Journal of Theoretical Biology 280 (2011) 159–166

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# Estimating phylogenetic trees from pairwise likelihoods and posterior probabilities of substitution counts

Mark T. Holder[a,*], Mike Steel[b]

[a] Department of Ecology and Evolutionary Biology, University of Kansas, 1200 Sunnyside Ave, Lawrence KS 66045, United States
[b] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

*Article history:*
Received 14 September 2010
Received in revised form
20 February 2011
Accepted 8 April 2011
Available online 17 April 2011

*Keywords:*
Phylogenetic inference
Distance-based phylogenetic inference
Counting processes
Maximum likelihood
Consistency

ABSTRACT

The field of phylogenetic tree estimation has been dominated by three broad classes of methods: distance-based approaches, parsimony and likelihood-based methods (including maximum likelihood (ML) and Bayesian approaches). Here we introduce two new approaches to tree inference: pairwise likelihood estimation and a distance-based method that estimates the number of substitutions along the paths through the tree. Our results include the derivation of the formulae for the probability that two leaves will be identical at a site given a number of substitutions along the path connecting them. We also derive the posterior probability of the number of substitutions along a path between two sequences. The calculations for the posterior probabilities are exact for group-based, symmetric models of character evolution, but are only approximate for more general models.

## 1. Introduction

The traditional approach to estimating a tree from the distances between leaves entails calculating a corrected distance estimate and then searching for a tree that displays path lengths between its leaves that are as close as possible to the corrected distances. Typically, maximum likelihood estimates of branch lengths are used as the distance corrections in the first step. Some form of weighted least squares (WLS) is often used for the second step (Fitch and Margoliash, 1967; Beyer et al., 1974). For example, if we have $M$ species then we may search for the tree that minimizes the weighted sum of squared errors, $E(T,\mu)$:

$$E(T,\mu)=\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} w_{ij}(D_{ij}-\mu_{ij})^2, \tag{1}$$

where $w_{ij}$ is a (non-negative) weight for the pairwise comparison between leaves $i$ and $j$, $D_{ij}$ is the estimate of the distance between leaf $i$ and $j$ (calculated from the data), and $\mu_{ij}$ is the path length between leaves $i$ and $j$ on the tree. $\mu_{ij}$ is simply the sum of the lengths of all the edges along the shortest path from leaf $i$ to leaf $j$ on the tree. See Chapter 11 of Felsenstein (2004) for a more detailed discussion of distance-based tree inference. This approach to inferring trees has a few obvious weaknesses (all of which have been pointed out by other workers):

1. each pairwise comparison of sequences in Eq. (1) is treated as if it were an independent observation that should contribute to the overall score;
2. the most appropriate weighting coefficient, $w_{ij}$, may not be obvious; and
3. pairwise comparisons do not use all of the relevant information because they do not enforce logical constraints that are implicit in an evolutionary reconstruction.

### 1.1. Non-independence of pairwise comparisons

As with most methods based on pairwise comparison of sequences, the two approaches that we introduce here do not address the first objection. That is, they ignore the statistical dependencies that arise because events that occur along an edge in the tree will affect all pairwise combinations that traverse that edge. Generalized least squares (GLS) methods, which account for covariances caused by shared history, have been developed by Hasegawa et al. (1985) and Bulmer (1991); interested readers should refer to Bryant and Waddell (1998) for an efficient algorithm for using GLS for tree inference, and to Susko (2003) and Sanjuan and Wrobel (2005) for discussions of topology testing in the context of GLS. Our assumption that each pairwise comparison is independent simplifies

[*] Corresponding author. Tel.: +1 7858645789.
*E-mail addresses:* mtholder@ku.edu (M.T. Holder), m.steel@math.canterbury.ac.nz (M. Steel).
*URLS:* http://phylo.bio.ku.edu/mark-holder (M.T. Holder), http://www.math.canterbury.ac.nz/~m.steel (M. Steel).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.04.005

the scoring of trees and it makes the methods that we introduce more directly comparable to the widely used WLS methods. Czarna et al. (2006) report that the theoretical advantages of the GLS approach over WLS may not translate to improvements in the domain of topology testing (at least, not with the implementations of GLS which are currently available).

### 1.2. Downweighting comparison involving large distances

A variety of systems have been proposed for weighting the terms of the least squares equation (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967; Beyer et al., 1974). Typically, lower weights are assigned to comparisons that have higher variances. Because the variance of a distance correction increases with the distance, these weighting coefficients (the $w_{ij}$ factors in Eq. (1)) are small for large distances. Here, we introduce an estimation scheme that uses the likelihood calculated from pairwise comparisons of leaves.

### 1.3. Enforcing constraints

Character-based approaches (likelihood and parsimony) calculate a score for a tree while enforcing the constraints that each site must have a valid substitutional history – a history which assigns a single state at each internal node. Distance-based approaches do not consider ancestral sequences and do not enforce any constraints on valid substitutional histories. The constraint that the internal nodes of the tree must be assigned a single state represents a source of information about the histories of the sequences that is not being used by distance methods. We can enforce some logical constraints about the paths between leaves by:

1. treating the number of substitutions (rather than simply the *expected* number of substitutions) as a parameter to be estimated, and
2. only considering numbers of substitutions along a path that are compatible with the sequences observed at the ends of the path.

Here we will develop an approach to tree estimation that enforces these conditions. The approach can be cast as a hierarchical model: the edges of the tree have an (unknown) expected number of substitutions, $\mu$. Even if we know the correct value of $\mu$, the actual number of substitutions is still a parameter to be estimated. Thus, the edge lengths commonly used in distance-based phylogenetics become hyper-parameters in our formulation.

## 2. Results

### 2.1. Pairwise maximum likelihood

The use of WLS scoring (Eq. (1)) for tree estimation can be justified if the errors in distance estimates are expected to be drawn from a normal distribution that has a mean of 0 and a variance of $w_{ij}$. Susko (2003) demonstrated that this assumption of normality was valid in the context for distances estimated from long sequences. When the data consist of short sequences, it may be more appropriate to use a likelihood function to evaluate path length, rather than assuming a normal distribution as an error model for our estimates. We can define a likelihood-based score, $S$, for a tree and edge lengths, $\mu$, as:

$$S(T,\mu)=\sum_{i=1}^{M-1}\sum_{j=1+i}^{M}\ln[\mathbb{P}(X_i,X_j|T,\mu_{ij})], \tag{2}$$

where $X_i$ and $X_j$ denote the character data for leaves $i$ and $j$, respectively, and $\mu_{ij}$ is the sum of edge lengths on the path from $i$ to $j$. Thus, $\mu_{ij}$ is the expected number of substitutions on the path in $T$ connecting leaves $i$ and $j$. In a later section, we will consider the actual number of substitutions on this path, which we will denote by $\xi_{ij}$ to avoid confusion with $\mu_{ij}$. If $\mathcal{P}(T,i,j)$ is the set of edges that are traversed when moving from $i$ to $j$ on tree $T$ and $\mu_e$ is the length of edge $e$, then $\mu_{ij}=\sum_{e\in\mathcal{P}(T,i,j)}\mu_e$. The tree and edge lengths with the highest score would be preferred. Eq. (2) can be factored:

$$S(T,\mu)=\sum_{i=1}^{M-1}\sum_{j=1+i}^{M}\ln[\mathbb{P}(X_i)\mathbb{P}(X_j|X_i,\mu_{ij})], \tag{3}$$

where one leaf (indexed by $i$) is arbitrarily chosen to serve as the ancestral sequence for each comparison (when using time reversible models, the score will not change if the directionality of the edge is reversed). If we assume that the $n$ sites involved in each comparison evolve independently, then we have:

$$S(T,\mu)=\sum_{i=1}^{M-1}\sum_{j=1+i}^{M}\sum_{k=1}^{n}\ln[\mathbb{P}(X_{i,k})\mathbb{P}(X_{j,k}|X_{i,k},\mu_{ij}^{k})], \tag{4}$$

where $X_{i,k}$ denotes the sequence for leaf $i$ at site $k$, and $\mu_{ij}^{k}$ is the expected number of substitutions at site $k$ on the path connecting leaves $i$ and $j$.

