ELSEVIER

# Inconsistency of Evolutionary Tree Topology Reconstruction Methods When Substitution Rates Vary Across Characters

JOSEPH T. CHANG
*Yale University Statistics Department, New Haven, Connecticut 06520-8290*

*Received August 30, 1994; accepted December 14, 1995*

ABSTRACT

A fundamental problem in reconstructing the evolutionary history of a set of species is to infer the topology of the evolutionary tree that relates those species. A statistical method for estimating such a topology from character data is called consistent if, given data from more and more characters, the method is sure to converge to the true topology. A number of popular methods are based on modeling the evolution of each character as a Markov process along the evolutionary tree. The standard models further assume that each character has in fact evolved according to the same Markov process. This homogeneity assumption is unrealistic; for example, different types of characters are known to experience substitutions at different rates. Certain distance and maximum likelihood methods for topology estimation have been shown to be consistent under the homogeneity assumption. Here we give examples showing that these methods can fail to be consistent when the homogeneity assumption is relaxed. The examples are very simple, requiring only four taxa, binary characters, and characters that evolve at two different rates.

## 1. INTRODUCTION

One of the basic aims of a phylogenetic study of a set of species is to infer the tree topology that most accurately summarizes the evolutionary relationships among those species. Markov models of character changes have been well studied and extensively used in analyzing character data such as DNA sequences. A simplifying assumption that has typically been made is that each character evolves according to precisely the same Markov model. In particular, this "homogeneity assumption" incorporates the assertion that, at any given point in the evolutionary tree, all of the different characters evolve at the same rate. Unfortunately, this assumption is known to be very unrealistic. At one extreme, a character might be so essential to the functioning of the organism that no change in the character could survive. Such characters are called "invariable," and we could describe the rate of evolutionary

*MATHEMATICAL BIOSCIENCES* 134:189–215 (1996)

655 Avenue of the Americas, New York, NY 10010
0025-5564/96/$15.00
SSDI 0025-5564(95)00172-7

change of such characters as zero. On the other hand, there are characters that can experience certain changes without producing any modification in any expressed amino acid sequence. Such characters, which may include characters in noncoding regions as well as the third codon positions in coding regions, are free to change at a relatively rapid rate.

It might be said that a close enough look at reality inevitably renders all standard statistical assumptions unrealistic. Our task then becomes to assess the harm that can come from making such assumptions. A fundamental statistical criterion in judging the adequacy of an inference method is that of consistency. A tree topology reconstruction method is said to be *consistent* under an assumed model of evolution if, when given more and more characters generated according to the assumed model, the reconstructed topology is certain to converge to the correct topology. Some commonly used distance methods and maximum likelihood methods are consistent under Markov models having homogeneous evolution across characters. The purpose of this paper is to present examples that show that these methods may fail to be consistent even under very simple models that allow for rate variation across characters.

Markov models have a significant history in phylogeny reconstruction, and they continue to play an important role. For example, the models of Jukes and Cantor [1], Cavender [2], Kimura [3], Barry and Hartigan [4], and many others are all Markov models, allowing different numbers of variable parameters in their probability transition matrices. We will not attempt a detailed review of this field here, as several excellent accounts are available: the articles [5–7] are recommended for an overview of Markov models and other methods in phylogeny reconstruction, as are the textbooks [8,9] for more general background. The Markov assumption, roughly speaking, asserts that character changes are independent in different branches of the tree. This has been viewed as an appealing compromise, giving theoretical and computational tractability while remaining relatively palatable scientifically. We will have more to say about the nature and suitability of the Markov assumption in the discussion in Section 6.

As mentioned above, however, the homogeneity assumption that each character evolves according to the same Markov model is worrisome. Beginning with the work of Fitch and Margoliash [10], a substantial literature has developed evidence for rate variation across characters as well as statistical approaches for taking such rate variation into account. A recent and convincing statement of evidence is provided by Shoemaker and Fitch [11], whose study may be consulted for further references. A variety of statistical models and methods for incorporat-

ing rate heterogeneity have been proposed recently; these include the work of Churchill et al. [12], Jin and Nei [13], Kelly and Rice [14], Lundstrom et al. [15], Sidow et al. [16], and Yang [17].

It is natural to ask how necessary it is to go to the effort of developing and using such techniques. If we use standard methods developed for models that assume homogeneous evolution across characters, how can we be led astray when the assumption is false? One form of deception that has been established is a systematic bias in the estimation of branch lengths; see Fitch and Margoliash [10] and Kelly and Rice [14]. In this paper we focus on another fundamental question: Can we consistently recover the tree topology? We present examples showing that distance and maximum likelihood methods that are consistent for the homogeneous Markov model can fail to be consistent when rates of evolution vary across characters. Our aim is to establish the existence of the phenomenon and give some understanding of it without extensive analytic calculations or computer simulations. A more extensive analysis of issues such as conditions under which the phenomenon occurs and how common it might be will be left for later work.

The distance method of Barry and Hartigan [4] is consistent for the homogeneous Markov model; see Chang and Hartigan [18]. Cavender and Felsenstein [19] gave a numerical example that showed that their closely related method may be inconsistent when different characters evolve at different rates. In Section 4 we show that in fact the presence of invariable characters is enough to admit examples for which the methods of Barry and Hartigan [4] and Cavender and Felsenstein [19] become inconsistent. Thus, the present work may be considered a continuation of the observation of Cavender and Felsenstein.

Not surprisingly, the maximum likelihood method of Felsenstein [20], is consistent for the homogeneous Markov model. One would anticipate this since, as shown by Wald [21], a maximum likelihood method developed for a given family of distributions is usually (that is, subject to mild regularity conditions) consistent when used for that family. In Section 5 we show that the maximum likelihood method of [20] may be inconsistent when invariable characters or more general forms of rate heterogeneity are present.

The study of the statistical consistency of topology reconstruction methods in phylogenetic analysis was launched in 1978 by Felsenstein's [22] demonstration that the popular parsimony method of Farris et al. [23] and Fitch [24] is not consistent under a simple Markov model. Felsenstein introduced the phrase "positively misleading" to convey the idea that as the number of observed characters grows to infinity, the tree topology chosen by the parsimony method may actually converge with probability 1 to an incorrect topology. Saying that a method is

positively misleading is logically stronger than saying it is inconsistent: The failure of an estimator to converge to the true topology does not imply that the estimator converges to an incorrect topology–it might not converge at all. All of the examples of inconsistency below will actually be positively misleading in this sense.

The recent studies of Tateno et al. [25] and Kuhner and Felsenstein [26] used simulation to investigate the performance of phylogenetic methods when substitution rates vary across sites. These are extensive studies that make progress toward quantifying the extent to which rate heterogeneity degrades the accuracy of several methods.

Bull et al. [27] considered the closely related question of whether or not heterogeneous data sets should be combined before analysis. They studied the parsimony method, examining both consistency and efficiency. With regard to consistency, they tried combining characters generated by two Markov models, one of which causes parsimony to be inconsistent. Calling the characters generated by the two models "consistent" and "inconsistent," they found that if the proportion of inconsistent characters in the mixture is sufficiently high, the combined data set can cause parsimony to be inconsistent. Expressed in this language, the examples below show that for the distance and maximum likelihood methods under investigation, it is in fact possible to combine two data sets, both of which are consistent, into a mixture that is inconsistent.

We are testing the performance of these methods on mixtures of Markov models, whereas the methods were originally derived by thinking about pure Markov models. Thus, it is not altogether surprising that the methods might be inconsistent under these circumstances. However, it does not seem obvious that they should do so either. In particular, all we ask of the methods is that they identify the tree topology, which is a rather minimal request in terms of detailed information; one might reasonably harbor some hope for the methods to be able to do this much. After all, even though we allow rates and branch lengths to vary across characters, we assume that all characters evolve according to one common tree topology. If each character contributes some information about the tree topology that they all share, it seems reasonable to hope that as more and more characters accumulate, the data could reveal to us the common topology with more and more certainty. The examples below show that the distance and maximum likelihood methods under consideration cannot reliably extract that information from the data.

## 2. TREES AND MARKOV MODELS

Throughout this paper, we make use of the simple subclass of Markov models considered by Cavender [2]. We consider four taxa, called $A$, $B$, $C$, and $D$; these are the species whose evolutionary history

we wish to infer. Characters are binary, taking states in the the set $\{0,1\}$. Different characters are assumed to evolve independently, with the same probabilistic behavior. Therefore, to specify a model for a collection of characters, it is sufficient to say how the model describes the evolution of a single character.

