 and T_3; i.e.,

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu)$$

Part 1 of Theorem 2 is the special case when $p = 1/2$, and a and b are rephrased as $a = 1/2 - (C + x)$ and $b = 1/2 - (C - x)$. Note, when $p = 1/2$, $\delta = \gamma$ and thus the internal edge has the same branch length in the two trees.

In Štefankovič and Vigoda (2006) (see Proposition 17) we give a relatively simple proof of the above theorem. That proof relies on some symmetry properties of the model which are introduced as a precursor to the proof of the duality theorem. Since we have not introduced these properties here, we instead present a more "brute-force" style proof.

Proof. Swapping the leaves 1 and 3 changes T_1 into T_3. Let σ denote an assignment of labels from $\{0, 1\}$ to the leaves, i.e., $\sigma : \{1, 2, 3, 4\} \to \{0, 1\}$. Let $\hat{\sigma}$ denote the assignment obtained from σ with the assignment for leaves 1 and 3 swapped (i.e., $\hat{\sigma}(1) = \sigma(3)$, $\hat{\sigma}(3) = \sigma(1)$ and $\hat{\sigma}(i) = \sigma(i)$ for $i = 2, 4$).

An assignment σ has the same probability in (T_1, w) as the assignment $\hat{\sigma}$ in (T_3, w). Hence,

$$\mu(\sigma) = \hat{\mu}(\hat{\sigma}). \tag{2}$$

Note, many assignments are fixed under swapping the labels for leaves 1 and 3. In particular, for σ as any of the following values:

$$0000, 0001, 0100, 0101, 1010, 1011, 1110, 1111$$

we have $\sigma = \hat{\sigma}$ and hence $\mu(\sigma) = \hat{\mu}(\sigma)$. Thus we will can ignore these assignments and prove that the probabilities of the other assignments remain the same when the labels of leaves (1) and (3) are swapped.

We will show that in the mixture distribution μ, the probabilities of assignments 0010 and 1000 are the same:

$$\mu(0010) = \mu(1000) \tag{3}$$

Since for $\sigma = 0010$, we have $\hat{\sigma} = 1000$, then from (3) and (2) it follows that

$$\mu(0010) = \hat{\mu}(0010) \quad \text{and} \quad \mu(1000) = \hat{\mu}(1000).$$

Moreover, since the CFN model is symmetric, we also have

$$\mu(1101) = \hat{\mu}(1101) \quad \text{and} \quad \mu(0111) = \hat{\mu}(0111).$$

Finally, we also show that

$$\mu(0110) = \mu(1100), \tag{4}$$

which implies

$$\mu(0110) = \hat{\mu}(0110), \quad \mu(1100) = \hat{\mu}(1100), \quad \mu(1001) = \hat{\mu}(1001),$$
$$\mu(0011) = \hat{\mu}(0011).$$

Hence, (3) and (4) imply that those assignments which are not fixed by swapping leaf labels 1 and 3, have the same probability in μ and $\hat{\mu}$. Thus to complete the proof, we need to show that (3) and (4) hold. These are straightforward to check in any symbolic algebra system, such as Maple.

2.2 Proof Tools for Maximum Likelihood Results

In this section we present the technical tools needed to prove Theorems 1 and 2. Before stating precise lemmas, we begin by explaining some of the intuition of the proofs and how the notation applies to the theorems.

Let μ_x denote the mixture distribution defined by the example in Figure 2 (in the CFN or Jukes-Cantor model). Our result is for $x > 0$, but the proof will use properties of the case $x = 0$. Let $\mu = \mu_0$. Note in the case $x = 0$, the two trees in the mixture are identical. Hence, $\mu = \mu_0$ is a pure distribution (i.e., generated by a single tree). Moreover, for $x = 0$, since the internal edge weight is 0 this distribution is achievable on every topology; simply set the terminal edges to taxa 1, 2, 3, and 4 to have weight C, and the other internal edge to have weight 0. Let v denote this assignment of edge weights. For every tree S, this assignment generates distribution μ (i.e., $\mu(S, v) = \mu$). This is also the unique such weight (see Štefankovič and Vigoda, 2006, for a proof). Thus this is the assignment of edge weights for S that maximizes the expected log-likelihood under μ.

