Bulletin of Mathematical Biology, Vol. 59, No. 3, pp. 581–607, 1997
Elsevier Science Inc.
© 1997 Society for Mathematical Biology
0092-8240/97 $17.00 + 0.00

S0092-8240(97)00001-3

LINKS BETWEEN MAXIMUM LIKELIHOOD AND MAXIMUM PARSIMONY UNDER A SIMPLE MODEL OF SITE SUBSTITUTION

■ CHRIS TUFFLEY and MIKE STEEL
Biomathematics Research Centre,
Department of Mathematics and Statistics,
University of Canterbury,
Christchurch, New Zealand

(*E.mail*: *m.steel@math.canterbury.ac.nz*)

Stochastic models of nucleotide substitution are playing an increasingly important role in phylogenetic reconstruction through such methods as maximum likelihood. Here, we examine the behaviour of a simple substitution model, and establish some links between the methods of maximum parsimony and maximum likelihood under this model. © 1997 Society for Mathematical Biology

1. Introduction. Stochastic models for nucleotide substitution are becoming increasingly important as a foundation for inferring phylogenetic trees from genetic sequence data. Such models allow for tree reconstruction through either maximum likelihood-based approaches or the fitting of transformed functions of the data to trees (see Swofford *et al.* (1996) for a recent survey). The models are also useful for analysing the performance of other, more conventional tree reconstruction methods, which are not explicitly based on such models, such as the popular maximum parsimony method (see, for example, Fitch (1971)). Such methods will indeed perform well (be "statistically consistent") for sequences that evolve under simple models with certain constraints (see, for example, Hendy and Penny (1989)), although without these constraints, the methods may be misled (Felsenstein, 1978). Thus, for certain data sets, maximum parsimony and maximum likelihood will agree, and in other cases, they will disagree. In this paper, we carry this analysis a little further for a simple model on any number of states, in which the rate of substitution is the same between any two states. In particular, we establish, for any tree and any number of states, an inequality between the probability of a character at a site and a function of the character's parsimony score on the underlying tree (Theorem 1). This bound becomes an equality for certain choices of parameters in the underlying model, and we completely characterise these choices when r (the number of states) is 2 (Theorem 3).

We then use these results in four applications to the theory of phylogenetic analysis in Section 6. We establish three further cases in which maximum parsimony will agree with certain versions of maximum likelihood in the selection of trees and the reconstruction of ancestral states on a given tree. One of these results, Theorem 5, extends a result of Penny *et al.* (1994) from 2 to r states; another offers insight into the observations in Lockhart *et al.* (1996). We also generalise the example of Steel (1994) to an arbitrary number of species, and thereby show that the maximum likelihood function can be maximised at many points in the underlying parameter space.

2. Preliminaries.

Definitions 1 (Phylogenetic trees, characters). A *phylogenetic tree* is a tree $T=(V(T),E(T))$ having no vertices of degree 2, and such that each *leaf* (degree 1 vertex) is given a unique label from $\{1,\dots,n\}$, where n is the number of leaves of T. We say that T is a tree on n leaves, and write $[n]$ for $\{1,\dots,n\}$. Where convenient, we identify each leaf with its label. If every internal (non-leaf) vertex of T has degree 3, we say that T is *binary*. In the case of rooted trees, we allow the root to have degree 2.

A function $\chi:[n]\mapsto\mathcal{S}$, where $\mathcal{S}$ is a set of r *states*, is an (*r-state*) *character*. When $r=2$, χ is said to be *binary*. A function $\hat{\chi}:V(T)\mapsto\mathcal{S}$ is called a *state function* for T; if $\hat{\chi}$ is such that $\hat{\chi}|_{[n]}=\chi$ (that is, $\hat{\chi}$ agrees with χ on the leaves of T), then $\hat{\chi}$ is called an *extension* of χ (on T).

With each character χ and phylogenetic tree T on n leaves, we may associate a non-negative integer (the "length" of χ on T) as follows.

Definitions 2 (Length of χ on T, minimal extensions). If $\hat{\chi}:V(T)\mapsto\mathcal{S}$, then the *changing number* of $\hat{\chi}, ch(\hat{\chi})$, is the number of edges $\{u,v\}$ such that $\hat{\chi}(u)\neq\hat{\chi}(v)$. We say that a *change occurs across* $\{u,v\}$ *under* $\hat{\chi}$.

If $\chi:[n]\mapsto\mathcal{S}$, then the *length of* χ *on the phylogenetic tree* T, $l(\chi,T)$, is the minimum of $ch(\hat{\chi})$ over all extensions $\hat{\chi}$ of χ on T. An extension of minimal changing number is called a *minimal extension of* χ (on T).

Figure 1 illustrates these definitions.

In practical applications, the length of a character on a given tree is found using Fitch's algorithm, which is an order n process for determining $l(\chi,T)$ and finding a minimal extension (Fitch (1971)). However, for theoretical purposes, $l(\chi,T)$ is usefully given in the two-state case by the following corollary of Menger's Theorem, a result that will be of use to us later. Although this is an often-quoted result (for example, Erdős and Székely (1993); Steel (1993b)), we include a proof as it does not follow directly from Menger's Theorem, and we believe a proof has yet to appear in the literature.

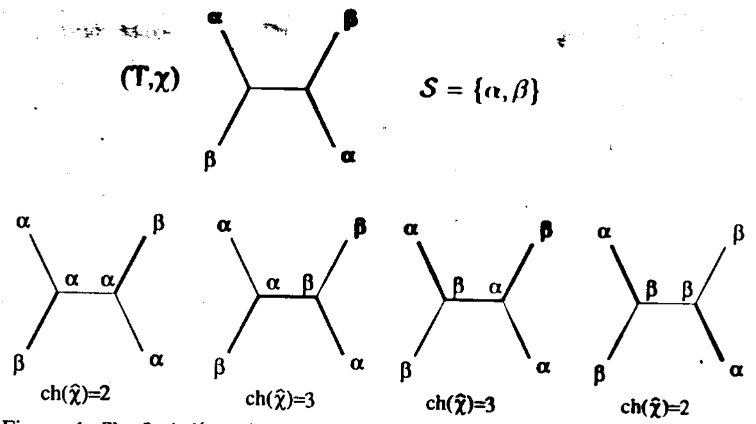

Figure 1. To find $l(\chi,T)$ for the tree and character shown with state set $\mathcal{S}=\{\alpha,\beta\}$, consider all possible assignments of states to the internal vertices. Since T has two internal vertices, there are $2^2=4$ such assignments. Edges on which changes occur are shown in bold. The minimum value of $ch(\hat{\chi})$ is 2, so that $l(\chi,T)=2$. There are two minimal extensions.

LEMMA 1. *If* χ *is a binary character, then* $\mathrm{l}(\chi,\mathrm{T})$ *equals the maximum number of edge-disjoint paths connecting leaves in different states.*

Proof. Given disjoint sets $X,Y\subseteq V(T)$, a ***cutset*** for X and Y is a subset $Z\subseteq E(T)$ such that any path in T joining a vertex in X to a vertex in Y crosses at least one edge in Z. A variation of Menger's Theorem (see Harary (1969)) then states that the cardinality of a cutset of minimal size is equal to the maximal size of sets of edge-disjoint paths joining vertices in X to vertices in Y.

If $\mathcal{S}=\{\alpha,\beta\}$, set $X=\chi^{-1}(\{\alpha\}),Y=\chi^{-1}(\{\beta\})$ and let l be the maximum number of edge-disjoint paths connecting leaves in different states. If $\hat{\chi}$ is an extension of χ on T, then the set of edges on which changes occur under $\hat{\chi}$ gives a cutset for X and Y, so by Menger's Theorem, $ch(\hat{\chi})\geqslant l$, and therefore $l(\chi,T)\geqslant l$.

Now, let Z be a cutset for X and Y of minimal size. Generate a state function $\hat{\chi}$ for T by putting $\hat{\chi}(v)=\chi(i)$ if there is a path from v to leaf i that does not cross any edges in Z. We claim that this is well defined. Firstly, there cannot be two such paths to leaves i and j such that $\chi(i)\neq\chi(j)$; otherwise, Z would not be a cutset for X and Y. Secondly, there must be at least one such path. If not, let e be a nearest edge in Z to v. Since there is no path from v to a leaf that does not cross an edge in Z, there can be no path P from a vertex in X to a vertex in Y such that e is the only edge in Z crossed by P. Thus, $Z\setminus\{e\}$ is a cutset for X and Y, contradicting the minimality of $|Z|$. It follows that $\hat{\chi}$ is indeed well defined.

The cutset given by the edges on which changes occur under $\hat{\chi}$ is contained in Z, so we have $l(\chi, T) \leqslant ch(\hat{\chi}) \leqslant |Z| = l$. Putting the two inequalities together gives $l(\chi, T) = l$, as required.

Although this result applies only to binary characters, an extension to r-state characters has been developed recently by Erdős and Székely (1994), the principal difference being that the paths are permitted to intersect provided certain conditions are met.

In biology, each vertex of a phylogenetic tree represents a species, with the edges denoting (immediate) ancestor–descendant relationships. The leaves represent extant species, the internal vertices ancestral species, and in rooted trees, the root represents a common ancestral species from which all other species on the tree are descended. Since we are primarily interested in speciation events, where the tree "branches," we do not allow vertices of degree 2, except possibly at the root.