If the model of sequence evolution is assumed to be identical for all sites (so $\mu_{ij}^{k}=\mu_{ij}$) then the dataset can be compressed by counting the number of times each unique pattern of states is observed in each pair of leaves. For a nucleotide model, only 16 possible patterns are possible. Compressing redundant patterns can be performed prior to tree searching; therefore the calculation of the scores for trees encountered during a search will be independent of the sequence length:

$$S(T,\mu)=\sum_{i=1}^{M-1}\sum_{j=1+i}^{M}\sum_{p=1}^{16}c_p\ln[\mathbb{P}(X_{i,p})\mathbb{P}(X_{j,p}|X_{i,p},\mu_{ij})], \tag{5}$$

where $c_p$ is the number of times the data pattern $p$ was observed for the pairwise comparison of taxa $i$ and $j$.

The use of the pairwise log-likelihood parameterized by the path length separating two leaves is inspired by the WLS error formulation. While the definition of the pairwise likelihood score (Eq. (2)) does not explicitly use a weighting coefficient to downweight long paths, the use of the likelihood function will naturally accommodate our intuition that the sequence comparisons associated with long paths are less reliable. Plotting $\mathbb{P}(X_{j,k}|X_{i,k},\mu_{ij})$ versus $\mu_{ij}$ for commonly used Markov models of sequence evolution reveals the familiar plateauing of the transition probability that is associated with saturation caused by multiple hits. The flattening of this curve illustrates that small changes to a path length will have a relatively small effect on the likelihood when the path is already long.

If all of the columns in the data matrix were scored for exactly two leaves, no column of data would be used in more than one pairwise comparison. In this case, the pairwise likelihood approach would be identical to a standard ML estimation procedure. When data columns are scored for more than two leaves, then the pairwise likelihood scoring system is effectively replicating the data to produce an expanded matrix (as shown in Fig. 1).

This duplication of data will not compromise the consistency of the method, but it does mean that the score calculated in the pairwise likelihood method cannot be treated as a valid likelihood for the purposes of likelihood ratio testing. The fact that this pairwise ML will be a consistent estimator of the tree and edge lengths is formalized in the following result, the proof of which is given in Appendix. Recall that a phylogenetic $\mathcal{L}$-tree is a phylogenetic tree with a leaf (taxon) set $\mathcal{L}$.

**Theorem 1.** *Consider a reversible Markov (or mixed-Markov) model of site evolution on r states for which the evolutionary distance* ($\mu_{ij}$) *between any pair* $(i,j)$ *of taxa can be uniquely determined from the* a

| Leaf | Sequence |
|---|---|
| 1 | $x_{1,1}x_{1,2}\ldots x_{1,n}$ |
| 2 | $x_{2,1}x_{2,2}\ldots x_{2,n}$ |
| 3 | $x_{3,1}x_{3,2}\ldots x_{3,n}$ |
| 4 | $x_{4,1}x_{4,2}\ldots x_{4,n}$ |

b

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | $x_{1,1}x_{1,2}\ldots x_{1,n}$ | $x_{1,1}x_{1,2}\ldots x_{1,n}$ | $x_{1,1}x_{1,2}\ldots x_{1,n}$ | ? | ? | ? |
| 2 | $x_{2,1}x_{2,2}\ldots x_{2,n}$ | ? | ? | $x_{2,1}x_{2,2}\ldots x_{2,n}$ | $x_{2,1}x_{2,2}\ldots x_{2,n}$ | ? |
| 3 | ? | $x_{3,1}x_{3,2}\ldots x_{3,n}$ | ? | $x_{3,1}x_{3,2}\ldots x_{3,n}$ | ? | $x_{3,1}x_{3,2}\ldots x_{3,n}$ |
| 4 | ? | ? | $x_{4,1}x_{4,2}\ldots x_{4,n}$ | ? | $x_{4,1}x_{4,2}\ldots x_{4,n}$ | $x_{4,1}x_{4,2}\ldots x_{4,n}$ |

**Fig. 1.** A depiction of the data duplication that is implicitly performed by the pairwise likelihood scoring system. (a) A depiction of the observed data; (b) the pairwise form of the data matrix in which each sequence is duplicated $M$ times. The '?' symbol denotes a set of $n$ cells of missing data.

*$r\times r$ joint probability distribution on pairs of states for the two taxa. Suppose $k$ sites evolve i.i.d. under this model on a binary (fully resolved) phylogenetic $\mathcal{L}$-tree $T_0$ with edge length assignment $\mu_0$ that is strictly positive and finite on every edge. Then a procedure that returns a phylogenetic $\mathcal{L}$-tree $\hat{T}_k$ and (possibly zero or infinite) edge length assignment $\hat{\mu}_k$ that maximizes the likelihood-based score defined by Eq. (2) has the property that, with probability1, as $k$ tends to infinity, $\hat{T}_k=T_0$ and $\hat{\mu}_k$ converges to $\mu_0$.*

### 2.2. Calculating the posterior probability of the number of substitutions

The pairwise likelihood criterion outlined in the previous section does not address the concern that pairwise methods do not enforce constraints on the number of substitutions that occurred over a tree. We can address this point by estimating the number of substitutions that have occurred along an edge. We will begin by calculating the posterior probability of the number of substitutions across an edge.

Suppose we have two aligned sequences, each of length $n$, which are separated on an evolutionary tree by a path for which the expected (prior) number of substitutions is $\mu$. We observe that $k$ of the $n$ sites are different in the two sequences. Let $P_{n,k}(j)$ denote the posterior probability that the actual number of substitutions that occurred on the path separating the two sequences was $j$. Clearly:

$$P_{n,k}(j)=0, \tag{6}$$

if $j<k$.

First, consider the very simple case where $n=1$ (i.e. a sequence of length 1), in which case $k=1$ or 0. So $P_{1,1}(j)$ is the posterior probability that $j$ changes have occurred along the path at the site, given that the site has undergone a net change of state between the two ends of the path.