Accordingly, consider a single character possessed by the four taxa. A Markov model for the evolution of that character is described by a tree. A tree consists of nodes and branches. Each external node (node of degree 1) of the tree corresponds to one of the the taxa; typically the taxa are current species observable today, while internal nodes (nodes of degree greater than 1) correspond to unobserved ancestral species. On each branch the character switches states at the times of a Poisson process with a rate $q$ that may depend on the branch. That is, $q$ may be different for different branches of the tree but is a constant over any given branch: $q = q(\text{branch})$. Thus, the expected number $\nu$ of character changes on a branch is given by the product

$$\nu(\text{branch}) = q(\text{branch}) \times t(\text{branch}),$$

where $t$(branch) is the time elapsed over the branch. We will think of $\nu$ as a "branch length," and in diagrams the branch lengths drawn will be intended to indicate the values of $\nu$ on the corresponding branches.

The Poisson processes governing the character changes in different branches of the tree are assumed to be independent. This implies the familiar *Markov property*, which says that the probability distribution of the character state at a given node, conditional on the character states at all of the other nodes in the tree, depends only on the character states at nodes that are joined to the given node by a branch.

Having specified the probability transition structure that governs how a character changes, we complete the description of a Markov model by specifying a marginal distribution at some point along the tree. We assume throughout the paper that $P\{X_A = 0\} = 1/2 = P\{X_A = 1\}$. Together with the assumed probability transition structure, this implies that $P\{X_I = 0\} = 1/2 = P\{X_I = 1\}$ for all species $I$ in the tree.

A Markov model is determined by the topology and the branch lengths of the corresponding tree. More precisely, when we speak of "the topology of the tree," we mean the topology of the tree together with the labels on the terminal nodes of the tree, as discussed formally in section 2 of [18]. Notice that one of the three species $B$, $C$, or $D$ is topologically closest to $A$, in the sense that it is connected to $A$ by a path consisting of two branches, while the remaining two species are joined to $A$ by paths of three branches. We specify a topology by saying which one of the species $B$, $C$, or $D$ is topologically closest to $A$, denoting the three corresponding topologies by $\mathcal{T}_{AB}$, $\mathcal{T}_{AC}$, and $\mathcal{T}_{AD}$.

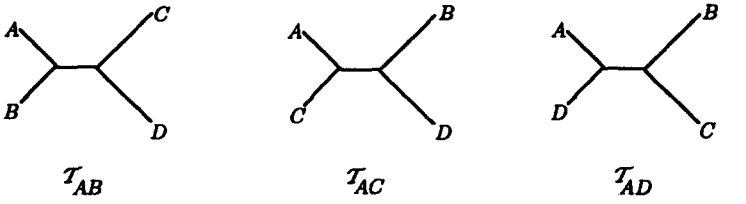


FIG. 1. Three possible tree topologies for four taxa.

Examples of trees in these three topologies are illustrated in Figure 1.

More formally, the space of Markov models we consider may be parametrized as follows. The parameter space $\Theta$ consists of elements $\theta$ of the form

$$\theta = (\text{closest}_A, \nu_A, \nu_B, \nu_C, \nu_D, \nu_{\text{int}}) \in \{B,C,D\} \times [0,\infty]^5,$$

where the first component, $\text{closest}_A$, specifies the tree topology by telling which species is topologically closest to $A$; the next four components give the lengths of the four branches that lead to the four external species $A$, $B$, $C$, and $D$; and the final component, $\nu_{\text{int}}$, gives the length of the internal branch. Note that we require branch lengths to be nonnegative. Also, we allow them to take on the value $\infty$; the resulting compactness of the parameter space will be convenient in Section 5. Now the topology $\mathcal{T}_{AB}$ may be defined as a subset of $\Theta$ by

$$\mathcal{T}_{AB} = \{\theta \in \Theta : \text{closest}_A = B\},$$

with analogous definitions for the topologies $\mathcal{T}_{AC}$ and $\mathcal{T}_{AD}$. These definitions are illustrated in Figure 2.

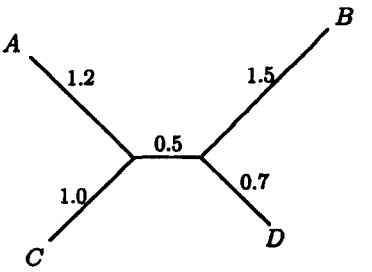


FIG. 2. Illustrating the parametrization of Markov models. Here $\theta = (C, 1.2, 1.5, 1.0, 0.7, 0.5) \in \mathcal{T}_{AC}$.

Since observed data consist of character states for the taxa $A$, $B$, $C$, and $D$, the feature of a probability model that is of most direct concern to us is the joint probability distribution that the model gives to the character vector $(X_A, X_B, X_C, X_D)$. For a Markov model $\theta \in \Theta$, let $P_\theta$ denote the joint distribution of $(X_A, X_B, X_C, X_D)$ induced by the model $\theta$.

If $P$ is a joint distribution for $(X_A, X_B, X_C, X_D)$, we abuse notation by writing, for example, "$P \in \mathcal{T}_{AB}$" to mean "$P = P_\theta$ for some $\theta \in \mathcal{T}_{AB}$." Note that in this sense the three topologies intersect in the "star phylogeny"; for example, if $\theta \in \mathcal{T}_{AB}$ has $\nu_{\text{int}} = 0$, then we could also write $P_\theta \in \mathcal{T}_{AC}$ and $P_\theta \in \mathcal{T}_{AD}$.

We have described a branch in terms of its "branch length" $\nu$. An important alternative quantity that characterizes the probabilistic behavior of a character on a branch is the probability that the character changes across the branch. We use the letter $p$ (perhaps with subscripts and so on) throughout this paper to denote this sort of probability. The relationship between $p$ and $\nu$ is simple: $p$ is the probability that the number of character changes—which has a Poisson distribution with mean $\nu$ on a branch of length $\nu$— is odd, which is

$$p(\nu) = \tfrac{1}{2}(1 - e^{-2\nu}). \tag{2.1}$$

Note that $p(\nu) = 0$ when $\nu = 0$, and $p(\nu)$ attains its maximum value of $1/2$ when $\nu = \infty$.

The methods for estimating tree topologies in Markov models that we will discuss include distance methods and maximum likelihood. Maximum likelihood, although computationally by far the more demanding of the two methods, is conceptually straightforward, finding the Markov model $P_{\hat{\theta}}$ that gives the data the highest probability. The desired tree topology is estimated by the topology of the maximum likelihood estimator $\hat{\theta}$. Note that the method entails estimating the branch lengths of the tree in addition to the topology of the tree.

To describe the measure of distance used in the distance methods, suppose that two species $A$ and $B$ are related by the transition matrix $P(A, B) = (P_{ij}(A, B))$, whose $(i, j)$th entry is the conditional probability $P\{X_B = j | X_A = i\}$. Barry and Hartigan [4] and Cavender and Felsenstein [19] independently introduced the distance measure

$$d(A, B) = -(1/2)\log\det P(A, B). \tag{2.2}$$

Since probability transition matrices multiply along a path in a Markov model, the corresponding determinants also multiply, so their logarithms add. Thus, the function defined by Equation (2.2) is additive in

Markov models: If $I$ is a species along the path from $A$ to $B$, then $d(A,B) = d(A,I) + d(I,B)$.

To specialize to the case of binary characters, consider a binary character in two species $A$ and $B$. Whether or not the model is Markovian–and we use this extra generality later when we consider mixtures of Markov models–we may write down a transition matrix $P(A,B)$. Suppose this matrix is of the form

$$P(A,B) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array},$$

so that $p$ may be interpreted as the probability that $A$ and $B$ differ in that character. Then, by definition, the distance from species $A$ to species $B$ is

$$\begin{aligned} d(A,B) &= -\tfrac{1}{2}\log\det P(A,B) = -\tfrac{1}{2}\log\left[(1-p)^2 - p^2\right] \\ &= -\tfrac{1}{2}\log(1-2p). \end{aligned} \tag{2.3}$$

Note that in the case where the species $A$ and $B$ are joined in a Markov model by a branch (or a path) of length $v$, by substituting the expression (2.1) for $p$ in (2.3), we see that the distance from $A$ to $B$ reduces to the branch length $v$.