Characters are obtained by gathering data such as DNA sequence information from present day species. For example, given a set of n aligned sequences, the rule "nucleotide at site i" gives a character with state set $\mathscr{S} = \{A, G, C, T\}$. Each extension of a character is a way that it could have evolved on the tree, and the changing number of an extension is the number of changes or mutations it involves. The length of a character is therefore the minimum number of mutations required for it to evolve on the tree, and is used in methods such as maximum parsimony to estimate the true phylogeny.

3. The Model. The model we will be considering is a generalisation to r states of the Cavender–Farris (Cavender, 1978; Farris, 1973; two-state case) and Jukes–Cantor (Jukes and Cantor, 1969; four-state case) models, and appears in Neyman (1971). We will refer to it as the *fully symmetric model* since it makes no distinction between any of the character states. Given a rooted phylogenetic tree T and a mutation probability p_e on each edge e of T, the state at the root "evolves" down the tree, assigning a state to each vertex of T and generating a state function $\hat{\chi}$ for T. We suppose that this evolution takes place such that:

- there is an even distribution of states at the root ρ, that is,

$$\mathbb{P}[\hat{\chi}(\rho) = \alpha] = \frac{1}{r} \qquad \forall \alpha \in \mathscr{S}; \tag{1}$$

- the probability of a net change of state occurring across an edge e (a "mutation event") is given by p_e, and if a net change occurs, each of the remaining $r-1$ states is equally likely;
- mutation events on different edges are independent;
- p_e satisfies $0 \leqslant p_e \leqslant (r-1)/r$ (see below).

The probability of generating a given state function $\hat{\chi}$ will, in general, depend on T and the vector $p = (p_e)_{e \in E(T)}$ of probabilities, and is given by

$$\mathbb{P}[\hat{\chi} \mid T, p] = \frac{1}{r} \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1}. \tag{2}$$

The probability of generating a character χ under the model is found by summing (2) over all extensions $\hat{\chi}$ of χ, so that

$$\mathbb{P}[\chi \mid T, p] = \sum_{\hat{\chi} : \hat{\chi}|_{[n]} = \chi} \mathbb{P}[\hat{\chi} \mid T, p]. \tag{3}$$

Figure 2 shows a calculation of $\mathbb{P}[\chi \mid T, p]$ for a simple tree and character.

This model may be formulated in terms of a continuous-time Markov process with $r \times r$ *rate matrix*

$$Q = \begin{pmatrix} 1-r & 1 & \cdots & 1 \\ 1 & 1-r & & \vdots \\ \vdots & & \ddots & 1 \\ 1 & \cdots & 1 & 1-r \end{pmatrix}, \tag{4}$$

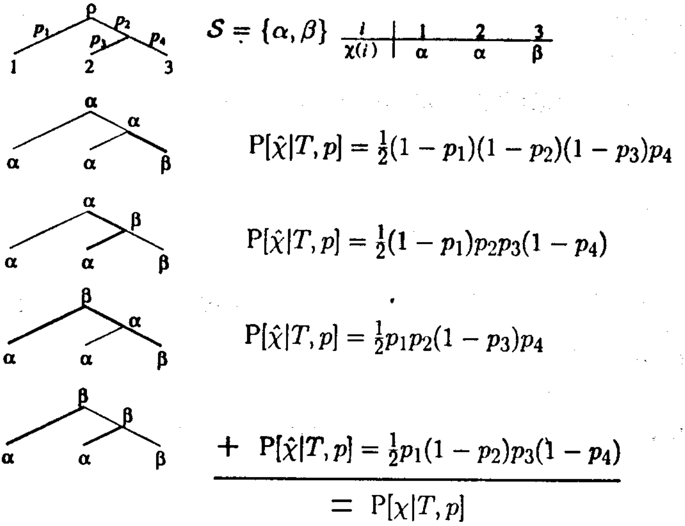
p_1
p_2
p_3
p_4
ρ
1
2
3
$S = \{\alpha, \beta\}$
i | 1 | 2 | 3
$\chi(i)$ | α | α | β
α α α α β
$\mathrm{P}[\hat{\chi}|T,p] = \frac{1}{2}(1-p_1)(1-p_2)(1-p_3)p_4$
α β α α β
$\mathrm{P}[\hat{\chi}|T,p] = \frac{1}{2}(1-p_1)p_2p_3(1-p_4)$
β α α α β
$\mathrm{P}[\hat{\chi}|T,p] = \frac{1}{2}p_1p_2(1-p_3)p_4$
β β α α β
$+\ \mathrm{P}[\hat{\chi}|T,p] = \frac{1}{2}p_1(1-p_2)p_3(1-p_4)$
$= \mathrm{P}[\chi|T,p]$

Figure 2. To calculate $\mathbb{P}[\chi \mid T, p]$ for the tree and character shown, sum $\mathbb{P}[\hat{\chi} \mid T, p]$ over all extensions $\hat{\chi}$ of χ. In the two-state case, each edge on which a change occurs contributes a factor of p_e to $\mathbb{P}[\hat{\chi} \mid T, p]$; all other edges contribute a factor of $1 - p_e$.

whose $\alpha\beta$-entry gives the rate at which α changes to β. With each edge e, we associate a positive "length" τ_e. The conditional probability of a change from α to β across e given that the vertex closer to the root is in state α is then given by the $\alpha\beta$-entry of the *transition matrix*

$$m^e = \exp(\tau_e Q). \tag{5}$$

Diagonalising Q, we obtain

$$m^e = \begin{pmatrix} 1-p_e & \frac{p_e}{r-1} & \cdots & \frac{p_e}{r-1} \\ \frac{p_e}{r-1} & 1-p_e & & \vdots \\ \vdots & & \ddots & \frac{p_e}{r-1} \\ \frac{p_e}{r-1} & \cdots & \frac{p_e}{r-1} & 1-p_e \end{pmatrix} \tag{6}$$

where, as in Neyman (1971),

$$p_e = \frac{r-1}{r}(1-\exp(-r\tau_e)) \tag{7}$$

and satisfies $0 < p_e < (r-1)/r$. For simplicity, we include the endpoints of this interval so that our set of possible mutation probability vectors is compact.

This Markov process with the initial distribution of states $\pi = (1/r, \ldots, 1/r)$ has some additional properties (namely, *stationarity* (that is, $\pi Q = 0$) and *reversibility* ($\pi_\alpha Q_{\alpha\beta} = \pi_\beta Q_{\beta\alpha}$ for all α, β)) that imply we may re-root T at any vertex (for example, leaf 1), keeping the same distribution of states at the new root and the same transition matrices on each edge. The transition matrix m^P for any path P is then the product of the matrices along the path, so that

$$m^P = \prod_{e\in P} \exp(\tau_e Q) = \exp\left(\sum_{e\in P} \tau_e Q\right), \tag{8}$$

and hence,

$$p_P = \frac{r-1}{r}\left(1-\exp\left(-r\sum_{e\in P}\tau_e\right)\right); \tag{9}$$

in terms of the mutation probabilities, this is

$$p_P = \frac{r-1}{r}\left(1-\prod_{e\in P}\left(1-\frac{r}{r-1}p_e\right)\right). \tag{10}$$

Since p_e is a monotonically increasing function of τ_e, we have also

$$p_P \geqslant \max_{e \in P} p_e, \tag{11}$$

an inequality that will be of use to us later.

4. Bounding $\mathbb{P}[\chi \mid T, p]$. Penny *et al.* (1994) have shown that in the two-state case,

$$\max_p \{\mathbb{P}[\chi \mid T, p]\} = 2^{-l(\chi,T)-1}. \tag{12}$$

The proof made use of equation (10) and the set of edge-disjoint paths given by Lemma 1 to establish $2^{-l(\chi,T)-1}$ as an upper bound. This method could not be readily generalised to r-states as the paths of an Erdős–Székely path system need not be edge-disjoint, which was an important requirement of the method. We extend (12) here to r-states via a different method of proof. A key step (Lemma 4) was originally proved using Erdős–Székely path systems; however, a simpler proof has since been found, which we give here.

THEOREM 1 (Upper bound for $\mathbb{P}[\chi \mid T, p]$). *If χ is an* r*-state character, then*

$$\max_{\mathrm{p}} \{\mathbb{P}[\chi \mid \mathrm{T}, \mathrm{p}]\} = \mathrm{r}^{-\mathrm{l}(\chi,\mathrm{T})-1}. \tag{13}$$

That is, for a given tree T, *the maximum value of* $\mathbb{P}[\chi \mid \mathrm{T}, \mathrm{p}]$ *over all valid choices of* p *is* $\mathrm{r}^{-\mathrm{l}(\chi,\mathrm{T})-1}$.

Note that this theorem applies only to a single character, and not to a set of characters. However, if X is a set of r-state characters that evolve identically and independently according to the fully symmetric model, Theorem 1 gives

$$\mathbb{P}[X \mid T, p]\left(= \prod_{\chi \in X} \mathbb{P}[\chi \mid T, p]\right) \leqslant r^{l(X,T)-|X|} \tag{14}$$

where $l(X,T) = \Sigma_{\chi \in X}\, l(\chi, T)$. In contrast to (13), it will not, in general, be possible to realise equality in (14).