For x small the following lemma bounds the expected maximum expected log-likelihood in terms of the log-likelihood at $x = 0$ and the Hessian and Jacobian of the likelihood function. The error term will be $o(x^2)$, which is a smaller order term when x is sufficiently small. This will then imply that for x sufficiently small, $\mathcal{L}_S(\mu_x)$ is close to $\mathcal{L}_S(\mu_0)$ and we simply need to compute the Hessian and Jacobian functions to bound $\mathcal{L}_S(\mu_x)$.

Before stating the general lemma, let us preview the terminology in our setting. We are interested in computing the likelihood for distribution μ_x where $x > 0$. We will use the distribution $\mu = \mu_0$ which corresponds to the tree with internal edge weight zero. (Note under μ each of the 4^4 assignments to the leaves has positive probability.) To bound the maximum likelihood of μ_x on some tree topology S (which can be any topology), we need that the distribution is achievable on S. In our case, this is clearly true, as discussed above.

If x is small, then

$$\mu_x = \mu + x\Delta\mu_x$$

where $\Delta\mu_x$ is a vector whose sum of coordinates is zero (thus, $\Delta\mu_x$ is a vector representing the change in μ_x from μ). As $x \to 0$ we will have $\Delta\mu_x \to \Delta\mu_0$. We will use $\Delta\mu$ for bounding the maximum likelihood of μ_x for x small. The maximum expected log-likelihood of μ_x on S will be expressed in terms of the expected log-likelihood of μ under the tree corresponding to μ_x, which is

$$(\mu_x)^T \ln(\mu) = \left[\mu^T + x(\Delta\mu_x)^T\right] \ln(\mu),$$

and the first few terms from the Taylor expansion.

Here is the formal statement of the lemma.

Lemma 6 (Štefankovič and Vigoda, 2006). *Let μ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n nodes. Suppose that there exists v in the closure of the model such that $\mu(S, v) = \mu$ and that v is the unique such weight. Let $\Delta\mu_x$ be such that $\Delta\mu_x^T 1 = 0$, and $\Delta\mu_x \to \Delta\mu_0$ as $x \to 0$.*

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

$$\mathcal{L}_S(\mu + x\Delta\mu_x) \le \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2} J_0 H^{-1} J_0^T + o(x^2). \tag{5}$$

If $(H^{-1}J^T)_i \le 0$ for all i such that $v_i = 0$ then the inequality in (5) can be replaced by equality.

Remark 7. *When $(H^{-1}J^T)_i < 0$ for all i such that $v_i = 0$ then the likelihood is maximized at non-trivial branch lengths. In particular, for the CFN model, the branch lengths are in the interval $(0, 1/2)$, and there are no branches of length 0 or $1/2$. Similarly for the Jukes-Cantor the lengths are in $(0, 1/4)$.*

We will need the following extension of the above result.

Lemma 8. *Let μ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n nodes. Suppose that there exists v in the closure of the model such that $\mu(S, v) = \mu$ and that v is the unique such weight. Let $\Delta\mu_x$ be such that $\Delta\mu_x^T 1 = 0$, and $\Delta\mu_x \to \Delta\mu_0$ as $x \to 0$. Let $g(w) = \mathcal{L}_{S,w}(\mu)$ and $h_x(w) = (\Delta\mu_x)^T \ln \mu_{S,w}$. Let H be the Hessian of g at v and J_x be the Jacobian of h_x at v. Assume that H has full rank. Then*

$$\mathcal{L}_S(\mu + x\Delta\mu_x)$$
$$= \mu^T \ln \mu + xh_x(v) + \max_{\Delta w} \left[\frac{1}{2}(\Delta w)^T H(\Delta w) + xJ_0(\Delta w)\right] + o(x^2). \tag{6}$$

where the maximization is restricted to Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$.