**Theorem 2.** *For the Jukes–Cantor model, the posterior probabilities $P_{n,k}(j)$ for $n=1$ and $j\geq 0$ are given by:*

$$\text{(i)}\ P_{1,0}(j)=\left(\frac{e^{-\mu}\mu^j}{j!}\right)\cdot\left(\frac{1+3(-1/3)^j)}{1+3e^{-4\mu/3}}\right), \tag{7}$$

$$\text{(ii)}\ P_{1,1}(j)=\left(\frac{e^{-\mu}\mu^j}{j!}\right)\cdot\left(\frac{1-(-1/3)^j}{1-e^{-4\mu/3}}\right). \tag{8}$$

Note that $P_{1,1}(j)=0$ for $j=0$ (by Eq. (6)) and $P_{1,0}(j)=0$ for $j=1$ (since if exactly one change has occurred along the path, the site cannot be in the same state).

**Proof of Theorem 2.** Let $N$ be the number of substitutions on the path connecting the two taxa, and let $\mathcal{E}_0$ and $\mathcal{E}_1$ be the events that the two states at the end points are the same or different, respectively. Then for $k=0,1$, $P_{1,k}(j)$ is the conditional probability $\mathbb{P}(N=j|\mathcal{E}_k)$. Bayes' formula gives:

$$\mathbb{P}(N=j|\mathcal{E}_k)=\frac{\mathbb{P}(\mathcal{E}_k|N=j)\mathbb{P}(N=j)}{\mathbb{P}(\mathcal{E}_k)}. \tag{9}$$

In the Jukes–Cantor model $N$ has a Poisson distribution and so $\mathbb{P}(N=j)=e^{-\mu}\mu^j/j!$. Moreover, in the Jukes–Cantor model, $\mathbb{P}(\mathcal{E}_0)=\frac{1}{4}(1+3e^{-4\mu/3})$ and $\mathbb{P}(\mathcal{E}_1)=\frac{3}{4}(1-e^{-4\mu/3})$. The remaining term in Eq. (9) to be determined is $\mathbb{P}(\mathcal{E}_k|N=j)$. Clearly, $\mathbb{P}(\mathcal{E}_1|N=j)=1-\mathbb{P}(\mathcal{E}_0|N=j)$ so it suffices to determine $\mathbb{P}(\mathcal{E}_0|N=j)$. In the Jukes–Cantor model, this is simply $(\frac{1}{3})^j$ times the number of walks of length $j$ from a state to itself (regarding the states as vertices of a complete graph on four vertices). The number of such walks is easily shown to be $\frac{1}{4}(3^j+3(-1)^j)$ by algebraic or recursive arguments (see, for example, Barry, 2007). Thus:

$$\mathbb{P}(\mathcal{E}_0|N=j)=\tfrac{1}{4}(1+3(-1/3)^j), \tag{10}$$

$$\mathbb{P}(\mathcal{E}_1|N=j)=\tfrac{3}{4}(1-(-1/3)^j). \tag{11}$$

The formulae in Theorem 1 now follow by substituting the various quantities into Eq. (9). □

#### 2.2.1. Extensions to other numbers of states

The results extend easily to the symmetric Poisson model on any number $r\geq 2$ of states by replacing 4 by $r$ and 3 by $r-1$ above. Thus, the analog of Theorem 2 for $r$ states is

$$P_{1,0}(j)=\left(\frac{e^{-\mu}\mu^j}{j!}\right)\cdot\left(\frac{1+(r-1)(-1/(r-1))^j)}{1+(r-1)e^{-r\mu/(r-1)}}\right);$$

$$P_{1,1}(j)=\left(\frac{e^{-\mu}\mu^j}{j!}\right)\cdot\left(\frac{1-(-1/(r-1))^j}{1-e^{-r\mu/(r-1)}}\right).$$

#### 2.2.2. Extension to more general models

Consider a general time reversible (GTR) Markov process with the rate matrix $Q$. For a path along which the expected (prior) number of substitutions is $\mu$ and, given a pair of states $\alpha$ and $\beta$ (where we may also let $\alpha=\beta$), let $\mathcal{E}_{\alpha\beta}$ be the event that we observe states $\alpha$ and $\beta$ at the endpoints of the path. As in the proof of Theorem 2, let $N$ be the number of substitutions on the path. We would like to calculate the conditional probability $\mathbb{P}(N=j|\mathcal{E}_{\alpha\beta})$. Once again, Bayes' formula gives

$$\mathbb{P}(N=j|\mathcal{E}_{\alpha\beta})=\frac{\mathbb{P}(\mathcal{E}_{\alpha\beta}|N=j)\mathbb{P}(N=j)}{\mathbb{P}(\mathcal{E}_{\alpha\beta})}. \tag{12}$$

The denominator $\mathbb{P}(\mathcal{E}_{\alpha\beta})$ is easily calculated. If $P(t)=\exp(Qt)$ is the transition matrix for the process operating for duration $t$, and $\pi$ denotes the stationary distribution of states for the process and $\Pi$ is the corresponding diagonal matrix, then:

$$\mathbb{P}(\mathcal{E}_{\alpha\beta}) = \pi_\alpha \exp\left(Q\cdot\frac{\mu}{-tr(\Pi Q)}\right)_{\alpha\beta}. \tag{13}$$

To calculate the first term in the numerator, namely $\mathbb{P}(\mathcal{E}_{\alpha\beta}|N=j)$, consider the 'embedded Markov chain' associated with the continuous-time Markov process. This discrete Markov chain, on the same set of states, has a matrix $S=[s_{\alpha\beta}]$ of transition probabilities that are described as follows: for $\alpha\neq\beta$ set: $s_{\alpha\beta}=q_{\alpha\beta}/\sum_{\beta\neq\alpha}q_{\alpha\beta}$, and for each state $\alpha$, $s_{\alpha\alpha}=0$. Thus $S$ is the transition matrix for the process that records changes of state. Accordingly, we have

$$\mathbb{P}(\mathcal{E}_{\alpha\beta}|N=j)=\pi_\alpha\cdot(S^j)_{\alpha\beta}. \tag{14}$$

Finally, consider the second term in the numerator, namely $\mathbb{P}(N=j)$. For the Jukes–Cantor model, this is described (exactly) by a Poisson distribution. A Poisson distribution also applies exactly to some other models, including 'group-based' models, such as the Kimura 3ST model. To see this, note that we can express $N$ as the sum of three independent random variables $N=N_1+N_2+N_3$, where $N_1$ is the number of transitions, and $N_2$ and $N_3$ are the number of transversions of type I and II, respectively, which occur along the path. $N_1$,$N_2$,$N_3$ each have a Poisson distribution, and, although they may have different means, their independence ensures that their sum is also Poisson.

For more complex (non-symmetric) models $\mathbb{P}(N=j)$ may fail to be exactly described by a Poisson distribution, though it may still provide a good approximation (in cases where it is not, one must simply modify the first bracketed term in the following equation or use the approaches introduced by Minin and Suchard (2008a,b)).