As a corollary to the proof of Theorem 1, we obtain the following result. With each extension $\hat{\chi}$ of a character χ, we may associate a subset $E(\hat{\chi})$ of the edges of T by taking the set of edges on which changes occur under $\hat{\chi}$. In general, for a given extension $\hat{\chi}$, there may be another extension $\bar{\chi}$ such that $E(\hat{\chi}) = E(\bar{\chi})$; however, in the case of minimal extensions, Theorem 2 says that this cannot occur.

THEOREM 2. *A minimal extension $\hat{\chi}$ of an* r-*state character χ on* T *is uniquely determined by χ and the set of edges on which changes occur under $\hat{\chi}$. That is, if χ_1 and χ_2 are minimal extensions of χ and the set of edges on which changes occur under χ_1 equals the set of edges on which changes occur under χ_2, then $\chi_1 = \chi_2$.*

The proof of Theorem 1 proceeds via a series of lemmas. We begin by reducing to the case where, for every edge e of T, p_e is either 0 or $(r-1)/r$. For notational convenience, we make the following definitions.

Definitions 3. Let

$$M(T) = \left\{p \in [0,(r-1)/r]^{|E(T)|} : p_e \in \{0,(r-1)/r\} \qquad \forall e \in E(T)\right\} \tag{15}$$

be the set of mutation probability vectors with each component either 0 or $(r-1)/r$, and for each $p \in M(T)$, define

$$E(p) = \{e \in E(T) : p_e = (r-1)/r\}, \tag{16}$$

the set of edges where $p_e = (r-1)/r$. For each state function $\hat{\chi}$ for T, let

$$E(\hat{\chi}) = \left\{\{u,v\} \in E(T) : \hat{\chi}(u) \neq \hat{\chi}(v)\right\} \tag{17}$$

be the set of edges on which changes occur under $\hat{\chi}$.

Let χ be an r-state character of length l on T. For the moment, we will consider χ and T to be fixed and write $\mathbb{P}(p)$ for $\mathbb{P}[\chi | T, p]$, to emphasise the view of $\mathbb{P}[\chi | T, p]$ as a function of p. Thus,

$$\mathbb{P}(p) = \frac{1}{r} \sum_{\hat{\chi} : \hat{\chi}|_{[n]} = \chi} \; \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}(u) = \hat{\chi}(v)}} (1 - p_e) \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}(u) \neq \hat{\chi}(v)}} \frac{p_e}{r-1}. \tag{18}$$

Note that for each edge e of T, p_e occurs in each term in the sum in (18) exactly once. Let $p \in [0,(r-1)/r]^{|E(T)|}$. Choosing $e' \in E(T)$ and fixing p_e for $e \in E(T) \setminus \{e'\}$ (so that we regard $\mathbb{P}(p)$ as a function of $p_{e'}$), we therefore obtain a polynomial of degree at most one in $p_{e'}$. On a closed interval, the extreme values of such a polynomial occur at the endpoints, so there is a vector p' of mutation probabilities such that

$$p'_e = \begin{cases} p_e & \text{if } e \neq e' \\ 0 \text{ or } \dfrac{r-1}{r} & \text{if } e = e' \end{cases} \tag{19}$$

and

$$\mathbb{P}(p) \leqslant \mathbb{P}(p'). \tag{20}$$

Carrying out this process for each edge of T in turn, we eventually arrive at a vector p'' such that $p'' \in M(T)$ and

$$\mathbb{P}(p) \leqslant \mathbb{P}(p''). \tag{21}$$

We have established the following lemma.

LEMMA 2. $\max_{\mathrm{p}}\{\mathbb{P}[\chi \mid \mathrm{T}, \mathrm{p}]\}$ *is realised by some* $\mathrm{p} \in \mathrm{M(T)}$.

Now, let $p \in M(T)$. Each extension $\hat{\chi}$ of χ contributes a term

$$\mathbb{P}[\hat{\chi} \mid T, p] = \frac{1}{r} \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} \tag{22}$$

to $\mathbb{P}(p)$. If there is an edge $e=\{u,v\}$ for which $\hat{\chi}(u) \neq \hat{\chi}(v)$ and $p_e = 0$, then a factor of zero occurs in the right-hand product in (22) and we have $\mathbb{P}[\hat{\chi} \mid T, p] = 0$. Hence, we need only sum over extensions $\hat{\chi}$ such that $E(\hat{\chi}) \subseteq E(p)$. Further, if $E(\hat{\chi}) \subseteq E(p)$, then each edge $e=\{u,v\}$ contributes a factor

$$m^{e}_{\hat{\chi}(u)\hat{\chi}(v)} = \begin{cases} \dfrac{p_e}{r-1} = \dfrac{1}{r} & \text{if } \hat{\chi}(u) \neq \hat{\chi}(v) \\ 1 - p_e = \dfrac{1}{r} & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \text{ and } p_e = \dfrac{r-1}{r} \\ 1 - p_e = 1 & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \text{ and } p_e = 0 \end{cases} \tag{23}$$

to $\mathbb{P}[\hat{\chi} \mid T, p]$. Thus, each edge for which $p_e = (r-1)/r$ contributes a factor of $1/r$ to $\mathbb{P}[\hat{\chi} \mid T, p]$, and all other edges a factor of 1, so we have the following.

LEMMA 3. *If* $\mathrm{p} \in \mathrm{M(T)}$, *then*

$$\mathbb{P}(\mathrm{p}) = \frac{1}{\mathrm{r}^{|\mathrm{E(p)}|+1}} \left|\left\{ \hat{\chi} : \hat{\chi}|_{[\mathrm{n}]} = \chi, \mathrm{E}(\hat{\chi}) \subseteq \mathrm{E(p)} \right\}\right|. \tag{24}$$

By Lemma 3, to calculate $\mathbb{P}(p)$ for $p \in M(T)$, we must count the number of extensions $\hat{\chi}$ of χ for which $E(\hat{\chi}) \subseteq E(p)$. With a view to proving Theorem 1, we would like to show that

$$\left|\left\{ \hat{\chi} : \hat{\chi}|_{[n]} = \chi, E(\hat{\chi}) \subseteq E(p) \right\}\right| \leqslant r^{|E(p)|-l}, \tag{25}$$

with this bound attained by some $p \in M(T)$. Since it may be the case that there are no extensions with $E(\hat{\chi}) \subseteq E(p)$ (this will certainly be the case if $|E(p)| < l$), we make the following definition.

Definitions 4 (χ-viable). $S \subseteq E(T)$ is *χ-viable* or *viable for χ on T* if there is an extension $\hat{\chi}$ of χ such that $E(\hat{\chi}) \subseteq S$.

Let S be viable for χ on $T, \hat{\chi}$ such that $E(\hat{\chi}) \subseteq S$, and put $k = |S|$. Deleting S from T, denoted $T \setminus S$, will divide T into $k+1$ connected components, and $\hat{\chi}$ must be constant on each of these since $E(\hat{\chi}) \subseteq S$. In particular, if v is a vertex belonging to a component containing a leaf i of T (an "external" component), then we must have $\hat{\chi}(v) = \chi(i)$. However, on components that do not contain a leaf of T ("internal" components), $\hat{\chi}$ may take any of the r states in $\mathscr{S}$. Since $\hat{\chi}$ is completely determined by the state of each connected component of $T \setminus S$, it follows that if there are λ internal components then there are precisely r^{λ} extensions of χ such that $E(\hat{\chi}) \subseteq S$. The inequality (25) follows from these arguments and the following lemma.

LEMMA 4. *Let* χ *be an* r-*state character of length* 1 *on* T. *If* S *is viable for* χ *on* T *and* |S| = k, *then* T \ S *has at most* k − 1 *internal components.*

Proof. We may regard $T \setminus S$ as a tree T_S in a natural way by viewing the connected components of $T \setminus S$ as vertices, with two vertices of T_S connected by an edge precisely if the corresponding components are joined by an edge in S. Note, however, that the internal vertices of T_S do not necessarily correspond to the internal components of $T \setminus S$.

Let $T \setminus S$ have s internal components. Since S is viable for χ on T, the value of χ on each external component gives us a partial state function for T_S (see Fig. 3). Assigning a state to each of the remaining s vertices of T_S (and thereby to each internal component of $T \setminus S$) will induce an extension $\hat{\chi}$ of χ on T having the same changing number as the assignment of states to T_S. Do this by rooting T_S arbitrarily at a vertex that has already been assigned a state and directing all edges away from the root. If v has not yet

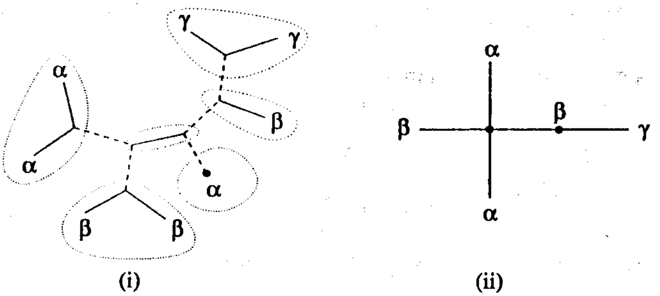

Figure 3. The tree T_S. (i) A tree T and character χ with a set S of χ-viable edges (shown dotted) deleted. The connected components are circled. (ii) The corresponding tree T_S with the partial state function induced by χ. Note that some of the internal vertices correspond to external components of $T \setminus S$.

been assigned a state but its immediate ancestor u has, assign v the same state as u so that there is no change on the edge $\{u, v\}$. Then, for each of the s vertices not assigned a state by χ, we get at least one edge on which no change occurs, so that there are at most $k-s$ changes. But $ch(\hat{\chi}, T) \geqslant l$ so that $k-s \geqslant l$, and hence $s \leqslant k-l$ as required.