## 3. RATE HETEROGENEITY

Toward the goal of modeling heterogeneous evolution across characters, we wish to relax the assumption that all characters evolve according to a fixed Markov model $P_\theta$. A *mixture* of Markov models assumes that there is a probability distribution $Q$ on the set of Markov models $\Theta$ and that each character evolves according to its own randomly chosen Markov model drawn from the distribution $Q$. That is, such a mixture model $P$ gives an event $F$ the probability

$$P(F) = \int_{\theta \in \Theta} P_\theta(F) Q(d\theta).$$

We model rate heterogeneity across characters by assuming that characters evolve independently according to such a mixture of Markov models.

Figure 3 depicts a gradation in generality of classes of models, from special to general. Row 1 shows an example of a homogeneous Markov model. Here the trees governing the evolution of different characters all

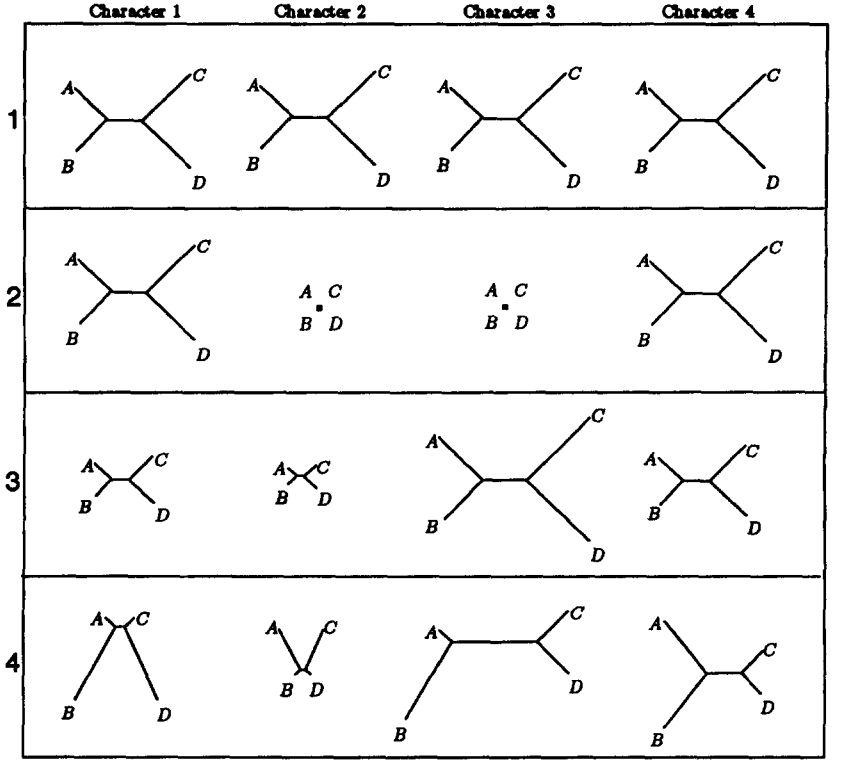


FIG. 3. Models with homogeneous and heterogeneous rates. Row 1: Homogeneous evolution across characters. Row 2: A mixture of two rates. One of the rates is positive; the other is zero, which corresponds to invariable characters. Row 3: A distribution of rates, "similar" trees. Row 4: General heterogeneous model.

have precisely the same set of branch lengths, so the trees are geometrically identical. One might also consider this to be a trivial case of the mixture model; here the mixing distribution $Q$ places probability 1 on a single model $\theta$. Row 4 is intended to depict a general mixture of Markov models, in which $Q$ spreads its mass over a variety of models $\theta$ contained in some fixed topology. Here each character evolves according to the topology $\mathcal{T}_{AB}$, but each character is allowed to have its own arbitrary set of edge lengths. Intermediate between the two extremes represented by rows 1 and 4 lie a variety of levels of generality for the mixing distribution $Q$. The third row depicts an example from a class of models that allows a certain structured sort of heterogeneity; this class was also considered by [14,28]. Here the trees for different characters are geometrically "similar": the sets of branch lengths may differ, but

only by proportionality constants. Continuing to specialize further, row 2 depicts an example of the class of models considered by [12,29] that is a very small subclass of the class of models in row 3. Here either a character is invariable or it evolves according to a single Markov model $P_\theta$, say. This is a subclass of the mixtures of "similar" models, because the invariable site model may be described in any topology by the branch lengths $\nu_A = \nu_B = \nu_C = \nu_D = \nu_{\text{int}} = 0$, so that the invariable site model is similar to $P_\theta$. Since the mixing distribution $Q$ places probability on only two models, the invariable site model and $P_\theta$, this is in fact a rather modest generalization of the homogeneous Markov model of row 1. However, we will show that even under this restricted sort of model of heterogeneous evolution across characters, one can produce examples where estimation of the tree will be in error if we mistakenly assume homogeneous Markov evolution across characters.

## 4. INCONSISTENCY OF DISTANCE METHODS

### 4.1. *THEORY*

Define

$$\begin{aligned} d(A,B) &= -\tfrac{1}{4}\log\det\{P(A,B)P(B,A)\} \\ &= -\tfrac{1}{4}(\log|\det P(A,B)| + \log|\det P(B,A)|), \end{aligned} \tag{4.4}$$

with $-\log(0)$ defined to be $\infty$; the equality of the two forms of the definition follows from the simple observation that the two determinants $\det P(A,B)$ and $\det P(B,A)$ must have the same sign. We think of $d(A,B)$ as the true distance between $A$ and $B$. Given data $X^1,\ldots,X^n$ for $n$ characters, we estimate the transition matrix $P(A,B)$ in the obvious way by the empirical transition matrix $\hat{P}_n(A,B)$ defined by

$$\left(\hat{P}_n(A,B)\right)_{i,j} = \frac{\sum_{k=1}^n \{X_A^k = i,\ X_B^k = j\}}{\sum_{k=1}^n \{X_A^k = i\}};$$

this is undefined if the denominator is zero. To avoid bothersome concerns about division by zero, let us assume that $P\{X_I = i\} > 0$ for all species $I$ and character states $i$; this will be the case for all of the examples we study in this paper, which in fact have $P\{X_I = 0\} = 1/2 = P\{X_I = 1\}$. Then by the strong law of large numbers,

$$\begin{aligned} \left(\hat{P}_n(A,B)\right)_{i,j} &= \frac{n^{-1}\sum_{k=1}^n \{X_A^k = i,\ X_B^k = j\}}{n^{-1}\sum_{k=1}^n \{X_A^k = i\}} \to \frac{P\{X_A = i,\ X_B = j\}}{P\{X_A = i\}} \\ &= (P(A,B))_{i,j} \end{aligned}$$

with probability 1 as $n \to \infty$. We estimate the true distance $d(A,B)$ by the empirical distance $\hat{d}_n(A,B)$, defined by replacing the transition matrices in definition (4.4) by the corresponding empirical transition matrices, so that

$$\hat{d}_n(A,B) = -\tfrac{1}{4}\log\det\{\hat{P}_n(A,B)\hat{P}_n(B,A)\}. \tag{4.5}$$

The definitions in (4.4) and (4.5) are slightly modified versions of those of Barry and Hartigan [4] and Cavender and Felsenstein [19]. One of the modifications is symmetrization, which was also discussed in [18,30,31]. The function $d$ gives a distance; in fact, it is an *additive* function on the nodes of the tree, in the sense that $d(A,B) = d(A,I) + d(I,B)$ whenever $I$ is a node on the path between $A$ and $B$. One aspect of our definitions differs slightly from previous definitions, which declare the true and empirical distances undefined if the corresponding probability transition matrices have negative determinant. In practice, it is reasonable to regard a transition matrix having negative determinant as a signal that caution may be in order because transition matrices arising from typical continuous-time Markov process models will have positive determinant. However, for our purposes, we have adopted definitions that are always meaningful whether the determinants are positive or negative. This is mathematically convenient; for example, it guarantees that $\hat{d}_n(A,B) \to d(A,B)$ with probability 1 even when $d(A,B) = \infty$.

