SIAM J. DISCRETE MATH.
Vol. 14, No. 1, pp. 36–48

SUFFICIENT CONDITIONS FOR TWO TREE RECONSTRUCTION TECHNIQUES TO SUCCEED ON SUFFICIENTLY LONG SEQUENCES*

MIKE STEEL†

Abstract. The reconstruction of evolutionary trees (phylogenies) from DNA sequence data is a central problem in biology. We describe simple sufficient conditions for two tree reconstruction methods (maximum parsimony and maximum compatibility) to correctly reconstruct a tree when applied to sufficiently many sequence sites generated under a simple stochastic model.

Key words. trees, genetic sequences, maximum parsimony method, stochastic models

AMS subject classifications. 05C05, 92D15

PII. S0895480198343571

1. Introduction.

In biology, (graph-theoretic) trees are widely used to represent the evolutionary relationship between a group of extant species. Such a tree is sometimes called a "phylogeny" or "evolutionary tree." The extant species comprise the set $\mathcal{L}$ of leaves (vertices of degree 1) and the tree T describes the evolutionary history of the species from some hypothetical ancestor (located on some edge of the tree). Ideally, each vertex of T that is not in $\mathcal{L}$ has degree 3, in which case T is said to be *fully resolved*.

An important task in biology is to reconstruct such trees from observed features or data describing the extant species. We can regard each item of data as a function from $\mathcal{L}$ into some set R of r states (where $r = |R|$). Such functions are called r-state *characters* and they correspond to characteristics (morphological, physiological, genetic) on which the extant species differ. For example, in genetics, each site in a collection of aligned DNA sequences (one for each extant species, and with the same number of aligned sites for each species) provides a 4-state (or 2-state) character. For further biological details the interested reader is referred to [13].

The *maximum parsimony* method (abbreviated $\mathcal{MP}$) is a very popular technique for reconstructing evolutionary trees from collections of characters. To each character f and each tree T we associate a value $L(f,T)$ which is the minimum number of edges that must be assigned different states to its endpoints in order to extend f to assign states from R to all the vertices of T (we call the corresponding function $g : V \to R$ an *extension* of f). These concepts are illustrated in Figure 1 and will be defined more rigorously in the next section.

The $\mathcal{MP}$ method selects the tree (or trees) T that minimizes the sum of $L(f,T)$ over the characters f in the data. Informally, such a tree minimizes the number of "mutations" (changes of state across the edges of the tree) that need to be hypothesized in order to explain how the characters could have all evolved on tree T from some ancestral vertex.

*Received by the editors August 14, 1998; accepted for publication (in revised form) September 5, 2000; published electronically December 28, 2000. This research was supported by the New Zealand Marsden Fund.

http://www.siam.org/journals/sidma/14-1/34357.html

†Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand (m.steel@math.canterbury.ac.nz).

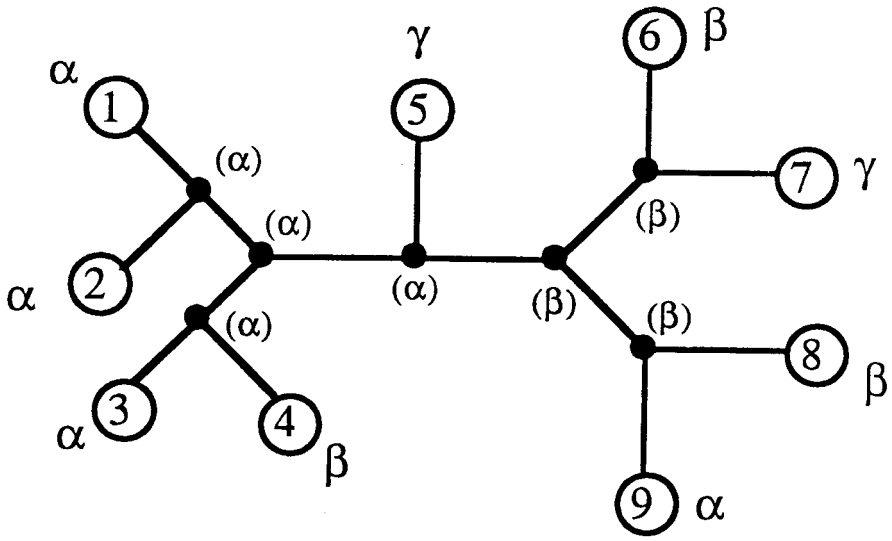

FIG. 1. *A fully resolved tree on leaf set* $\mathcal{L} = \{1, \ldots, 9\}$*, together with a* 3*-state character* $f : \mathcal{L} \to \{\alpha, \beta, \gamma\}$ *having* $L(f, T) = 5$*. An example of an extension* g *of* f *with* $\mathrm{ch}(g, T) = 5$ *is given by the additional assignment of states shown in brackets.*

The problem of finding an $\mathcal{MP}$ tree for a given sequence of characters is NP-hard and is a special case of the Steiner problem in graph theory (see [5] for details).