Theorem 2 follows from an application of Lemma 4.

COROLLARY 1 (Theorem 2). *Let* χ_1, χ_2 *be minimal extensions of an* r*-state character* χ *on* T. *Then the set of edges* $\mathrm{E}(\chi_1)$ *on which changes occur under* χ_1 *equals the set of edges* $\mathrm{E}(\chi_2)$ *on which changes occur under* χ_2 *if and only if* $\chi_1 = \chi_2$.

Proof. If χ_1 is a minimal extension of χ, then $E(\chi_1)$ is a χ-viable set of cardinality $l(\chi, T)$. Then, by Lemma 4, $T \backslash E(\chi_1)$ has no internal components so that there are exactly $r^0 = 1$ extensions $\hat{\chi}$ of χ such that $E(\hat{\chi}) \subseteq E(\chi_1)$, namely, $\hat{\chi} = \chi_1$. Hence, if $E(\chi_1) = E(\chi_2)$, then $\chi_1 = \chi_2$.

Lemma 4 establishes the inequality (25), proving $\mathbb{P}[\chi \mid T, p] \leqslant r^{-l(\chi, T)-1}$. To complete the proof of Theorem 1, we must exhibit a vector of probabilities p such that $\mathbb{P}(p) = r^{-l-1}$. The vector $p^{\hat{\chi}}$ defined by

$$p^{\hat{\chi}}_{\{u,v\}} = \begin{cases} \dfrac{r-1}{r} & \text{if } \hat{\chi}(u) \neq \hat{\chi}(v) \\ 0 & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \end{cases} \tag{26}$$

is easily seen to be such a vector whenever $\hat{\chi}$ is a minimal extension of χ, and we have our result.

5. Realising the Upper Bound in the Two-State Case.

Having found an upper bound for $\mathbb{P}[\chi \mid T, p]$, it is natural to ask under what circumstances this bound is achieved. Here, we give a partial answer to this question, answering it in the case $r = 2$. If $\hat{\chi}$ is a minimal extension of χ, then $\mathbb{P}[\chi \mid T, p^{\hat{\chi}}] = r^{-l(\chi, T)-1}$, where $p^{\hat{\chi}}$ is as defined above, and for $r = 2$, this turns out to be a complete characterisation of the vectors p maximising $\mathbb{P}[\chi \mid T, p]$ provided the tree has no vertices of degree 2. Where the tree does have a vertex of degree 2 (note that we only allow this to occur at the root), the two edges incident with this vertex behave as a single edge with the path mutation probability

$$p_P = \tfrac{1}{2}(1-(1-2p_1)(1-2p_2)) \tag{27}$$

where p_1 and p_2 are the mutation probabilities on the two edges and P is the path across them (see Fig. 4). The condition then becomes

$$p^{\hat{\chi}}_P = \begin{cases} \tfrac{1}{2} & \text{if a change occurs across } P \text{ under } \hat{\chi} \\ 0 & \text{if no change occurs across } P \text{ under } \hat{\chi}. \end{cases} \tag{28}$$

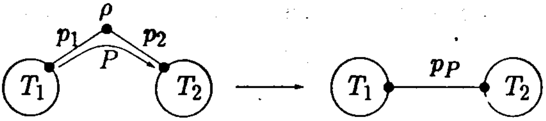

Figure 4. Two edges incident with a root ρ of degree 2 behave as a single edge with the path mutation probability p_P. The circles marked T_1, T_2 denote rooted subtrees.

In terms of the edge parameters, we have $p_P = 0$ if and only if both $p_1 = p_2 = 0$, and $p_P = 1/2$ if and only if at least one of $p_1, p_2 = 1/2$. Thus, although we state the following result only for trees with no vertices of degree 2, it still enables us to characterise the vectors p maximising $\mathbb{P}[\chi \mid T, p]$ in the case T has a root of degree 2.

THEOREM 3. *If χ is a binary character and* T *has no vertices of degree* 2, *then* p *maximises* $\mathbb{P}[\chi \mid \mathrm{T}, \mathrm{p}]$ *if and only if* $\mathrm{p} = \mathrm{p}^{\hat{\chi}}$ *for some minimal extension* $\hat{\chi}$ *of* χ, *where*

$$\mathrm{p}^{\hat{\chi}}_{\{\mathrm{u},\mathrm{v}\}} = \begin{cases} \frac{1}{2} & \text{if } \hat{\chi}(\mathrm{u}) \neq \hat{\chi}(\mathrm{v}) \\ 0 & \text{if } \hat{\chi}(\mathrm{u}) = \hat{\chi}(\mathrm{v}). \end{cases} \tag{29}$$

Proof. The backward direction of Theorem 3 is already established; we prove the forward direction in two stages, first establishing it for binary trees, and then reducing the general case to that where T is binary.

The proof in the binary case is by induction on n, the number of leaves of T. Consider $n = 2$, for which there are two possible characters χ up to permutation, $\chi(1) = \chi(2)$ and $\chi(1) \neq \chi(2)$. Clearly, $\mathbb{P}[\chi \mid T, p]$ is maximised in the first case only if $p_e = 0$, and in the second only if $p_e = 1/2$.

Suppose the result is true for binary trees on $n - 1$ leaves, where $n \geqslant 3$. Let T be a binary tree on n leaves, χ a character of length l on T, and suppose that p is such that $\mathbb{P}[\chi \mid T, p]$ is maximised. Since T is binary, it has a pair of adjacent pendant edges, that is, a pair of edges $\{u, v\}$ and $\{u, v'\}$ such that v and v' are leaves of T. We consider two cases: $\chi(v) = \chi(v')$ and $\chi(v) \neq \chi(v')$.

Case 1. $\chi(v) = \chi(v')$.

Without loss of generality, $\chi(v) = \chi(v') = \alpha$. Let T' be the tree on $n - 1$ leaves obtained by deleting $\{u, v\}$ and $\{u, v'\}$ from T, χ_α the character on the leaves of T' such that χ_α agrees with χ on their common leaves and $\chi_\alpha(u) = \alpha$, and define χ_β similarly. For convenience, put $e = \{u, v\}$, $e' = \{u, v'\}$, and let the vertex w and edge e'' be as shown in Fig. 5.

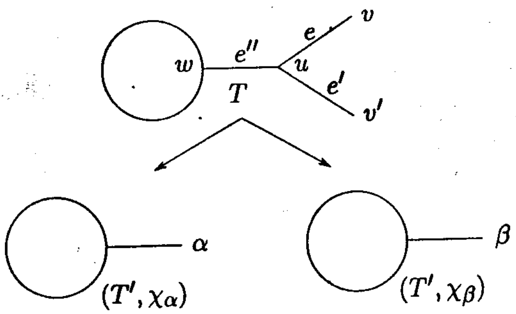

Figure 5. The trees T, T' and characters χ_α, χ_β. The circle denotes a rooted subtree.

Then

$$\mathbb{P}[\chi \mid T, p] = (1-p_e)(1-p_{e'})\mathbb{P}[\chi_\alpha \mid T', p] + p_e p_{e'} \mathbb{P}[\chi_\beta \mid T', p]. \quad (30)$$

Now, if $\hat{\chi}$ is a minimal length extension of χ on T, then $\hat{\chi}(u) = \alpha$; for if $\hat{\chi}(w) = \alpha$, we get no changes on e, e' and e'' if $\hat{\chi}(w) = \alpha$, and one change if $\hat{\chi}(w) = \beta$, while if $\hat{\chi}(u) = \beta$, we get two or three changes depending on whether $\hat{\chi}(w)$ equals α or β. It follows that χ_α has length l on T'.

However, χ_β may have length less than l. For if $\bar{\chi}$ is an extension of χ such that $\bar{\chi}(u) = \beta$, then $\bar{\chi}$ is not a minimal length extension of χ, and so has changing number at least $l+1$. But two of these changes occur on e and e', which are deleted in forming T' and χ_β, so that $ch(\bar{\chi} \mid_{V(T')}) \geqslant l-1$. Hence, χ_β may have length less than l, but the decrease is by at most one.

By Theorem 1, $\mathbb{P}[\chi_\alpha \mid T', p] \leqslant 2^{-l-1}$ and $\mathbb{P}[\chi_\beta \mid T', p] \leqslant 2^{-l}$ so that

$$\begin{aligned}\mathbb{P}[\chi \mid T, p] &\leqslant (1-p_e)(1-p_{e'})2^{-l-1} + p_e p_{e'} 2^{-l} \\ &= 2^{-l-1}((1-p_e)(1-p_{e'}) + 2p_e p_{e'}) \\ &= 2^{-l-1}(1 - p_e - p_{e'} + 3p_e p_{e'}). \qquad (31)\end{aligned}$$

Consider $1 - p_e - p_{e'} + 3p_e p_{e'} = 1 - p_{e'} + p_e(3p_{e'} - 1)$. If $p_e = 0$, then

$$1 - p_e - p_{e'} + 3p_e p_{e'} = 1 - p_{e'} \leqslant 1, \quad (32)$$

with equality if and only if $p_{e'} = 0$. If $p_e > 0$ and $0 \leqslant p_{e'} < 1/3$, then $p_e(3p_{e'} - 1) < 0$ so $1 - p_e - p_{e'} + 3p_e p_{e'} < 1$. Finally, if $1/3 \leqslant p_{e'} \leqslant 1/2$, then $1 - p_{e'} \leqslant 2/3$ and $p_e(3p_{e'} - 1) \leqslant 1/4$ so that

$$1 - p_e - p_{e'} + 3p_e p_{e'} \leqslant \tfrac{2}{3} + \tfrac{1}{4} = \tfrac{11}{12} < 1. \quad (33)$$

Hence, $1 - p_e - p_{e'} + 3p_e p_{e'} \leqslant 1$ with equality if and only if $p_e = p_{e'} = 0$.