Assuming an exact (or approximate) Poisson distribution for $N$ then, from Eqs. (13) and (14), we have:

$$\mathbb{P}(N=j|\mathcal{E}_{\alpha\beta})=\left(\frac{e^{-\mu}\mu^j}{j!}\right)\cdot\left(\frac{(S^j)_{\alpha\beta}}{\exp\left(Q\cdot\frac{\mu}{-tr(\Pi Q)}\right)_{\alpha\beta}}\right). \tag{15}$$

### 2.2.3. Exact calculations for $n>1$ for the Jukes–Cantor model

Let $\mu_{n,k}$ and $\sigma^2_{n,k}$ denote the mean and variance of the distribution $P_{n,k}(j)$, and let $\hat{p_1}=k/n$ (the proportion of sites that show different states at the ends of the path) and $\hat{p_0}=1-\hat{p_1}$.

**Theorem 3.** *Suppose n sites evolve i.i.d. under the Jukes–Cantor model. Then,*

(i) $\mu_{n,k}=\mu n\cdot(\alpha_0\hat{p_0}+\alpha_1\hat{p_1})$, *and* $\sigma^2_{n,k}=\mu n\cdot(\beta_0\hat{p_0}+\beta_1\hat{p_1})$, *where*

$$\alpha_0=\frac{1-\theta}{1+3\theta},\quad \alpha_1=\frac{1+\theta/3}{1-\theta};$$

$$\beta_0=\frac{16\mu\theta}{3(1+3\theta)^2}+\frac{1-\theta}{1+3\theta};\quad \beta_1=\frac{-16\mu\theta}{9(1-\theta)^2}+\frac{1+\theta/3}{1-\theta}$$

$$\text{and}\quad \theta=e^{-4\mu/3}.$$

(ii) *Moreover*:

(a) *For large values of n, we can approximate $P_{n,k}(j)$ by a normal distribution with mean $\mu_{n,k}$ and variance $\sigma^2_{n,k}$ as described in Part* (i).

(b) *Regarding k as a random variable that counts the number of sites that have different states at the endpoints of the path, $\mu_{n,k}/n$ is an unbiased and consistent estimator of $\mu$; that is: $\mathbb{E}[\mu_{n,k}/n]=\mu$, and $\mu_{n,k}/n$ converges in probability to $\mu$ as $n\to\infty$.*

**Proof of Theorem 3.** For $k=0$ and 1, consider the probability generating function:

$$P_k(x)=\sum_{j\geq 0}P_{1,k}(j)x^j.$$

From (2), we have:

$$P_0(x)=\frac{e^{-\mu}}{1+3e^{-4\mu/3}}(e^{\mu x}+3e^{-\mu x/3})\quad\text{and}$$

$$P_1(x)=\frac{e^{-\mu}}{1-e^{-4\mu/3}}(e^{\mu x}-e^{-\mu x/3}). \tag{16}$$

Now:

$$\mu_{1,k}=\left.\frac{dP_k(x)}{dx}\right|_{x=1}$$

and

$$\sigma^2_{1,k}=\left.\frac{d^2P_k(x)}{dx^2}\right|_{x=1}+\mu_{1,k}-\mu^2_{1,k}.$$

Applying these identities in Eq. (16) and performing routine calculations shows that, for $k=0{,}1$:

$$\mu_{1,k}=\alpha_k\mu, \tag{17}$$

$$\sigma^2_{1,k}=\beta_k\mu. \tag{18}$$

Now, consider the total number $N_{\rm tot}$ of substitutions along the path connecting the two sequences and $N_i$, the number of substitutions along the path for site $i$. Thus $N_{\rm tot}=\sum_{i=1}^n N_i$. For $k=0{,}1$ let $\mathcal{E}^i_k$ be the event that $\mathcal{E}_k$ occurs at site $i$ ($\mathcal{E}_k$ defined as in the proof of Theorem 2). Then, by the i.i.d. assumption:

$$\mu_{n,k}=\mathbb{E}[N_{\rm tot}|\mathcal{E}^1_1\cdots\mathcal{E}^k_1,\mathcal{E}^{k+1}_0\cdots\mathcal{E}^n_0]=k\mathbb{E}[N|\mathcal{E}_1]+(n-k)\mathbb{E}[N|\mathcal{E}_0],$$

and so, by (17):

$$\mu_{n,k}=k\mu\alpha_1+(n-k)\mu\alpha_0=\mu n\cdot(\alpha_0\hat{p_0}+\alpha_1\hat{p_1}).$$

A similar identity applies for the variance, based on (18). This establishes Part (i).

For Part (ii) of the theorem, Part (a) follows from the central limit theorem applied to the sums of independent random variables (in two classes). For Part (b), note that $k$ has a binomial distribution with mean $np$ and variance $np(1-p)$ where $p=\frac{3}{4}(1-e^{-4\mu/3})$. Thus, $\mathbb{E}[\mu_{n,k}/n]=\mu(\alpha_0(1-p)+\alpha_1 p)=\mu$ and the variance of $\mu_{n,k}/n$ is of order $n^{-1}$. This justifies the claims. □

### 2.2.4. Extension to sequences with $n>1$ under general independent-sites models

We can calculate the posterior probabilities of different values of $j$ when we have multiple sites ($n>1$) using recursion. When the sequences are identical ($k=0$), we have:

$$P_{n,0}(j)=\sum_{i=0}^{j}P_{1,0}(i)P_{n-1,0}(j-i). \tag{19}$$

For the Jukes–Cantor model with $n>1$ and $0<k\leq n$, the recursive formulation is

$$P_{n,k}(j)=\sum_{i=0}^{j}P_{1,1}(i)P_{n-1,k-1}(j-i). \tag{20}$$

However, we know that some of the terms in this summation will be 0. In particular, the term associated with $i=0$ will be 0 because $P_{1,1}(0)=0$. Similarly, whenever $j-i<k-1$ then terms will drop out because $P_{n-1,k-1}(j-i)$ will be 0. Thus, we can truncate the

summation to:

$$P_{n,k}(j) = \sum_{i=1}^{j+1-k} P_{1,1}(i)P_{n-1,k-1}(j-i). \tag{21}$$

Under any i.i.d. model, the same probability statements will be encountered many times during the recursive calculations (Eqs. (19) and (21)). Caching the probabilities in a data-structure (such as a hash-table) that allows for efficient lookup of a value can dramatically improve the running time. As written, the equations decompose the calculations for $n$ sites into calculations for one site and $n-1$ sites. The computational time of algorithms based these equations will scale on the order $\mathcal{O}(jn)$. This scaling can be improved to $\mathcal{O}(j\log(n))$ by splitting the $n$ sites in half (or as close to half as possible):

$$P_{n,0}(j) = \sum_{i=0}^{j} P_{\lfloor n/2\rfloor,0}(i)P_{n-\lfloor n/2\rfloor,0}(j-i). \tag{22}$$

When $k \geq 1$, we must ensure that all of the terms that we consider account for exactly $k$ differences between the leaves. If $\lfloor n/2\rfloor \leq k$ then:

$$P_{n,k}(j) = \sum_{i=1}^{j-k+\lfloor n/2\rfloor} P_{\lfloor n/2\rfloor,\lfloor n/2\rfloor}(i)P_{n-\lfloor n/2\rfloor,k-\lfloor n/2\rfloor}(j-i). \tag{23}$$

However, if $\lfloor n/2\rfloor > k$ then:

$$P_{n,k}(j) = \sum_{i=1}^{j} P_{\lfloor n/2\rfloor,k}(i)P_{n-\lfloor n/2\rfloor,0}(j-i). \tag{24}$$

Applying this recursive style of calculation to models that are more complex than the Jukes–Cantor model requires more complicated bookkeeping. Rather than simply tallying the total number of sites $n$ and the number of sites that differ between the leaves, we must record the number of times each type of data pattern is observed in the pair of leaves. Nevertheless, the general approach outlined in Eqs. (19) and (21) would still apply. Note that it would even be possible to use this form of calculation if the different sites evolved according to different models of evolution. For example, different sites could be allowed to evolve at differing rates.