We will also consider a related method, called *maximum compatibility* denoted $\mathcal{MC}$. Informally, this method selects a tree that maximizes the number of characters in the data that could have evolved on the tree from the hypothetical ancestor without any parallel or reverse mutations (a more precise definition is given in the next section).

A fundamental theoretical question is to determine conditions under which $\mathcal{MP}$ or $\mathcal{MC}$ would recover a tree when applied to a large number of characters that evolved independently on that tree, according to some stochastic model. A variety of Markov-style models have been proposed for modeling and analyzing the evolution of DNA sequences (see [13]). The simplest such model, which we will call the N_r model, assigns equal probability to all possible transitions amongst the r states in R. This is as familiar to geneticists as the *Jukes–Cantor model* in the case $r = 4$ [7].

In a landmark paper [4], Felsenstein investigated whether $\mathcal{MP}$ and $\mathcal{MC}$ would tend to select the correct tree if the underlying characters were generated by this model. He showed that there exist various parameter settings in the N_2 model for which $\mathcal{MP}$ and $\mathcal{MC}$ will select an incorrect tree with probability tending to 1 as the sequence length tends to infinity. Indeed, this occurs even when T has just four leaves (in which case $\mathcal{MP}$=$\mathcal{MC}$). This statistical inconsistency phenomena has subsequently been refined and extended by others [6], [8], [11]. Sufficient conditions for the statistically consistency of $\mathcal{MP}$ or $\mathcal{MC}$ have only been described when either T has just four leaves [10], or for special cases [6], [8], [11], [12].

In this paper we provide the first explicit sufficient conditions for the statistical consistency of $\mathcal{MP}$ (resp., $\mathcal{MC}$) that are applicable to any tree on any number of leaves under the N_r (resp., N_2) models. Essentially, our results are of the form that

if the mutation probabilities associated to the edges of a tree under these models are sufficiently small, and not too unequal across the tree, then the two methods are statistically consistent. We now formalize some of the concepts described above.

2. Preliminaries. We begin by formalizing the definition of the parsimony score of a character $f : \mathcal{L} \to R$ on a tree $T = (V, E)$, where $\mathcal{L}$ is set of leaves (degree 1 vertices) of T.

Given a function $g : V \to R$, the *changing number* of g is defined by $\mathrm{ch}(g, T) := |\{e = \{u, v\} \in E : g(u) \neq g(v)\}|$. Given a character $f : \mathcal{L} \to R$, the *parsimony score* of f on T is defined by

$$L(f, T) := \min_g \{\mathrm{ch}(g, T) : g|\mathcal{L} = f\},$$

where we adopt the convention throughout this paper that $g|\mathcal{L}$ denotes the restriction of g to $\mathcal{L}$. It is easily shown that $L(f, T) \geq |f(\mathcal{L})| - 1$ and thus the quantity

$$H(f, T) := L(f, T) - |f(\mathcal{L})| + 1 \tag{2.1}$$

is nonnegative. $H(f, T)$ is sometimes called the *homoplasy* (or number of "extra steps") of f relative to T. This quantity turns out to be useful in the analysis of maximum parsimony, and it is also the basis of another method we describe now.

First, note that $H(f, T) = 0$ precisely if there is an extension g of f to V for which the subset of V that is mapped to any state in R by g forms a connected subtree of T (in biology this corresponds to being able to describe the evolution of f on T without any reverse or parallel changes of states). The maximum compatibility method, $\mathcal{MC}$, selects the tree (or trees) T that maximizes the number of characters f in that data for which $H(f, T) = 0$.

To investigate the statistical properties of these two methods we also need to specify a model by which characters can "evolve" on trees. In this paper we consider the simplest such model, namely, the symmetric r-state model, due to Neyman [9] (see also, [1], [3], [4]) and abbreviated hereafter as the N_r model. In this model R has cardinality r and the model has as its underlying parameters a fully resolved tree $T = (V, E)$ and a map $p : E \to (0, \frac{r-1}{r})$ that associates to each edge e of T a corresponding *mutation probability* $p(e)$. We now describe how this model assigns states to the vertices of T.

First we select an arbitrary fixed vertex v_0 of T and direct all edges of T away from v_0. We then randomly assign, with uniform probability, an element of R to v_0, and then assign states to the remaining vertices recursively as follows: for any arc (u, v) for which u has been assigned a state but v has not yet been assigned a state, randomly assign v the same state as u with probability $1 - p(e)$ (where $e = \{u, v\}$) or assign v one of the other states in R with equal probability (viz, $\frac{p(e)}{r-1}$).

In this way we generate a random function $G : V \to R$. Under the N_r model, with parameters (T, p), let $\mathbb{P}(G = g)$ be the probability that $G = g$, and let $\mathbb{P}(f, T)$ denote the probability that $G|\mathcal{L} = f$ (i.e., that G restricted to $\mathcal{L}$ equals f). By definition, and the assumptions of the model,

$$\mathbb{P}(f, T) = \sum_{\{g: V \to R: g|\mathcal{L} = f\}} \mathbb{P}(G = g)$$

and

$$\mathbb{P}(G = g) = \frac{1}{r} \prod_{\{e=\{u,v\}: g(u) \neq g(v)\}} \frac{p(e)}{r-1} \prod_{\{e=\{u,v\}: g(u) = g(v)\}} (1 - p(e)),$$

from which we immediately see that the probability distribution on characters f (and extensions g) is independent of our choice of v_0.