Since $\max_p\{\mathbb{P}[\chi \mid T, p]\} = 2^{-l-1}$ and p maximises $\mathbb{P}[\chi \mid T, p]$, we must have $p_e = p_{e'} = 0$. By the induction hypothesis, $\mathbb{P}[\chi_\alpha \mid T', p] = 2^{-l-1}$ if and only if $p = p^{\hat{\chi}_\alpha}$ on T' for a minimal extension $\hat{\chi}_\alpha$ of χ_α. A minimal extension of χ_α extends naturally to a minimal extension of χ and $p_e = p_{e'} = 0$ so that $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T.

Case 2. $\chi(v) \neq \chi(v')$.

Without loss of generality, $\chi(v) = \alpha$ and $\chi(v') = \beta$. Let T', χ_α and χ_β again be as in Fig. 5. If $\hat{\chi}$ is a minimal extension of χ, then $\hat{\chi}$ involves a change on exactly one of e, e' regardless of the state assigned to u, so that $l(\chi_\alpha, T'), l(\chi_\beta, T') \geqslant l - 1$. Hence,

$$\begin{aligned}\mathbb{P}[\chi \mid T, p] &= (1 - p_e)p_{e'}\mathbb{P}[\chi_\alpha \mid T', p] + p_e(1 - p_{e'})\mathbb{P}\big[\chi_\beta \mid T', p\big] \\ &\leqslant 2^{-l}((1 - p_e)p_{e'} + p_e(1 - p_{e'})) \\ &= 2^{-l-1}(1 - (1 - 2p_e)(1 - 2p_{e'})). \end{aligned} \tag{34}$$

Since $1 - (1 - 2p_e)(1 - 2p_{e'}) \leqslant 1$ with equality if and only if at least one of $p_e, p_{e'} = 1/2$, either

1. $p_e = 0$, $p_{e'} = 1/2$ and $\mathbb{P}[\chi_\alpha \mid T', p] = 2^{-l}$;
2. $p_e = 1/2$, $p_{e'} = 0$ and $\mathbb{P}[\chi_\beta \mid T', p] = 2^{-l}$;

or if $p_e p_{e'} \neq 0$, then

3. $\mathbb{P}[\chi_\alpha \mid T', p] = \mathbb{P}[\chi_\beta \mid T', p] = 2^{-l}$ and at least one of $p_e, p_{e'} = 1/2$.

Under the induction hypothesis, 1) and 2) have $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$, so it remains to show that 3) cannot occur. By the induction hypothesis, $\mathbb{P}[\chi_\alpha \mid T', p] = \mathbb{P}[\chi_\beta \mid T', p] = 2^{-l}$ occurs if and only if $E(p) = E(\hat{\chi}_\alpha) = E(\hat{\chi}_\beta)$ for minimal extensions $\hat{\chi}_\alpha$ and $\hat{\chi}_\beta$ of χ_α and χ_β, respectively. Let i be a leaf of T' other than u, and without loss of generality, assume $\chi(i) = \alpha$. Consider the number of changes that occur on the path P from i to u. Since $\hat{\chi}_\alpha(i) = \hat{\chi}_\alpha(u)$, an even number of changes must take place on this path under χ_α; but $\hat{\chi}_\beta(i) \neq \hat{\chi}_\beta(u)$ so that an odd number of changes must take place under χ_β. Hence, $E(\hat{\chi}_\alpha) = E(\hat{\chi}_\beta)$ is not possible, so that 3) cannot occur and the theorem is proved for binary trees.

In order to reduce the general case to that where T is binary, we require an auxiliary theorem.

Definitions 5 (Refinement). T_2 is said to refine T_1 (written $T_1 \leqslant T_2$) if T_1 may be obtained from T_2 by contracting a number of edges.

The order given by $\leqslant$ is a partial order on the set of phylogenetic trees on n leaves, the maximal elements being the binary phylogenetic trees.

some of the internal vertices correspond to

THEOREM 4. *Let* T *be a tree and* χ *a binary character. There is a binary tree* T′ *refining* T *such that* l(χ,T′) = l(χ,T). T′ *may be chosen in such a way that the minimal extensions of* χ *on* T′ *are in a natural bijective correspondence with the minimal extensions of* χ *on* T.

Proof. Let $\mathscr{P}$ be a set of $l = l(\chi, T)$ edge-disjoint paths joining leaves in different states, the existence of which is guaranteed by Lemma 1. Form the sequence $T = T_1 < T_2 < \cdots$ refining T inductively as follows. Given T_i, choose $v \in V(T_i)$ of degree greater than or equal to 4. If there is a path $P \in \mathscr{P}$ passing through v, choose e_1 and e_2 incident with v and lying on P; otherwise, choose e_1 and e_2 incident with v arbitrarily. Create T_{i+1} by inserting a new edge e separating e_1 and e_2 from the remaining edges incident with v. Then $T_i < T_{i+1}$, and no path in $\mathscr{P}$ lies on e so that $\mathscr{P}$ remains edge-disjoint in T_{i+1} (see Fig. 6).

The new edge in T_{i+1} splits v into two vertices, one of degree 3 and one of degree 1 less than that of v, so this process must eventually terminate in a binary tree $T_m = T'$. $\mathscr{P}$ remains edge-disjoint in T', so by Lemma 1, we have $l(\chi, T') \geqslant l$. If $\hat{\chi}$ is a minimal extension of χ on T, then we may obtain an extension $\bar{\chi}$ of χ on T' by identifying each vertex of T' with the vertex of T it was created from during the refinement process and requiring $\hat{\chi}$ and $\bar{\chi}$ to agree under this identification. A change occurs across an edge of T' if and only if it is an edge of T on which a change occurs under $\hat{\chi}$, so that $ch(\bar{\chi}) = ch(\hat{\chi}) = l$, implying $l(\chi, T') \leqslant l$ and hence equality.

Futhermore, every minimal extension of χ on T' arises in this way. Let $\bar{\chi}$ be such an extension. Since each path in $\mathscr{P}$ joins leaves in different states, there must be at least one change on each path. Moreover, $\mathscr{P}$ has cardinality $l(\chi, T')$, so there is exactly one change on each path and no changes on edges not on paths. Since none of the newly created edges lies on any of the paths, $\bar{\chi}$ must be constant on the set of vertices identified with a given vertex v of T, and we obtain a minimal extension $\hat{\chi}$ of χ on T by putting $\hat{\chi}(v)$ equal to this common state.

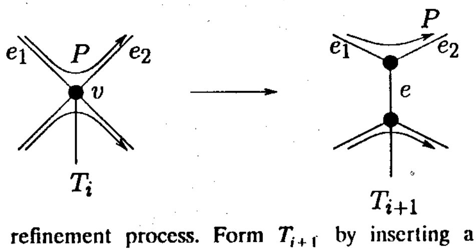

Figure 6. The refinement process. Form T_{i+1} by inserting a new edge e separating e_1 and e_2 from the remaining edges incident with v. None of the paths in $\mathscr{P}$ lies on e, so that $\mathscr{P}$ remains edge disjoint in T_{i+1}. v splits into two vertices, one of degree 3 and the other of degree 1 less than the degree of v.

We now complete the proof of Theorem 3 in the general case.

Let T be a phylogenetic tree, χ a binary character and suppose p maximises $\mathbb{P}[\chi \mid T, p]$. Let T' be a binary tree refining T as constructed in Theorem 4, and put $p'_e = 0$ if e is an edge of T' inserted during refinement, and $p'_e = p_e$ if e is an edge of T. Then

$$\mathbb{P}[\chi \mid T', p'] = \tfrac{1}{2} \sum_{\hat{\chi} : \hat{\chi}|_{[n]} = \chi} \; \prod_{\substack{e=(u,v): \\ \hat{\chi}(u) = \hat{\chi}(v)}} (1 - p'_e) \prod_{\substack{e=(u,v): \\ \hat{\chi}(u) \neq \hat{\chi}(v)}} p'_e . \qquad (35)$$

On newly created edges of T', $p'_e = 0$ so we need only sum over extensions for which no changes occur on newly created edges. Such an extension corresponds to an extension of χ on T, and it follows that

$$\mathbb{P}[\chi \mid T', p'] = \mathbb{P}[\chi \mid T, p] = 2^{-l(\chi, T)-1} = 2^{-l(\chi, T')-1}. \qquad (36)$$

By the result proved for the binary tree case, $p' = p^{\bar{\chi}}$ for a minimal extension $\bar{\chi}$ of χ on T', and it follows from the construction of T' that $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T.