Remark 9. *Now we remark on how the optimization problem in* (6) *can be solved. Let I be the set of indices with $v_i = 0$. Let u be an optimal solution of $(1/2)u^T Hu + xJ_0 u$ subject to the restriction $u_i \geq 0$ for $i \in I$. Note that u exists because the optimization happens on a continuous function over a compact set. Let $I' \subseteq I$ be a set of indices that are zero in u. Let J' be J_0 restricted to indices that are not in I' and let H' be H restricted to indices not in I'. Then, by the optimality of u, we have that u restricted to indices not in I' is $(H')^{-1}(xJ')^T$.*

Thus the maximization problem in (6) *can be solved as follows. For each $I' \subseteq I$ do the following. Let J' be J_0 with entries at indices in I' zeroed out and let H' be H with entries at indices in I' zeroed out. Compute $(H')^{-1}(xJ')^T$ and check whether all entries at positions in I are non-negative. If they are then $xJ'(H')^{-1}(xJ')^T$ is a candidate solution. Pick the largest candidate solution. Note that only solutions that satisfy the restriction $u_i \geq 0$ for $i \in I$ are candidate, and by the previous paragraph an optimal solution is candidate. Thus using the procedure described we obtain an optimal solution of* (6).

Proof of Lemma 8. In the proof of Lemma 6 from Štefankovič and Vigoda (2006), it is proved that

$$\mathcal{L}_{T,v+\Delta v}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh_x(v) + \frac{1}{2}(\Delta v)^T H(\Delta v) + xJ_x(\Delta v) + O(||\Delta v||^3 + x||\Delta v||^2), \tag{7}$$

and $\mathcal{L}_{v+\Delta v}(\mu + x\Delta\mu_x) < \mathcal{L}_v(\mu + x\Delta\mu_x)$ for $||\Delta v|| = \omega(x)$. Thus (7) is maximized for $||\Delta v|| = O(x)$.

The edge weights $v + \Delta v$ are invalid if, for some i, $v_i = 0$ and $(\Delta v)_i = 0$. On the other hand, if for all i where $v_i = 0$ we have $(\Delta v)_i \geq 0$ and Δv are sufficiently small, then then $v + \Delta v$ are edge weights. Note that we are considering $||\Delta v|| = O(x)$. Hence, we have Δv is sufficiently small by choosing x sufficiently small. Therefore,

$$\mathcal{L}_S(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh(v) + \max_{\Delta w} \left[\frac{1}{2}(\Delta w)^T H(\Delta w) + xJ_x(\Delta w)\right] + O(x^3),$$

where the maximum is taken over Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$. From $\Delta\mu_x \to \Delta\mu_0$ it follows that

$$\max_{\Delta w} \left[\frac{1}{2}(\Delta w)^T H(\Delta w) + xJ_x(\Delta w)\right] = \max_{\Delta w} \left[\frac{1}{2}(\Delta w)^T H(\Delta w) + xJ_0(\Delta w)\right][1 + o(1)],$$

where the maxima are taken over Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$. □

2.3 Proof of Theorem 2 for the CFN model

We begin with the simpler proof for the CFN model.

Let $\alpha = \beta \cdot x^2$ be the weight of the middle edge.

We will need the following polynomials in C:

$$Z := (1 - 4C + 4C^2)(1 - 2C + 2C^2)^2$$
$$U1 := 24C^6 - 72C^5 + 86C^4 - 52C^3 + 16C^2 - 2C$$
$$U2 := -16C^4 + 32C^3 - 16C^2 + 2$$
$$U3 := -8C^4 + 16C^3 - 16C^2 + 8C - 2$$
$$U4 := -2(2C^2 - 2C + 1)^2$$

We will need the following property of the polynomials.

Observation 10. *The polynomials Z, U_2 and $-U_1$, $-U_3$, $-U_4$ are positive for $C \in (0, 1/2)$.*

The observation is easily proved by plugging in $C = 1/4$, and then checking that the polynomials do not have roots on $(0, 1/2)$; e. g., using the method of Sturm sequences.

Let H_i be the Hessian and J_i the Jacobian for T_i. Because of symmetry we have that the Hessians are the same. We let $H := H_1 = H_2 = H_3$ and let H' be H with the last (fifth) row and column replaced by zeros. The Jacobians differ only in the last coordinate. Let J' be the vector which agrees with J_1, J_2, J_3 on the first four coordinates and is zero on the last (fifth) coordinate.