In the case of four taxa, Cavender and Felsenstein [19] proposed choosing the tree topology as follows. Suppose that the six empirical distances among the four taxa $I$, $J$, $K$, and $L$ satisfy the relation

$$\hat{d}_n(I,J) + \hat{d}_n(K,L) < \min\left[\hat{d}_n(I,K) + \hat{d}_n(J,L), \hat{d}_n(I,L) + \hat{d}_n(J,K)\right]. \tag{4.6}$$

Then the method chooses the topology $\mathcal{T}_{IJ}$. If there is not a unique minimum among the three sums in (4.6), but rather two or more of those sums are tied, the method is indifferent between the corresponding tied topologies. Since the empirical distances converge to the true distances with probability 1 as $n \to \infty$, it is easy to see what is involved in generating an example for which the method of Cavender and Felsenstein is inconsistent. It would be sufficient, for example, to find a model having the topology $\mathcal{T}_{AB}$ whose true distances satisfy

$$d(A,C) + d(B,D) < \min\left[d(A,B) + d(C,D), d(A,D) + d(B,C)\right].$$

The distance method of Barry and Hartigan [4] applies to any number of taxa. To describe it for four taxa, let $d$ and $\hat{d}_n$ denote the vectors whose components list the true and empirical distances between the various pairs of taxa, so that

$$d = (d(A,B), d(C,D), d(A,C), d(B,D), d(A,D), d(B,C)),$$

and similarly for $\hat{d}_n$. For a given Markov model $\theta$, let $d(\theta)$ denote the vector of six pairwise distances under $P_\theta$. For simplicity of exposition, suppose for the remainder of the description of this method that the true distances are all finite, so that $d \in [0,\infty)^6$. Although generally there will be no Markov model $\theta$ such that $d(\theta) = \hat{d}_n$ exactly, Barry and Hartigan [4] propose finding a Markov model $\theta$ whose distances $d(\theta)$ give a best fit to $\hat{d}_n$ in the least squares sense. They then estimate the unknown tree topology by the topology of this best-fitting Markov model. More formally, for a given topology $\mathcal{T}$, define

$$\mathcal{D}_\mathcal{T} := \{d(\theta): \theta \in \mathcal{T}\}$$

and let $\mathcal{P}_\mathcal{T}: [0,\infty)^6 \to \mathcal{D}_\mathcal{T}$ denote the projection into $\mathcal{D}_\mathcal{T}$ defined by

$$\|\mathcal{P}_\mathcal{T} x - x\| = \min\{\|y - x\|: y \in \mathcal{D}_\mathcal{T}\},$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^6$. Thus, $\mathcal{P}_\mathcal{T}\hat{d}_n$ is the distance vector that gives the closest fit to the empirical distances $\hat{d}_n$ over all distance vectors generated by nonnegative branch lengths in the tree topology $\mathcal{T}$. The topology estimate $\hat{\mathcal{T}}_n$ is chosen to minimize $\|\hat{d}_n - \mathcal{P}_\mathcal{T}\hat{d}_n\|$ over choices of $\mathcal{T}$.

Since $\mathcal{P}_\mathcal{T}$ is continuous, the convergence $\hat{d}_n \to d$ of the empirical distances to the true distances as $n \to \infty$ implies that

$$\|\hat{d}_n - \mathcal{P}_\mathcal{T}\hat{d}_n\| \to \|d - \mathcal{P}_\mathcal{T} d\| \text{ with probability 1.}$$

Thus, a sufficient condition for the method of Barry and Hartigan to be inconsistent may be stated as follows. Denote the true topology by $\mathcal{T}$ and the true distance vector by $d$. Suppose there is a topology $\mathcal{T}_1$ such that

$$\|d - \mathcal{P}_{\mathcal{T}_1} d\| < \|d - \mathcal{P}_\mathcal{T} d\|.$$

Then, under the assumed model $P$, the distance method gives an inconsistent estimate of the tree topology; in fact, $P(\hat{\mathcal{T}}_n \neq \mathcal{T}) \to 1$.

The next result is special enough to be proved easily but general enough for the purposes of this paper. It tells how to find the best-fit-

ting topology called for by the above condition for inconsistency, given distances that satisfy certain relations.

*LEMMA 4.1*

*Suppose that the six pairwise distances among the four species I, J, K, and L are finite and satisfy the relations*

$$\frac{d(I,J)+d(K,L)}{2} < d(I,K) = d(J,L) \leqslant d(I,L) = d(J,K).$$

*Then the best least-squares fit to the given distances has topology $\mathcal{T}_{IJ}$, with branch lengths*

$$v_I = v_J = \tfrac{1}{2}d(I,J), \qquad v_K = v_L = \tfrac{1}{2}d(K,L),$$

*and*

$$v_{\text{int}} = \tfrac{1}{2}[d(I,K)+d(I,L)-d(I,J)-d(K,L)] > 0.$$

The above remarks may be used to give conditions for both of the distance methods we have discussed to be inconsistent. The following simple sufficient condition will be used in our examples.

*PROPOSITION 4.2*

*If a model having the topology $\mathcal{T}_{AB}$ has true distances satisfying $d(A,B) = d(C,D)$, $d(A,D) = d(B,C)$, and*

$$d(A,C)+d(B,D) < \min\{d(A,B)+d(C,D), d(A,D)+d(B,C)\},$$

*then both the methods of Barry and Hartigan* [4] *and Cavender and Felsenstein* [19] *will be inconsistent for that model.*

4.2. *EXAMPLE WITH TWO NONSIMILAR TREES*

If we allow ourselves the generality depicted in row 4 of Figure 3, that is, general heterogeneous models with no "similarity" restriction, it is easy to display a simple and intuitively clear example of inconsistency. Suppose each character has probability $1/2$ of evolving according to the two trees in Figure 4, where the $v$ values are shown on the edges. The branch lengths are expressed in terms of a parameter $\varepsilon$, which we will usually think of as a small number. The particular form of the function $1/\varepsilon$ is not important; the purpose of specifying a particular functional form was simply to have a concrete example with just one free parameter. Likewise, again for concreteness, we consider mixtures of two models, each having probability $1/2$. The same sort of phenomena

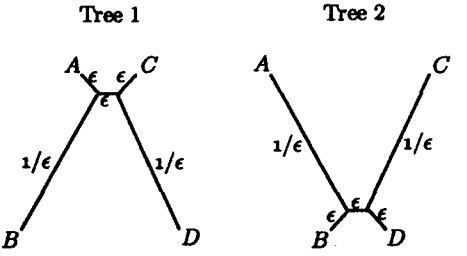


FIG. 4. Example of inconsistency. Characters evolve according to the two "non-similar" trees shown with probability 1/2 each.

would be exhibited by more general mixing probabilities; there is nothing special about the value 1/2.

The idea is simply this. In the mixture, the pair $A$ and $C$ will be separated by a rather short distance, since the inclusion of Tree 1 in the mixture causes $P\{X_A \neq X_C\}$ to be less than 1/2. Similarly, the pair $B$ and $D$ will be equally close. However, all four other pairs are separated by a large distance in both Tree 1 and Tree 2 and hence also in the mixture. The topology compatible with these requirements is $\mathcal{T}_{AC}$, which is different from the true topology $\mathcal{T}_{AB}$.

More formally, considering the pair of species $A$ and $C$, for example, we have

$$P\{X_A \neq X_C \mid \text{Tree } 1\} = p(3\varepsilon)$$

and

$$P\{X_A \neq X_C \mid \text{Tree } 2\} = p\left(\frac{2}{\varepsilon} + \varepsilon\right),$$

where the function $p(\cdot)$ is as in (2.1). Thus, letting $p_{IJ}$ denote the probability that the characters $X_I$ and $X_J$ differ in two species $I$ and $J$, we have

$$\begin{aligned} p_{AC} &= \frac{1}{2}p(3\varepsilon) + \frac{1}{2}p\left(\frac{2}{\varepsilon} + \varepsilon\right) \\ &= \frac{1}{4}(1 - e^{-6\varepsilon}) + \frac{1}{4}(1 - e^{-(4/\varepsilon + 2\varepsilon)}) \\ &= \frac{1}{4} + \frac{3}{2}\varepsilon + O(\varepsilon^2) \end{aligned}$$

as $\varepsilon \to 0$. Using the relation (2.3) to convert this probability back to a distance yields

$$d(A,C) = -\tfrac{1}{2}\log(1-\tfrac{1}{2}-3\varepsilon+O(\varepsilon^2)) = \tfrac{1}{2}\log 2+3\varepsilon+O(\varepsilon^2).$$

By symmetry, $d(B,D)=d(A,C)$. The remaining four distances are obvious: for example, $d(A,B)$ is clearly $1/\varepsilon+\varepsilon$, as $A$ and $B$ are separated by a path of that length in both Tree 1 and Tree 2. Thus, we see that

$$\begin{aligned} d(A,C) = d(B,D) &= (1/2)\log 2+3\varepsilon+O(\varepsilon^2), \\ d(A,B) = d(C,D) &= 1/\varepsilon+\varepsilon, \\ d(A,D) = d(B,C) &= 1/\varepsilon+2\varepsilon. \end{aligned}$$

From this, by Proposition 4.2, we conclude that the distance methods are inconsistent on this example for small enough $\varepsilon > 0$. The best fitting model given by Proposition 4.1 is shown, up to terms of order $\varepsilon$, in Figure 5.

