Journal of Theoretical Biology 256 (2009) 467–472

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A basic limitation on inferring phylogenies by pairwise sequence comparisons

Mike Steel *

Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 19 August 2008
Received in revised form
9 October 2008
Accepted 9 October 2008
Available online 22 October 2008

MSC:
05C05
92D15

Keywords:
Phylogenetic tree
Distance-based methods
Gamma distributed rates
Identifiability

ABSTRACT

Distance-based approaches in phylogenetics such as Neighbor-Joining are a fast and popular approach for building trees. These methods take pairs of sequences, and from them construct a value that, in expectation, is additive under a stochastic model of site substitution. Most models assume a distribution of rates across sites, often based on a gamma distribution. Provided the (shape) parameter of this distribution is known, the method can correctly reconstruct the tree. However, if the shape parameter is not known then we show that topologically different trees, with different shape parameters and associated positive branch lengths, can lead to exactly matching distributions on pairwise site patterns between all pairs of taxa. Thus, one could not distinguish between the two trees using pairs of sequences without some prior knowledge of the shape parameter. More surprisingly, this can happen for *any* choice of distinct shape parameters on the two trees, and thus the result is not peculiar to a particular or contrived selection of the shape parameters. On a positive note, we point out known conditions where identifiability can be restored (namely, when the branch lengths are clocklike, or if methods such as maximum likelihood are used).

1. Introduction

Stochastic models that describe the evolution of aligned DNA sequence sites are fundamental to most modern approaches to phylogenetic tree reconstruction (Felsenstein, 2003). Making these models more realistic usually requires introducing additional parameters. However, this raises the prospect that one might lose the ability to estimate a tree if one has to rely on the data to estimate all the parameters in the model. This could occur for various reasons—for example, it may be that two different trees could produce exactly the same probability distribution on site patterns for two appropriately selected settings of the other parameters in the model. Such a scenario would be a problem for any method of tree reconstruction (including maximum likelihood and Bayesian methods) as it would mean that in some cases, one could not distinguish between two trees even with infinitely long sequences. This loss of statistical 'identifiability' has been demonstrated for certain types of DNA substitution models, including rates-across-sites models (Steel et al., 1994) and, more recently, simple mixture models (Matsen and Steel, 2007). On the positive side, a number of identifiability results have also been established for suitably constrained models (see, for example, Allman et al., 2008; Allman and Rhodes, 2006; Allman and Rhodes, 2008a, b; Chang, 1996; Steel, 1994; Steel and Penny, 2000).

In this paper, we are interested in a phenomenon that is related to, but different from the loss of statistical identifiability, since it is method-dependent. We will describe a situation where pairwise sequence comparison methods can fail to distinguish between trees, even though more sophisticated methods such as ML can. Thus, the models are statistically identifiable, as far as the tree parameter is concerned, but only if one uses the full matrix of aligned sequence information and not just pairwise sequence comparisons. Specifically we consider tree reconstruction when sequences evolve under a model in which site rates have a gamma distribution, but where the (shape) parameter of the gamma distribution is not known. In this case, if one uses all the aligned sequence data, or at least 3–way sequence comparisons then, for DNA sequences one can recover the shape parameter in a statistically consistent way, and thereby the underlying phylogenetic tree, by a recent result of Allman et al. (2008). However, if one just uses pairwise sequence comparisons we show that two different trees can produce exactly the same pairwise sequence comparisons; moreover, this can happen for any different choice of shape parameters for the two trees (by selecting the branch lengths on the two trees appropriately).

The intuition behind this limitation on pairwise sequence comparisons has been nicely summarized by Felsenstein (2003,

* Tel.: +64 3 366 7001; fax: +64 3 364 2587.
E-mail address: m.steel@math.canterbury.ac.nz

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.10.010

p. 175): the rate at which a site is evolving affects all the taxa, but this constraint is not reflected by a method that is based on pairwise comparisons, and so, for example, "once one is looking at changes within rodents it will forget where changes were seen among primates."

Before describing our results we mention some earlier papers that described related but different phenomena. Baake (1998) considered a model in which half the sites are invariable and the remaining sites evolve under a general Markov model. Although this model (and the tree) is generically identifiable using all the sequence information (as recently shown in Allman and Rhodes, 2008a) Baake showed that two trees can produce identical pairwise sequence comparisons. The non-indentifiability of divergence times on a fixed tree under various rates-across-sites models has also been recently investigated by Evans and Warnow (2004). Finally we note that our result that distance-based methods can be misleading for tree inference complements some earlier work (Bandelt and Fischer, 2008; Huson and Steel, 2004) which highlighted a different result in which distances can perfectly 'fit' one phylogenetic tree when the full sequence data support a different tree.

