ELSEVIER

Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency

JOSEPH T. CHANG
Department of Statistics, Yale University, Box 208290 Yale Station; New Haven, CT 06520-8290

Received November 20, 1995; revised April 30, 1996; accepted May 1, 1996.

ABSTRACT

A Markov model of evolution of characters on a phylogenetic tree consists of a tree topology together with a specification of probability transition matrices on the edges of the tree. Previous work has shown that, under mild conditions, the tree topology may be reconstructed, in the sense that the topology is identifiable from knowledge of the joint distribution of character states at pairs of terminal nodes of the tree. Also, the method of maximum likelihood is statistically consistent for inferring the tree topology. In this article we answer the analogous questions for reconstructing the full model, including the edge transition matrices. Under mild conditions, such full reconstruction is achievable, not by using pairs of terminal nodes, but rather by using triples of terminal nodes. The identifiability result generalizes previous results that were restricted either to characters having two states or to transition matrices having special structure. The proof develops matrix relationships that may be exploited to identify the model. We also use the identifiability result to prove that the method of maximum likelihood is consistent for reconstructing the full model.

1. INTRODUCTION

Evolutionary relationships among species are commonly conceptualized in terms of an "evolutionary tree." Statistical approaches to phylogeny reconstruction use observations on the terminal nodes of the tree to estimate parameters of a model of the evolutionary process. The class of Markov models on evolutionary trees has proved to be a useful compromise between biological realism and analytical tractability. Such

Address correspondence to Joseph T. Chang, Department of Statistics, Yale University, Box 208290 Yale Station; New Haven, CT 06520-8290.

MATHEMATICAL BIOSCIENCES 137:51–73 (1996)

655 Avenue of the Americas, New York, NY 10010
0025-5564/96/$15.00
PII S0025-5564(96)00075-2

a Markov model consists of a tree topology together with a specification of probability transition matrices on the edges of the tree. The topology summarizes in a qualitative sense certain aspects of the evolutionary relationships, such as which species are the most closely related. Quantitative information about such matters as the timing of the speciation events and the rates of evolution is contained in the edge transition matrices. Thus, the problem of inferring these matrices is of considerable scientific interest.

Previous work on the theory of the general Markov model has shown that under mild conditions, the tree topology may be reconstructed in the sense that the topology is identifiable from knowledge of the joint distribution of character states at the terminal nodes of the tree. Also, the method of maximum likelihood has the statistical property of consistency for inferring the tree topology. However, the analogous questions for reconstructing the full model, including the edge transition matrices, have remained unresolved.

This article shows that under some mild conditions on the transition matrices, such full reconstruction is achievable. The proof develops matrix relationships that may be exploited to identify the model. In addition, we show as a consequence of the identifiability result that the method of maximum likelihood is consistent for reconstructing the full model.

The general class Markov models contains as special cases the most popular and most studied models in phylogeny construction. For example, the models of Cavender [1], Jukes and Cantor [2], Kimura [3], and Tajima and Nei [4], among others, are all Markov models that contain different numbers of parameters in their edge transition matrices. In the case of DNA sequence data, the general Markov model can accommodate 12 parameters per edge transition matrix. This general Markov model was studied by Barry and Hartigan [5], who introduced a distance measure whose use has been increasing steadily. As data sets become larger, the simpler Markov models often do not achieve an adequate fit, and more general models are required, as pointed out by Rzhetsky and Nei [6].

The issue addressed by identifiability is whether or not the inference problem is well posed, in the following sense. A set of data consists of observations of character states at the terminal nodes of the tree; character states at the internal nodes are hidden from view. The question is whether or not such observations at the terminal nodes potentially contain enough information to reveal the details of the internal structure of the Markov model. We hope that the model is revealed more and more accurately as the sample size increases, so we ask whether we would in fact know the model precisely if we were given

an "infinite sample." Having such a hypothetical sample would mean having exact knowledge of the joint distribution of character states at the terminal nodes. Thus, our question is whether knowledge of this joint distribution is sufficient to determine the model; if so, we say the model is *identifiable*. Identifiability fails if two different models (differing in topology or edge transition matrices or both) give rise to the same joint distribution of character states at the terminal nodes; in this case there would be no hope of using data to distinguish between those two models.

Identifiability of the tree topology from the distribution of character states at terminal nodes was established by Chang and Hartigan [7]; it was also found independently by Steel, Hendy, and Penny [8, 9]. This work is reviewed in Section 3. Previous results on the identifiability of the full model, including the edge transition matrices, were obtained by Pearl and Tarsi [10]. Their results were limited to the case of two-state characters; the general case of a finite state space requires somewhat more elaborate methods. We give a general, unified treatment that identifies the rows of the edge transition matrices as eigenvectors of certain matrices that can be formed from joint distributions of triples of terminal nodes. In this level of generality certain conceptual issues appear that do not arise for two-state characters. For example, whereas in the case of two-state characters conditions on the determinants of the transition matrices are sufficient to establish identifiability, this is not so in the general case. A related complication is that in the general case it becomes useful to deal with classes of matrices that are not closed under multiplication. In the two-state case it is sufficient to consider the class of matrices that have a positive determinant, which is closed under multiplication.

In view of the nature of DNA, the case of four-state characters is of particular interest for phylogenetic analysis. Here Steel, Hendy, and Penny [8, 9] have obtained the most general identifiability results to date, using the Hadarmard inversion technique introduced by Hendy [11]. However, these results are limited to models incorporating strong symmetry assumptions. The most general model they considered for four-state characters is a generalized Kimura 3ST model, in which each edge transition matrix is a member of the three-parameter family of transition matrices of the form

$$\begin{pmatrix} 1-a-b-c & a & b & c \\ a & 1-a-b-c & c & b \\ b & c & 1-a-b-c & a \\ c & b & a & 1-a-b-c \end{pmatrix}.$$

Although they create the analytic simplifications exploited by Hadamard inversion, the symmetries assumed by restricting the general 12-parameter family of 4×4 Markov transition matrices to such a three-parameter family do not allow the modeling of known biological phenomena such as nonuniformities in frequencies of the four nucleotides.

An important part of the foundation for this work is Paul Lazarsfeld's development of latent structure analysis, begun around 1950 and described in detail by the book of Lazarsfeld and Henry [12]. The idea of latent structure analysis is to model observed random variables as conditionally independent outcomes, given the state of an unobserved "latent" variable. Thus, to identify the parameters in a star phylogeny may be viewed as a problem of latent structure analysis. We give a short, self-contained development involving simple matrix manipulations inspired by the treatment of Lazarsfeld, although they seem somewhat simpler and they apply to random variables taking any finite number of values. One feature here that appears to be novel is a perturbation argument that proves the general case by first handling a special case where a certain matrix has distinct eigenvalues.

