Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences

September 1, 1998

Mike Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand

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
Abbreviated title. Consistency conditions for tree reconstruction

1 Preliminaries

The *Maximimum Parsimony* method (abbreviated MP) is a very popular technique for reconstructing evolutionary trees from biological data. Formally, we are given a collection of functions, each of which maps the set $\mathcal{L}$ of leaves of a tree $T = (V, E)$ into some set S of states - such functions are called *characters*. In evolutionary studies, the leaves (degree 1 vertices) of the T correspond to extant species, and the tree T describes the evolutionary history of these species from some hypothetical ancestor (located on some edge of the tree). Characters correspond to characteristics (morphological, physiological, genetic) on which the extant species differ. For example, in genetics, aligned DNA sequences (one for each extant species) provides a 4-state (or 2-state) characters - one for each site in the aligned sequences. For further biological details the interested reader is referred to [10]. Given a character f we seek to extend f to a function $h : V \to S$ in such a way as to minimize the changing number of h defined by $ch(h,T) := |\{e = \{u,v\} \in E : h(u) \neq h(v)\}|$. Let $L(f,T)$ denote this minimal value of $ch(h,T)$ (over all extensions h), sometimes called the *parsimony score* of f on T. These concepts are illustrated in Fig. 1.

The 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

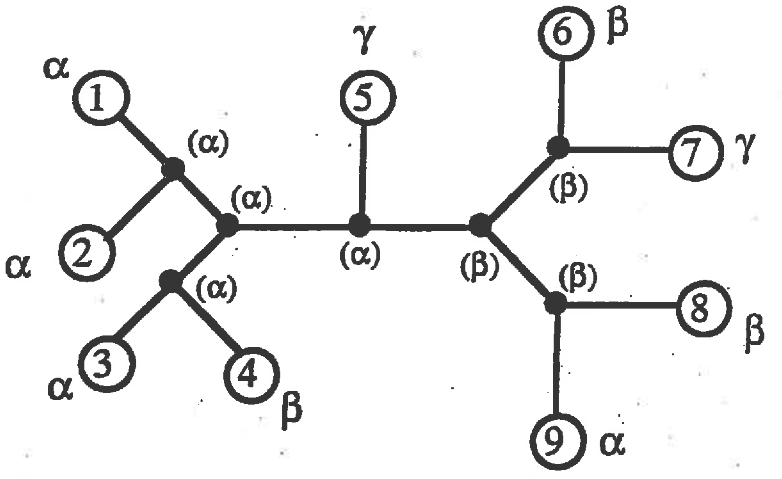

Figure 1: A fully resolved tree on leaf set $L = \{1, \ldots, 9\}$, together with a 3-state character $f : L \rightarrow \{\alpha, \beta, \gamma\}$ having $L(f,T) = 5$. An example of an extension h of f with $ch(h,T) = 5$ is given by the the additional assignment of states shown in brackets

of state across the edges of the tree) that need to be hypothesised in order to explain how the characters could have all evolved on tree T from some ancestral vertex.

A related method, *Maximum Compatibility*, abbreviated MC, selects the tree (or trees) T which maximizes the number of characters f for which $L(f,T) = |f(L)| - 1$. Informally, such a tree maximizes the number of characters which could have evolved from some ancestral vertex without any parallel or reverse mutations.

A fundamental theoretical question is to determine conditions under which MP or 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 modelling and analysing the evolution of DNA sequences (see,[10]) - in this paper we consider only the simplest such example - the symmetric r-state model, due to Neyman [8] (see also, [1], [3], [4]) and abbreviated hereafter as the *$N_r$model* (the case $r = 4$ is known in genetics as the *Jukes-Cantor* model).

In this model, $S = \{0, 1, \ldots r-1\}$, the underlying tree T is *fully resolved* - that is, each vertex has degree 1 or 3, and we have an associated function $p : E \rightarrow (0, \frac{r-1}{r})$. Randomly assign, with uniform probability, an element of S to any one fixed vertex, v_0 of T and then assign states to the remaining vertices recursively as follows: for each vertex v which has an adjacent vertex closer to v_0 that has been assigned a state, randomly assign v this same state with probability $1 - p(e)$ (where $e = \{v_0, v\}$) or assign v one of the other states with equal probability (viz, $\frac{p(e)}{r-1}$). In this way we generate a random function $F : V \rightarrow S$. Under the N_r model, with parameters (T,p), let $\mathbb{P}(F = h)$ be the probability that $F = h$, and let $\mathbb{P}(f,T)$ denote the probability that $F_{|L} = f$.