2. Definitions and observations

In sequence-based approaches to phylogenetics, the data usually consists of a collection of n sequences $s^1, s^2, \ldots, s^n$, each of length N, where each sequence site takes values in some state space. We will suppose that there are r states, and denote them by greek letters μ, ν throughout—for example, for aligned DNA sequence data $r = 4$ and the state space is the four DNA bases (A,C,G,T). Given the aligned sequences, biologists seek to infer a phylogenetic tree $\mathscr{T}$, whose leaves are labeled by $\{1, \ldots, n\}$ and which describes the evolution of the sequences from some unknown common ancestral sequence (leaf i corresponds to the extant taxon from which sequence s^i has been obtained). For further background on phylogenetics, the reader may consult Felsenstein (2003) and Semple and Steel (2003).

Given two sequences $s^i = (s_1, s_2, \ldots, s_N)$ and $s^j = (s'_1, s'_2, \ldots, s'_N)$ let $\hat{J}_{ij}$ be the $r \times r$ matrix whose $\mu\nu$-entry is the proportion of sites where sequence s^i is in state μ and sequence s^j is in state ν. The proportion of sites where sequence s^i and s^j differ, δ_{ij} (the normalized sequence dissimilarity) is therefore the sum of the off-diagonal entries of $\hat{J}_{ij}$; more formally, $\delta_{ij} = (1/N)\sum_{k=1}^{N} \{k : s^i_k \neq s^j_k\} = 1 - \mathrm{tr}(\hat{J}_{ij})$, where tr refers to matrix trace (the sum of the diagonal entries).

Given a collection of sequences $s^1, s^2, \ldots, s^n$, each of length N, one can easily derive the collection of pairwise $\hat{J}$—matrices $\hat{J}_{ij} : i, j \in \{1, \ldots, n\}$. This reduction process, from aligned sequences to pairwise comparisons, is highly redundant (for typical values of n) since it reduces the frequencies of r^n site patterns to $\binom{n}{2}$ comparisons of r^2 sites pattern frequencies. The further reduction to the δ values involves even more redundancy (Steel et al., 1988). Despite this, it is well known that these reduced matrices (and sometimes just the δ values) provide a statistically consistent way to estimate the underlying tree, under simple models of DNA site substitution. This follows by combining two well-known facts.

Fact one: Under the assumption that the aligned sequence sites evolve i.i.d., the law of large numbers tells us that the $\hat{J}_{ij}$ matrices (and thereby the δ_{ij} values) converge in probability to their expected values as the sequence length N becomes large.

To explain this further we introduce two key definitions: For $i, j \in X := \{1, 2, \ldots, n\}$, let J_{ij} be the expected value of $\hat{J}_{ij}$—thus, J_{ij} is an $r \times r$ matrix whose $\mu\nu$-entry is

$$J^{\mu\nu}_{ij} := \mathbb{P}(s^i_k = \mu, s^j_k = \nu),$$

for each pair of states μ, ν, and any given k; and let

$$d_{ij} := \mathbb{P}(s^k_i \neq s^k_j),$$

for any given k. In words, J_{ij} is the matrix whose entries describe the joint probability that at any given site the sequences s^i and s^j are in specified states, while d_{ij} is simply the probability that these states are different at a given site. By definition, $d_{ij} = 1 - \mathrm{tr}(J_{ij})$.

With this notation, Fact one can be restated as the condition that, for all $i, j \in X$:

$$\hat{J}_{ij} \xrightarrow{p} J_{ij} \quad \text{and} \quad \delta_{ij} \xrightarrow{p} d_{ij},$$

where $\xrightarrow{p}$ denotes convergence in probability as $N \to \infty$.