A feature special to the star phylogeny estimation problem that is not present in general latent structure problems is that the number of classes (or "states") is the same for each of the observed, as well as the latent, variables; for example, for DNA data there are four latent classes: A, C, G, T. One mathematically convenient implication of this special structure is that we work entirely with square matrices. In this problem the latent classes have a more definite meaning than is typical in other uses of latent structure analysis in the social sciences, for example, where the latent classes may have no particular interpretation, at least initially, and one might try to give them an interpretation after looking at the values of the fitted parameters. This difference has a burdensome aspect: whereas typically in latent structure analysis one might not be concerned with giving correct labels to the states of the hidden random variable, in phylogeny estimation this becomes important.

A statistical estimation method is said to he *consistent* if the estimated quantity is certain to converge to the true quantity as the number of observations used in forming the estimate tends to infinity. Felsenstein [13] sparked the considerable and continuing interest in the issue of consistency in phylogenetic analysis with his striking observation that the popular method of parsimony is not consistent for estimating the tree topology over the class of Markov models.

For models of the type we are considering, identifiability considerations are the principal difficulty in establishing the consistency of maximum likelihood. Claims have appeared (e.g., [14]), without proof,

that maximum likelihood is consistent for estimating the tree topology. Symptoms of the resulting unsatisfactory state of affairs include explicitly stated doubts in the literature (e.g., chapter 5 of [15]) as to whether consistency of the maximum likelihood topology estimate has been established and, indeed, whether or not it is true. Such issues deserve a careful treatment. In fact, the identifiability results required to establish consistency have not been in place until recently ([7–9]) for the problem of topology estimation and until the present article for the estimation of the full model. In Section 5 we apply the identifiability result from Section 4 to prove the consistency of the maximum likelihood estimators of the edge transition matrices.

2. MARKOV MODELS ON TREES: DEFINITIONS AND NOTATION

Let T denote a finite set of taxa; this is typically the set of current species whose phylogenetic history we wish to infer. A tree, defined as a connected graph without cycles, consists of nodes and edges. The *degree* of a node is the number of edges incident to the node. Nodes of degree one are *terminal nodes*, and nodes of higher degree are *internal nodes*. Thus, S may be partitioned into a union $S = T \cup N$ of the set T of terminal nodes and the set N of nonterminal nodes; the notational overlap between the two uses of T is intentional, because each taxon in T is identified with a terminal node in the graph. The terminal nodes are labeled by names of taxa in T. We assume that speciation events occur at internal nodes, so that the tree has no nodes of degree 2. Internal nodes may have degree 3 or greater; a node of degree d corresponds to the splitting of one species into $d-1$ species. An edge $e \in E$ may be thought of as a subset $e = \{r, s\}$ containing two distinct nodes r, $s \in S$. These edges are undirected, so that $\{s, r\} = \{r, s\}$; this is consistent with conceptualizing an edge as a set of two nodes rather than an ordered pair of nodes.

Let $\mathscr{C}$ denote a finite set of character states, and let C denote $|\mathscr{C}|$, the cardinality of $\mathscr{C}$. For example, $\mathscr{C}$ might be the set of four nucleotides. The evolution of a character is modeled as a random process, as follows. For each $s \in S$ there is a corresponding random variable X_s taking values in $\mathscr{C}$; for example, X_s might identify the nucleotide occupying a particular site in the DNA of a representative of species s. Let $\pi^s(\cdot)$ denote the marginal distribution of X_s, that is, $\pi^s(i) = \mathbb{P}\{X_s = i\}$ for $i \in \mathscr{C}$. We assume that $\{X_s : s \in S\}$ is a Markov random field on $\mathscr{T}$, which means that for each $s \in S$, the conditional distribution of X_s given all of the other values $\{X_r : r \neq s\}$ is the same as the conditional distribution of X_s given just the values $\{X_r : \{r, s\} \in E\}$ at the

"neighbors" of s. For each edge $\{r,s\}$ there are two $C \times C$ edge transition matrices P^{rs} and P^{sr} whose entries are given by

$$P^{rs}(i,j) = \mathbb{P}\{X_s = j \mid X_r = i\};$$

these conditional probabilities are well defined if the marginal probabilities $\pi^r(i)$ are all positive.

This completes the description of the probabilistic model for the evolution of a single character X. To model n characters for each species, the standard assumption—usually made grudgingly for the sake of simplicity—is that distinct characters are independent and identically distributed (*iid*); that is, $X^1, \ldots, X^n$ are *iid*, where each $X^i = \{X_s^i : s \in S\}$ is a Markov random field on $\mathcal{T}$.

For each character X^i, we observe the states $X_T^i = (X_t^i : t \in T)$ at the terminal nodes, but not the states $X_N^i = (X_s^i : s \in N)$ at the nonterminal nodes. The statistical problem is to use the observations $X_T^1, \ldots, X_T^n$ at the terminal nodes to infer the tree topology and the edge transition matrices. Inference of the edge transition matrices is the main problem considered in this article. Before turning to this problem in Section 4, in Section 3 we briefly review a result on the identification of the topology.

3. IDENTIFIABILITY OF THE TOPOLOGY

For future reference, in this section we review a result found by Chang and Hartigan [7] and, independently, by Steel et al. [8]. This result states that, under mild conditions, the tree topology is identifiable in the general Markov model from the joint distribution of states at the terminal nodes. In fact, much less information than the full joint distribution of $(X_t : t \in T)$ is required. It turns out that knowing the joint distributions of *pairs* of terminal nodes (X_t, X_u) for $t, u \in T$ is sufficient to determine the topology.

To formalize the notion of identifying a topology, we define an equivalence relation for evolutionary tree topologies as follows. Let $\mathcal{T}_1 = (S_1, E_1)$ and $\mathcal{T}_2 = (S_2, E_2)$ be trees with the same set of terminal nodes T. We say that $\mathcal{T}_1$ and $\mathcal{T}_2$ are *equivalent* if there is a bijective "relabeling" function $\rho : S_1 \to S_2$ such that $\rho(t) = t$ for all $t \in T$ and $E_2 = \{\{\rho(r), \rho(s)\} : \{r,s\} \in E_1\}$. That is, the topologies $\mathcal{T}_1$ and $\mathcal{T}_2$ are equivalent if they are the same up to a possible relabeling of nonterminal nodes.

PROPOSITION 3.1

Consider a family of Markov models satisfying the following conditions:

1. *The edge transition matrices are invertible and not equal to a permutation matrix.*

2. *There is a node r with $\pi^r(i) > 0$ for each $i \in \mathscr{C}$, that is, each character state has positive marginal probability at r.*

Then the topology is identifiable from the joint distributions of character states at pairs of terminal nodes. That is, if two models in the family induce the same pairwise distributions of character states at their terminal nodes, then the topologies of those two models must be equivalent.