By definition, and the assumptions of the model,

$$\mathbb{P}(f,T) = \sum_{\{h:V \rightarrow \{0,1,\ldots,r-1\}:h_{|L}=f\}} \mathbb{P}(F = h)$$

and

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

from which we immediately see that the probability distribution on characters f (and extensions h) is independent of our choice of v_0. There is also an elegant representation of $\mathbb{P}(f, T)$ using unitary matrices (see [5]; [11]), but we do not use this here. 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 generalisations thereof).

In a landmark paper, [4] Felsenstein showed that, under the N_2 model, MP is not statistically consistent for certain choices of p. Indeed this was demonstrated even when T has just four leaves (in which case MP=MC). This inconsistency phenomena has been refined and extended by others [5], [6], [9]. Sufficient conditions for the statistically consistency of MP or MC have only been described when either T has just four leaves [7], or for special cases [5], [6],[9].

Here we provide the first explict sufficient conditions for statistical consistency of MP (resp. MC) that are applicable to any tree on any number of leaves under the N_r (resp. N_2) model.

2 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.

Let $\mathring{E}$ denote the interior edges of T - that is, the edges of T that are not incident with a leaf. Deleting an edge $e \in E$ from T produces a partition β_e of the leaves of T into two subsets. If $\{f^{-1}(s) : s \in S\}$ is a bipartition, and equals β_e we say that f *corresponds to* edge e. Thus, 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 $c(T) = \cup_{e \in \mathring{E}} c(e)$. If $\{f^{-1}(s) : s \in S\} = \{g^{-1}(s) : s \in S\}$ we write $f \sim g$.

Suppose we are given a sequence C of characters. Let $n(C, f)$ denote the number of times character f occurs in this sequence. Let $n[C, f] := \sum_{g \sim f} n(C, g)$, and for an edge e of T let $n_e(C) = \sum_{f \in c(e)} n(C, f) = n[C, f_e]$, where f_e is any character that corresponds to edge e. Let:

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

$$n_+(C, T) := max\{n[C, f] : L(f, T) > 1\},$$

and

$$N(C, T) := \sum_{\{f : L(f,T) \geq 2\}} n(C, f) L(f, T).$$

The following result gives sufficient conditions for MP and 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.*

1. *MP selects tree T for a sequence C of r-state characters if $n_-(C, T) > N(C, T)$.*

2. *MC selects tree T for a sequence C of 2-state characters if $n_-(C, T) > n_+(C, T)$.*

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

Part 1: 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 $T_1 = T$. Firstly note that:

$$L(T) = \sum_{\{f:L(f,T)=1\}} n(f) + N,$$

and for any tree $T_1 \neq T$ we have:

$$L(T_1) \geq \sum_{\{f:L(f,T)=1\}} L(f,T_1)n(f),$$

Now, since $T_1 \neq T$ and T is fully resolved, T has at least one internal edge e for which, for all $f \in c(e)$ we have $L(f,T_1) \geq 2$. Consequently, $\sum_{\{f:L(f,T)=1\}} L(f,T_1)n(f) \geq \sum_{\{f:L(f,T)=1\}} n(f) + n_-$, which combined with the previous inequalities establishes the claim.

Part 2: Let $\nu(T_1)$ denote the number of occurences of a characters f with $L(f,T_1) = 1$. It suffices to show that $\nu(T_1)$ is strictly maximized when $T_1 = T$. Firstly 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 $T_1 \neq T$ on leaf set $\mathcal{L}$ let S denote the subset of internal edges e of T for which $\beta_e \neq \beta_{e'}$ for any edge e' of T_1. Since T is fully resolved, $S \neq \emptyset$, and the number of edges e in T_1 for which $\beta_e \cap \{\beta_{e'} : e' \in E(T)\} = \emptyset$ is at most $s = |S|$ and for each such edge $n_e(C) \leq n_+$. Thus,

$$\nu(T) - \nu(T_1) \geq \sum_{e \in S} n_e(C) - sn_+ > 0$$

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 MC and MP (Corollary 1). First we introduce the following terminology.