How small does $\varepsilon$ have to be in order to get inconsistency? Not particularly small: the boundary between consistency and inconsistency is at $\varepsilon = 0.628$ (so that $1/\varepsilon = 1.592$). In fact, for $\varepsilon = 0.628$ (see the middle row of Table 1), the mixture has distance 2.220 separating each of the four pairs $(A,B)$, $(C,D)$, $(A,C)$, and $(B,D)$ and distance 2.848 for the two pairs $(A,D)$ and $(B,C)$. In that case the distance methods are indifferent between the correct tree topology $\mathcal{T}_{AB}$ and the incorrect topology $\mathcal{T}_{AC}$. When $\varepsilon > 0.628$ (e.g., see the bottom row of Table 1), the mixture distances are ordered as

$$d(A,B) = d(C,D) < d(A,C) = d(B,D) < d(A,D) = d(B,C),$$

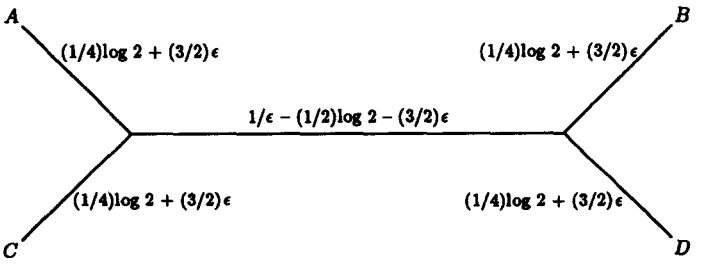


FIG. 5. The Markov model that gives the best fit in the example of Figure 4. Branch lengths are given up to terms of order $\varepsilon$.

TABLE 1

Consistency in the Example of Figure 4

| $\varepsilon$ | $d(A,B)$, $d(C,D)$ | $d(A,C)$, $d(B,D)$ | $d(A,D)$, $d(B,C)$ | Best-fitting topology |
|---|---|---|---|---|
| 0.600 | 2.267 | 2.140 | 2.867 | $\mathcal{T}_{AC}$ |
| 0.628 | 2.220 | 2.220 | 2.848 | $\mathcal{T}_{AB}$ or $\mathcal{T}_{AC}$ |
| 0.650 | 2.188 | 2.282 | 2.838 | $\mathcal{T}_{AB}$ |

For the example of Figure 4, the distance methods are inconsistent when $\varepsilon < 0.628$ and consistent when $\varepsilon > 0.628$.

so that the methods choose the correct topology $\mathcal{T}_{AB}$. On the other hand, when $\varepsilon < 0.628$ (e.g., see the top row of Table 1), the mixture distances satisfy

$$d(A,C) = d(B,D) < d(A,B) = d(C,D) < d(A,D) = d(B,C),$$

so that the methods choose the incorrect topology $\mathcal{T}_{AC}$. Thus, the distance method is "positively misleading" when $\varepsilon < 0.628$.

### 4.3. *EXAMPLE WITH INVARIABLE CHARACTERS*

The previous example provided a transparent example of how we can be fooled by evolution that is heterogeneous across characters. However, it incorporated rather artificial, biologically implausible behavior: a character state was likely to be shared by either the species pair $(A,C)$ or the pair $(B,D)$ but not by both pairs. This section presents an example that strains credulity much less severely, requiring only that some characters be invariable. More generally, it will be easy to see that the same phenomenon can be exhibited if some characters evolve very slowly compared to others.

We begin by examining the effect of invariable sites on distances. Throughout this section and the next, let $r$ be a number strictly between 0 and 1. Consider two species $A$ and $B$. Suppose that sites are invariable with probability $r$, and that, with the remaining probability $1-r$, species $A$ and $B$ are separated by a branch of length $\nu$ (or, equivalently, a sequence of branches of total length $\nu$). What is the distance $d(A,B)$ in the mixture? Since the species are separated by a distance of zero with probability $r$ and a distance of $\nu$ with probability $1-r$, it is natural to hope that the answer might be $(1-r)\nu$. However, this speculation is false; in fact, its failure contains the root of the difficulties rate heterogeneity can cause for the distance methods. As has been observed by Fitch and Margoliash [10], Kelly and Rice [14], and others, distance measures that provide reliable estimates of num-

bers of substitutions in homogeneous rate models generally give underestimates when substitution rates vary across sites. In our setting, we have

$$\begin{aligned} P\{X_A \neq X_B\} &= P\{X_A \neq X_B \,|\text{site invariable}\}P\{\text{site invariable}\} \\ &\quad + P\{X_A \neq X_B \,|\text{site variable}\}P\{\text{site variable}\} \\ &= 0 + (1-r)P\{X_A \neq X_B \,|\text{site variable}\} \\ &= \tfrac{1}{2}(1-r)(1-e^{-2\nu}), \end{aligned}$$

where the last equality uses (2.1). Therefore, by (2.3), the distance $d_{\text{mix}}(\nu)$ between $A$ and $B$ in the mixture is

$$d_{\text{mix}}(\nu) = -\tfrac{1}{2}\log[1-2P\{X_A \neq X_B\}] = -\tfrac{1}{2}\log[r+(1-r)e^{-2\nu}]. \quad (4.7)$$

Jensen's inequality implies that $d_{\text{mix}}(\nu) < (1-r)\nu$ for all $\nu > 0$. For small $\nu$, we have $d_{\text{mix}}(\nu) \sim (1-r)\nu$, so that the "natural guess" of $(1-r)\nu$ is nearly correct. On the other hand, as $\nu \to \infty$, we have $d_{\text{mix}}(\nu) \to -(1/2)\log r < \infty$. Thus, since the expected number of changes in the mixture model is indeed $(1-r)\nu$, we see that the distance $d_{\text{mix}}(\nu)$ always underestimates the expected number of changes, with the underestimation being slight when $\nu$ is small and severe when $\nu$ is large. The function $d_{\text{mix}}$ is strictly concave, so that

$$\tfrac{1}{2}d_{\text{mix}}(2\nu) < d_{\text{mix}}(\nu) \quad (4.8)$$

for all $\nu > 0$.

For an example of inconsistency, suppose that each character has probability $r$ of being invariable and probability $1-r$ of evolving according to the Markov model $P_\varepsilon$ illustrated in Figure 6, where $\nu$ is a

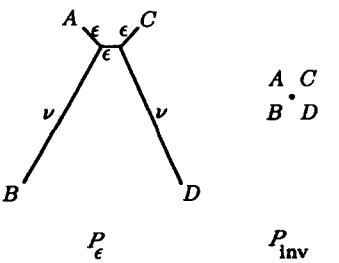


FIG. 6. Illustrating the ingredients in the definition of the probability $P^{\varepsilon}_{\text{mix}}$ as the mixture $(1-r)P_\varepsilon + rP_{\text{inv}}$.

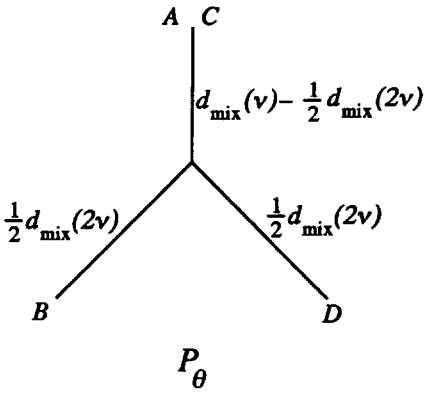


FIG. 7. The probability $P_\theta$ from Proposition 4.3.

fixed positive branch length. Denote this mixture model by $P^{\varepsilon}_{\text{mix}}$; that is, $P^{\varepsilon}_{\text{mix}} = (1-r)P_{\varepsilon} + rP_{\text{inv}}$. We will show that if $\varepsilon > 0$ is chosen small enough, then the distance methods are inconsistent when the characters are generated from the true distribution $P^{\varepsilon}_{\text{mix}}$.

We begin toward the goal of analyzing the probability $P^{\varepsilon}_{\text{mix}}$ for small positive $\varepsilon$ by first considering the case $\varepsilon = 0$. The next result says that the mixture model $P^{0}_{\text{mix}}$ yields distances that agree exactly with those in the Markov model $P_\theta$ illustrated in Figure 7. In fact, we will see in the next section that $P^{0}_{\text{mix}}$ and $P_\theta$ agree not only in their pairwise distances but also in their full joint distributions of character states of taxa.