The second result required to show that the $\hat{J}_{ij}$ values estimate the tree consistently is that for many models the J_{ij} values can be transformed to obtain a function on pairs of leaves that is additive. Recall that a function l_{ij} on pairs of leaves of a tree is said to be *additive* on a tree $\mathscr{T}$ if one can assign a positive real number l_e to each edge e of $\mathscr{T}$ so that l_{ij} is the sum of the numbers assigned to the edges on the path connecting the two leaves on the tree. That is

$$l_{ij} = \sum_{e \in p(\mathscr{T}; i, j)} l_e, \tag{1}$$

where $p(\mathscr{T}; i, j)$ denotes the edges on the path in $\mathscr{T}$ connecting i and j. This additivity condition implies that the tree $\mathscr{T}$ can be uniquely recovered from the l_{ij} values (see e.g. Semple and Steel, 2003). With this in mind we have:

Fact two: Under various models of sequence evolution, a distance function l on X that is additive on the underlying tree can be computed from the J matrices (and sometimes just the d values).

The two main models for which Fact two is known to apply are (i) the general Markov process, for which the function $\{i, j\} \mapsto -\log(\det(J_{ij}))$ is additive, and (ii) the general time-reversible (GTR) model with any known distribution of rates across sites. In this latter case—which is the one of interest in this paper—one can transform the J matrices to obtain an distance function l on X that corresponds to the expected number of substitutions ('evolutionary distance') between i and j—and which is therefore additive. For a GTR model, with a distribution $\mathscr{D}$ of rates across sites this transformation (Waddell and Steel, 1997) is

$$l_{ij} = -\mathrm{tr}(\Pi M^{-1}_{\mathscr{D}}(\Pi^{-1} J_{ij})),$$

where $M_{\mathscr{D}}$ is the moment generating function of the distribution of rates across sites, and where $\Pi = \mathrm{diag}(\boldsymbol{\pi})$ is the diagonal matrix whose leading diagonal is the vector $\boldsymbol{\pi} = [\pi_\mu]$ of the frequencies of the r states. For the GTR model (or any submodel) the matrix J_{ij} is symmetric (Waddell and Steel, 1997) and $J_{ii} = \Pi$ for each i.

Combining Facts one and two gives:

$$-\mathrm{tr}(\Pi M^{-1}_{\mathscr{D}}(\Pi^{-1} \hat{J}_{ij})) \xrightarrow{p} l_{ij}$$

and so the $\hat{J}_{ij}$ values allow us to reconstruct the underlying tree from sufficiently long sequences. Indeed even if we do not know the stationary frequencies of the states (the matrix Π) we can still recover the tree, since Π is determined by (the row sums of) J_{ij}, and so if we let $\hat{\Pi}_{ij}$ denote the corresponding empirical state frequencies (determined by the corresponding row sums of $\hat{J}_{ij}$) then we have

$$-\mathrm{tr}(\hat{\Pi} M^{-1}_{\mathscr{D}}(\hat{\Pi}^{-1} \hat{J}_{ij})) \xrightarrow{p} l_{ij}.$$

Thus, if for each pair i, j we derive an estimate $\hat{l}_{ij}$ of evolutionary distance (l_{ij}) by either maximum likelihood estimation or by the 'corrected distance' formula:

$$\hat{l}_{ij} = -\mathrm{tr}(\hat{\Pi} M^{-1}_{\mathscr{D}}(\hat{\Pi}^{-1} \hat{J}_{ij})) \tag{2}$$

then these estimated values will converge to the true l_{ij} values as the sequence length N grows, allowing for statistically consistent reconstruction of the tree by using fast distance-based tree reconstruction methods.

For some GTR models it is also possible to transform just the δ_{ij} to obtain l_{ij}—for example, under the simple symmetric 4-state model (the Jukes–Cantor model) the transformation is

$$l_{ij} = -\frac{3}{4}M_{\mathscr{D}}^{-1}\left(1-\frac{4}{3}d_{ij}\right). \tag{3}$$

For models in which l_{ij} can be expressed as a function of d_{ij} one can use δ in place of d to estimate l_{ij} (for certain models, such as the Jukes–Cantor model, this leads to the same l_{ij} estimates as a pairwise maximum likelihood estimate, but for more complex models this need not be the case).

The snag in this otherwise appealing story is that it assumes that we know the distribution $\mathscr{D}$ of rates across sites—what happens if $\mathscr{D}$ is unknown or has parameters that require estimation? If no constraints are placed upon $\mathscr{D}$ then identifiability of the tree can be completely lost (Steel et al., 1994). It is therefore fortunate that in molecular systematics $\mathscr{D}$ is typically described by a simple parametric distribution. In particular, the gamma distribution has a long and popular history in models that describe the variation of substitution rates across DNA sequence sites (Yang, 1993). Today, a common default option is the 'GTR + Γ + I' model in which each site is either invariant (with some probability), or it evolves according to a GTR Markov process that proceeds at a rate selected randomly from a gamma distribution. In this paper we will ignore the invariable sites, since our main result (Theorem 3.1) will automatically imply a corresponding result when invariable sites are present. Moreover, we may (without loss of generality) assume that the gamma distribution is normalized so that its mean is equal to 1 and so there remains just one parameter—the 'shape' parameter, k.