Proposition 3.1 follows by combining a result of Buneman [16] with the fact that the function

$$f\{r,s\} = -\log\det P^{rs} - \log\det P^{sr}$$

is *additive*, in the following sense. For a function f defined on subsets of S consisting of two nodes, we say that f is additive if $f\{r,s\} = \sum_{e \in \text{path}\{r,s\}} f(e)$ for each $r, s \in S$. Here "path$\{r,s\}$" denotes the set of edges on the path joining the nodes r and s; this is empty if $r = s$. Buneman [16] showed that the values of an additive function on pairs of terminal nodes determine the topology of the tree, as well as the function on all pairs of nodes. Thus, knowledge of the function f is sufficient to determine the topology. The function f is clearly determined by the joint distribution of pairs of terminal nodes. The log determinant was first used as a measure of distance by Barry and Hartigan [5] and by Cavender and Felsenstein [17].

4. IDENTIFIABILITY OF THE FULL MODEL

4.1. *PAIRWISE DISTRIBUTIONS DO NOT DETERMINE THE FULL MODEL*

Although knowledge of the joint distributions of pairs of terminal nodes determines the tree topology, it is only in certain restricted classes of Markov models incorporating symmetry assumptions that such pairwise distributions are enough to determine the full model. That is, in the general class of Markov models, we will see that pairwise distributions do not determine the edge transition matrices P^{rs} for $\{r,s\} \in E$.

In fact, as shown by the next proposition, examples of this phenomenon can be found in the smallest nontrivial case: a tree with three terminal nodes $T = \{a,b,c\}$ and one nonterminal node $N = \{m\}$, say. The result is illustrated in Figure 1. Let 1 denote a vector of ones.

PROPOSITION 4.1

Consider a Markov model $\mathbb{P}$ having marginal probability vector π^m at node m and edge transition matrices P^{ms} for $s = a, b, c$. Let R be an

invertible matrix satisfying the conditions

1. $R1 = 1$,
2. *the matrices* $R^{-1}P^{ms}$ *and* $P^{sm}R$ *have nonnegative entries for* $s = a, b, c$, *as does the vector* $\tilde{\pi}^m := \pi^m R$, *and*
3. $R^{\mathrm{T}}\Pi^m R$ *is a diagonal matrix.*

Then the Markov model $\tilde{\mathbb{P}}$ *having marginal probability vector* $\tilde{\pi}^m$ *at node* m *and edge transition matrices* $\tilde{P}^{ms} = R^{-1}P^{ms}$ *for* $s = a, b, c$ *has the same pairwise distributions over the terminal nodes as* $\mathbb{P}$ *does; that is,* $\mathbb{P}\{X_s = i, X_t = j\} = \tilde{\mathbb{P}}\{X_s = i, X_t = j\}$ *for each* $i, j \in \mathscr{C}$ *and each pair of terminal nodes* $s, t \in \{a, b, c\}$.

Proof. First note that the specification of a marginal probability vector π^m at node m and edge transition matrices P^{ma}, P^{mb}, and P^{mc} determines the full joint distribution of a Markov random field $\{X_a, X_b, X_c, X_m\}$, and, in particular, edge transition matrices P^{am}, P^{bm}, and P^{cm}. For example, observing that

$$\pi^a(i)P^{am}(i,j) = \mathbb{P}\{X_a = i, X_m = j\} = \pi^m(j)P^{ma}(j,i)$$

and defining the diagonal matrices

$$\Pi^s = \operatorname{diag}(\pi^s(1), \ldots, \pi^s(C)) \qquad \text{for } s \in S,$$

we see that

$$\Pi^a P^{am} = (\Pi^m P^{ma})^{\mathrm{T}} = (P^{ma})^{\mathrm{T}}\Pi^m,$$

or

$$P^{am} = (\Pi^a)^{-1}(P^{ma})^{\mathrm{T}}\Pi^m. \tag{1}$$

For $s \in \{a, b, c\}$, we have $\tilde{\pi}^s = \tilde{\pi}^m \tilde{P}^{ms} = \pi^s$, so that the marginal distributions of $\mathbb{P}$ and $\tilde{\mathbb{P}}$ agree on the terminal nodes. Therefore, as indicated in Figure 1, the proof becomes a matter of verifying that the reversed transition matrices $\tilde{P}^{sm}$ in the model $\tilde{\mathbb{P}}$ are given by $\tilde{P}^{sm} = P^{sm}R$, since it would then follow that the conditional distributions $P^{st} = P^{sm}P^{mt} = \tilde{P}^{sm}\tilde{P}^{mt} = \tilde{P}^{st}$ agree for the pair of taxa $s, t \in \{a, b, c\}$. Writing $R^{\mathrm{T}}\Pi^m R = \operatorname{diag}(v)$, say, the assumed conditions give

$$v = 1^{\mathrm{T}}(R^{\mathrm{T}}\Pi^m R) = (R1)^{\mathrm{T}}\Pi^m R = 1^{\mathrm{T}}\Pi^m R = \pi^m R = \tilde{\pi}^m,$$

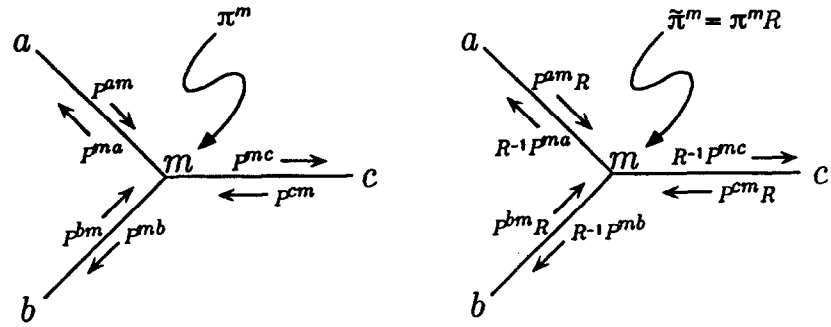

FIG. 1. Two different Markov models that have the same pairwise joint distributions on the terminal nodes.

so that in fact $R^{\mathrm{T}}\Pi^m R = \tilde{\Pi}^m$. Using this, we obtain

$$\begin{aligned}
\tilde{P}^{sm} &= (\tilde{\Pi}^s)^{-1}(\tilde{P}^{ms})^{\mathrm{T}}\tilde{\Pi}^m \qquad \text{by the relation (1)} \\
&= (\Pi^s)^{-1}(R^{-1}P^{ms})^{\mathrm{T}}(R^{\mathrm{T}}\Pi^m R) \\
&= (\Pi^s)^{-1}(P^{ms})^{\mathrm{T}}\Pi^m R \\
&= (P^{sm})R \qquad \text{again by (1)}
\end{aligned}$$

as desired. ■

It is easy to find such examples; in fact, two character states are enough. For example, if $\pi^m = (.5 \quad .5)$,

$$P^{ms} = P^{sm} = \begin{pmatrix} .75 & .25 \\ .25 & .75 \end{pmatrix} \qquad \text{for } s \in \{a, b, c\},$$

and

$$R = \begin{pmatrix} .75 & .25 \\ -.161438 & 1.161438 \end{pmatrix},$$

then

$$\tilde{P}^{sm} = P^{sm}R = \begin{pmatrix} .5221 & .4779 \\ .0664 & .9336 \end{pmatrix},$$

$$\tilde{P}^{ms} = R^{-1}P^{ms} = \begin{pmatrix} .88715 & .11285 \\ .3386 & .6614 \end{pmatrix},$$

and

$$\tilde{\Pi}^m = R^{\mathrm{T}}\Pi^m R = \frac{1}{2}R^{\mathrm{T}}R = \begin{pmatrix} .29428 & 0 \\ 0 & .70572 \end{pmatrix}$$

$$= \operatorname{diag}\{(.5 \quad .5)R\} = \operatorname{diag}(\pi^m R)$$

satisfy all of the conditions in the proposition.