Theorem 3 does not extend to $r \geqslant 3$ as the following counter-examples show. In part, this appears to be because r may be greater than or equal to the maximum degree of the internal vertices of T, making it easy to create an internal component from $T \backslash E(\hat{\chi})$, $\hat{\chi}$ a minimal extension of χ, by deleting a single additional edge. Since phylogenetic trees are assumed to have no vertices of degree 2, this does not occur for binary characters. However, if this requirement is dropped, then Theorem 3 no longer holds, as evidenced by our need to treat a degree 2 root separately.

Example. A counter-example to the extension of Theorem 3 to $r = 3$ is illustrated by the star shaped tree in Fig. 7. We have $l(\chi, T) = 2$ since

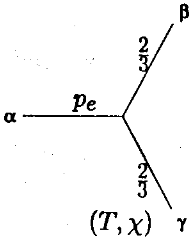

Figure 7. A counter-example to the extension of Theorem 3 to $r = 3$.

$ch(\hat{\chi}) = 2$ for all three possible extensions of χ on T. With p as shown, we have

$$\begin{aligned}\mathbb{P}[\chi \mid T, p] &= \frac{1}{3}\left((1-p_e)\frac{1}{3}\frac{1}{3} + \frac{p_e}{2}\left(1-\frac{2}{3}\right)\frac{1}{3} + \frac{p_e}{2}\frac{1}{3}\left(1-\frac{2}{3}\right)\right) \\ &= \frac{1}{27} - \frac{1}{27}p_e + \frac{1}{54}p_e + \frac{1}{54}p_e \\ &= \frac{1}{27} = 3^{-3}, \end{aligned} \tag{37}$$

so that $\mathbb{P}[\chi \mid T, p] = 3^{-l(\chi, T)-1}$ regardless of the value of p_e.

This example generalises readily to a counter-example for any $r \geqslant 3$ by considering the star-shaped tree on r leaves. This is the tree with vertices $\{0, 1, \ldots, r\}$ and edges $\{\{0,1\}, \{0,2\}, \ldots, \{0,r\}\}$.

6. Applications to Phylogenetic Analysis.

6.1. *Equivalence of maximum parsimony and maximum likelihood with no common mechanism.* Theorem 1 may be used to demonstrate the equivalence of the inference methods of maximum parsimony and maximum likelihood with no common mechanism under the fully symmetric model. By "no common mechanism," we mean that we may choose a different vector of mutation probabilities for each character, rather than requiring all of them to evolve according to a single vector of mutation probabilities, as is usually the case. This approach has the drawback that it is not necessarily statistically consistent; see below.

Penny et al. (1994) state this result for the $r = 2$ case, and related results appear elsewhere (Goldman, 1990; Farris, 1973). For a discussion of various methods of phylogenetic inference, see Swofford *et al.* (1996).

6.1.1. *Maximum parsimony inference.* There are many different methods of parsimony; we consider here only the simplest, Fitch parsimony. This method of inference may be stated as follows.

Given a set $\mathrm{X} = \{\chi_\mathrm{i}\}$ *of* k r-*state characters, choose the unrooted tree or trees* T (*the "maximum parsimony tree*(*s*)") *minimising*

$$\mathrm{l}(\mathrm{X}, \mathrm{T}) = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{l}(\chi_\mathrm{i}, \mathrm{T}). \tag{38}$$

Interpreting $l(\chi_i, T)$ as the minimum number of mutations required for χ_i to evolve on T, $l(X, T)$ is the minimum total number of mutations required for the χ_i to evolve on T. Thus, maximum parsimony chooses the trees on which the χ_i may evolve with as few mutations as possible overall.

6.1.2. *Maximum likelihood inference.* Edwards (1972) defines likelihood as follows.

> The *likelihood*, $\mathbb{L}[H \mid R]$, of the hypothesis H given data R and a specific model, is proportional to $\mathbb{P}[R \mid H]$, the constant of proportionality being arbitrary.

A maximum likelihood method of inference chooses the hypothesis H maximising the likelihood function for the data R. For the fully symmetric model, the hypothesis is the tree and mutation probability vector pair (T, p). The maximum likelihood method is then as follows.

Given a set $\mathrm{X} = \{\chi_\mathrm{i}\}$ *of* k r-*state characters, choose the unrooted tree and vector pair or pairs* (T, p) *maximising*

$$\mathbb{L}[(\mathrm{T},\mathrm{p}) \mid \mathrm{X}] = \mathbb{P}[\mathrm{X} \mid \mathrm{T},\mathrm{p}] = \prod_{\mathrm{i}=1}^{\mathrm{k}} \mathbb{P}[\chi_\mathrm{i} \mid \mathrm{T},\mathrm{p}]. \tag{39}$$

The tree inferred is the "maximum likelihood tree(s)."

This is the usual form of maximum likelihood, and "maximum likelihood" on its own will refer to this form unless stated otherwise. However, by relaxing some assumptions, we may arrive at a number of variations. The mutation probability vector is the part of the model that represents the substitution process and selection mechanism operating at a given site. By requiring this to be the same for each character, we are asserting that an equivalent mechanism is operating at each site, so the characters may be said to evolve under a "common mechanism." By allowing a different vector p for each character, we are allowing different mechanisms to operate at each site, and the characters may be said to evolve with "no common mechanism." In this case, the hypothesis becomes $(T, \{p_i\})$, and the method of *maximum likelihood with no common mechanism* is as follows.

Given a set $\mathrm{X} = \{\chi_\mathrm{i}\}$ *of* k r-*state characters, choose the unrooted tree and vector set pair or pairs* $(\mathrm{T}, \{\mathrm{p}_\mathrm{i}\})$ *maximising*

$$\mathbb{L}[(\mathrm{T},\{\mathrm{p}_\mathrm{i}\}) \mid \mathrm{X}] = \mathbb{P}[\mathrm{X} \mid \mathrm{T},\{\mathrm{p}_\mathrm{i}\}] = \prod_{\mathrm{i}=1}^{\mathrm{k}} \mathbb{P}[\chi_\mathrm{i} \mid \mathrm{T},\mathrm{p}_\mathrm{i}]. \tag{40}$$

Olsen (see Swofford *et al.* (1996, p. 443)) considers a third variation lying between these two, where the ratios of the underlying edge lengths τ_e are kept constant across characters, but the lengths themselves are scaled between characters. In this version, the hypothesis becomes $(T, \tau = (\tau_e)_{e \in E(T)}, \{\lambda_i\})$, and we seek to maximise

$$\mathbb{L}[(T, \tau, \{\lambda_i\}) \mid X] = \prod_{i=1}^{k} \mathbb{P}[\chi_i \mid T, p(\lambda_i \tau)]. \tag{41}$$

Note that in these last two methods, the number of parameters being estimated grows linearly with the number of characters, so the statistical consistency of these two methods is not guaranteed by standard results. Indeed, the former method can be provably statistically inconsistent, by Theorem 5 (with Felsenstein (1978)).

Under the fully symmetric model, we have the following result.

THEOREM 5. *Maximum parsimony and maximum likelihood with no common mechanism are equivalent in the sense that both choose the same tree or trees.*

Proof. The proof is the same as for the $r = 2$ case since it follows directly from Theorem 1. On any given tree T, we have $\max_p \mathbb{P}[\chi_i \mid T, p] = r^{-l(\chi_i, T)-1}$ so that

$$\max_{\{p_i\}} \mathbb{L}[(T, \{p_i\}) \mid X] = \prod_{i=1}^{k} r^{-l(\chi_i, T)-1} = r^{-\sum_{i=1}^{k}(l(\chi_i, T)+1)} = r^{-l(X,T)-k}, \quad (42)$$

and therefore the maximum likelihood with no common mechanism trees are precisely the maximum parsimony trees.

6.2. *Reconstruction of ancestral states.* Given a character χ and a tree T on which χ is assumed to have evolved, a problem of interest is to reconstruct character states at the internal vertices (see, for example, Maddison (1995) and references therein). Here, we consider a simpler problem in which we seek to reconstruct only the state that occurred at a given vertex v. By re-rooting T at v if necessary, we may assume that v is the root ρ with no loss of generality. Under a parsimony approach, the state at ρ may be estimated by the set of states α for which there is a least one minimal extension $\hat{\chi}$ of χ on T with $\hat{\chi}(\rho) = \alpha$. A maximum likelihood method in which the edge parameters are assumed unknown seeks to maximise

$$\mathbb{L}[\hat{\chi}(\rho) = \alpha, p \mid T, \chi] = \mathbb{P}[\chi \mid \hat{\chi}(\rho) = \alpha, T, p] \quad (43)$$

over α and p. We show that under the fully symmetric model, these two approaches agree.