A tree reconstruction method is *statistically consistent* under this N_r model with underlying parameters (T, p) if the probability that the method reconstructs T when applied to k independently generated characters converges to 1 as k tends to infinity. An example of such a method is the maximum likelihood technique, as Chang [2] recently established (for the N_r model and generalizations thereof).

3. Sufficient conditions for correct tree reconstruction. We begin with some definitions leading to a simple combinatorial sufficient condition for the two methods described to return a given tree.

An *interior edge* of T is an edge that is not incident with a leaf. We will let $\mathring{E}$ denote the set of interior edges of T. Deleting an edge $e \in E$ from T produces a partition π_e of the leaves of T into two subsets. Note also that each character f induces a partition of $\mathcal{L}$ by grouping together those leaves that are assigned the same state by f. If the partition induced by f equals π_e, we say that f *corresponds to* edge e. Note that f corresponds to some edge of T if and only if $L(f, T) = 1$. Let $c(e)$ denote the set of those $r(r-1)$ characters that correspond to edge e and let $c(T) = \cup_{e \in \mathring{E}} c(e)$.

Suppose we are given a sequence C of characters and a character f. Let $n(C, f)$ denote the number of occurrences of character f in C, and let $n(C, \hat{f})$ denote the total number of occurrences in C of all characters that induce the same partition of $\mathcal{L}$ as f. For an edge e of T let $n_e(C) = \sum_{f \in c(e)} n(C, f) = n(C, \hat{f}_e)$, where f_e is any character that corresponds to edge e. Let

$$n_-(C, T) := \min_e \{n_e(C) : e \in \mathring{E}\},$$

$$n_+(C, T) := \max_f \{n(C, \hat{f}) : L(f, T) > 1\},$$

and

$$H(C, T) := \sum_f n(C, f) H(f, T),$$

where $H(f, T)$ is described by (2.1).

We pause to briefly provide some interpretation of these definitions. The quantity $n_-(C, T)$ is a measure of the minimum support for any edge of T by the characters in C, while $n_+(C, T)$ is the total number of characters that support some split of the species set $\mathcal{L}$ that does not correspond to any edge of T. The quantity $H(C, T)$ is sometimes called the *homoplasy* (or "number of extra steps") of C relative to T. Note that $H(C, T) \geq 0$.

The following result gives sufficient conditions for $\mathcal{MP}$ and $\mathcal{MC}$ to return a given tree from some sequence C of characters, regardless of how these characters arise.

LEMMA 1. *Let T be any fully resolved tree.*

(A) *$\mathcal{MP}$ selects tree T for a sequence C of r-state characters if*

$$n_-(C, T) > H(C, T).$$

(B) *$\mathcal{MC}$ selects tree T for a sequence C of 2-state characters if*

$$n_-(C, T) > n_+(C, T).$$

Proof. For brevity, we let $n(f) = n(C, f), n_- = n_-(C, T), n_+ = n_+(C, T)$.

Part (A). For a tree T_1, let $L(T_1) := \sum_f L(f, T_1)n(f)$. It suffices to show that $L(T_1)$ is strictly minimized when $T_1 = T$. First note that

$$L(T_1) = \Delta + H(C, T_1),$$

where $\Delta = \sum_f (|f(\mathcal{L})| - 1)n(f)$. Now, if $T_1 \neq T$ and since T is fully resolved, T has at least one interior edge e for which, for all $f \in c(e)$, we have $L(f, T_1) \geq 2$, and since $|f(\mathcal{L})| = 2$ for each $f \in c(e)$, this implies that $H(f, T_1) = L(f, T_1) - |f(\mathcal{L})| + 1 \geq 1$. Consequently, $H(C, T_1) \geq n_-$, and so, by the assumption in the lemma,

$$L(T_1) \geq \Delta + n_- > \Delta + H(C, T) = L(T),$$

which establishes the claim.

Part (B). Let $\nu(T_1)$ denote the number of occurrences in C of a character f with $L(f, T_1) = 1$. It suffices to show that $\nu(T_1)$ is strictly maximized when $T_1 = T$. First note that

$$\nu(T_1) = \sum_{\{f: L(f,T_1)=1\}} n(f) = \sum_{e \in E(T_1)} n_e(C).$$

Now, for any tree $T_1 \neq T$ let E' denote the subset of interior edges e of T for which $\pi_e \neq \pi_{e'}$ for any edge e' of T_1. Since T is fully resolved, $E' \neq \emptyset$, and the number of edges e in T_1 for which $\pi_e \cap \{\pi_{e'} : e' \in E(T)\} = \emptyset$ is at most $|E'|$ and for each such edge $n_e(C) \leq n_+$. Thus,

$$\nu(T) - \nu(T_1) \geq \sum_{e \in E'} n_e(C) - |E'|n_+ > 0$$

as required. □

When C is generated under the N_r model, we can apply this lemma to obtain sufficient conditions for the statistical consistency of $\mathcal{M}C$ and $\mathcal{M}P$ (Corollary 1). First we introduce the following terminology.