Notice that, although the joint distributions of character states at pairs of terminal nodes are identical under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ in this example, the full joint distribution of character states (X_a, X_b, X_c) under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are different. For example,

$$\mathbb{P}\{X_a = 0, X_b = 0, X_c = 0\} = (.5)(.75)^3 + (.5)(.25)^3 = .21875,$$

but

$$\tilde{\mathbb{P}}\{X_a = 0, X_b = 0, X_c = 0\} = (.29428)(.11285)^3 + (.70757)(.6614)^3$$

$$= .23287.$$

Thus, although the two models $\mathbb{P}$ and $\tilde{\mathbb{P}}$ cannot be distinguished on the basis of the joint distributions they give to pairs of terminal nodes, they can be distinguished by the joint distributions they give to the full set of terminal nodes. The next section will show that this is a general phenomenon.

4.2. *DISTRIBUTIONS OF TRIPLES DETERMINE THE FULL MODEL*

In this section we show that under quite general conditions, two different Markov models can be distinguished by the joint distributions that they assign to their terminal nodes. In fact, the magic number is 3: joint distributions of triples of terminal nodes are enough to determine the full model.

The conditions guaranteeing that the edge transition matrices can be recovered involve the following concept.

DEFINITION

We say that a class of matrices $\mathcal{M}$ is *reconstructible from rows* if for each $M \in \mathcal{M}$ and each permutation matrix $R \neq I$, we have $RM \notin \mathcal{M}$.

That is, $\mathcal{M}$ is reconstructible from rows if a matrix in $\mathcal{M}$ is uniquely determined from its (unordered) set of rows: given this set, we can determine which row is the top row, which is second, and so on. For

example, here is one simple but useful class of matrices that is reconstructible from rows.

EXAMPLE

We say that a square matrix P satisfies the condition *diagonal largest in column* (DLC) if $P(j, j) > P(i, j)$ for all $i \neq j$. Clearly, the class of matrices satisfying the DLC condition is reconstructible from rows: if a matrix satisfies DLC, then no nontrivial row permutation of that matrix can also satisfy DLC.

THEOREM 4.1

Suppose that the evolutionary tree has no nodes of degree 2. Assume that there is a node m such that $\pi^m(i) > 0$ for all $i \in \mathscr{C}$. Assume also that for each edge $\{r, s\}$, the transition matrix P^{rs} is invertible, P^{rs} is not a permutation matrix, and $P^{rs} \in \mathscr{M}$, where $\mathscr{M}$ is a class of matrices that is reconstructible from rows. Then the full model is identifiable. That is, the topology and all of the transition matrices are uniquely determined by the joint distribution of character states at the terminal nodes of the tree.

We begin the proof of the theorem by considering the case of three terminal nodes in the next lemma. The general case will then be treated by an induction argument.

LEMMA 4.1

Consider a Markov model on a tree with three terminal nodes $T = \{a, b, c\}$ and a single nonterminal node $N = \{m\}$. Suppose that $\pi^m(i) > 0$ for all $i \in \mathscr{C}$, that the edge transition matrices are invertible, and that the matrix P^{mb} is a member of a class $\mathscr{M}$ that is reconstructible from rows. Then the full model is identifiable.

NOTATION

For a matrix P, let $P_{i\cdot}$ and $P_{\cdot j}$ denote the ith row and jth column, respectively.

Proof of Lemma 4.1. We present the proof first under the assumption that the matrix P^{mc} has a column whose entries are distinct from each other. That is, suppose there is a $\gamma \in \{1, \ldots, C\}$ such that $P^{mc}(k, \gamma) \neq P^{mc}(l, \gamma)$ for all $k \neq l$.

Our assumptions imply that all of the marginal distributions π^s must be positive; indeed, because π^m is positive by assumption and each column of the invertible matrix P^{ms} must contain at least one positive entry, it follows that each entry of $\pi^s = \pi^m P^{ms}$ must be positive. In particular, since π^a is positive, the conditional probabilities $\mathbb{P}\{X_c = \gamma, X_b = j \mid X_a = i\}$ are well defined; they are uniquely determined by the

joint distribution of (X_a, X_b, X_c). These conditional probabilities satisfy

$$\begin{aligned}
&\mathbb{P}\{X_c = \gamma, X_b = j \mid X_a = i\} \\
&\quad = \sum_k \mathbb{P}\{X_m = k, X_c = \gamma, X_b = j \mid X_a = i\} \\
&\quad = \sum_k \mathbb{P}\{X_m = k \mid X_a = i\} \times \mathbb{P}\{X_c = \gamma \mid X_m = k, X_a = i\} \\
&\qquad \times \mathbb{P}\{X_b = j \mid X_c = \gamma, X_m = k, X_a = i\},
\end{aligned}$$

which, by the Markov property, is the same as

$$\mathbb{P}\{X_c = \gamma, X_b = j \mid X_a = i\} = \sum_k P^{am}(i,k) P^{mc}(k,\gamma) P^{mb}(k,j).$$

Written in matrix form, defining a matrix $P^{ab,\gamma}$ by

$$P^{ab,\gamma}(i,j) = \mathbb{P}\{X_c = \gamma, X_b = j \mid X_a = i\},$$

this becomes

$$P^{ab,\gamma} = P^{am} \operatorname{diag}\left(P^{mc}_{\cdot\gamma}\right) P^{mb}.$$

Thus, multiplying by $(P^{ab})^{-1} = (P^{mb})^{-1}(P^{am})^{-1}$, we obtain

$$\left(P^{ab}\right)^{-1} P^{ab,\gamma} = \left(P^{mb}\right)^{-1} \operatorname{diag}\left(P^{mc}_{\cdot\gamma}\right) P^{mb}. \tag{2}$$