*PROPOSITION 4.3*

*For each* $v \in [0,\infty]$, *the Markov model* $P_\theta$ *described in Figure 7 has distances between taxa that coincide exactly with those from the probability* $P^{0}_{\text{mix}}$.

*Proof.* Direct verification. Notice that the branch length $d_{\text{mix}}(v) - (1/2)d_{\text{mix}}(2v)$ is positive, by (4.8). ■

*THEOREM 4.4*

*There exists* $\varepsilon > 0$ *such that the distance methods of* [4] *and* [19] *are inconsistent when the true model is the mixture* $P^{\varepsilon}_{\text{mix}} = (1-r)P_{\varepsilon} + rP_{\text{inv}}$.

*Proof.* This follows simply by continuity considerations. Letting $d^{\varepsilon}$ denote the distance vector of the model $P^{\varepsilon}_{\text{mix}}$ for $\varepsilon \geqslant 0$, we have $d^{\varepsilon} \to d^{0}$ as $\varepsilon \to 0$. The symmetries $d^{\varepsilon}(A,B) = d^{\varepsilon}(C,D)$ and $d^{\varepsilon}(A,D) = d^{\varepsilon}(B,C)$ for all $\varepsilon \geqslant 0$ follow from the specification of the models $P^{\varepsilon}_{\text{mix}}$.

Thus, since

$$\tfrac{1}{2}\left[d^0(A,C)+d^0(B,D)\right] < d^0(A,B) = d^0(C,D)$$
$$= d^0(A,D) = d^0(B,C),$$

clearly the sufficient conditions of Proposition 4.2 are satisfied by the models $P^{\varepsilon}_{\text{mix}}$ for small enough $\varepsilon > 0$. ■

Finally, as in the previous section, we consider how small $\varepsilon$ must be in order to cause inconsistency. Here, the answer depends on $r$ and $\nu$. The distance methods are positively misleading when

$$d_{\text{mix}}(3\varepsilon) + d_{\text{mix}}(2\nu + \varepsilon) < 2d_{\text{mix}}(\nu + \varepsilon).$$

For example, suppose that the probability of invariable characters is $r = 1/2$. Then for $\nu = 1$, a numerical calculation shows that the inconsistency condition becomes $\varepsilon < 0.337$. From a Taylor expansion of the logarithm, we can see that for large $\nu$ the boundary separating inconsistency from consistency is at $\varepsilon \sim \nu/2$, whereas for small $\nu$ the boundary is at $\varepsilon \sim \nu^2/2$. Thus, when $\nu$ is large, $\varepsilon$ need not be small to cause inconsistency; in fact, inconsistency occurs when $\varepsilon$ is about half as large as $\nu$. On the other hand, when $\nu$ is very small, $\varepsilon$ must be much smaller still to force inconsistency. This makes qualitative sense: When the distances in $P_{\varepsilon}$ are small, the "underestimation" phenomenon discussed after (4.7) is only slight, so it is more difficult for this effect to undermine consistency.

## 5. MAXIMUM LIKELIHOOD

For an optimist, the demonstrations of the previous section would not eliminate all hope that maximum likelihood might consistently estimate the tree topology. Presumably, maximum likelihood can, in some sense, "make use of more of the information in the data" than distance methods can. For example, distance methods are restricted to doing calculations with marginal distributions of pairs of taxa, while maximum likelihood can use the joint distribution of all taxa. Unfortunately, however, the same simple examples presented in the previous section can be shown to cause maximum likelihood to be inconsistent. In this section we will discuss the example with invariable sites in detail.

An intuitive outline of the reasoning is as follows. We imagine that we are observing a sequence of characters from the mixture model $P^{\varepsilon}_{\text{mix}}$, which has the topology $\mathcal{T}_{AB}$. Proposition 5.2 will show that the model $P^{0}_{\text{mix}}$ induces precisely the same distribution of terminal character states as does the model $P_{\theta}$ depicted in Figure 7, which has the topology $\mathcal{T}_{AC}$.

From this it follows that for small $\varepsilon > 0$, the distribution of data generated from the model $P^{\varepsilon}_{\text{mix}}$ is nearly the same as that from the model $P_\theta$. Thus, in this sense, assuming that $\varepsilon$ is sufficiently small, it turns out that there is a pure Markov model $P_\theta$ in the incorrect topology $\mathcal{T}_{AC}$ that is closer to $P^{\varepsilon}_{\text{mix}}$ than any pure Markov model in the true topology $\mathcal{T}_{AB}$ is. Proposition 5.1 formalizes the notion that this will cause the maximum likelihood estimator to fail to belong to the true topology with probability 1 for sufficiently large sample sizes.

To begin, we state without proof the following simple variation of similar results established by Wald [21]; the proof follows the method of Wald.

*PROPOSITION 5.1*

*Suppose the random variables $X_1, X_2, \ldots$ taking values in a finite set $\mathcal{X}$ are independent and identically distributed with probability distribution $P$, and let $\{P_\theta\colon\ \theta \in \Theta\}$ be a family of probability distributions on $\mathcal{X}$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ maximize $\sum_{i=1}^{n} \log P_\theta(X_i)$ over $\theta \in \Theta$. Suppose that $\mathcal{T}$ is a compact subset of $\Theta$, the function $\theta \mapsto P_\theta(x)$ is continuous for all $x$, and*

$$\sup_{\theta \in \mathcal{T}} \sum_x P(x) \log P_\theta(x) < \sup_{\theta \in \Theta} \sum_x P(x) \log P_\theta(x).$$

*Then $P\{\hat{\theta}_n \notin \mathcal{T}$ eventually$\} = 1$; that is, there is a random variable $N$, finite with probability 1, such that $\hat{\theta}_n \notin \mathcal{T}$ for all $n \geqslant N$.*

Note that the family $\{P_\theta\colon\ \theta \in \Theta\}$ need not contain the true distribution $P$ that generated the data. This will be important because we will be taking $\Theta$ to be the set of Markov models described in Section 2, whereas the true distribution will be a mixture of the type described in Figure 6. Also observe that since we have allowed branch lengths to take on the value $\infty$, the tree topologies $\mathcal{T}_{AB}$, $\mathcal{T}_{AC}$, and $\mathcal{T}_{AD}$ are compact subsets of $\Theta$.

The next result is the strengthening of Proposition 4.3 promised in Section 4.

*PROPOSITION 5.2*

*For each $\nu \in [0, \infty]$, the Markov model $P_\theta$ described in Figure 7 and the probability $P^{0}_{\text{mix}}$ have exactly the same joint distributions for $(X_A, X_B, X_C, X_D)$.*

*Proof.* Proposition 4.3 tells us that the two models $P^{0}_{\text{mix}}$ and $P_\theta$ agree in their distances separating pairs of taxa. Therefore, $P^{0}_{\text{mix}}$ and $P_\theta$ must agree in the marginal distributions they give to the six pairs

$(X_A, X_B), (X_A, X_C), \ldots, (X_C, X_D)$. We want to show that $P^0_{\text{mix}}$ and $P_\theta$ agree in the full joint distributions they give to the vector $(X_A, X_B, X_C, X_D)$. For convenience, let $P_0$ denote $P^0_{\text{mix}}$ and let $P_1$ denote $P_\theta$. Note that $P_i\{X_A = X_C\} = 1$ for both $i = 0, 1$, so that we need consider only the joint distributions of $(X_A, X_B, X_D)$. Next observe that both probabilities $P_i$ for $i = 0, 1$ satisfy

$$\begin{aligned} P_i\{X_A = 0, X_B = 0, X_D = 0\} &= P_i\{X_A = 1, X_B = 1, X_D = 1\} =: a_i, \\ P_i\{X_A = 0, X_B = 0, X_D = 1\} &= P_i\{X_A = 0, X_B = 1, X_D = 0\} \\ &= P_i\{X_A = 1, X_B = 0, X_D = 1\} \\ &= P_i\{X_A = 1, X_B = 1, X_D = 0\} =: b_i, \end{aligned}$$

and

$$P_i\{X_A = 0, X_B = 1, X_D = 1\} = P_i\{X_A = 1, X_B = 0, X_D = 0\} := c_i,$$

where the symbol " =: " indicates that the expression on the right-hand side is being defined by the expression on the left-hand side. Thus, the equality of the pairwise distributions gives as a particular case that

$$\begin{aligned} a_0 + c_0 &= P_0\{X_A = 0, X_B = 0, X_D = 0\} + P_0\{X_A = 1, X_B = 0, X_D = 0\} \\ &= P_0\{X_B = 0, X_D = 0\} = P_1\{X_B = 0, X_D = 0\} \\ &= P_1\{X_A = 0, X_B = 0, X_D = 0\} + P_1\{X_A = 1, X_B = 0, X_D = 0\} \\ &= a_1 + c_1, \end{aligned}$$

which, upon making the substitution $c_i = 1/2 - a_i - 2b_i$ for $i = 0, 1$, yields that $b_0 = b_1$. Similarly, another calculation like that in the last display gives $a_0 + b_0 = a_1 + b_1$, so that we obtain $a_0 = a_1$, and we are done. ■

Next we prove the main inconsistency result for the maximum likelihood topology estimator.