Definitions. Under the N_r model with parameters (T,p), let

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

and

$$\mu := \sum_f \mathbb{P}(f,T)L(f,T);\ M := \sum_{\{f:L(f,T)\geq 2\}} \mathbb{P}(f,T)L(f,T).$$

Corollary 1 *1. MP is statistically consistent under the N_r model if $m_- > M/r$.*

2. MC is statistically consistent under the N_2 model if $m_- > m_+$.

Proof Note that, under the N_r model, if $f \sim g$ then, $\mathbb{P}(f,T) = \mathbb{P}(g,T)$. 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,

$$\frac{n_-(C,T)}{c} \rightarrow_p rm_-,$$

$$\frac{n_+(C,T)}{c} \rightarrow_p rm_+,$$

and

$$\frac{N(C,T)}{c} \rightarrow_p M,$$

where $\rightarrow_p$ denotes convergence in probability. 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*

$$s := \sum_e p(e),$$

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

and

$$x := p_- + p_+$$

Then,

1. *MP is statistically consistent if*
$$p_- \geq \frac{(r-1)s^2}{1-s}$$
2. *When $r = 2$, MC is statistically consistent if*

$$p_- \geq p_+ x + 2x^2(1+x) \tag{1}$$

Remark: Informally, for 2-state characters, the condition for the consistency of MP is that the probability of a mutation on any internal edge (i.e. $p(e)$) 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 MC the condition described states, informally, that the probability of mutation on any internal edge is at least some multiple c of the square of the largest mutation probability p_+ in the tree. This places an upper bound on p_+ as follows: let t denote the largest possible value of p_+ subject to the condition (1). Then, since $x \leq 2t$, and $t \geq p_-$, condition (1) shows that $t \geq 2t^2 + 8t^2(1+2t)$ from which $t \leq \frac{-5+\sqrt{41}}{16} \approx 0.087$ (and indeed we can achieve this by putting all $p(e)$ values equal). Thus, $12p_+^2 \geq 10p_+^2 + 16p_+^3 \geq p_+x + 2x^2(1+x)$ and so we see that $p_- \geq cp_+^2$ implies condition (1) for $c = 12$. However if p_+ is less than t, c can be reduced, and indeed c converges to 3 as p_+ converges to 0. More comments concerning the interpretation and consequences of this Theorem are contained in Section 5.

Proof of Theorem 1 *Part 1:* Let us first recall the definition of μ introduced just before Corollary 1, and let $\mathbb{E}$ denote expectation. Then, $\mu = \mathbb{E}(L(F_{|L}, T))$ for an extension F randomly generated on T under the N_r model. Now, $L(F_{|L}, T) \leq ch(F, T)$, and so,

$$\mu \leq \mathbb{E}(ch(F, T)) \tag{2}$$

However, $ch(F, T)$ is simply a sum of independent $0/1$ random variables - one for each edge, and taking the value 1 on edge $e = \{u, v\}$ iff $F(u) \neq F(v)$ which has probability $p(e)$. Consequently,

$$\mathbb{E}(ch(F, T)) = s \tag{3}$$

Let

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

Now, $M = \mu - \mathbb{P}(L(F_{|L}, T) = 1)$ and $\mathbb{P}(L(F_{|L}, T) = 1) \geq \mathbb{P}(ch(F, T) = 1)$. Furthermore, $\mathbb{P}(ch(F, T) = 1) = \sum_e p(e) \prod_{e' \neq e} (1 - p(e)) \geq sQ$. Thus, $M \leq \mu - sQ$ while (in)equalities (2), (3) give $\mu \leq s$, and hence

$$M \leq s - sQ \tag{4}$$

Now, for a character $f \in c(T)$, let e_0 denote the edge of T that f corresponds to. An extension h 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_0, and for each edge $e \neq e_0$ assigning the same state to each end of e_0. Consequently,

$$\mathbb{P}(f, T) \geq \mathbb{P}(F = h) = \frac{1}{r} \times \frac{p(e_0)}{r - 1} \prod_{e \neq e_0} (1 - p(e)) > \frac{1}{r} Q\delta,$$

where $\delta = \frac{p_-}{r-1}$ and so:

$$m_- \geq \frac{1}{r} Q\delta \tag{5}$$

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

$$Q > 1 - s,$$

implies that $Q > \frac{s}{s+\delta}$. Rearranging gives $Q\delta > s - sQ$ and thus, in view of the inequalities (4), (5) we have

$$m_- > M/r.$$

Part (1) of the Theorem now follows from Corollary 1(1).

Part 2: 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 vertex of T is in state $\mu \in \{0, 1, \dots, r-1\}$ is precisely $r\mathbb{P}(f, T)$.
- Let t_1 and t_2 be two subtrees of T which share one non-leaf 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.