Let G denote the matrix $(P^{ab})^{-1}P^{ab,\gamma}$, and observe that G is determined by the joint distribution of (X_a, X_b, X_c). The identity (2) is an eigenvalue-eigenvector decomposition of G. In particular, the eigenvalues of G are the entries in the column $P^{mc}_{\cdot\gamma}$, which are distinct, by assumption. Writing these eigenvalues as $\lambda_1, \ldots, \lambda_C$, there are C corresponding linearly independent eigenspaces $\Lambda_1, \ldots, \Lambda_C$ defined by $\Lambda_i = \{v^{\mathrm{T}} : v^{\mathrm{T}} G = \lambda_i v^{\mathrm{T}}\}$, and each such subspace is one-dimensional. The set of eigenspaces $\{\Lambda_1, \ldots, \Lambda_C\}$ is uniquely determined by G. Because the rows $P^{mb}_{1\cdot}, \ldots, P^{mb}_{C\cdot}$ form a basis of eigenvectors of G, each such row must belong to one of the eigenspaces, and each eigenspace must contain one of the rows. So each subspace Λ_i contains a unique vector v_i^{T} that satisfies the normalization condition $v_i^{\mathrm{T}}\mathbf{1} = 1$, and we have the equality $\{v_1^{\mathrm{T}}, \ldots, v_C^{\mathrm{T}}\} = \{P^{mb}_{1\cdot}, \ldots, P^{mb}_{C\cdot}\}$ as sets. Thus, because the set of normalized eigenvectors $\{v_1^{\mathrm{T}}, \ldots, v_C^{\mathrm{T}}\}$ is determined by G, the set of rows of P^{mb} is determined by G. Therefore, by the assumption that $P^{mb} \in \mathcal{M}$, the matrix P^{mb} itself is determined by G and, hence, by the joint distribution of (X_a, X_b, X_c).

Having recovered the matrix P^{mb}, we may deduce the marginal probabilities $\pi^m = \pi^b(P^{mb})^{-1}$ and hence also the transition matrix $P^{bm} = (\Pi^b)^{-1}(P^{mb})^{\mathrm{T}}\Pi^m$. At this point we further obtain $P^{mc} = (P^{bm})^{-1}P^{bc}$ and, similarly, all of the remaining transition matrices.

It remains to prove the result without the assumption that P^{mc} has a column with distinct entries. We want to show that, under the assumed conditions, there is still only one choice of the matrix P^{mb} that can lead to the given joint distribution for (X_a, X_b, X_c). Using the assumed invertibility of P^{mc}, it is easy to see that there exists an invertible transition matrix Q such that $P^{mc}Q_{\cdot 1}$ has distinct entries. Consider the model produced by adding a new node d and a new edge $\{c, d\}$ to the given model, taking the transition matrix P^{cd} to be Q. The joint distribution of (X_a, X_b, X_d) is uniquely determined from that of (X_a, X_b, X_c) by

$$\mathbb{P}\{X_a = i, X_b = j, X_d = l\} = \sum_k \mathbb{P}\{X_a = i, X_b = j, X_c = k\}Q(k,l).$$

Clearly $P^{md} = P^{mc}Q$ is invertible, $\pi^d = \pi^c Q$ has positive entries (as π^c has positive entries and the invertible matrix Q cannot have a column of zeroes), and $P^{dm} = (\Pi^d)^{-1}(P^{md})^{\mathrm{T}}\Pi^m$ is invertible. Thus, as the matrix P^{md} has a column with distinct entries, we may apply what we have just shown to the model for (X_a, X_b, X_d) to conclude that the matrix P^{mb} is uniquely determined. That is, if there were two different models with two different matrices P^{mb} that both give the same distribution for (X_a, X_b, X_c), then there would be two different models having two different matrices P^{mb} that both give the same distribution for (X_a, X_b, X_d), which we know is not possible. ■

Proof of Theorem 4.1. The proof is by induction on the number of terminal nodes. There is a nonterminal node m, say, that is joined to at least two terminal nodes; this follows from the "pigeonhole principle" in conjunction with the simple observation that the number of terminal nodes must be greater than the number of nonterminal nodes. The degree of m is at least 3, so there are two cases.

Case 1. The degree of m is greater than 3. (See Figure 2.) The idea is to strip off one of the terminal nodes attached to m; the induction hypothesis may then be applied, because the remaining tree will still have no nodes of degree 2. So by induction we will know the model for this remaining tree. For the details, let a and b be terminal nodes joined to m. From the given tree $\mathcal{T} = (S, E)$, strip off the node a and the edge $\{m, a\}$, leaving the tree $\mathcal{T}' = (S', E')$, where $S' = S - \{a\}$ and $E' = E - \{\{m, a\}\}$. Clearly the tree $\mathcal{T}'$ satisfies all of the conditions in

the theorem, so that by the induction hypothesis we know all of the matrices P^{rs} for $\{r, s\} \in E'$. Thus, we need show only that we may determine the matrices P^{ma} and P^{am}. But this is easy: for example, $P^{ma} = (P^{bm})^{-1} P^{ba}$, P^{bm} is known by the induction hypothesis, and P^{ba} is known by the assumption that we know the joint distribution on the terminal nodes of $\mathcal{T}$.

Case 2. The degree of m is 3. (See Figure 3.) Assuming that the number of terminal nodes in $\mathcal{T}$ is greater than 3 (since otherwise we are done by Lemma 4.1), there are exactly two terminal nodes, say a and b, joined to m. Let c be any terminal node other than a or b. Then clearly the submodel for (X_a, X_b, X_c, X_m) satisfies the conditions of Lemma 4.1. Therefore, we can deduce the matrices P^{ma}, P^{am}, P^{mb}, and P^{bm}.

If we stripped off just one of the nodes a and b, we would be left with a node m of degree 2, which would hinder the application of the induction hypothesis. So strip off both a and b, getting the tree $\mathcal{T}' = (S', E')$, where $S' = T' \cup N'$, $T' = T \cup \{m\} - \{a, b\}$, $N' = N - \{m\}$, and $E' = E - \{\{m, a\}, \{m, b\}\}$. By the induction hypothesis, we will be done if we can show that we know the joint distribution of the vector X, say, of character states at the terminal nodes of $\mathcal{T}'$.

Let V denote the terminal nodes of $\mathcal{T}'$ other than m, that is, $V = T' - \{m\} = T - \{a, b\}$. We use the notation X_V for the vector of character states $(X_t : t \in V)$ at the nodes of V. Let us consider fixing X_V at some arbitrary states x_V, say. Then, to complete the induction, our goal is to show that we can determine the function

$$\varphi(i) = \mathbb{P}\{X_V = x_V, X_m = i\}.$$

Define a $C \times C$ matrix F by

$$F(j,k) = \mathbb{P}\{X_V = x_V, X_a = j, X_b = k\};$$

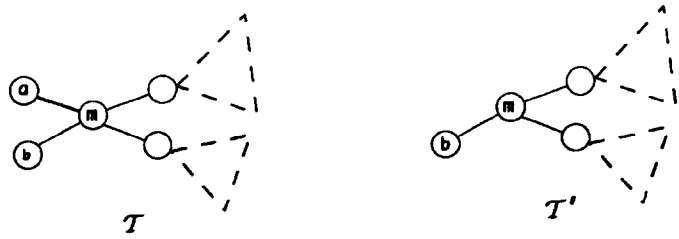

FIG. 2. Case 1 of the proof of Theorem 4.1.