#### 2.2.5. *Tree searching with estimates of the number of substitutions per edge*

We have implemented the methods into software that can score trees using the Jukes–Cantor model. The software uses numerical optimization techniques (specifically the simplex method) to search for edge lengths that maximize the pairwise likelihood score (Eq. (2)). Alternatively, the software can jointly estimate the edge lengths (specified as the expected number of changes) and the number of changes along an edge by maximizing a pairwise scoring scheme related to a posterior density. Specifically, for each edge $e$ on a tree, $T$, we search for a value of the expected number of substitutions per site, $\mu_e$, and an actual number of changes, $\xi_e$. During this estimation procedure, we attempt to maximize the score, $B$:

$$B(T,\boldsymbol{\mu},\boldsymbol{\xi}) = \sum_{e\in\mathcal{E}(T)} \ln[\mathbb{P}(\mu_e)\mathbb{P}(\xi_e|\mu_e)] + \sum_{i=1}^{M-1}\sum_{j=1+i}^{M} \ln[\mathbb{P}(X_j|T,\xi_{ij},X_i)\mathbb{P}(X_i)], \tag{25}$$

where $\mathcal{E}(T)$ is the collection of indices for edges in the tree and $\xi_{ij}$ is simply the sum of all substitution counts for all edges along the path from leaf $i$ to leaf $j$. Note that $\xi_{ij}$ is not a free parameter that is separate from the counts of substitutions assigned to the edges; it is simply the sum for the path from $i$ to $j$:

$\xi_{ij} = \sum_{e\in\mathcal{P}(T,i,j)}\xi_e.$

Note that Eq. (25) is very similar to Eq. (3), except that it calculates the probability of the end state given the number of substitutions along the path ($\xi_{ij}$), and it contains a summation of the log of the prior density of $\mathbb{P}(\mu_e,\xi_e)$ over all edges. As in the previous sections, we assume that $\mathbb{P}(\xi_e|\mu_e)$ is the probability from a Poisson distribution with mean $\mu_e$. Using Theorem 2 and the recursions in Eqs. (19) and (21) directly would require us to have a prior probability for the paths between different leaves. Here we chose to specify the prior probability for edge lengths by treating each edge length parameter as an independent parameter and placing an identical prior distribution on each parameter (as is commonly done in Bayesian inference of unrooted trees). Calculation of $\mathbb{P}(X_j|T,\xi_{ij},X_i)$ can be done with recursions similar to those described in Eqs. (19) and (21). When considering a single site, we use:

$$\mathbb{P}(X_j|\xi_{ij},X_i) = \tfrac{1}{4}(1+3(-1/3)^j) \quad \text{if } X_{i,k} = X_{j,k} \tag{26}$$

$$\mathbb{P}(X_j|\xi_{ij},X_i) = \tfrac{1}{4}(1-(-1/3)^j) \text{ otherwise}. \tag{27}$$

These formulae follow directly from Eqs. (10) and (11).

Although we have not concentrated on an optimized software implementation, we note that if the optimization routine in the software was to change a single edge length, then a small subset of the $\binom{M}{2}$ pairwise comparisons would be affected. Thus, it would be possible to implement the efficient nearest-neighbor-interchange branch-swapping techniques similar to the ones used by Desper and Gascuel (2002) in their FastME software.

### 2.3. *Example*

Farris (1981) provided an example of a dataset that is the clearest demonstration of the objection that distance-based approaches do not enforce all of the constraints that are implied by descent from a common ancestor. He pointed out that on a dataset such as the one shown in Fig. 2a, a distance-based approach will infer a star tree with equal edge lengths (see Fig. 2b). The path length between any pair of leaves shown in Fig. 2b will be 0.25, which corresponds exactly with the observed (uncorrected) distance based on the data matrix. This result seems unsatisfying because the common ancestor of all of the sequences must have had a nucleotide at the first position. No sequence has a distance of 0.125 from all of the four leaves, and thus it seems counterintuitive that the inferred tree would display these edge lengths and be considered to have a perfect fit to the data. Put another way: the star tree shown in Fig. 2b shows that distance methods do not impose the constraint that the interior node of the tree must be assigned a sequence. Felsenstein (1984) referred to this constraint as “realizability.”

Intuitively, one might expect that an inference method would judge the four trees shown in Fig. 2c to be tied as optimal solutions. In each of the four trees shown in Fig. 2c, we would infer the ancestral sequence to be identical to one of the leaf sequences; this leads to one edge with length 0 and three edges with length 0.25.

We conducted analyses of the dataset shown in Fig. 2a using the Jukes–Cantor model under WLS (Fitch and Margoliash, 1967) and ML implemented in PAUP* (Swofford, 2003), as well as the two new criteria introduced here: pairwise likelihood scoring, and joint estimation of the substitution counts and edge lengths. The edge lengths optimized on the star tree and a fully resolved tree are summarized in Table 1. The approximate edge lengths returned by simplex numerical optimization routines were modified slightly to assure that the score was optimized completely. We used an (improper) uniform prior distribution on the expected edge length in those distance analyses that used