Definition 1. *Under the N_r model with parameters (T, p), let*

$$m_- := \min_f \{\mathbb{P}(f, T) : f \in c(T)\};\ m_+ := \max_f \{\mathbb{P}(f, T) : L(f, T) > 1\}$$

and

$$\overline{H} := \sum_f \mathbb{P}(f, T)H(f, T).$$

The quantity $\overline{H}$ is the expected homoplasy of the (random) character f on the underlying tree T.

Corollary 1.

(A) *$\mathcal{M}P$ is statistically consistent under the N_r model if*

$$m_- > \frac{\overline{H}}{r(r-1)}.$$

(B) *$\mathcal{M}C$ is statistically consistent under the N_2 model if $m_- > m_+$.*

Proof. Note that, under the N_r model, if f and f' induce the same partition of $\mathcal{L}$, then $\mathbb{P}(f,T) = \mathbb{P}(f',T)$. Suppose that we have a sequence C of c characters which evolve identically and independently under the N_r model. Then, by the weak law of large numbers, as c tends to infinity,

$$\frac{n_-(C,T)}{c} \to_p r(r-1)m_-,$$

where $\to_p$ denotes convergence in probability, since for each $f \in c(T)$ there are precisely $r(r-1)$ r-state characters that induce the same partition of $\mathcal{L}$ as f.

Also, we have

$$\frac{H(C,T)}{c} \to_p \overline{H},$$

while, for $r = 2$, we have

$$\frac{n_+(C,T)}{c} \to_p 2m_+.$$

The result now follows by Lemma 1. □

We can now state our main result.

THEOREM 1. *Under the N_r model with parameters (T,p) let*

$$p_{\text{sum}} := \sum_e p(e),$$

$$p_- := \min\{p(e) : e \in \mathring{E}\}; p_+ := \max\{p(e) : e \in E\},$$

and

$$p_\pm := p_- + p_+.$$

Then,

(A) *$\mathcal{M}P$ is statistically consistent if $p_{\text{sum}} < 1$ and*

$$p_- \geq \frac{p_{\text{sum}}^2}{1 - p_{\text{sum}}}; \tag{3.1}$$

(B) *when $r = 2$, $\mathcal{M}C$ is statistically consistent if*

$$p_- \geq p_+ p_\pm + 2p_\pm^2(1 + p_\pm), \tag{3.2}$$

and, in particular, this is satisfied whenever

$$p_- \geq 12p_+^2. \tag{3.3}$$

Remarks.

- Informally, the condition for the consistency of $\mathcal{M}P$ is that the mutation probability $p(e)$ associated to any interior edge should be at least (approximately) the square of the total expected number of mutations in the tree. Thus it assumes the $p(e)$ values are small and not too unequal.
  For $\mathcal{M}C$ the condition described states, informally, that the smallest mutation probability on any interior edge is of the order of the square of the largest

mutation probability. In particular, $\mathcal{MC}$ is statistically consistent under the N_2 model whenever $p(e)$ is a constant, p, across the edges of the tree and p takes a value at most $\frac{-5+\sqrt{41}}{16} \approx 0.087$, since in that case $p \geq p(2p) + 2(2p)^2(1+2p)$ and so (noting that $p_- = p_+ = p$ and $p_\pm = 2p$) we see that inequality (3.2) is satisfied. More generally for any bound on the ratio of the $p(e)$ values, Theorem 1(B) implies that there exists an upper bound on p_+ (dependent on that bound) for which $\mathcal{MC}$ is statistically consistent under the N_2 model.

- Note also that the size r of the state space does not enter into inequality (3.2). In fact it can be shown that, for any fixed values of $p_- > 0$ and $p_+ < 1$, if r is sufficiently large, then $\mathcal{MP}$ will be consistent (this is a special case of Theorem 3 of [12]).

Proof of Theorem 1 (A). Let

$$\overline{L} := \sum_f \mathbb{P}(f,T)L(f,T).$$

Letting $\mathbb{E}$ denote expectation, we have $\overline{L} = \mathbb{E}[L(G|\mathcal{L},T)]$ for an extension G randomly generated on T under the N_r model. Now, $L(G|\mathcal{L},T) \leq \text{ ch}(G,T)$, and so,

$$\overline{L} \leq \mathbb{E}[\text{ch}(G,T)]. \tag{3.4}$$

However, $\text{ch}(G,T)$ is simply a sum of independent 0/1 random variables as follows: to each edge $e = \{u,v\}$ independently assign the value 1 if and only if $G(u) \neq G(v)$ (which has probability $p(e)$), and assign 0 otherwise. Consequently,

$$\mathbb{E}[\text{ch}(G,T)] = p_{\text{sum}}. \tag{3.5}$$

Let

$$Q := \prod_e (1-p(e)),$$

which is the probability that that there is no mutation on any edge e of T (i.e., for each edge $e = \{u,v\}$ we have $G(u) = G(v)$). Now,

$$\overline{H} = \overline{L} - \sum_f \mathbb{P}(f,T)(|f(\mathcal{L})| - 1), \tag{3.6}$$

and

$$\sum_f \mathbb{P}(f,T)(|f(\mathcal{L})| - 1) \geq \mathbb{P}(L(G|\mathcal{L},T) = 1) \geq \mathbb{P}(\text{ch}(G,T) = 1). \tag{3.7}$$