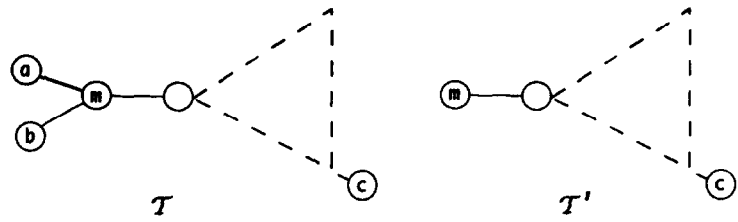

FIG. 3. Case 2 of the proof of Theorem 4.1.

observe that F is known from the joint distribution of states at the terminal nodes of $\mathcal{T}$. We have

$$F(j,k) = \sum_i \mathbb{P}\{X_V = x_V, X_m = i, X_a = j, X_b = k\}$$
$$= \sum_i \varphi(i) P^{ma}(i,j) P^{mb}(i,k).$$

Written in matrix form, defining the matrix $\Phi = \operatorname{diag}(\varphi(1), \ldots, \varphi(C))$, this becomes

$$(P^{ma})^{\mathrm{T}} \Phi P^{mb} = F,$$

so that

$$\Phi = (P^{ma})^{-\mathrm{T}} F (P^{mb})^{-1}.$$

Thus, as F, P^{ma}, and P^{mb} are all known, so is Φ. This completes the proof. ■

We conclude this section with remarks about the conditions of the theorem.

1. For characters that may take more than two states, conditions about determinants are not sufficient for identifiability. To see this, consider the case of three terminal nodes, as in Proposition 4.1, and take

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

for example. Then clearly replacing π^m by $\tilde{\pi}^m = \pi^m R$ and replacing P^{mi} by $\tilde{P}^{mi} = RP^{mi}$ for $i = a, b, c$ gives a model with exactly the same

joint distribution of terminal nodes (X_a, X_b, X_c), but with different probability transition matrices; this modification simply corresponds to interchanging labels of character states at the internal node. Note that because $\det(R) = 1$, none of the determinants of the transition matrices is changed by this multiplication. In the case $C = 2$ of binary characters, the set of probability transition matrices with positive determinants is reconstructible from rows; in fact it is the same as the set of transition matrices satisfying the DLC condition. This does not extend to higher values of C, because for $C > 2$ there are nontrivial permutation matrices that have determinant 1.

2. Another example of interest is the class of transition matrices that arise from continuous-time Markov chains, that is, matrices of the form $P = e^Q$, where $Q(i,j) \geqslant 0$ for $i \neq j$ and $Q\mathbf{1} = \mathbf{0}$. As

$$\det(e^Q) = e^{\mathrm{Trace}(Q)} > 0,$$

this class is reconstructible from rows in the case of $C = 2$. However, for $C > 2$, this class is not reconstructible from rows. For example, if

$$Q = \begin{pmatrix} -4 & 2 & 2 \\ 2 & -4 & 2 \\ 2 & 2 & -4 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

then

$$R\exp(Q) = \begin{pmatrix} .332507 & .334986 & .332507 \\ .332507 & .332507 & .334986 \\ .334986 & .332507 & .332507 \end{pmatrix} = \exp(\tilde{Q}),$$

where

$$\tilde{Q} = \begin{pmatrix} -4 & 3.2092 & .7908 \\ .7908 & -4 & 3.2092 \\ 3.2092 & .7908 & -4 \end{pmatrix}.$$

3. Extra care was taken in the proof to accommodate classes of matrices that are reconstructible from rows but not closed under multiplication. Such classes include a number of cases of interest; for instance, the previous example shows that the class DLC is not closed

under multiplication, as, clearly, $\exp(\tilde{Q}/n)$ satisfies the DLC condition for sufficiently large n, whereas $\exp(\tilde{Q}) = [\exp(\tilde{Q}/n)]^n$ does not satisfy the DLC condition. Thus, it is fortunate that we need only assume the reconstructibility from rows condition for edge transition matrices and do not need it to hold for transition matrices over paths, which are products of edge transition matrices. This was made possible by proving Lemma 4.1 under the assumption that just one of the edge transition matrices is in a class that is reconstructible from rows.

4. The theorem could be generalized by allowing, for each pair r, $s \in S$, a different class of matrices $\mathcal{M}_{rs}$ that is reconstructible from rows. Also, because the proof of the Lemma 4.1 required only one of the edge transition matrices to be reconstructible from rows, the conditions of the theorem could be further weakened.

5. The condition $P^{rs} \neq I$ avoids trivialities. Without it, even the topology is not identifiable; for example, a star phylogeny with four terminal nodes could also be considered to be any of the three possible bifurcating topologies with an internal branch having the identity matrix as its transition matrix.

6. The invertibility condition plays a similar role; noninvertibility corresponds to infinite distance. For a trivial example, if the edge transition matrices leading to the terminal nodes were

$$P^{st} = \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix} \qquad \text{for } t \in T,$$

then the character states at the terminal nodes would be independent coin flips and neither the topology nor the edge transition matrices would be identifiable.

7. The condition that the marginal distribution be positive at some node m (and hence at all nodes, as observed in the proof of Lemma 4.1) is required in order to have the transition matrices uniquely determined. For example, if the probability $\pi^r(i)$ of state i at node r were zero, then the ith row of any matrix P^{rs} for $s \in S$ could be changed arbitrarily with no effect on the joint distribution.

8. The prohibition on nodes of degree 2 is clear. For example, if node s has exactly two neighbors r and t, it is not possible to identify the matrices P^{rs} and P^{st} individually; one can hope only to identify their product.

9. The DLC condition seems scientifically more appealing than the analogous diagonal largest in row (DLR) condition would be. For example, nonuniformities in base compositions of DNA are well known—the nucleotides are generally not well modeled as having probability .25 each. If a process has a stationary distribution that is not

uniform, then the transition matrix for any sufficiently long edge would not satisfy the DLR condition. However, transition matrices that do not satisfy the DLC condition seem biologically implausible.

5. RECONSTRUCTION FROM DATA: CONSISTENCY OF MAXIMUM LIKELIHOOD

The previous identifiability result says that the true model may be inferred from the probability distribution of character states at the terminal nodes. In the actual inference problem, what we are given is not this distribution, but rather some data, which we assume constitute a sample of n independent observations from the unknown distribution. The method of maximum likelihood estimates the true model by the Markov model that maximizes the probability of the observed data. Methods for computing maximum likelihood estimators in phylogenetic estimation are discussed in [18] and [19].