*THEOREM 5.3*

*Let $r \in (0, 1)$ and consider the maximum likelihood topology estimator, where the likelihood is maximized over all pure Markov models. There exists $\varepsilon > 0$ such that the maximum likelihood topology estimator is inconsistent when the true model is the mixture $P^\varepsilon_{\text{mix}} = (1 - r)P_\varepsilon + rP_{\text{inv}}$.*

*Proof.* We want to establish the existence of a positive $\varepsilon$ such that the maximum likelihood estimator $\hat{\mathcal{T}}_n$ based on $n$ observations from the mixture distribution $P^\varepsilon_{\text{mix}}$ has probability 1 of eventually lying outside

the true tree topology $\mathcal{T}_{AB}$ for large enough $n$. In fact, we will show that for sufficiently large $n$, $\hat{\mathcal{T}}_n$ lies in $\mathcal{T}_{AC}$, that is,

$$\hat{\mathcal{T}}_n \in \mathcal{T}_{AC} - (\mathcal{T}_{AB} \cup \mathcal{T}_{AD}) = \Theta - (\mathcal{T}_{AB} \cup \mathcal{T}_{AD}).$$

For $P$ and $Q$ joint distributions on $(X_A, X_B, X_C, X_D)$, define

$$g(P,Q) = \sum_x P(x) \log Q(x);$$

this takes the value $-\infty$ if there is an $x$ such that $P(x) > 0$ and $Q(x) = 0$. A standard property of the function $g$ is that

$$g(P,Q) < g(P,P) \qquad \text{whenever } Q \neq P; \tag{5.9}$$

see, for example, Lemma 1.4.1 of [32].

We claim that

$$\max_{\theta \in \mathcal{T}_{AB} \cup \mathcal{T}_{AD}} g(P^0_{\text{mix}}, P_\theta) < \max_{\theta \in \Theta} g(P^0_{\text{mix}}, P_\theta). \tag{5.10}$$

To verify this, observe that, by (5.9), it is sufficient to show that $P^0_{\text{mix}} \in \Theta$ while $P^0_{\text{mix}} \notin \mathcal{T}_{AB} \cup \mathcal{T}_{AD}$. However, Proposition 5.2 shows that $P^0_{\text{mix}} \in \mathcal{T}_{AC} \subset \Theta$. To show that $P^0_{\text{mix}} \notin \mathcal{T}_{AB} \cup \mathcal{T}_{AD}$, note that the distances $d$ corresponding to any model in $\mathcal{T}_{AB}$ satisfy the inequality $d(A,B) + d(C,D) \leqslant d(A,C) + d(B,D)$. However, inspection of Figure 7 together with (4.8) shows that the distances $d^0_{\text{mix}}$ under $P^0_{\text{mix}}$ satisfy

$$\begin{aligned} d^0_{\text{mix}}(A,B) + d^0_{\text{mix}}(C,D) &= 2d_{\text{mix}}(\nu) > d_{\text{mix}}(2\nu) \\ &= d^0_{\text{mix}}(A,C) + d^0_{\text{mix}}(B,D). \end{aligned}$$

Thus, $P^0_{\text{mix}} \notin \mathcal{T}_{AB}$; similarly, $P^0_{\text{mix}} \notin \mathcal{T}_{AD}$.

Now by Proposition 5.1, to complete the proof of the theorem it suffices to establish that, as $\varepsilon \to 0$,

$$\max_{\theta \in \mathcal{T}_{AB} \cup \mathcal{T}_{AD}} g(P^\varepsilon_{\text{mix}}, P_\theta) \to \max_{\theta \in \mathcal{T}_{AB} \cup \mathcal{T}_{AD}} g(P^0_{\text{mix}}, P_\theta) \tag{5.11}$$

and

$$\max_{\theta \in \Theta} g(P^\varepsilon_{\text{mix}}, P_\theta) \to \max_{\theta \in \Theta} g(P^0_{\text{mix}}, P_\theta), \tag{5.12}$$

because these will combine with (5.10) to give

$$\max_{\theta \in \mathcal{T}_{AB} \cup \mathcal{T}_{AD}} g(P^\varepsilon_{\text{mix}}, P_\theta) < \max_{\theta \in \Theta} g(P^\varepsilon_{\text{mix}}, P_\theta)$$

for sufficiently small $\varepsilon$. To circumvent concerns about $g$ taking on the value $-\infty$, define $h(P,Q) = \exp[g(P,Q)]$ if $g(P,Q) > -\infty$ and $h(P,Q) = 0$ if $g(P,Q) = -\infty$. Letting $\mathscr{T}$ denote a compact subset of $\Theta$, it suffices to show that

$$\sup_{\theta \in \mathscr{T}} h(P^{\varepsilon}_{\text{mix}}, P_\theta) \to \sup_{\theta \in \mathscr{T}} h(P^{0}_{\text{mix}}, P_\theta) \tag{5.13}$$

as $\varepsilon \to 0$ to establish the desired assertions (5.11) and (5.12). However, letting $\mathscr{B}$ denote a closed ball around $P^0_{\text{mix}}$, note that $h$ is a continuous function on the compact set $\mathscr{B} \times \mathscr{T}$, so that $h$ is uniformly continuous there. From this, the fact that $P^{\varepsilon}_{\text{mix}} \to P^0_{\text{mix}}$ as $\varepsilon \to 0$ easily implies (5.13). ■

## 6. DISCUSSION

The existence of site-to-site heterogeneity in rates of evolution is well established. The examples in this paper used mild, unexotic forms of rate heterogeneity together with the type of branch length patterns found in the standard example of the inconsistency of parsimony. In this sense, the distance and maximum likelihood methods considered in this paper join parsimony in being susceptible to the charge of inconsistency on certain simple models. How serious the implications of this type of theoretical result are for phylogenetic practice is an issue that has long been debated; see Sober [33], for example.

The examples presented here do not constitute an indictment of maximum likelihood as a general principle for deriving estimators. In our context, they show a failure of standard methods that maximize the likelihood over the class of pure Markov models; more elaborate methods that maximize the likelihood over more general mixtures of Markov models, such as some of the methods cited in the Introduction, can be expected to give improved performance. Establishing the extent to which such methods are successful under various levels of generality in heterogeneous models is an area for further research.

The basic mathematical idea behind the counterexamples is this. We find a pair of pure Markov models that both belong to the intersection of two tree topologies, say $\mathscr{T}_{AB}$ and $\mathscr{T}_{AC}$, such that their mixture lies well within $\mathscr{T}_{AC}$, say, and outside $\mathscr{T}_{AB}$. Then by slightly perturbing the pair of Markov models, we can move them slightly inside the topology $\mathscr{T}_{AB}$, while their mixture remains closer to $\mathscr{T}_{AC}$ than to $\mathscr{T}_{AB}$. In particular, for the example with invariable characters described in Figure 6, the original pair of Markov models is $P_0$ and $P_{\text{inv}}$. These both lie in $\mathscr{T}_{AB} \cap \mathscr{T}_{AC}$, but by Proposition 5.2 their mixture lies within $\mathscr{T}_{AC}$. The slight perturbation is to change $P_0$ to $P_\varepsilon$ for a small positive $\varepsilon$.

The counterexamples were produced in the very simple setting described in Section 2, with binary characters, four taxa, and so on. We adopted this simple setting not only for ease of exposition but also, at least as important, because of the nature of the study of counterexamples. If an example of inconsistency exists within a special class of models, then an example obviously exists within any more general class of models that contains the special class. For example, if we had displayed an example of inconsistency using four-valued characters, then we would still wonder whether an example exists using only binary characters. However, finding such an example using binary characters implies in a trivial way that an example using four-valued characters exists. Similarly, having produced an example that is simply a mixture of two "rates," one of which is zero, we see that inconsistency does not require more than two rates and in particular does not require a continuum of rates having a continuous distribution, for example.