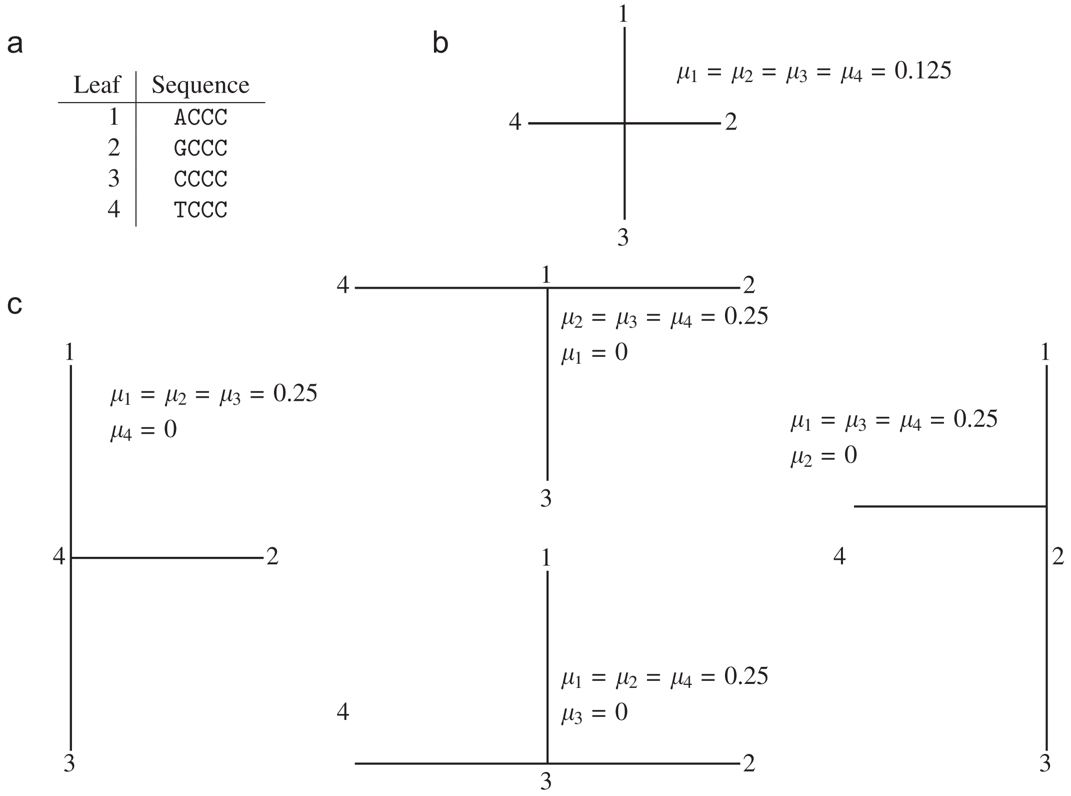
a
Leaf | Sequence
1 ACCC
2 GCCC
3 CCCC
4 TCCC
b
1
2
3
4
$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.125$
c
$\mu_2 = \mu_3 = \mu_4 = 0.25$
$\mu_1 = 0$
$\mu_1 = \mu_2 = \mu_3 = 0.25$
$\mu_4 = 0$
$\mu_1 = \mu_3 = \mu_4 = 0.25$
$\mu_2 = 0$
$\mu_1 = \mu_2 = \mu_4 = 0.25$
$\mu_3 = 0$

**Fig. 2.** A version of the dataset discussed by Farris (1981) and Felsenstein (1984): (a) the data matrix, (b) a tree and edge length that have a perfect fit to the data (using uncorrected distances and a least squares criterion), (c) four possible combinations of edge lengths that one might expect to be the optimal solutions because they consider the sequences for the internal node.

**Table 1**
Summary of optimal edge lengths under different criteria for the dataset shown in Fig. 2a.

| Criterion | Tree topology | Optimal | Approximate optimal edge lengths |
|---|---|---|---|
| WLS | Star | Yes | 0.152 for all edges |
| | Any binary | Yes | 0.152 for terminal edges; 0 for internal |
| ML | Star | No | 0.216 for all edges |
| | Any binary | Yes | 0.194 for terminal edges; 0.078 for internal |
| Pairwise ML | Star | Yes | 0.152 for all edges |
| | Any binary | Yes | 0.152 for terminal edges; 0 for internal |
| With subst. counts | Star | Yes | $\mu=0.25,\chi=1$ for three edges;<br>$\mu=0,\chi=0$ for one edge |
| | Any binary | Yes | $\mu=0.25,\chi=1$ for three terminal edges;<br>$\mu=0,\chi=0$ for the internal and one terminal edge |

substitution counts. Note that the edge lengths estimates from pairwise ML are identical to the estimates from WLS for this dataset. This is unsurprising because, as Farris pointed out, the star tree can result in trees with a perfect fit for this dataset when perfect fit is assessed via pairwise comparisons of leaves.

Our distance analysis which jointly estimates the number of substitutions favored the set of edge lengths ($\boldsymbol{\mu}$) identical to those shown in Fig. 2c. Thus the method does seem to address, to some degree, the concern raised by Farris (1981) that the edge lengths in a phylogenetic analysis should return a tree estimate that could be realized by actual sequence evolution. However, it is not clear that estimating the number of substitutions will improve the accuracy of the tree estimation procedure. Felsenstein (1984) pointed out that the behavior of traditional distance methods (which estimate the tree shown in Fig. 2b) is indeed appropriate if one views the edge lengths as parameters. These parameters express the expected number of changes along an edge instead of a count of the number of changes that must have occurred. More importantly, Felsenstein (1984) points out that Farris' objections do not undercut the validity of distance-based methods – indeed, distance-based phylogenetic methods have been proven to be statistically consistent (Atteson, 1999) and surprisingly powerful (Roch, 2010).

One can ask whether our system for jointly considering substitution counts and edge lengths is sufficient to guarantee that only realizable solutions are considered. In other words, is it true that every set of substitution counts that has a score greater than $-\infty$ according to Eq. (25) will be realizable by assigning sequences to internal nodes? The answer to this question is “no,” as can be seen by considering the data shown in Fig. 2a and the tree shown in Fig. 2b with one substitution assigned to every edge. The path between each pair of sequences will contain two substitutions, and this scenario is compatible with one observed difference between each pair of leaves (and the score of $B(T,\boldsymbol{\mu},\xi)$ will be not be $-\infty$ if the branch lengths are not 0). However, no sequence can be assigned to the internal node that is one substitution away from all four sequences shown at the leaves. Thus, consideration of the substitutional counts during pairwise comparisons can eliminate some impossible scenarios from consideration (because $\mathbb{P}(X_j|\xi_{ij},X_i)$ will be zero for some scenarios that are obviously incompatible with the data), but this is not sufficient to enforce all of the conditions that would address Farris' critique. It is not clear if the maximum *a posteriori* (MAP) estimate of $(\boldsymbol{\mu},\boldsymbol{\xi})$ will ever correspond to an unrealizable scenario.

Also note that the estimates of the expected numbers of changes per site are precisely equal to the count of the number of changes divided by the number of sites. This will occur if the prior distribution used for the edge length hyper-parameters ($\mu$) is a uniform distribution, and we use the MAP estimate of all parameters and hyper-parameters.

## 3. Conclusions

We have developed two new optimality criteria for evaluating evolutionary trees. These criteria are closely tied to distance-based approaches to tree estimation because they calculate a score based on all possible pairwise comparisons of the leaves. However, they also have properties in common with character-based tree inference approaches. The pairwise ML procedure uses the likelihood of a pairwise sequence spectrum under a model of sequence evolution to assess the fit of a set of edge lengths to the data. It is possible that this approach will be preferable to WLS estimation in cases of short sequences (where assuming the normality of errors in distance estimates may not be justified). Tamura et al. (2004) reported that a pairwise ML approach for distance corrections results in more accurate neighbor-joining trees.

The other tree estimation procedure that we introduced is the joint estimation of the number of substitutions and edge lengths. This method was inspired by the observation that distance-based approaches assume additivity of the edge lengths when expressed as the expected number of substitutions, but they do not attempt to utilize the fact that the actual count of substitutions should be additive.

The results reported here may be useful in other contexts. For example, Markov chain Monte Carlo techniques that use data-augmentation (e.g. Lartillot, 2006) require draws from the substitutional history along a branch of fixed length and end points; our results on the posterior probability of the number of substitutions along a branch would allow for efficiently drawing realizations of substitutional histories. Minin and Suchard (2008a,b) have developed methods for calculating the moments of counting processes (such as counting the number of non-synonymous substitutions in a codon model) over a phylogenetic tree. Unlike our results, their approaches are not restricted to the domain of models that have a Poisson prior distribution on the number of substitutions. As a result of the generality of their methods, their calculations of the probability of the number of substitutions, given the end states, are more expensive – unlike our results for $P_{1,0}(j)$ and $P_{1,1}(j)$ (see Theorem 2), the more general approach entails a summation of $j$ terms. O'Brein et al. (2009) used the approaches introduced by Minin and Suchard (2008a) to develop robust distance estimators based on the expected number of changes between two sequences. However, they did not investigate distance-based tree inference where the number of substitutions along each branch is treated as a parameter.