An estimator is said to be *consistent* if it is certain to converge to the true quantity as the sample size grows. More formally, given a parametrized family of distributions $\{P_\theta : \theta \in \Theta\}$, let $X_1, X_2, \ldots$ be independent and identically distributed observations from P_θ. For each n let $\hat{\theta}_n$ be an estimator; $\hat{\theta}_n$ is a function of $(X_1, \ldots, X_n)$. We say that the sequence $\hat{\theta}_1, \hat{\theta}_2, \ldots$ is consistent if $P_\theta\{\lim_{n\to\infty} \hat{\theta}_n = \theta\} = 1$ holds for all θ.

Identifiability is a key prerequisite for consistency. The idea is this: if identifiability failed to hold, that is, if there were two different Markov models in the class under consideration that produced the same joint distribution on the observed nodes, then we could not distinguish between those models on the basis of the observed data, so that no method could be sure to converge to the correct model. Here we use the identifiability result from the previous section to show that, under mild conditions, maximum likelihood can consistently recover the full Markov model.

For simplicity, let us assume that the true topology is bifurcating, or "nondegenerate," in the terminology of Bandelt and Dress [20]. Otherwise, there are a number of points of view that one could take. For example, if our definition of consistency requires a correct estimate of the topology for all sufficiently large sample sizes (as well as having estimated transition probabilities that approach the true values), then maximum likelihood cannot be consistent when the true model is degenerate, because the likelihood will be maximized by some nondegenerate model for arbitrarily large sample sizes. On the other hand, we could modify the definition of consistency by defining an appropriate metric on the space of all models, including those that are degenerate.

The idea would be to consider a degenerate model (i.e., a "multifurcating model") to be equivalent to a number of different nondegenerate topologies with identity transition matrices (i.e., zero branch lengths) on the appropriate branches. Then the multifurcating model may be defined to be close to any nondegenerate model whose topology is the same as, and whose edge transition matrices are close to, any of the models equivalent to the multifurcating model. For example, a nondegenerate model on four taxa with a very short internal branch would be considered to be close to a star phylogeny. With this type of metric, it can be shown using the result that follows that maximum likelihood is consistent.

In addition to assuming that the true topology is nondegenerate, we assume the following conditions on the edge transition matrices. Two of these are as before: for each $\{r, s\} \in E$, the transition matrix P^{rs} is not a permutation matrix and $\det(P^{rs}) \neq 0$. We also use a slight strengthening of the reconstructibility from rows condition. For the definition, it is useful to consider a $C \times C$ matrix as a point in C^2-dimensional Euclidean space, so that notions such as the closure $\bar{\mathcal{M}}$ of a set of matrices $\mathcal{M}$ are defined in an obvious way.

DEFINITION

We say that a set of matrices $\mathcal{M}$ is *strongly reconstructible from rows* if, for each $M \in \mathcal{M}$ and each permutation matrix $R \neq I$, we have $RM \notin \bar{\mathcal{M}}$.

EXAMPLES

The DLC class is strongly reconstructible from rows. To construct an artificial example of a class $\mathcal{M}$ that is reconstructible from rows but not strongly reconstructible from rows, let DSC denote the class of transition matrices whose diagonal entries are the smallest in their respective columns, and define

$$\mathcal{M} = \{M : M_{ij} \text{ rational for all } i, j \text{ and } M \text{ satisfies DLC}\}$$
$$\cup \{M : M_{ij} \text{ irrational for all } i, j \text{ and } M \text{ satisfies DSC}\}.$$

Formally, let us think of the unknown parameter θ as a vector consisting of the tree topology (which could be specified by a number—e.g., "topology #7"—ranking the topology in a list of all bifurcating topologies), the marginal probabilities at some specified node, and the entries in the edge transition matrices, given in some specified order. Because we are considering only bifurcating topologies, so that each tree has the same number of edges, the parameter space Θ

is a subset of a single Euclidean space. Notions such as the closure of a set of models or parameter values are defined accordingly.

To state the main consistency result, let $X_T = (X_t : t \in T)$ and $X_N = (X_t : t \in N)$ denote the character states at the terminal and nonterminal nodes, respectively. We assume that X, X^1, X^2,... are independent and identically distributed, where $X^i = (X_T^i, X_N^i)$.

THEOREM 5.1

Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a class of Markov models on trees that have a fixed set of terminal nodes. Suppose that the models satisfy each of the following conditions:

1. *Each tree has a nondegenerate topology.*
2. *The marginal distribution π^s is positive at some node s.*
3. *The edge transition matrices are invertible, not equal to a permutation matrix, and belong to a class of matrices $\mathcal{M}$ that is strongly reconstructible from rows.*

Then the method of maximum likelihood consistently recovers the topology and the edge transition matrices. That is, letting $\hat{\theta}_n$ denote the maximum likelihood estimate based on n independent observations $X_T^1, \ldots, X_T^n$ of character states at the terminal nodes of the tree, we have

$$\mathbb{P}_\theta\{\hat{\theta}_n \to \theta \text{ as } n \to \infty\} = 1 \qquad \text{for all } \theta \in \Theta.$$

The theorem will be established using the next lemma, which is a customized variant of the fundamental consistency result of Wald [21]. One issue it deals with is the fact that, although the parameter space Θ is bounded, it is not closed. For example, identity matrices are not allowed as edge transition matrices, whereas matrices that are arbitrarily close to the identity are allowed. Also conditions for reconstructibility from rows, such as DLC, may involve strict inequalities, which do not specify closed sets.

LEMMA 5.1

Let $\mathcal{X}$ be a finite set and let $\{\mathcal{P}_\theta : \theta \in \Theta\}$ be a family of probability distributions on $\mathcal{X}$, where the closure $\overline{\Theta}$ of Θ is a compact subset of a metric space. Let X_1, X_2,... be independent and identically distributed random variables (or vectors) with probability distribution $\mathcal{P}_{\theta_0}$ for some $\theta_0 \in \Theta$. Assume the identifiability condition

$$\mathcal{P}_\theta \neq \mathcal{P}_{\theta_0} \text{ for each } \theta \in \overline{\Theta} \text{ with } \theta \neq \theta_0.$$

Suppose that for each $x \in \mathscr{X}$ the function $\theta \mapsto \mathscr{P}_\theta(x)$ is continuous on $\overline{\Theta}$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ maximize the log likelihood $\sum_{i=1}^n \log \mathscr{P}_\theta(X_i)$ over $\theta \in \overline{\Theta}$. Then $\mathscr{P}_{\theta_0}\{\hat{\theta}_n \to \theta_0\} = 1$.