We will show that *any* two different shape parameters can provide exactly the same J matrices on a pair of topologically distinct trees (with appropriately assigned branch lengths). Consequently, using just pairwise comparisons (the $\hat{J}$ matrices) to infer phylogeny from the resulting data, without prior knowledge of the shape parameter is potentially problematic—either of the two trees could describe the data much better than the other if one were to select the shape parameter appropriate for that tree. Thus, a biologist exploring data by seeing the effect of varying k might note that for one value of k his/her data fit a tree perfectly. The result described here shows that it could be dangerous to stop at this point and report the tree, as there may well be another value of k for which the pairwise sequence data (or distance data) fit a different tree perfectly. Using all the data (i.e. not reducing to pairwise comparisons) will overcome this problem for a gamma distribution as established recently by Allman et al. (2008) (who also pointed out errors in an earlier approach from Rogers, 2001).

3. Results

In this paper we consider a particular type of reversible stationary Markov process, called the *equal input model*. In this model, the rate of substitution does not depend on the current state, and when a substitution event occurs, the new state is selected according to the stationary distribution of states, which we encode by the vector $\boldsymbol{\pi}$. Thus the rate matrix R is defined by the condition $R_{\mu\nu} = \pi_\nu$ for all $\nu \neq \mu$. In the case of $r = 4$ states, this model has been called the 'Tajima-Nei equal input model' or the 'Felsenstein 1981 model'; when, in addition, $\boldsymbol{\pi}$ is uniform, it is the known as the 'Jukes–Cantor' model. For more mathematical background on the equal input model, see, for example, Semple and Steel (2003). Although the equal-input model is a special case of the GTR model, we have chosen it because it is simple enough to allow tractable exact calculations, yet without being overly simplistic (for example, it allows arbitrary stationary frequencies for the states).

Under the equal input model, and with constant-rate site evolution we have

$$J_{ij}^{\mu\nu} = \begin{cases} \pi_\mu\pi_\nu(1-\exp(-l_{ij}/\gamma)) & \text{if } \mu \neq \nu; \\ \pi_\mu(\pi_\mu + (1-\pi_\mu)\exp(-l_{ij}/\gamma))) & \text{if } \mu = \nu, \end{cases} \tag{4}$$

where l_{ij} is the expected number of substitutions on the path connecting i and j in $\mathscr{T}$ (an additive distance) and $\gamma = 1 - \sum_\mu \pi_\mu^2$ (this number is the expected normalized sequence dissimilarity for stationary random sequences—for example, in the Jukes–Cantor model, it takes the value $1 - 4\cdot(\frac{1}{4})^2 = \frac{3}{4}$. More briefly we can write

$$J_{ij}^{\mu\nu} = a_{\mu\nu} + b_{\mu\nu}\exp(-l_{ij}/\gamma), \tag{5}$$

where $a_{\mu\nu}, b_{\mu\nu}$ are constants that depend on the pair μ, ν and the vector π.

If we now impose an associated distribution $\mathscr{D}$ of rates across sites on this equal-input model, in which case each site evolves according to the same equal input model, but with a rate selected randomly according to $\mathscr{D}$. In this case (5) becomes

$$J_{ij}^{\mu\nu} = a_{\mu\nu} + b_{\mu\nu}M_{\mathscr{D}}(-l_{ij}/\gamma), \tag{6}$$

where $M_{\mathscr{D}}(x)$ is the moment generating function for $\mathscr{D}$. When $\mathscr{D}$ is a gamma distribution of rates across sites with shape parameter k and mean 1 we have

$$M_{\mathscr{D}}(x) = \left(1-\frac{x}{k}\right)^{-k},$$

and so Eq. (6) becomes

$$J_{ij}^{\mu\nu} = a_{\mu\nu} + b_{\mu\nu}\left(1+\frac{l_{ij}}{k\gamma}\right)^{-k}. \tag{7}$$

Now, suppose we have two topologically distinct binary phylogenetic trees $\mathscr{T}$ and $\mathscr{T}'$ on the same leaf set, where $\mathscr{T}$ has branch length l and gamma distribution of rates across sites (with mean 1) with shape parameter k, while $\mathscr{T}'$ has branch lengths l' and gamma distribution of rates across sites (with mean 1) with shape parameter k', where $k' \neq k$. We can now state the main result of this paper.