Future work would include thorough simulation studies to determine the contexts in which the approaches introduced here perform differently from ML- and WLS-based inference. Distance-based approaches to tree estimation are usually much faster than ML inference, but the likelihood principle implies that they make less efficient use of the data. At least two promising avenues for improving the accuracy of distance-based methods appear to be available: developing better error models for the disagreement between pairwise distance and path lengths, and developing methods that primarily not only rely on distances but also utilize some information about plausible substitutional histories. It is unclear which of these avenues is most promising.

The criteria we introduce can be thought of as intermediates between ML- and WLS-based approaches. We do not expect these methods to outperform ML estimation, but we hope that they will be helpful in diagnosing the causes for disagreements between ML and WLS estimates. Cases in which the tree returned by WLS and the pairwise ML approach disagree may highlight contexts in which the sum of squared errors criterion is an inappropriate error function. Similarly, future work may identify cases in which jointly estimating the edge lengths and the number of changes along a path improves tree inference. Such cases could inform us about conditions in which statistical power requires us to enforce some constraints on valid substitutional histories.

It would also be interesting to a study a method that estimates the number of changes *at each site* for each branch. Because such a procedure would reveal sites which have a large number of changes, it would allow for the joint estimation of the rate of evolution for a site (at the expense of introducing a large number of parameters to be estimated). Traditional distance-based approaches have difficulty dealing with among-site rate variation, because such variation cannot be accurately detected from pairwise comparisons alone.

## Acknowledgments

We thank two anonymous reviewers for several helpful suggestions. MTH thanks the U.S. National Science Foundation (DEB-ATOL-0732920 grant) for financial support. MS thanks the Allan Wilson Centre for Molecular Ecology and Evolution, and the NZ Royal Society (James Cook Fellowship) for funding support.

## Appendix A. Proof of Theorem 1

We apply the following result from Chang (1996).

**Lemma A1** (Chang, 1996). *Let $\mathcal{X}$ be a finite set and let $\{\mathcal{P}_\theta : \theta \in \Theta\}$ be a family of probability distributions on $\mathcal{X}$, where the closure $\overline{\Theta}$ of $\Theta$ is a compact subset of a metric space. Let $Y_1,Y_2,\ldots$ be independent and identically distributed random variables (or vectors) with probability distribution $\mathcal{P}_{\theta_0}$ for some $\theta_0 \in \Theta$. Assume the identifiability condition*

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

*Suppose that for each $x \in \mathcal{X}$, the function $\theta \mapsto \mathcal{P}_\theta(x)$ is continuous on $\overline{\Theta}$, and let $\hat{\theta}_k = \hat{\theta}_k(Y_1,\ldots,Y_k)$ maximize the log-likelihood $\sum_{s=1}^{k} \log\mathcal{P}_\theta(Y_s)$ over $\theta \in \overline{\Theta}$. Then $\mathcal{P}_{\theta_0}(\hat{\theta}_k \to \theta_0) = 1$.*

Let $\mathcal{L} = \{1,\ldots n\}$ denote the leaf (taxon) set, $\mathcal{L}^{(2)}$ denote the set of pairs $(i,j) \in \mathcal{L} \times \mathcal{L}$ with $i<j$, and $\mathcal{A}$ denote the alphabet of character states. Consider the set $\mathcal{T}_\mathcal{L} = \{(T,\mu)\}$ of phylogenetic $\mathcal{L}$-trees with finite edge lengths, and the map:

$$\phi : \mathcal{T}_\mathcal{L} \to [0,1]^{\mathcal{L}^{(2)}} \text{ defined by } (T,\mu) \mapsto \theta^{(T,\mu)},$$

where, for each $(i,j)$ in $\mathcal{L}^{(2)}$, $\theta_{ij}^{(T,\mu)} = \exp(-\sum_{e \in P(T,i,j)} \mu_e)$. Thus, $-\log\theta_{ij}^{(T,\mu)}$ is the length of the path connecting the leaves $i,j$ in $T$ under the edge weighting $\mu$. To apply Chang's lemma in the present setting, let:

$$\Theta = \phi(\mathcal{T}_\mathcal{L}) = \{\theta^{(T,\mu)} : (T,\mu) \in \mathcal{T}_\mathcal{L}\}.$$

The closure $\overline{\Theta}$ of $\Theta$ is the set of limit points in $\Theta$ obtained by letting various sets of edge lengths in phylogenetic trees tend to infinity. Note that $\overline{\Theta}$ is a closed bounded subset of $[0,1]^{\mathcal{L}^{(2)}}$ and hence is a compact subset of a metric space.

Now suppose we have a stochastic process on site patterns, which is either a reversible, continuous-time Markov process (e.g. Jukes–Cantor or General time reversible (GTR)) or a mixture of such processes (by allowing a distribution of site specific rates, e.g. GTR + $\Gamma$). All such models have the property that for any pair of leaves $i,j$ separated by a path of length $\mu_{ij}$, for any ordered pair of states $\alpha,\beta \in \mathcal{A}$, the joint distribution $\mathbb{P}(X_i=\alpha,X_j=\beta)$ can be expressed as a linear function of non-negative (but not necessarily integer) powers of $\theta_{ij}=e^{-\mu_{ij}}$, where the coefficients are independent of the tree topology or edge lengths. Allowing one or more branch lengths to be infinite on this path corresponds to allowing $\theta_{ij}=0$, in which case $\mathbb{P}(X_i=\alpha,X_j=\beta)=\mathbb{P}(X_i=\alpha)\cdot\mathbb{P}(X_j=\beta)=\pi_\alpha\cdot\pi_\beta$ (the product of the equilibrium probabilities for the states $\alpha$ and $\beta$). Thus, for any $\theta=[\theta_{ij}:(i,j)\in\mathcal{L}^{(2)}]\in\overline{\Theta}$, we can write $\mathbb{P}(X_i=\alpha\cap X_j=\beta)$ as a continuous function of $\theta_{ij}$, which we make explicit by writing: $\mathbb{P}(X_i=\alpha,X_j=\beta|\theta_{ij})$.