Sketch of Proof. Let $\theta_0 \in \Theta$, and consider any open neighborhood $N(\theta_0)$ of θ_0; we want to show that with probability 1, we have $\hat{\theta}_n \in N(\theta_0)$ for sufficiently large n. Let L denote the log likelihood function. Using the identifiability condition, positivity of the Kullback–Leibler distance, and continuity, for each $\theta \in \overline{\Theta}$ distinct from θ_0, there is an open neighborhood $N(\theta)$ of θ with $E_{\theta_0} \sup\{L(\theta') : \theta' \in N(\theta)\} < E_{\theta_0} L(\theta_0)$. So by the strong law of large numbers, with $\mathscr{P}_{\theta_0}$-probability 1, $\hat{\theta}_n$ will eventually lie outside $N(\theta)$. Using compactness of $\overline{\Theta} - N(\theta_0)$, take a finite collection $\theta_1, \ldots, \theta_k$ such that $\cup_{i=1}^k N(\theta_i)$ covers $\overline{\Theta} - N(\theta_0)$. So with probability 1, $\hat{\theta}_n$ eventually stays outside the set $\overline{\Theta} - N(\theta_0)$, that is, within the neighborhood $N(\theta_0)$, as desired. ■

The statement of this lemma was tailored for simplicity and convenience in the particular setting of phylogenetic inference from discrete characters. Certain technical considerations required in Wald's general treatment become trivial in the case of discrete probability mass functions, as considered here. For example, one of Wald's conditions requires that the expected value of the supremum of the log likelihood over the parameter space be finite. This is automatic here—for a probability mass function the log likelihood is never greater than zero. The existence of a maximum likelihood estimator is not an issue here either because the maximum of the continuous log likelihood function over the compact set $\overline{\Theta}$ must be attained.

Proof of Theorem 5.1. We apply Lemma 5.1 with $\mathscr{X} = \mathscr{C}^{|T|}$ and with $\mathscr{P}_\theta$ taken to be the distribution of the states of the terminal nodes under $\mathbb{P}_\theta$, that is, $\mathscr{P}_\theta(A) = \mathbb{P}_\theta(X_T \in A)$ for $A \subseteq \mathscr{C}^{|T|}$. By the lemma, the proof will be completed by showing that if $\theta_0 \in \Theta$ and $\theta \in \overline{\Theta}$ with $\theta \neq \theta_0$, then $\mathscr{P}_\theta \neq \mathscr{P}_{\theta_0}$. Equivalently, letting $\theta_0 \in \Theta$ and $\theta \in \overline{\Theta}$, and assuming that the measures $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_\theta$ induce the same joint distributions on the terminal nodes, we want to show that in fact $\theta = \theta_0$. Let $\{P_0^{rs} : r, s \in S\}$ and $\{P^{rs} : r, s \in S\}$ be the transition matrices between pairs of nodes in models θ_0 and θ, respectively. Our assumptions imply that $P_0^{tu} = P^{tu}$ for terminal nodes t and u. It follows that, because we have also assumed that the edge transition matrices P_0^{rs} are invertible for edges $\{r, s\} \in E_0$, the same must be true of the edge transition matrices P^{rs} for $\{r, s\} \in E$—a product of transition matrices is invertible if and only if each of the matrices is invertible. Next, because the topologies of models θ_0 and θ are bifurcating and P_0^{rs} is not a permutation for each edge $\{r, s\} \in E_0$ by assumption, it is easy to see that P^{rs} is not a permutation for each edge $\{r, s\} \in E$. For example, if an edge transition

matrix P^{rs} were a permutation matrix, then we could produce a model that is equivalent to θ by coalescing the nodes r and s, eliminating the branch $\{r, s\}$ from the topology (creating a multifurcation), and modifying some edge transition matrices in an obvious way. The resulting model would have no branches of length 0, a topology that is different from that of θ_0, but pairwise joint distributions on terminal nodes that are the same as those of θ_0, which would contradict Proposition 3.1. Thus, although we required only that θ be in the closure $\overline{\Theta}$ rather than Θ, the assumption that $\mathbb{P}_\theta\{X_T \in \cdot\} = \mathbb{P}_{\theta_0}\{X_T \in \cdot\}$ in fact implies that the model θ still satisfies $\det P^{rs} \neq 0$ and P^{rs} is not a permutation for all edges $\{r, s\} \in E$. By Proposition 3.1 we conclude that the topologies of models θ_0 and θ are the same, so our problem is to show that the corresponding edge transition matrices are the same. This follows from the same inductive reasoning as was used in the proof of Theorem 4.1; we need only check that the same reasoning can still be carried through under the weaker assumption that $\theta \in \overline{\Theta}$. For example, when we strip off an edge $\{m, a\}$, say, as in the proof of Theorem 4.1, we conclude just as before that we can recover the sets of rows of the matrices P_0^{ma} and P^{ma} and these must be the same; the only question is the ordering of those rows. That is, we must rule out the possibility that P^{ma} is a nontrivial row permutation of P_0^{ma}. However, because the edge transition matrices of the models θ_0 and θ are members of $\mathcal{M}$ and $\bar{\mathcal{M}}$, respectively, the required conclusion follows from the assumption that the class $\mathcal{M}$ is strongly reconstructible from rows. The rest of the reasoning in case 2 of the proof also carries through without difficulty, as we have established that the edge transition matrices P^{rs} are invertible. ■

ACKNOWLEDGMENTS

I am grateful to John Hartigan, Junhyong Kim, and Michael Steel for their helpful comments.

REFERENCES

1. J. A. Cavender. Taxonomy with Confidence. *Math. Biosci.*, 40:271–280 (1978).
2. T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, Academic Press, New York, 1969, pp. 21–132.
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. F. Tajima and M. Nei. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, 1:269–285 (1984).
5. D. Barry and J. A. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43:261–276 (1987).

6. A. Rzhetsky and M. Nei. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.*, 12:131–151 (1995).
7. J. T. Chang and J. A. Hartigan. Reconstruction of evolutionary trees from pairwise distributions on current species. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, E. M. Keramidas, ed., Interface Foundation, Fairfax Station, VA, 1991, pp. 254–257.
8. M. Steel, M. D. Hendy, and D. Penny. Invertible models of sequence evolution. Mathematical and Information Science report 93/02, Massey University, Palmerston North, New Zealand, September 1993.
9. M. Steel, M. D. Hendy, and D. Penny. Reconstructing evolutionary trees from nucleotide pattern probabilities. Forschungsscherpunkt Mathematisierung-Strukturbildungsporozesse, Preprint XCIV, Bielefeld University, 1995.
10. J. Pearl and M. Tarsi. Structuring causal trees. *J. Complex.*, 2:60–77 (1986).
11. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Syst. Zool.*, 38:310–321 (1989).
12. P. Lazarsfeld and N. Henry. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.
13. J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410 (1978).
14. J. Felsenstein. Maximum likelihood and minimum-steps method for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, 22:240–249 (1973).
15. E. Sober. *Reconstructing the Past: Parsimony, Evolution, and Inference*. MIT Press, Cambridge, 1988.
16. P. Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
17. J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. Class.*, 4:57–71 (1987).
18. J. Felsenstein. Statistical inference of phylogenies. *J. Roy. Statist. Soc. Ser. A*, 146:246–272 (1983).
19. D. Barry and J. A. Hartigan. Statistical analysis of hominoid molecular evolution. *Statist. Sci.*, 2:191–210 (1987).
20. H. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.*, 7:309–343 (1986).
21. A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, 20:595–601 (1949).