Furthermore, $\mathbb{P}(\text{ch}(G,T) = 1) = \sum_e p(e) \prod_{e' \neq e}(1-p(e')) \geq p_{\text{sum}}Q$. Thus, by (3.6) and inequality (3.7) we have $\overline{H} \leq \overline{L} - p_{\text{sum}}Q$, while inequality (3.4) and (3.5) give $\overline{L} \leq p_{\text{sum}}$, and hence

$$\overline{H} \leq p_{\text{sum}}(1-Q). \tag{3.8}$$

Now, for a character $f \in c(T)$, let e_f denote the edge of T to which f corresponds. An extension g of f to V can be obtained by assigning a leaf v_0 the value $f(v_0)$ (with

probability $\frac{1}{r}$), assigning (appropriate) different states to the ends of e_f, and for each edge $e \neq e_f$ assigning the same state to each end of e. Consequently,

$$\mathbb{P}(f,T) \geq \mathbb{P}(G=g) = \frac{1}{r} \times \frac{p(e_0)}{r-1} \prod_{e \neq e_f} (1-p(e)) > \frac{1}{r(r-1)} Q p_-.$$

Thus,

$$m_- \geq \frac{1}{r(r-1)} p_-. \tag{3.9}$$

Now, our hypothesis is that $p_- \geq \frac{p_{\text{sum}}^2}{1-p_{\text{sum}}}$. Then, $1 - p_{\text{sum}} \geq \frac{p_{\text{sum}}}{p_{\text{sum}}+p_-}$ which together with the purely algebraic inequality

$$Q > 1 - p_{\text{sum}}$$

implies that $Q > \frac{p_{\text{sum}}}{p_{\text{sum}}+p_-}$. Rearranging gives $Qp_- > p_{\text{sum}}(1-Q)$ and thus, in view of the inequalities (3.8), (3.9) we have

$$m_- > \frac{\overline{H}}{r(r-1)Q} > \frac{\overline{H}}{r(r-1)}.$$

Part (A) of the theorem now follows from Corollary 1(A).

Part (B). Throughout this proof we will make extensive use of the following two properties of the N_r model with underlying tree T:

- The conditional probability of generating a character f given that a leaf $l \in \mathcal{L}$ of T is in a particular state $\mu \in R$ is precisely $r\mathbb{P}(f,T)$ if $f(l) = \mu$ (and is 0 otherwise).
- Let t_1 and t_2 be two subtrees of T that share one nonleaf vertex, v. Let f_1 and f_2 denote the restrictions of f to the leaves of t_1 and t_2, respectively. Then f_1 and f_2 are conditionally independent once the state of vertex v is specified.

Throughout the rest of this proof we will take (without loss of generality) $R = \{0,1\}$. Let P_0 denote the probability of generating the character that maps all leaves to state 0. We first establish the following inequality. Suppose that $f \in c(T)$ corresponds to edge $e \in \mathring{E}$. Then,

$$\mathbb{P}(f,T) > \frac{p(e)}{1-p(e)} P_0. \tag{3.10}$$

To establish (3.10) let T_1, T_2 denote the two rooted subtrees of T whose roots are the ends of edge e. Without loss of generality we may suppose all the leaves of T_1 (resp., T_2) are mapped by f to 0 (resp., 1). For $i = 1, 2$, and $\mu, \nu \in \{0,1\}$, let $P_i(\mu,\nu)$ denote the conditional probability, under the N_2 model (restricted to T_i) that all the leaves in T_i are in state μ given that the root vertex (which we take as our v_0) is in state ν. Let $\alpha = \frac{1}{2}(P_1(0,1)P_2(1,0) + P_1(0,0)P_2(1,1)); \beta = \frac{1}{2}(P_1(0,0)P_2(1,0) + P_1(0,1)P_2(1,1))$. Then,

$$\mathbb{P}(f,T) = \alpha p(e) + \beta(1-p(e)), \tag{3.11}$$

and by virtue of the symmetry in the N_2 model which implies that

$$P_i(0,0) = P_i(1,1) \text{ and } P_i(0,1) = P_i(1,0),$$

we see that

$$P_0 = \alpha(1 - p(e)) + \beta p(e).$$

Straightforward algebraic manipulation then shows that (since $p(e) < \frac{1}{2}$), $\frac{\mathbb{P}(f,T)}{P_0} > \frac{p(e)}{1-p(e)}$, as required to establish (3.10). Actually all we shall require is the following corollary of inequality (3.10), namely for any $f \in c(T)$,

$$\mathbb{P}(f, T) > P_0 p_-. \tag{3.12}$$

Most of the remainder of the proof is devoted to establishing the following upper bounds on $\mathbb{P}(f, T)$.