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,

$$\mathbb{P}(f, T) > \frac{p(e)}{1-p(e)} P_0 \tag{6}$$

To establish (6) 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,

$$\mathbb{P}(f, T) = \alpha p(e) + \beta(1 - p(e)) \tag{7}$$

and by virtue of the symmetry in the N_2 model which implies that $P_i(0,0) = P_i(1,1); P_i(0,1) = P_i(1,0)$ 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 (6). Actually all we shall require is the following corollary of inequality (6): for any $f \in c(T)$,

$$\mathbb{P}(f, T) > P_0 p_- \tag{8}$$

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,

$$\mathbb{P}(f, T) < x P_0 \tag{9}$$

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

$$\mathbb{P}(f, T) < 2x^2(1 + x) P_0 < x P_0 \tag{10}$$

The proof of inequality (9) 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 $p(e) \in (0, 0.5)$, $\mathbb{P}(f, T) = \frac{1}{2}p(e) < p(e)(1 - p(e)) = xP_0$ as required. Now, suppose that $n > 2$. We distinguish two subcases.

- $f \in c(T)$
- f corresponds to a non-interior 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 Fig. 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

$$P_i(\mu, \nu) = P_i^a(\mu, \nu)P_i^b(\mu, \nu). \tag{11}$$

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,

$$P_i^\theta(\mu, \nu) < xP_i^\theta(0, 0), (\mu \neq \nu) \tag{12}$$

Now, recalling Equation (7) we have $\mathbb{P}(f, T) = \alpha p(e) + \beta(1 - p(e))$ and so, substituting in Equations (11, 12) into the definitions of α and β we deduce that:

$$\mathbb{P}(f, T) < R[p(e) + 2x^2(1 - p(e)) + p(e)x^4]. \tag{13}$$

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

Also, we have,

$$P_0 \geq (1 - p(e))R \geq (1 - p_+)R \tag{14}$$

Now, we can bound the term in brackets in (13) by noting that $p(e) + 2x^2(1 - p(e)) + p(e)x^4 < p_+ + 2x^2 + x^3$ (since $x < 1$), and then, by our assumption (1), we have: $p_+ + 2x^2 + x^3 \leq x(1 - p_+)$. Substituting this into (13) and comparing the result to (14) establishes inequality (9) 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:

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

and thus, since $p(e) + x^2(1 - p(e)) < p_+ + x^2 < x(1 - p_+)$, we have:

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

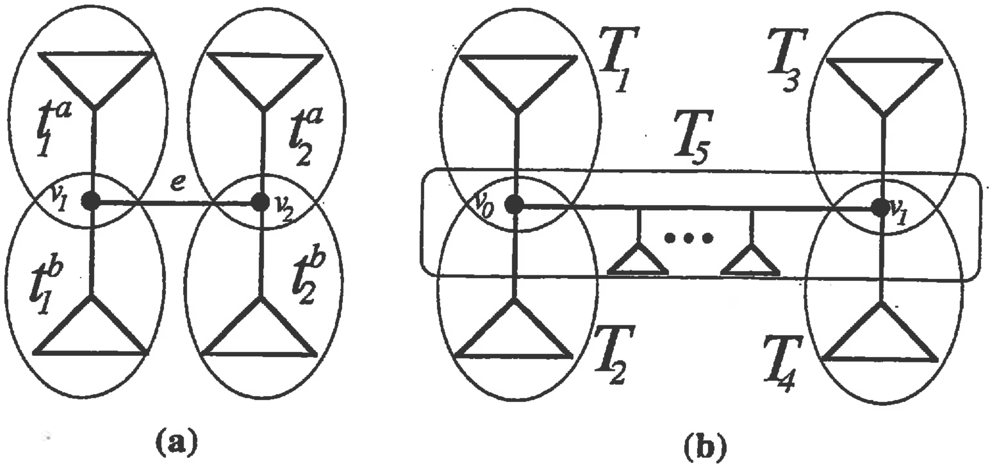

Figure 2: Representations of T for the proof of the upper bounds (9) and (10).

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 (9) holds in this subcase also.

We now establish inequality (10). We first observe that our condition (1) forces $2x^2(1+x) < x$ so it is only the first inequality in (10) 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 ([12], Lemma 1). Thus, we may represent T as in Fig. 2(b), with five subtrees trees $T_1, ..., T_5$ as shown, each pair of which is disjoint or overlaps at one of two (generally non-adjacent) 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_μ^i 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

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

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_0^i equal twice the probability of generating on T_i the character which maps all leaves to 0. Then,

$$\mathbb{P}(f,T) = \frac{1}{2}\sum_{\mu,\nu} P_5(\mu,\nu) \prod_{i=1,2} P_i(\mu) \prod_{i=3,4} P_i(\nu). \tag{15}$$

Now, by induction, we may assume that inequality (10) holds for all five subtrees.