Let ℓ_i be the last coordinate of $-H^{-1}J_i$. Recall that if ℓ_i is positive then, by Lemma 6,

$$\mathcal{L}_{T_i}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2}(J_i)_0 H^{-1}(J_i)_0^T + o(x^2). \tag{8}$$

On the other hand, if ℓ_i is not positive then, by Remark 9,

$$\mathcal{L}_{T_i}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2}(J')(H')^{-1}(J')^T + o(x^2). \tag{9}$$

Note that when ℓ_i is positive then the right hand side of (8) is strictly bigger than the right-hand side of 9.

A tedious computation yields

$$\ell_1 = \beta + U_4/Z,$$
$$\ell_2 = (U_1\beta + U_3)/Z,$$
$$\ell_3 = (U_1\beta + U_2)/Z.$$

Since U_1 and U_3 are negative and Z is positive, then ℓ_2 is always negative.

For small β we have that ℓ_3 is positive and ℓ_1 is negative. Thus, tree T_3 has higher likelihood than T_2 and T_1 (for sufficiently small x). In addition, by Remark 7, since $\ell_3 > 0$, the maximum likelihood for T_3 is achieved with non-zero branch lengths. For T_1 and T_2 the optimization procedure outlined in Remark 9 may give zero length edges. But, since we only want to upper bound the likelihood for these trees, we can allow zero length edges. This proves Part (b) of Theorem 2.

We know from Theorem 5 that there exists an α_c' where the likelihood of T_1 and T_3 are equal. We want to show there exists α_c where the likelihoods of T_1 and T_3 are the same, and, for $\alpha < \alpha_c$, T_3 is the maximum likelihood tree. This requires showing there exists α_c where the following hold: at α_c we have $\ell_1 = \ell_3$ (and both are positive so that the maximum likelihoods on these trees are achieved at non-zero branch lengths); and for $\alpha < \alpha_c$ we have $\ell_1 > \ell_3$ (and ℓ_1 is positive).

For large β (i.e., large α) we have that ℓ_3 is negative and ℓ_1 is positive. Thus for large β, tree T_1 has higher likelihood than T_2 and T_3.

We first argue that for any α, $\ell_1 > 0$ and/or $\ell_3 > 0$. Then the result follows easily by considering the smallest α where $\ell_1 = \ell_3$.

Multiply ℓ_1 by the positive polynomial $(-U_1)$ and add it to ℓ_3. Since $-U_1$ is always positive, then if $-\ell_1 U_1 + \ell_3 > 0$ then at least one of ℓ_1 and ℓ_3 must be positive. Note

$$-\ell_1 U_1 + \ell_3 = \frac{1}{Z}(U_2 - U_3 U_1).$$

We obtain

$$\frac{1}{Z}(U_2 - U_3U_1) = \frac{1}{Z}(192C^{10} - 960C^9 + 2224C^8 - 3136C^7$$
$$+2960C^6 - 1936C^5 + 890C^4 - 232C^3 + 32C^2 - 4C + 2),$$

which is positive for $C \in [0, 1/2]$ (this is proved using Sturm sequences as before). Thus, at least one of ℓ_1 or ℓ_3 is always positive.

Take the smallest β_c where $\ell_1 = \ell_3$. We know there exists at least one such β_c since for small β we have $\ell_1 > 0$ and $\ell_3 < 0$, whereas for large β we have $\ell_1 < 0$ and $\ell_3 > 0$. Since at least one of ℓ_1 and ℓ_3 is positive, we have that both ℓ_1 and ℓ_3 are positive when $\ell_1 = \ell_3$. Thus the maximum likelihoods on T_1 and T_3 are the same at β_c (since $\ell_1 = \ell_3$), and are achieved on non-zero branch lengths (since $\ell_1, \ell_3 > 0$).

All the formulas are continuous in β and hence for all $\beta < \beta_c := \alpha_c/x^{-2}$ the likelihood of T_3 is higher than the likelihood of T_1, and for all $\beta > \beta_c$ the likelihood of T_3 is smaller than the likelihood of T_1.