Claim.

- *If $L(f, T) = 1$, then*

$$\mathbb{P}(f, T) < p_\pm P_0, \tag{3.13}$$

- *while, if $L(f, T) > 1$, then*

$$\mathbb{P}(f, T) < 2p_\pm^2(1 + p_\pm)P_0 < p_\pm P_0. \tag{3.14}$$

(Note that inequality (3.13) is required purely to justify the proof of inequality (3.14).) The proof of inequality (3.13) is by induction on the number n of leaves of T. The inequality holds for $n = 2$ since then there is just one edge e and $p(e) = p_- = p_+; P_0 = \frac{1}{2}(1 - p(e))$ and so, since $p(e) \in (0, 0.5)$, $\mathbb{P}(f, T) = \frac{1}{2}p(e) < p(e)(1 - p(e)) = p_\pm P_0$, as required. Now, suppose that $n > 2$. We distinguish two subcases.

- $f \in c(T)$,
- f corresponds to a noninterior edge of T.

In the first subcase let T_1, T_2 be as described above. Consider the two subtrees $\{t_i^a, t_i^b\}$ of T_i that intersect precisely on the vertex v_i, where $e = \{v_1, v_2\}$ is the edge associated with f (see Figure 2(a)). For $\theta = a, b$ let $P_i^\theta(\mu, \nu)$ denote the conditional probability, under the N_2 model restricted to t_i^θ, that the leaves t_i^θ are all in state μ given that v_i is in state ν. Then

$$P_i(\mu, \nu) = P_i^a(\mu, \nu)P_i^b(\mu, \nu). \tag{3.15}$$

Now, if $\mu \neq \nu$, then the restriction of f to the leaves of each subtree has L value of 1 on each subtree, so by the inductive hypothesis,

$$P_i^\theta(\mu, \nu) < p_\pm P_i^\theta(0, 0), (\mu \neq \nu). \tag{3.16}$$

Now, recalling (3.11) we have $\mathbb{P}(f, T) = \alpha p(e) + \beta(1 - p(e))$ and so, substituting (3.15) and inequality (3.16) into the definitions of α and β, we deduce that

$$\mathbb{P}(f, T) < K[p(e) + 2p_\pm^2(1 - p(e)) + p(e)p_\pm^4], \tag{3.17}$$

where $K = \frac{1}{2}P_1^a(0, 0)P_1^b(0, 0)P_2^a(0, 0)P_2^b(0, 0)$.

Also, we have

$$P_0 \geq (1 - p(e))K \geq (1 - p_+)K. \tag{3.18}$$

Now, we can bound the term in brackets in (3.17) by noting that $p(e) + 2p_\pm^2(1 - p(e)) + p(e)p_\pm^4 < p_+ + 2p_\pm^2 + p_\pm^3$ (since $p_\pm < 1$), and then, by our assumption (3.2),

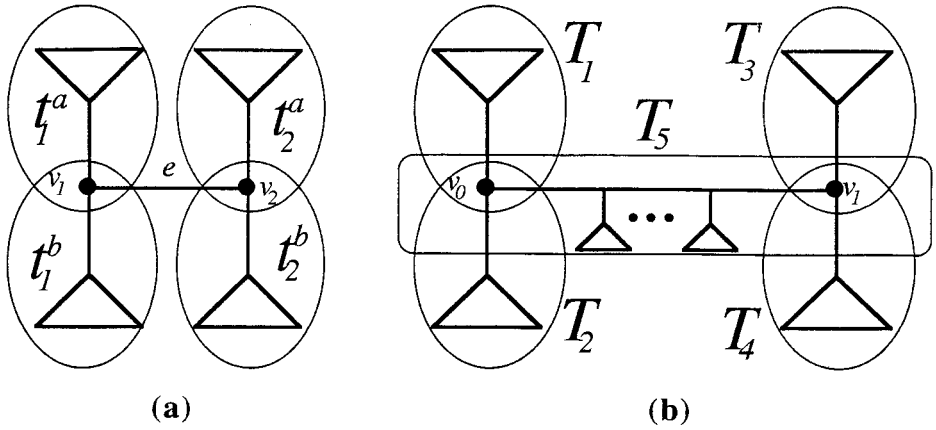

FIG. 2. *Representations of T for the proof of the upper bounds* (3.13) *and* (3.14).

we have $p_+ + 2p_\pm^2 + p_\pm^3 \leq p_\pm(1-p_+)$. Substituting this into (3.17) and comparing the result to (3.18) establishes inequality (3.13) in the first subcase.

For the second subcase, we may assume that f maps some leaf, incident with an edge e, to state 0, and all other leaves of T to state 1.