Theorem 3.1. *Consider a fixed equal input model on $r \geqslant 2$ states. Then for any $k, k' > 0$ with $k \neq k'$ and for any binary phylogenetic tree $\mathscr{T}$ on a set X of four or more leaves there exists a topologically distinct binary phylogenetic tree $\mathscr{T}'$ on leaf set X, and strictly positive branch lengths l for $\mathscr{T}$ and l' for $\mathscr{T}'$, respectively, so that the matrices of joint pairwise distributions J_{ij} and J'_{ij} agree for all $i, j \in X$.*

Remarks. The significance of this result for phylogenetic reconstruction is that it shows that even if one uses pairwise sequence comparisons, the choice of the correct shape parameter for the gamma distribution is essential—if we selected shape parameter k, the corrected distances (obtained by pairwise ML estimation or by (2)) would fit $\mathscr{T}$ perfectly as the sequence lengths become large; while if we selected shape parameter k', the corrected distances would fit $\mathscr{T}'$ perfectly for sufficiently long sequences. Notice that the pair $(\mathscr{T}, k)$ and $(\mathscr{T}', k')$ fit the data produced by either tree (with its associated shape parameter) equally well (i.e. perfectly in the limit as the sequence lengths become large). Moreover, our result assumes that the base frequency vector ($\boldsymbol{\pi}$) is known and the same for both trees. Notice also that Theorem 3.1 automatically implies that any distance correction method that transforms the sequences dissimilarities (the δ values) will be

unable to distinguish between $\mathscr{T}$ and $\mathscr{T}'$ if the shape parameter is unknown.

Proof of Theorem 3.1. For a given assignment of branch lengths l and gamma shape parameter k for $\mathscr{T}$ let J_{ij} denote the induced pairwise distribution matrix, defined by Eq. (7), for each i,j. Similarly, for a given assignment of branch lengths l' and gamma shape parameter k' for $\mathscr{T}'$ let J'_{ij} denote the induced pairwise distribution matrix for each i,j. By symmetry, we may assume (without loss of generality) that $k>k'$. Let

$$\tau_{ij} := 1 + \frac{l_{ij}}{k\gamma}, \tag{8}$$

$$\tau'_{ij} := 1 + \frac{l'_{ij}}{k'\gamma}, \tag{9}$$

and let

$$\rho = \frac{k}{k'} > 1.$$

From Eq. (7) and the notation of (8) and (9), we have the following fundamental identity:

$$J_{ij} = J'_{ij} \quad \text{if and only if} \quad \tau'_{ij} = (\tau_{ij})^{\rho}. \tag{10}$$

We will first prove Theorem 3.1 in the case where $|X| = 4$, and then extend the proof to the general case.

The case $|X| = 4$: Consider the tree $\mathscr{T}$ with branch lengths given in Fig. 1(a), and the tree $\mathscr{T}'$ with branch lengths given in Fig. 1(b). By (1) we have, for example, $l_{12} = l_1 + l_2$, and $l_{13} = l_1 + l_3 + l_5$. Let l'_{ij} be the corresponding l' values induced by $\mathscr{T}'$.

Notice that if we set

$$x_i := \frac{1}{2} + \frac{l_i}{k\gamma} \quad \text{for } i = 1, \ldots, 4, \tag{11}$$

and set

$$\varepsilon := \frac{l_5}{k\gamma}, \tag{12}$$

then for each distinct pair i,j we have

$$\tau_{ij} = \begin{cases} x_i + x_j & \text{if } \{i,j\} = \{1,2\} \text{ or } \{3,4\}, \\ x_i + x_j + \varepsilon & \text{otherwise.} \end{cases} \tag{13}$$

thus τ_{ij} is additive on $\mathscr{T}$ (similarly, τ'_{ij} defined by (9) is additive on $\mathscr{T}'$).

We will describe an assignment of positive branch lengths for $\mathscr{T}$, and then an assignment of branch lengths for $\mathscr{T}'$. Firstly, however we state a convexity lemma; for completeness a proof is provided in Appendix A.