THEOREM 6. *The maximum parsimony and maximum likelihood estimates of the root state agree under the fully symmetric model.*

Proof. Let ρ be incident with the rooted subtrees $T_1, T_2, \ldots, T_k$, and form the trees $T_1', T_2', \ldots, T_k'$ by attaching an additional leaf labelled $n+1$ to the root of each T_i (see Fig. 8). For each $\alpha \in \mathcal{S}$, let $\chi_\alpha^{(i)}$ be the restriction of χ

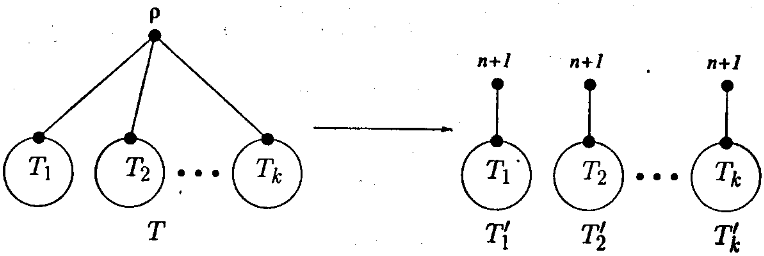

Figure 8. Decompose T into the trees $T_1', \ldots, T_k'$ by joining a leaf $n+1$ to the root of each of the rooted subtrees $T_1, \ldots, T_k$ of T.

to the leaves of T_i, with $\chi_\alpha^{(i)}(n+1) = \alpha$. Then we have

$$\begin{aligned}
\mathbb{L}[\hat{\chi}(\rho) = \alpha, p \mid T, \chi] &= \mathbb{P}[\chi \mid \hat{\chi}(\rho) = \alpha, T, p] \\
&= \sum_{\substack{\hat{\chi} : \hat{\chi}|_{[n]} = \chi \\ \hat{\chi}(\rho) = \alpha}} \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}(u) = \hat{\chi}(v)}} (1 - p_e) \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}(u) \neq \hat{\chi}(v)}} \frac{p_e}{r-1} \\
&= \prod_{i=1}^{k} \left(\sum_{\substack{\hat{\chi}_i : V(T_i') \to \mathscr{S} \\ \hat{\chi}_i|_{[n+1]} = \chi_\alpha^{(i)}}} \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}_i(u) = \hat{\chi}_i(v)}} (1 - p_e) \prod_{\substack{e = \{u,v\}: \\ \hat{\chi}_i(u) \neq \hat{\chi}_i(v)}} \frac{p_e}{r-1} \right) \\
&= \prod_{i=1}^{k} r\mathbb{P}\left[\chi_\alpha^{(i)} \mid T_i', p^{(i)}\right] \qquad (44)
\end{aligned}$$

where $p^{(i)}$ is the restriction of p to $E(T_i')$. By Theorem 1,

$$\max_{p^{(i)}} r\mathbb{P}\left[\chi_\alpha^{(i)} \mid T_i', p^{(i)}\right] = r^{-l(\chi_\alpha^{(i)}, T_i')}, \qquad (45)$$

so that

$$\max_{p} \mathbb{L}[\hat{\chi}(\rho) = \alpha, p \mid T, \chi] = r^{-\sum_{i=1}^{k} l(\chi_\alpha^{(i)}, T_i')}. \qquad (46)$$

But

$$\sum_{i=1}^{k} l\left(\chi_\alpha^{(i)}, T_i'\right) = \min\{ch(\hat{\chi}) : \hat{\chi}(\rho) = \alpha\} \geqslant l(\chi, T), \qquad (47)$$

and hence $\max_p \mathbb{L}[\hat{\chi}(\rho) = \alpha, p \mid T, \chi] \leqslant r^{-l(\chi, T)}$, with equality if and only if there is a minimal extension $\hat{\chi}$ of χ on T such that $\hat{\chi}(\rho) = \alpha$. The result follows.

6.3. *The maximum likelihood point is not unique.* Maximum likelihood algorithms using a hill-climbing method to maximise over the edge parameters on a given tree are effective in locating a maximal point of the likelihood function. The question then arises as to whether the maximal point found is a global or only a local maximum. Fukami and Tateno (1989) claimed to have answered this question for one-parameter models of substitution by showing that, under a certain condition, the likelihood function has a unique maximal point. Steel (1994) gave a simple counter-example to this claim, using a tree on four leaves for which the likelihood function had two extrema at widely separated points. More recently, Tillier (1994) claimed to give a sufficient condition for a multi-parameter model to have a single maximum. However, this condition is satisfied by the fully symmetric model, so Steel's counter-example applies in this case also. The results in this paper may be used to construct further counter-examples.

We have seen that $p^{\hat{\chi}}$ as defined in equation (26) maximises $\mathbb{P}[\chi \mid T, p]$ whenever $\hat{\chi}$ is a minimal extension of χ. By Theorem 2, these vectors are distinct, so that there are at least as many vectors maximising $\mathbb{P}[\chi \mid T, p]$ as there are minimal extensions of χ on T (by Theorem 3, exactly as many when $r = 2$). Since a character may have many minimal extensions on a given tree (Steel (1993a) constructs a tree and character pair on $2n$ leaves having a number of minimal extensions equal to the nth Fibonacci number, so that the number of minimal extensions may, in fact, grow exponentially with n), data consisting only of many copies of such a character will also have multiple optima on that tree.

6.4. *Many constant characters implies maximum likelihood equals maximum parsimony.* In this section, we demonstrate a further connection between the methods of maximum parsimony and maximum likelihood under the fully symmetric model. It has recently been suggested that the existence, in some sequences, of large numbers of sites which are invariant (unable to undergo a mutation that will fix in the population) for functional or structural reasons can mislead phylogeny reconstruction using maximum likelihood; see Lockhart *et al.* (1996). In particular, Lockhart *et al.* and Chang (1996) showed that, for a tree on four species, applying the maximum likelihood method to all sites (under the incorrect assumption that they were equally free to undergo mutation) could select the (incorrect) maximum parsimony tree. Here, we show that this will always happen for *any* input on *any* number of taxa—that is, if one adjoins a number k of constant sites to any data, and applies the maximum likelihood method under the assumption that the sites are independent and identically distributed according to the fully symmetric model, then one will necessarily select a maximum parsimony tree if k is sufficiently large (dependent on the input data and n). It follows that maximum likelihood under this model and assuming all sites are free to vary will select an incorrect tree on any dataset for which maximum parsimony selects an incorrect tree if the dataset contains a sufficiently large number of invariant sites.

Intuitively, we may understand this result as follows. Adding constant characters will lower the maximum likelihood estimates of the mutation probabilities, and, as has been argued elsewhere (for example, Felsenstein (1981)), the contributions from minimal extensions become the dominant terms in $\mathbb{P}[\chi \mid T, p]$ as the mutation probabilities tend to zero with their ratios bounded.

Given data $X = \{\chi_i\}$ consisting of r-state characters, for each non-constant character χ, let n_χ be the number of times χ occurs in X, and let n_0 be the number of constant characters in X (we do not distinguish between the r different constant characters). For a given tree T and mutation probability vector p, let $f_\chi = f_\chi(T, p) = \mathbb{P}[\chi \mid T, p]$ for each $\chi : [n] \mapsto \mathcal{S}$, and let $f_0 = f_0(T, p) = \mathbb{P}[\chi \text{ is constant} \mid T, p] = r\mathbb{P}[\chi(i) = \alpha \ \forall i \in [n] \mid T, p]$. Then the likelihood of (T, p) given the data X can be written

$$\mathbb{L}[(T,p) \mid X] = f_0^{n_0} \prod_{\chi \neq 0} f_\chi^{n_\chi} \tag{48}$$

where we have used 0 to denote the constant characters. Now let

$$\Phi(p,k) = -\ln\left(f_0^{n_0+k} \prod_{\chi \neq 0} f_\chi^{n_\chi}\right), \tag{49}$$

that is, Φ is the minus log-likelihood of (T, p) given X with k additional constant characters. Then we have the following.

LEMMA 5. *If* $\mathrm{p_e = 1/k \ \forall e \in E(T)}$, *then*

$$\lim_{k \to \infty} \frac{\Phi(\mathrm{p,k})}{\ln \mathrm{k}} = \sum_{\chi} \mathrm{n}_\chi \mathrm{l}(\chi, \mathrm{T}) = \mathrm{l(X,T)}. \tag{50}$$

Proof. We have

$$\Phi(p,k) = -(n_0 + k)\ln f_0 - \sum_{\chi \neq 0} n_\chi \ln f_\chi. \tag{51}$$

If $\bar{\chi}$ is a minimal length extension of χ on T, then

$$\begin{aligned} f_\chi &\geq \mathbb{P}[\bar{\chi} \mid T, p] \\ &= \frac{1}{r} \frac{1}{((r-1)k)^{l(\chi,T)}} \left(1 - \frac{1}{k}\right)^{|E(T)| - l(\chi,T)} \end{aligned}$$

$$
\begin{aligned}
&\geqslant \frac{1}{r}\,\frac{1}{((r-1)k)^{l(\chi,T)}}\left(1-\frac{1}{k}\right)^{|E(T)|}\\
&\geqslant \frac{1}{2r}\,\frac{1}{((r-1)k)^{l(\chi,T)}} \qquad (52)
\end{aligned}
$$

for k sufficiently large that $(1-1/k)^{|E(T)|} > 1/2$. Also,

$$
\begin{aligned}
f_\chi &= \frac{1}{r}\sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi}\frac{1}{((r-1)k)^{ch(\hat{\chi})}}\left(1-\frac{1}{k}\right)^{|E(T)|-ch(\hat{\chi})}\\
&\leqslant \frac{1}{r}\sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi}\frac{1}{((r-1)k)^{l(\chi,T)}}\\
&= r^{|V(T)|-n-1}\frac{1}{((r-1)k)^{l(\chi,T)}} \qquad (53)
\end{aligned}
$$

so that for k sufficiently large,

$$
\frac{c_1}{((r-1)k)^{l(\chi,T)}} \leqslant f_\chi \leqslant \frac{c_2}{((r-1)k)^{l(\chi,T)}} \qquad (54)
$$

for constants c_1, c_2 dependent only on r, n and $|V(T)|$.