Let t^a, t^b denote the other two subtrees of T which intersect precisely on the vertex at the other end of edge e from the leaf. Then, defining $P^a(\mu,\nu), P^b(\mu,\nu)$ analogously as before, we have

$$\mathbb{P}(f,T) = \frac{1}{2}(P^a(1,1)P^b(1,1)p(e) + P^a(1,0)P^b(1,0)(1-p(e))),$$

and so

$$\mathbb{P}(f,T) < \frac{1}{2}P^a(0,0)P^b(0,0)[p(e) + p_\pm^2(1-p(e))].$$

Consequently, since $p(e) + p_\pm^2(1-p(e)) < p_+ + p_\pm^2 < p_\pm(1-p_+)$, we have

$$\mathbb{P}(f,T) < \frac{1}{2}P^a(0,0)P^b(0,0)p_\pm(1-p_+).$$

Furthermore, since $P_0 \geq \frac{1}{2}P^a(0,0)P^b(0,0)(1-p(e)) \geq \frac{1}{2}P^a(0,0)P^b(0,0)(1-p_+)$, we deduce that inequality (3.13) holds in this subcase also.

We now establish inequality (3.14). We first observe that our condition (3.2) forces $2p_\pm^2(1+p_\pm) < p_\pm$ so it is only the first inequality in (3.14) we need to establish.

The proof again is by induction on n. For $n = 2, 3$ there is nothing to prove. Suppose $n \geq 4$ and $L(f,T) > 1$. Then, a standard application of Menger's theorem from graph theory shows that there are two edge-disjoint paths in T, each of which connects leaves assigned different states by f [14, Lemma 1]. Thus, we may represent T as in Figure 2(b), with five subtrees trees $T_1, \ldots, T_5$ as shown, each pair of which is disjoint or overlaps at one of two (generally nonadjacent) vertices v_0 and v_1 as shown. We call these two vertices *reference vertices*, and note that each of $T_1, \ldots, T_4$

has exactly one reference vertex (and it is a leaf of that subtree) while T_5 has both reference vertices as leaves.

For $i = 1, \ldots, 4$ and $\mu \in \{0, 1\}$ let f^i_μ be the character defined on the leaf set of T_i which maps its reference vertex to μ and all other leaves of T_i to the element that f specifies. Let $f_{\mu,\nu}$ be the character defined on the leaf set of T_5 which maps v_0 to μ, v_1 to ν, and every other leaf in T_5 to the element that f specifies. For $i = 1, \ldots, 4$ let

$$P_i(\mu) := 2\mathbb{P}(f^i_\mu, T_i); L_i(\mu) := L(f^i_\mu, T_i)$$

and

$$P_5(\mu, \nu) := 2\mathbb{P}(f_{\mu,\nu}, T_5); L_5(\mu, \nu) := L(f_{\mu,\nu}, T_5).$$

For $i = 1, \ldots, 5$ let P^i_0 equal twice the probability of generating on T_i the character which maps all leaves to 0. Then,

$$\mathbb{P}(f, T) = \frac{1}{2} \sum_{\mu,\nu} P_5(\mu, \nu) \prod_{i=1,2} P_i(\mu) \prod_{i=3,4} P_i(\nu). \tag{3.19}$$

Now, by induction, we may assume that inequality (3.14) holds for all five subtrees, and invoking inequality (3.13) if necessary we deduce that, for $i = 1, \ldots, 4$, $P_i(\mu) < p_\pm P^i_0$ when $L_i(\mu) > 0$ (while $P_i(\mu) = P^i_0$ otherwise). Similarly, $P_5(\mu, \nu) < p_\pm P^5_0$ when $L_5(\mu, \nu) > 0$, and $P_i(\mu) = P^i_5$ otherwise. Consequently for when $\mu = \nu \in \{0, 1\}$ we introduce at least two powers of $p_\pm$ into the product terms of (3.19), while for $\mu \neq \nu$ we introduce at least three powers of $p_\pm$. Thus, we deduce that

$$\mathbb{P}(f, T) < 2p^2_\pm(1 + p_\pm) \cdot \frac{1}{2} \cdot \prod_{i=1,\ldots,5} P^i_0$$

and inequality (3.14) now follows by observing that

$$P_0 > \frac{1}{2} \cdot \prod_{i=1,\ldots,5} P^i_0.$$

Finally we establish part (B) of the theorem. In view of inequalities (3.12) and (3.14) we have

$$m_- > P_0 p_-$$

and

$$m_+ < 2p^2_\pm(1 + p_\pm)P_0.$$

Now, by our condition (3.2), $p_- > p_+p_\pm + 2p^2_\pm(1 + p_\pm) > 2p^2_\pm(1 + p_\pm)$, we see that $m_- > m_+$. The first claim in part (B) of the theorem now follows from Corollary 1(B).