It is also clear that examples of inconsistency exist that do not involve strictly invariable characters but still lie within the class of models of row 3 of Figure 3. For instance, the example of Figure 6 could be perturbed slightly by choosing a small number $\delta > 0$ and replacing the invariant character model $P_{\text{inv}}$ by a tree geometrically similar to that of $P_\varepsilon$, with branch lengths all multiplied by $\delta$. By continuity considerations, it is easy to see that there is a positive $\delta$ that is sufficiently small that the resulting perturbed example would still be an example of inconsistency for the distance and maximum likelihood methods. In the resulting example, both of the distributions in the mixture strictly favor the tree topology $\mathcal{T}_{AB}$; if they were not mixed, they would both cause the estimation method to converge to the correct topology $\mathcal{T}_{AB}$. Under the mixture, however, the methods converge with probability 1 to the incorrect topology $\mathcal{T}_{AC}$.

A recent investigation by Steel et al. [28] presents a remarkable example showing that the tree topology is in general not identifiable from the joint distribution of character states at the terminal nodes of the tree when different positions are allowed to evolve at different rates. Their example reveals that there are fundamental limits, the full nature and extent of which are not yet clear, on what may be achieved by any method in such heterogeneous models. This result is complementary to the results presented here. The example of [28] exhibits two mixture models from different topologies that have precisely the same joint distribution of terminal character states, with both mixture models being of the type depicted in row 3 of Figure 3. Our main example also displays two models from different topologies. The models are simpler than those of [28], with one of the models being a pure Markov model (row 1 of Figure 3) and the other being from row 2 of Figure 3. Here

the pair of models have joint distributions of terminal character states that are arbitrarily close to each other, but not identical. This is enough to cause the methods under consideration to become positively misleading.

Finally, how can we reconcile the counterexamples with the positive statements of Cavender and Felsenstein [19] and Chang and Hartigan [18] for the distance methods and Felsenstein [20] for maximum likelihood? Those results state that if characters are truly generated by a single Markov model, then the distance methods and maximum likelihood will all consistently recover the tree topology. The resolution is this: A mixture of Markov models is generally not a Markov model. This is easy to see. Consider, for example, a character generated by the model $P^{\varepsilon}_{\text{mix}}$ of Figure 6, with $\nu > 0$. Referring to Figure 6, let $E$ and $F$ denote the internal nodes adjacent to $A$ and $C$, respectively, so that the neighbors of $E$ are $A$, $B$, and $F$. Supposing that we know the character states $X_A$, $X_B$, and $X_F$, to show that the Markov property fails we ask: Could knowledge of any character states other than those neighboring states have any effect on probabilistic statements about $X_E$? Clearly the answer is yes. For example, suppose we know that $X_A = X_B = X_F = 0$. If we then further learn that $X_C = 1$, we have gained the additional knowledge that the character is not invariant, so that it must have been generated by the model $P_{\varepsilon}$. Thus, using the Markov property of $P_{\varepsilon}$,

$$P^{\varepsilon}_{\text{mix}}\{X_E = 0 | X_A = X_B = X_F = 0,\ X_C = 1\} = P_{\varepsilon}\{X_E = 0 | X_A = X_B = X_F = 0\}.$$

The last probability is clearly smaller than the mixture probability $P^{\varepsilon}_{\text{mix}}\{X_E = 0 | X_A = X_B = X_F = 0\}$. Thus, given the neighboring states $X_A = X_B = X_F = 0$, the additional knowledge that $X_C = 1$ decreases the conditional probability under $P^{\varepsilon}_{\text{mix}}$ that $X_E = 0$. This violation of the Markov property calls into question what might otherwise seem to be a rather innocuous Markov assumption.

*I am grateful to Colleen Kelly for introducing me to the issue of rate heterogeneity and for a number of stimulating discussions about this work. Thanks also to the referees for their helpful comments.*

REFERENCES

1 T. H. Jukes and C. R. Cantor, Evolution of protein molecules, in *Mammalian Protein Metabolism*, H. N. Munro, ed., Academic, New York, 1969, pp. 21–132.

2 J. A. Cavender, Taxonomy with confidence, *Math. Biosci.* 40:271–280 (1978).

3 M. Kimura, A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences, *J. Mol. Evol.* 16:111–120 (1980).

4 D. Barry and J. A. Hartigan, Asynchronous distance between homologous DNA sequences, *Biometrics* 43:261–276 (1987).

5 J. Felsenstein, Statistical inference of phylogenies, *J. Roy. Stat. Soc. A* 146:264–272 (1983).

6 J. Felsenstein, Phylogenies from molecular sequences: inference and reliability, *Annu. Rev. Genet.* 22:521–565 (1988).

7 D. Barry and J. A. Hartigan, Statistical analysis of hominoid molecular evolution, *Stat. Sci.* 2:191–210 (1987).

8 W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution*, Sinauer, Sunderland, MA, 1991.

9 M. Nei, *Molecular Evolutionary Genetrics*, Columbia Univ. Press, New York, 1987.

10 W. M. Fitch and E. Margoliash, A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case, *Biochem. Genet.* 1:65–71 (1967).

11 J. S. Shoemaker and W. M. Fitch, Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated, *Mol. Biol. Evol.* 6:270–289 (1989).

12 G. A. Churchill, A. von Haeseler, and W. C. Navidi, Sample size for a phylogenetic inference, *Mol. Biol. Evol.* 9:753–769 (1992).

13 L. Jin and M. Nei, Limitations of the evolutionary parsimony method of phylogenetic analysis, *Mol. Biol. Evol.* 7:82–102 (1990).

14 C. Kelly and J. Rice, Modeling molecular evolution: a heterogeneous rate analysis, *Math. Biosci.* 133:85–109 (1996).

15 R. Lundstrom, S. Tavaré, and R. H. Ward, Modeling the evolution of the human mitochondrial genome, *Math. Biosci.* 112:319–335 (1992).

16 A. Sidow, T. Nguyen, and T. P. Speed, Estimating the fraction of invariable codons with a capture–recapture method, *J. Mol. Evol.* 35:253–260 (1992).

17 Z. Yang, Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates vary over sites, *Mol. Biol. Evol.* 10:1396–1401 (1993).

18 J. T. Chang and J. A. Hartigan, Reconstruction of evolutionary trees from pairwise distributions on current species, in *Computing Science and Statistics*, Proc. 23rd Symp. on the Interface, E. M. Keramidas, ed., Interface Foundation, Fairfax Station, VA, 1991, pp. 254–257.

19 J. A. Cavender and J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, *J. Classif.* 4:57–71 (1987).

20 J. Felsenstein, Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters, *Syst. Zool.* 22:240–249 (1973).

21 A. Wald, Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stat.* 20:595–601 (1949).

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

23 J. S. Farris, A. G. Kluge, and M. J. Eckardt, A numerical approach to phylogenetic systematics, *Syst. Zool.* 19:172–191 (1970).

24 W. M. Fitch, Toward defining the course of evolution: minimum change for a specific tree topology, *Syst. Zool.* 20:406–416 (1971).

25 Y. Tateno, N. Takezaki, and M. Nei, Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods where substitution rate varies with site, *Mol. Biol. Evol.* 11:261–277 (1994).
26 M. K. Kuhner and J. Felsenstein, A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates, *Mol. Biol. Evol.* 11:459–468 (1994).
27 J. J. Bull, J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell, Partitioning and combining data in phyogenetic analysis, *Syst. Biol.* 42:384–397 (1993).
28 M. A. Steel, L. A. Szekely, and M. D. Hendy, Reconstructing trees when sequence sites evolve at variable rates, *J. Comput. Biol.* 1:153–163 (1994).
29 M. Hasegawa, H. Kishino, and T. Yano, Dating of the human–ape splitting by a molecular clock of mitrochondrial DNA, *J. Mol. Evol.* 22:160–174 (1985).
30 P. J. Lockhart, M. A. Steel, M. D. Hendy, and D. Penny, Recoving evolutionary trees under a more realistic model of sequence evolution, *Mol. Biol. Evol.* 11:605–612 (1994).
31 J. A. Lake, Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances, *Proc. Natl. Acad. Sci. USA* 91:1455–1459 (1994).
32 R. B. Ash, *Information Theory*, Wiley, New York, 1965.
33 E. Sober, *Reconstructing the Past: Parsimony, Evolution, and Inference*, MIT Press, Cambridge, MA, 1988.