Lemma 3.2. *Suppose f is twice-differentiable, and that f'' is strictly positive on the positive reals. If $u' < u \leqslant v < v'$ and $u + v = u' + v'$ then: $f(u') + f(v') > f(u) + f(v)$.*

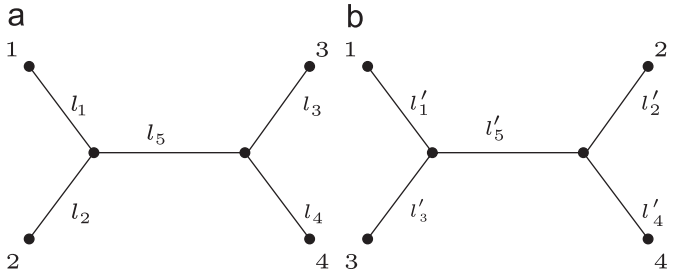

Fig. 1. (a) Tree $\mathscr{T}$ with branch lengths l; (b) tree $\mathscr{T}'$ with branch lengths l'.

We will apply Lemma 3.2 twice during the proof, using the function $f(x) = x^{\rho}$ which satisfies the hypotheses of this lemma, since $f''(x) = \rho(\rho-1)x^{\rho-2}$ and $\rho > 1$.

Returning to the assignment of branch lengths for $\mathscr{T}$, let $L > \frac{1}{2}$ and $s \in (0,1]$, and select $l_1, \ldots, l_5$ so that $(x_1, x_2, x_3, x_4, \varepsilon)$ defined by (11) and (12) satisfy the following system of inequalities:

$$\min\{x_1, x_2, x_3, x_4\} = L, \tag{14}$$

$$x_3 \geqslant x_4 + s, \tag{15}$$

$$x_2 < x_4, \tag{16}$$

$$x_1 + x_3 \leqslant x_2 + x_4, \tag{17}$$

$$x_1 + x_4 \leqslant x_2 + x_3, \tag{18}$$

$$|x_i - x_j| \leqslant 1 \quad \text{for all} \quad i,j, \tag{19}$$

and

$$(x_1 + x_4 + \varepsilon)^{\rho} + (x_2 + x_3 + \varepsilon)^{\rho} = (x_1 + x_2)^{\rho} + (x_3 + x_4)^{\rho}. \tag{20}$$

We pause to observe that this system (for the five l_i values) is feasible for arbitrarily large values of L. For example, we can take $x_1 = L, x_2 = L + \frac{1}{3}, x_3 = L + 1, x_4 = L + \frac{2}{3}$ and $s = \frac{1}{3}$, to satisfy (14)–(19), and then for $i = 1, \ldots, 4$ let $l_i = k\gamma(x_i - \frac{1}{2})$, which is strictly positive since $x_i \geqslant L > \frac{1}{2}$; then for l_5 there exists a positive value of ε satisfying (20). To see this last claim regarding ε, let

$$u = x_1 + x_4, \quad v = x_2 + x_3,$$

and

$$u' = x_1 + x_2, \quad v' = x_3 + x_4.$$

Notice that the inequalities (16) and (18) imply that $u' < u \leqslant v < v'$ and, since $u + v = u' + v'$, Lemma 3.2 applied to $f(x) = x^{\rho}$ gives $f(u) + f(v) < f(u') + f(v')$. Since f is strictly increasing, this implies that there is a finite and strictly positive value of $\varepsilon > 0$ (and thereby of l_5 by (12)) for which $f(u + \varepsilon) + f(v + \varepsilon) = f(u') + f(v')$, as claimed.

Next we show that the branch lengths we have assigned for $\mathscr{T}$ allows us to assign positive branch lengths to $\mathscr{T}'$ so $J_{ij} = J'_{ij}$ holds for all i,j. Define $\lambda_{ij} := f(\tau_{ij})$ where $f(x) = x^{\rho}$. We will show that there exists an assignment of positive branch lengths l' to $\mathscr{T}'$ for which the associated vector τ' defined by (9) satisfies:

$$\lambda_{ij} = \tau'_{ij}. \tag{21}$$

In view of (10) this will establish the theorem in the case $|X| = 4$. Let

$$S'_{12|34} := \lambda_{12} + \lambda_{34}, S'_{13|24} := \lambda_{13} + \lambda_{24} \quad \text{and} \quad S'_{14|23} := \lambda_{14} + \lambda_{23}. \tag{22}$$