Finally we show that inequality (3.3) implies inequality (3.2). Suppose that $p_- \geq 12p^2_+$. Then, $p_+ \leq \frac{1}{12}$, and so, we have $p_- \geq 12p^2_+ \geq 2p^2_+ + 8p^2_+(1 + 2p_+) \geq p_+p_\pm + 2p^2_\pm(1 + p_\pm)$ (since $p_+ \geq \frac{1}{2}p_\pm$), as required. □

4. Remarks. An interesting theoretical question is whether $\mathcal{MP}$ is statistically consistent under the N_r model when $p(e) = p$ (for all edges e) and p is less than some value $p^{(r)} > 0$ that is independent of n. This question is open even for the case $r = 2$ (however, from [11], if such a positive value of $p^{(2)}$ exists, it must be less than $\frac{1}{8}$). Note that the sufficient condition (3.2) described in Theorem 1(B) requires that the $p(e)$ values to converge to 0 at least as fast as n^{-2}, where $n = |\mathcal{L}|$, so in a certain sense the sufficient condition described for $\mathcal{MC}$ is much stronger than that for $\mathcal{MP}$ (in the case $r = 2$).

It is also instructive to compare the strengths of the two parts of Theorem 1 for the case when $n = 4$. In [4] Felsenstein considered the N_2 model on a resolved tree on four leaves, with two nonadjacent edges having $p(e) = s$, and the remaining three edges having $p(e) = q$. He showed that $\mathcal{MP}$ is statistically inconsistent precisely when $q(1-q) < s^2$, which, for q small, amounts, approximately, to $q < s^2$. By contrast, the sufficient condition described in the above theorem for $\mathcal{MP}$ would require $q \geq \frac{(2s+3q)^2}{1-2s-3q}$ which for $q \ll s \ll 1$ amounts, approximately, to $q > 4s^2$. For $\mathcal{MC}$ (which agrees with $\mathcal{MP}$ on trees with four leaves) the analogous sufficient condition described by inequality (3.2) reduces, approximately, to $q > 3s^2$. In either case we see a gap between sufficiency and necessity conditions for statistical consistency. In fact, for the case of four leaves it is possible to characterize precisely the conditions on the five $p(e)$ values for the statistical consistency of $\mathcal{MP}$ (see [10]), however, in general, this appears to be difficult. Thus a challenge for the future would be to narrow the gap between necessary and sufficient conditions for the statistical consistency of $\mathcal{MP}$ and $\mathcal{MC}$. An extension of Theorem 1(B) to $r > 2$ would also be interesting.

Acknowledgments. The author thanks the Isaac Newton Institute (Cambridge) for its hospitality, Jotun Hein for offering helpful comments on an earlier version of this manuscript, and the anonymous referee for several suggestions for improving the presentation of this paper.

REFERENCES

[1] J.A. Cavender, *Taxonomy with confidence*, Math. Biosci., 40 (1978), pp. 271–280.
[2] J.T. Chang, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, Math. Biosci., 134 (1996), pp. 189–215.
[3] J.S. Farris, *A probability model for inferring evolutionary trees*, Syst. Zool., 22 (1973), pp. 250–256.
[4] J. Felsenstein, *Cases in which parsimony or compatibility will be positively misleading*, Syst. Zool., 27 (1978), pp. 401–410.
[5] L.R. Foulds and R.L. Graham, *The Steiner problem in phylogeny is NP-complete*, Adv. in Appl. Math., 3 (1982), pp. 43–49.
[6] M.D. Hendy and D. Penny, *A framework for the quantitative study of evolutionary trees*, Syst. Biol., 38 (1986), pp. 297–309.
[7] J.T. Jukes and C.R. Cantor, *Evolution of protein molecules*, in Mammalian Protein Metabolism, H.N. Munro ed., Academic Press, New York, 1996, pp. 21–132.
[8] J. Kim, *General inconsistency conditions for maximum parsimony: Effects of branch length and increasing the number of taxa*, Syst. Biol., 45 (1996), pp. 363–374.
[9] J. Neyman, *Molecular studies of evolution: A source of novel statistical problems*, in Statistical Decision Theory and Related Topics, S.S. Gupta and J. Yackel, eds., Academic Press, New York, 1971, pp. 1–27.
[10] D. Penny, M.D. Hendy, and M.A. Steel, *Testing the theory of descent*, in Phylogenetic Analysis of DNA sequences, M.M. Miyamoto and J. Cracraft, eds., Oxford University Press, Oxford, UK, 1991, pp. 155–183.
[11] M.A. Steel, *Distributions on Bicoloured Evolutionary Trees*, PhD thesis, Massey University, Palmerston North, New Zealand, 1989.

[12] M.A. Steel and D. Penny, *Parsimony, likelihood, and the role of models in molecular phylogenetics*, Mol. Biol. Evol., 17 (2000), pp. 839–850.
[13] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, *Phylogenetic inference*, in Molecular Systematics, 2nd ed., D.M. Hillis, C. Moritz, and B.K. Marble, eds., Sinauer Associates, Sunderland, MA, 1996, pp. 407–514.
[14] C. Tuffley and M. Steel, *Links between maximum likelihood and maximum parsimony under a simple model of site substitution*, Bull. Math. Biol., 59 (1997), pp. 581–607.