We now define $\mathcal{P}_\theta$ for $\theta\in\overline{\Theta}$. Let $\mathcal{X}=(\mathcal{A}\times\mathcal{A})^{\mathcal{L}^{(2)}}$, the set of assignments of an ordered pair of character states to each pair of taxa $(i,j):i<j$. For each specific vector $y=[y_{ij}:(i,j)\in\mathcal{L}^{(2)}]$ where each $y_{ij}\in\mathcal{A}\times\mathcal{A}$, let:

$$\mathcal{P}_\theta(Y=y):=\prod_{(i,j)\in\mathcal{L}^{(2)}}\mathbb{P}((X_i,X_j)=y_{ij}|\theta_{ij}).$$

Thus, $\mathcal{P}_\theta$ is the probability distribution of the states for pairs of leaves, if one was to treat these as independent events across different choices of pairs.[1] We can recover the marginal distribution for $(X_i,X_j)$ from $\mathcal{P}_\theta$ for all $(i,j)\in\mathcal{L}^{(2)}$. Thus, if $\theta_0\in\Theta$ and $\theta\in\overline{\Theta}$ satisfy the identity $\mathcal{P}_{\theta_0}=\mathcal{P}_\theta$ then, for each $(i,j)\in\mathcal{L}^{(2)}$ and all choices of $y_{ij}$, we have: $\mathbb{P}((X_i,X_j)=y_{ij}|(\theta_0)_{ij})=\mathbb{P}((X_i,X_j)=y_{ij}|\theta_{ij})$. It now follows, by the assumption in the statement of the theorem (and the fact that $\theta_0$ lies in $\Theta$ and not in $\overline{\Theta}-\Theta$) that $(\theta_0)_{ij}=\theta_{ij}$ for all $(i,j)\in\mathcal{L}^{(2)}$. Thus, $\theta_0=\theta$, and the required identifiability assumption holds. The continuity assumption also holds, i.e. $\theta\mapsto\mathcal{P}_\theta$ is continuous, since $\mathbb{P}((X_i,X_j)=y_{ij}|\theta_{ij})$ depends continuously on $\theta_{ij}$.

Let $Y_1,\ldots$ be i.i.d. random variables with the same distribution as $Y$. Notice that each $Y_s$ is a vector $[(X_i^s,X_j^s):(i,j)\in\mathcal{L}^{(2)}]$ where $(X_i^s,X_j^s)$ records the state of leaf $i$ and of leaf $j$ at site $s$. Note also that for each $s$, the random variables $(X_i^s,X_j^s)$ are not independent as $(i,j)$ varies; however, the vectors $Y_1,Y_2,\ldots$ are.

Now, finding a phylogenetic tree $(T,\mu)$ with branch lengths (possibly infinite) to maximize the score $S(T,\mu)$ in Eq. (2) corresponds to finding a value $\hat{\theta}_k\in\overline{\Theta}$ that maximizes $\sum_{s=1}^k\log\mathcal{P}_\theta(Y_s)$ and, by Chang's lemma, $\mathcal{P}_{\theta_0}(\hat{\theta}_k\to\theta_0)=1$. Thus if we let $\varepsilon=\min\{(\mu_0)_{ij}:(i,j)\in\mathcal{L}^{(2)}\}>0$, then with probability 1 there exists $N$ sufficiently large so that for all $k\geq N$ we have $(\hat{\theta}_k)_{ij}>\varepsilon/2$ for all $(i,j)\in\mathcal{L}^{(2)}$, which implies that $\hat{\theta}_k=\phi(\hat{T}_k,\hat{\mu}_k)$ for some phylogenetic tree $\hat{T}_k$ and edge lengths $\hat{\mu}_k$ that are uniformly bounded above by a constant dependent only on $\varepsilon$.

Taking logarithms, it follows that for all $(i,j)\in\mathcal{L}^{(2)}$ and as $k(>N)$ tends to infinity, the following limit holds with probability 1: $|(\hat{\mu}_k)_{ij}-(\mu_0)_{ij}|\to0$. Finally, if $\delta>0$ is the smallest interior edge length of $T_0$, then when $k$ is sufficiently large that $|(\hat{\mu}_k)_{ij}-(\mu_0)_{ij}|<\delta/2$ for all $(i,j)\in\mathcal{L}^{(2)}$, we have $\hat{T}_k=T$, by Theorem 2.1 (Part (ii)) of Moulton and Steel (1999), as required.

[1] Thus, $\mathcal{P}_\theta(Y=y)$ is different from the probability of the conjunctive event $\bigcap_{(i,j)\in\mathcal{L}^{(2)}}\{(X_i,X_j)=y_{ij}\}$ under the original stochastic process.

## References

Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25 (2–3), 251–278.

Barry, P., 2007. On integer sequences associated with the cyclic and complete graphs. Journal of Integer Sequences 10 07.4.8.

Beyer, W., Stein, M., Smith, T., Ulam, S., 1974. A molecular sequence metric and evolutionary trees. Mathematical Biosciences 19, 9–25.

Bryant, D., Waddell, P., 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. Molecular Biology and Evolution 15 (10), 1346–1359.

Bulmer, M., 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Molecular Biology and Evolution 8 (6), 868–883.

Cavalli-Sforza, L., Edwards, A., 1967. Phylogenetic analysis: models and estimation procedures. American Journal of Human Genetics 19 (3), 233–257.

Chang, J., 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Mathematical Biosciences 137, 51–73.

Czarna, A., Sanjuán, R., González-Candelas, F., Wróbel, B., 2006. Topology testing of phylogenies using least squares methods. BMC Evolutionary Biology 6 (1), 105. doi:10.1186/1471-2148-6-105.

Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. Journal of Computational Biology 9 (5), 687–705.

Farris, J.S., 1981. Distance data in phylogenetic analysis. In: Funk, V.A., Brooks, D.R. (Eds.), Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society, vol. 1. New York Botanical Garden, Bronx, New York, pp. 3–23.

Felsenstein, J., 1984. Distance methods for inferring phylogenies: a justification. Evolution 38 (1), 16–24.

Felsenstein, J., 2004. Inferring Phylogenies, first ed. Sinauer Associates, Sunderland, Massachusetts.

Fitch, W., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155 (3760), 279–284.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating the human–ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22, 160–174.

Lartillot, N., 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. Journal of Computational Biology 13 (10), 1701–1722. doi:10.1089/cmb.2006.13.1701.

Minin, V.N., Suchard, M.A., 2008a. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology 56, 391–412.

Minin, V.N., Suchard, M.A., 2008b. Fast, accurate and simulation-free stochastic mapping. Philosophical Transactions of the Royal Society of London, B, Biological Sciences 363, 3985–3995. doi:10.1098/rstb.2008.0176.

Moulton, V., Steel, M., 1999. Retractions of finite distance functions onto tree metrics. Discrete Applied Mathematics 91, 215–233.

O'Brein, J.D., Minin, V., Suchard, M., 2009. Learning to count: robust estimates for labeled distances between molecular sequences. Molecular Biology and Evolution 26 (4), 801–814. doi:10.1093/molbev/msp003.

Roch, S., 2010. Toward extracting all phylogenetic information from matrices of evolutionary distances. Science 327 (5971), 1376–1379. doi:10.1126/science.1182300.

Sanjuan, R., Wrobel, B., 2005. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. Systematic Biology 54 (2), 218–229. doi:10.1080/10635150590923308.

Susko, E., 2003. Confidence regions and hypothesis tests for topologies using generalized least squares. Molecular Biology and Evolution 20 (6), 862–868. doi:10.1093/molbev/msg093.

Swofford, D.L., 2003. PAUP∗. Phylogenetic Analysis Using Parsimony (∗and Other Methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor–joining method. Proceedings of the National Academy of Science 101 (30), 11030–11035.