If we let

$$u = x_1 + x_3 + \varepsilon, \quad v = x_2 + x_4 + \varepsilon,$$

and

$$u' = x_1 + x_4 + \varepsilon, \quad v' = x_2 + x_3 + \varepsilon,$$

then (15) and (17) imply that $u' < u \leqslant v < v'$ and, since $u + v = u' + v'$, Lemma 3.2 gives $f(u) + f(v) < f(u') + f(v')$. In view of (22) and (13) this implies that:

$$S'_{13|24} < S'_{14|23}. \tag{23}$$

Moreover, Eq. (20) implies that

$$S'_{12|34} = S'_{14|23}. \tag{24}$$

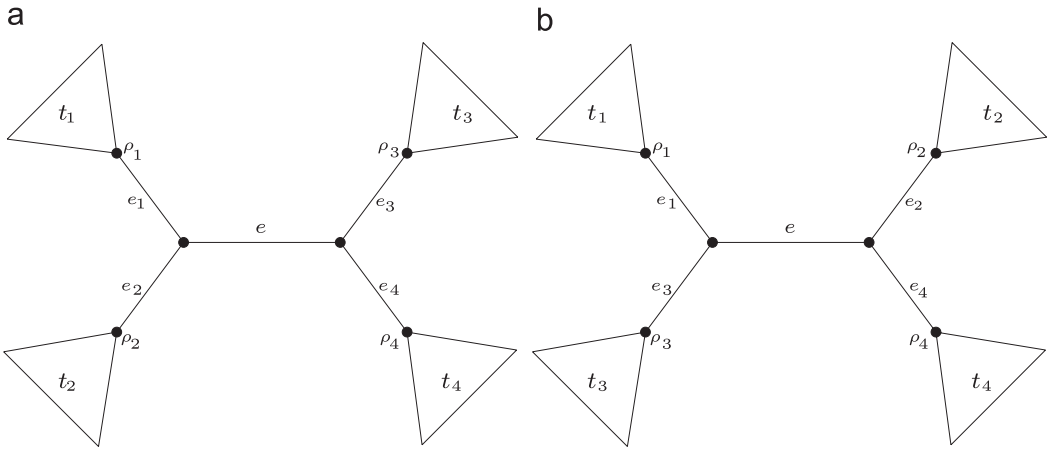

Fig. 2. Representation of $\mathscr{T}$ in (a), and of $\mathscr{T}'$ in (b), when $|X|>4$.

Eqs. (23) and (14) imply that λ_{ij} can be realized as a sum of real-valued branch lengths on $\mathscr{T}'$ by assigning a positive interior branch length (call it ε'), and real-valued (possibly negative) pendant branch lengths (by Hakimi and Patrinos, 1972). We will first show that these four pendant branch lengths are not only positive, but also strictly greater than $\frac{1}{2}$ provided L is chosen sufficiently large. For $i\in\{1,2,3,4\}$ if we let λ_i denote the branch length of the edge incident with leaf i, then $\lambda_i=\frac{1}{2}(\lambda_{ij}+\lambda_{ik}-\lambda_{jk})$ for any choice j,k for which $|\{i,j,k\}|=3$. Now, from (14) and (19), we have

$$\lambda_{ij}+\lambda_{ik}-\lambda_{jk}\geqslant f(2L)+f(2L)-f(2L+2+\varepsilon),$$

and so we can select a value of L that is sufficiently large to ensure that $\lambda_i>\frac{1}{2}$ for $i=1\ldots,4$. We can now assign the positive branch lengths to $\mathscr{T}'$ as follows. Let $l'_5=k'\gamma\varepsilon'$ and for $i\in\{1,\ldots,4\}$ let

$$l'_i=k'\gamma(\lambda_i-\tfrac{1}{2})>0.$$

With these branch lengths we have (from (9) and (21)),

$$\tau'_{ij}=\lambda_{ij}=(\tau_{ij})^{\rho},$$

for all i,j. By Eq. (10), this establishes the theorem in the case where $|X|=4$.

The case $|X|>4$: To extend the proof to larger trees we require a further lemma, which is based on the following definition. Given a rooted phylogenetic tree, t with root vertex ρ (which we assume is a vertex of degree at least two) and associated branch lengths l, we say that the branch-lengths on t are *clock-like* if the sum of the branch lengths from ρ to any leaf takes the same value for each leaf, which we will denote by $h(t,l)$ (the 'height' of ρ). We will use the following lemma, for which a proof is provided in Appendix A.