This completes the proof of the theorem.

2.4 Proof of Theorem 1 for the Jukes-Cantor model

The proof for the Jukes-Cantor model will follow the same lines as the argument for the CFN model; however, some of the quantities are more complicated.

Let $\alpha = \beta \cdot x^2$ be the weight of the middle edge.

We will need the following polynomials in C:

$$\begin{aligned}
Z &= (-1 + 2C) \\
&\quad \times (2{,}359{,}296C^{15} + 11{,}010{,}048C^{14} - 42{,}385{,}408C^{13} \\
&\quad + 55{,}336{,}960C^{12}) - 33{,}972{,}224C^{11} + 4{,}602{,}880C^{10} \\
&\quad + 9{,}055{,}232C^9 - 8{,}473{,}344C^8 + 4{,}156{,}416C^7 - 1{,}372{,}704C^6 \\
&\quad + 326{,}728C^5 - 57{,}200C^4 + 7{,}308C^3 - 654C^2 + 37C - 1) \\
U_1 &= 21{,}233{,}664C^{16} - 98{,}697{,}216C^{15} + 203{,}702{,}272C^{14} \\
&\quad - 247{,}480{,}320C^{13} + 196{,}929{,}536C^{12} - 107{,}718{,}656C^{11} \\
&\quad + 41{,}333{,}888C^{10} - 11{,}267{,}328C^9 + 2{,}317{,}632C^8 - 465{,}360C^7 \\
&\quad + 123{,}492C^6 - 33{,}482C^5 + 6{,}644C^4 - 846C^3 + 62C^2 - 2C \\
U_2 &= 393{,}216C^{12} - 1{,}515{,}520C^{11} + 2{,}001{,}920C^{10} - 917{,}504C^9 \\
&\quad - 441{,}728C^8 + 830{,}176C^7 - 517{,}696C^6 + 183{,}344C^5 \\
&\quad - 39{,}128C^4 + 4{,}604C^3 - 164C^2 - 22C + 2 \\
U_3 &= 49{,}152C^{12} + 548{,}864C^{11} - 1{,}646{,}080C^{10} + 2{,}025{,}984C^9 \\
&\quad - 1{,}440{,}896C^8 + 672{,}800C^7 - 226{,}064C^6 + 60{,}968C^5 \\
&\quad - 14{,}600C^4 + 3{,}044C^3 - 476C^2 + 46C - 2
\end{aligned}$$

We will need the following property of the polynomials.

Observation 11. *The polynomials Z, U_2 and $-U_1, -U_3$ are positive for $C \in (0, 1/4)$.*

Once again the observation is proved by plugging in $C = 1/8$ and then using Sturm sequences to prove that the polynomials do not have roots on $(0, 1/4)$.

Define H, J_1, J_2, J_3, and J as in the proof for the CFN model. More precisely, let H_i be the Hessian and J_i the Jacobian for T_i. Because of symmetry we have that the Hessians are the same, thus let $H := H_1 = H_2 = H_3$. The Jacobians differ only in the last coordinate. Let J be the vector which agrees with J_1, J_2, J_3 on the first four coordinates and is zero on the last (fifth) coordinate. Let ℓ_i be the last coordinate of $-H^{-1}J_i$. Recall that if ℓ_i is positive then (8) holds, and if ℓ_i is not positive then (9) holds.

Note that when ℓ_i is positive then the right-hand side of (8) is strictly bigger than the right-hand side of (9). We now have

$$\begin{aligned}
\ell_1 &= \beta + U_3/Z, \\
\ell_2 &= (U_1\beta + U_3)/Z, \\
\ell_3 &= (U_1\beta + U_2)/Z.
\end{aligned}$$

Note that ℓ_2 is always negative. As in the CFN model, for small β we have that ℓ_3 is positive and ℓ_1 is negative. Thus tree T_3 has higher likelihood than T_2 and T_1 (for sufficiently small x). The proof of Part (a) is along identical lines as for the CFN model. However the polynomial of interest, i.e., $(1/Z)(U_2 - U_3 \cdot U_1)$ is significantly more complicated.

This completes the proof of the theorem.