Now, if S is the event that every vertex of T is in the same state, then $1 \geqslant f_0 \geqslant \mathbb{P}[S \mid T, p] = (1-1/k)^{|E(T)|}$. By Taylor's Theorem,

$$
(1-x)^E = 1 - Ex + \tfrac{1}{2}E(E-1)(1-c)^{E-2}x^2 \qquad (55)
$$

for some c between 0 and x, so that

$$
1 - \frac{E}{k} \leqslant f_0 \leqslant 1. \qquad (56)
$$

Again, by Taylor's Theorem,

$$
\ln(1-x) = -x - \frac{1}{2(1-c)^2}x^2 \qquad (57)
$$

for some c between 0 and x, so for $k > 2|E(T)|$, we have

$$
0 \leqslant -\ln f_0 \leqslant \frac{|E(T)|}{k} + \frac{2|E(T)|^2}{k^2}. \qquad (58)
$$

Since

$$-\sum_{\chi\neq 0} n_\chi \ln \frac{c_i}{((r-1)k)^{l(\chi,T)}} = \sum_{\chi\neq 0} n_\chi l(\chi,T)\ln k - \sum_{\chi\neq 0} n_\chi \ln \frac{c_i}{(r-1)^{l(\chi,T)}}$$

$$= l(X,T)\ln k + C_i, \tag{59}$$

we have, from (51), (54), (58) and (59),

$$l(X,T)\ln k + C_2$$

$$\leqslant \Phi(p,k) \leqslant l(X,T)\ln k + C_1 + (n_0+k)\left(\frac{|E(T)|}{k} + \frac{2|E(T)|^2}{k^2}\right), \tag{60}$$

and on dividing by $\ln k$ and letting $k \to \infty$, we obtain the result.

Now, let $p^* = p^*(k,T)$ be the vector of mutation probabilities minimising $\Phi(p,k)$, and therefore maximising the likelihood of T given X with k additional constant characters. Then we have the following.

LEMMA 6.

$$\lim_{\mathrm{k}\to\infty} \frac{\Phi(\mathrm{p}^*,\mathrm{k})}{\ln \mathrm{k}} = \mathrm{l}(\mathrm{X},\mathrm{T}). \tag{61}$$

Proof. In view of Lemma 5, it suffices to show that

$$\lim_{k\to\infty} \frac{\Phi(p^*,k)}{\ln k} \geqslant l(X,T). \tag{62}$$

Let $p(k) = \max_{e\in E(T)} p_e^*$. Arguing as in Lemma 5, for $\chi \neq 0$, we have

$$f_\chi(p^*) \leqslant \frac{1}{r} \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \left(\frac{p(k)}{r-1}\right)^{ch(\hat{\chi})}$$

$$\leqslant \frac{1}{r} \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \left(\frac{p(k)}{r-1}\right)^{l(\chi,T)}$$

$$= r^{|V(T)|-n-1}\left(\frac{p(k)}{r-1}\right)^{l(\chi,T)} \tag{63}$$

Also,

$$f_0(p^*) \leqslant 1 - p(k) \tag{64}$$

since if leaves i and j are separated by an edge e_0 for which $p^*_{e_0} = p(k)$, then by inequality (11), the event E that leaves i and j are in the same state has probability less than or equal to $1 - p_{e_0}$, and $f_0(p^*) \leqslant \mathbb{P}[E]$.

Thus

$$\begin{aligned}\Phi(p^*,k) &\geqslant -(n_0+k)\ln(1-p(k)) - \sum_{\chi\neq 0} n_\chi \ln\left(r^{|V(T)|-n-1}\left(\frac{p(k)}{r-1}\right)^{l(\chi,T)}\right)\\ &\geqslant kp(k) - l(X,T)\ln p(k) + c \end{aligned} \tag{65}$$

where we have used $-\ln(1-p) \geqslant p, n_0 + k \geqslant k$, and c does not depend on k or $p(k)$. Now, minimise $h = kp - l(X,T)\ln p$ as a function of p. Setting

$$\frac{\partial h}{\partial p} = k - \frac{l(X,T)}{p} = 0 \tag{66}$$

we obtain $p = l(X,T)/k$, and the second derivative condition $\partial^2 h/\partial p^2 = l(X,T)/p^2 \geqslant 0$ shows that this is a minimum. Hence,

$$\Phi(p^*,k) \geqslant l(X,T) - l(X,T)\ln l(X,T) + l(X,T)\ln k + c, \tag{67}$$

and so

$$\lim_{k\to\infty} \frac{\Phi(p^*,k)}{\ln k} \geqslant l(X,T) \tag{68}$$

as claimed.

We are now in a position to prove the following.

THEOREM 7. *For data containing enough constant characters, the maximum likelihood tree under the fully symmetric model is a maximum parsimony tree.*

Proof. Write X_k for data X with an additional k constant characters, and put

$$\Phi_T(k) = -\ln \mathbb{L}[(T, p^*(k,T)) \mid X_k]. \tag{69}$$

By Lemma 6, there is K such that for $k \geqslant K$ and all trees T,

$$\left| \frac{\Phi_T(k)}{\ln k} - l(X,T) \right| < \frac{1}{2}. \tag{70}$$

It follows that whenever $k \geqslant K$ and $l(X,T_1) < l(X,T_2)$, then $\Phi_{T_1}(k) < \Phi_{T_2}(k)$, so that maximum likelihood given data X with K or more additional constant characters will choose a maximum parsimony tree.

7. Summary. We have demonstrated a relationship (Theorem 1) between the maximal probability of generating any given character on a tree T under a simple model, and the parsimony score of that character on T. In addition, we have characterised (Theorem 3) when this maximal probability will be realised in terms of the underlying edge parameters in the case of two states. We then derived some new results involving the maximum likelihood and maximum parsimony methods in phylogenetic analysis.

Possible future work would include a complete classification of the vectors which maximise $\mathbb{P}[\chi \mid T, p]$ for r-states (thereby generalising Theorem 3), although, clearly, such a classification must allow for the extra complication of a continuum of solutions, as the example at the end of Section 5 shows. Regarding the non-uniqueness of the maximum likelihood point, it would be interesting to construct multiple local optima strictly within the interior of the space of edge parameters.

It would also be useful to see if Theorem 7 could be extended so that the number of constant sites that need to be added to a set of characters in order to force maximum likelihood to return the same tree as maximum parsimony can be bounded above by a polynomial in n and $|X|$ (the number of characters). If so, it would follow that finding a maximum likelihood tree for character data, under the model described here, is an NP-hard problem.

Thanks go to Péter Erdős and László Székely for suggesting a simpler proof of Lemma 4, and to David Penny for conjecturing Theorem 7 while at the "Moir of Ord," Arthur's Pass. We also thank Joe Felsenstein for his helpful comments.

REFERENCES

Cavender, J. A. 1978. Taxonomy with confidence. *Mathematical Biosci.* **40**, 270–280.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosci.* **134**, 189–215.

Edwards, A. W. F. 1972. *Likelihood.* Cambridge: Cambridge University Press.

Erdős, P. L. and L. A. Székely. 1993. Counting bichromatic evolutionary trees. *Discrete Appl. Math.* **47**, 1–8.

Erdős, P. L. and L. A. Székely. 1994. On weighted multiway cuts in trees. *Mathematical Programming* **65**, 93–105.

Farris, J. S. 1973. A probability model for inferring evolutionary trees. *Systematic Zoology* **22**, 250–256.

Felsenstein, J. 1978. Cases in which parsimony or compatibility will be positively misleading. *Systematic Zoology* **27**, 401–410.

Felsenstein, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological J. Linnean Soc.* **16**, 183–196.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.

Fukami, K. and Y. Tateno. 1989. On the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point. *J. Molecular Evolution* **28**, 460–464.

Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology* **39**, 345–361.

Harary, F. 1969. *Graph Theory. Series in Mathematics*. Reading, MA: Addison-Wesley.

Hendy, M. D. and D. Penny. 1989. A framework for the qualitative study of evolutionary trees. *Systematic Zoology* **38**, 297–309.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro (Ed), pp. 21–132. New York: Academic Press.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93**, 1930–1934.

Maddison, W. P. 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Systematic Biol.* **44**, 474–481.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, S. S. Gupta and J. Yackel (Eds), pp. 1–27. New York: Academic Press.

Penny, D., P. J. Lockhart, M. A. Steel and M. D. Hendy. 1994. The role of models in reconstructing evolutionary trees. In *Models in Phylogeny Reconstruction*, R. W. Scotland, D. J. Siebert and D. M. Williams (Eds), Systematics Association Special Vol. 52, pp. 211–230. Oxford: Clarendon Press.

Steel, M. A. 1993a. Decompositions of leaf-colored binary trees. *Advances in Appl. Math.* **14**, 1–24.

Steel, M. A. 1993b. Distributions on bicoloured binary trees arising from the principle of parsimony. *Discrete Appl. Math.* **41**, 245–261.

Steel, M. A. 1994. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biol.* **43**, 560–564.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. 1996. Phylogenetic inference. In *Molecular Systematics*, 2nd ed., D. M. Hillis, C. Moritz and B. K. Marble (Eds), ch. 11, pp. 407–514. Sinauer Associates.

Tillier, E. R. M. 1994. Maximum likelihood with multiparameter models of substitution. *J. Molecular Evolution* **39**, 409–417.

Received 26 June 1996
Revised version accepted 11 December 1996