Thus, for $i = 1, \ldots, 4$, $P_i(\mu) < xP_0^i$ when $L_i(\mu) > 0$, and $P_i(\mu) = P_0^i$ otherwise. Similarly, $P_5(\mu,\nu) < xP_0^5$ when $L_5(\mu,\nu) > 0$, and $P_i(\mu) = P_5^i$ otherwise. Consequently for when $\mu = \nu \in$

$\{0,1\}$ we introduce at least two powers of x into the product terms of Equation (15), while for $\mu \neq \nu$ we introduce at least three powers of x. Thus, we deduce that:

$$\mathbb{P}(f,T) < 2x^2(1+x) \times \frac{1}{2} \cdot \prod_{i=1,\dots,5} P_0^i$$

and inequality (10) now follows by observing that:

$$P_0 > \frac{1}{2} \prod_{i=1,\dots,5} P_0^i.$$

Finally we establish part (2) of the theorem. In view of inequalities (9) and (10) we have:

$$m_- > P_0 p_-$$

and

$$m_+ < 2x^2(1+x)P_0$$

Now, by our condition (1), $p_- > p_+x + 2x^2(1+x) > 2x^2(1+x)$ we see that $m_- > m_+$. The theorem now follows from Corollary1(2). □

3 Remarks

An interesting consequence of part (2) of the previous theorem is that MC is statistically consistent under the N_2 model whenever $p(e)$ is constant ($= p$) across the edges of the tree, and p takes a value at most 0.087 (regardless of the number of leaves of T). An interesting theoretical question is whether a similar result holds for MP when $p(e) = p$ (from [9] it would be necessary, under the N_2 model, that $p < \frac{1}{8}$). Note that the sufficient condition described in Theorem 1(2) requires that the $p(e)$ values to converge to 0 at least as fast as n^{-2}, where $n = |L|$, so in a certain sense the sufficient condition described for *MC* is much stronger than that for MP.

More generally, for any bound B on the ratio of the $p(e)$ values, Theorem 1(2) implies that there exists some upper bound on p_+ (dependent on that bound) for which MC is statistically consistent under the N_2 model.

It is also instructive to compare the strengths of the two parts of Theorem 1, for the simplest case $n = 4$. In [4] Felsenstein considered the N_2 model on a resolved tree on four leaves, with two non-adjacent edges having $p(e) = r$, and the remaining three edges having $p(e) = q$. He showed that MP is statistically inconsistent precisely when $q(1-q) < r^2$, which, for q small, amounts, approximately, to $q < r^2$. By contrast, the sufficient condition described in the above theorem for MP would require $q \geq \frac{(2r+3q)^2}{1-2r-3q}$ which for $q \ll r \ll 1$ amounts, approximately, to $q > 4r^2$. For MC (which agrees with MP on trees with four leaves) the analogous sufficient condition reduces, approximately, to $q > 3r^2$. 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 MP (see [7]), 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 MP and MC. An extension of Theorem 1(2) to $r > 2$ would also be interesting.

4 Acknowledgement

The author thanks the New Zealand Marsden Fund for supporting this research, the Isaac Newton Institute (Cambridge) for its hospitality, and Jotun Hein, for offering helpful comments on an earlier version of this manuscript.

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), 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] M.D. HENDY, and D. PENNY, *A framework for the quantitative study of evolutionary trees,* Syst. Biol., 38 (1986), pp. 297-309.

[6] J. KIM, *General inconsistency conditions for maximum parsimony: effects of branch length and increasing the number of taxa,* Syst. Biol., 45(3) (1996), pp. 363-374.

[7] 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 Univ. Press, 1991, pp. 155-183.

[8] 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., New York, Academic Press, 1971, pp. 1-27.

[9] M.A. STEEL, Distributions on bicoloured evolutionary trees, PhD thesis, Massey University, Palmerston North, New Zealand, 1989.

[10] D.L. SWOFFORD, G.J. OĹSEN, 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, 1996, pp. 407-514.

[11] L.A. SZÉKELY, M.A. STEEL, and P.L. ERDŐS, *Fourier calculus on evoltuionary trees,* Adv. Appl. Math., 14 (1993), pp. 200-216.

[12] C. TUFFLEY and M. STEEL *Links between maximum likelihood and maximum parsimony under a simple model of site substitution,* Bull. Math. Biol., 59(3) (1997), pp. 581-607.