Lemma 3.3. *Let t be a rooted phylogenetic tree with at least two leaves. Suppose that the branch-lengths for t are clock-like, and that we have a gamma distribution of rates across sites (with mean 1) and with shape parameter k. For any other shape parameter k' there exists a unique associated vector of branch lengths l' for t that are clock-like and such that the induced J' matrices satisfy the condition:*

$$J'_{ij}=J_{ij}\quad \textit{for all leaves } i,j \textit{ of } t. \tag{25}$$

Moreover, for this vector l', we have

$$h(t,l')=\frac{1}{2}k'\gamma\left(-1+\left(1+\frac{2h(t,l)}{k\gamma}\right)^{\rho}\right). \tag{26}$$

Returning to the proof of the theorem, let $\mathscr{T}$ be any binary phylogenetic tree with more than four leaves, and select any interior edge e of $\mathscr{T}$. Consider the four subtrees (each of which could be an isolated leaf) t_1,t_2,t_3,t_4 of $\mathscr{T}$ that result from deleting this edge and its two endpoints, as shown in Fig. 2(a). Let $\mathscr{T}'$ be the tree obtained from $\mathscr{T}$ by interchanging the subtrees t_2 and t_3, as shown in Fig. 2(b). Let l and l' be strictly positive branch lengths for the two quartet trees of Fig. 1 for which we have $J_{ij}=J'_{ij}$ for all $i,j\in\{1,2,3,4\}$ (by the case of the theorem established already for $|X|=4$).

We now assign branch lengths to $\mathscr{T}$ and $\mathscr{T}'$. For tree $\mathscr{T}$ assign length l_5 to edge e indicated in Fig. 2(a), and for the tree $\mathscr{T}'$ assign length l'_5 to edge e of $\mathscr{T}'$ indicated in Fig. 2(b). If t_i consists of just a single leaf, we assign length l_i in $\mathscr{T}$ and l'_i in $\mathscr{T}'$ to edge e_i. Thus it remains to specify how to assign branch lengths to t_i and to the edge e_i that connects t_i to e whenever t_i contains more than one leaf. In that case, if we regard t_i as a rooted binary phylogenetic tree (for which the root ρ_i is the vertex adjacent to an endpoint of e, as shown in Fig. 2), we assign branch lengths to t_i that are clock-like and for which $h(t_i,\rho)=\xi_i$, where ξ_i is any strictly positive number that is less than l_i and which satisfies the condition:

$$\frac{1}{2}k'\gamma\left(-1+\left(1+\frac{2\xi_i}{k\gamma}\right)^{\rho}\right)<l'_i. \tag{27}$$

Then assign edge e_i length $l_i-\xi_i>0$. Note that we can select ξ_i to satisfy (27) since the left-hand side of (27) converges to zero as $\xi_i\to 0$. For tree $\mathscr{T}'$ assign t'_i branch lengths that are clock-like and satisfy (25) of Lemma 3.3 (for $t=t_i,t'=t'_i$), and assign edge e_i length $l'_i-h(t'_i,\rho)$ which is strictly positive by (27). We claim that $J_{ij}=J'_{ij}$ for all i,j. We have just shown that this holds whenever $\{i,j\}$ are leaves in the same subtree (t_1,t_2,t_3 or t_4), thus it remains to check the claim when i and j lie in different subtrees, say t_r,t_s. In this case the condition that $(\mathscr{T},l)$ and $(\mathscr{T}',l')$ satisfy the theorem in the case $|X|=4$ and the fact that the distance between i and j in $\mathscr{T}$ is l_{rs} and in $\mathscr{T}'$ is l'_{rs} (according to the way the branch lengths have been assigned) establishes case (ii). This completes the proof. □

4. Concluding comments

Our result shows that rate variation across sites can indeed provide an "inherent limitation that is worrisome" (Felsenstein, 2003) for methods that rely solely on pairwise sequence comparisons. Despite the limitation of distance-based phylogenetic reconstruction imposed by Theorem 3.1, there is one situation where distances suffice to recover a tree under a gamma rate distribution across sites, even when the shape parameter is

unknown. This is when the underlying branch lengths on the tree are clock-like (i.e. obey a 'molecular clock'). The reason the d-values allow us to reconstruct the tree in this setting is as follows: the clock-like condition is equivalent to requiring that the l_{ij} values are ultrametric and additive on the underlying tree, and, since $d_{ij} = \gamma(1 - M_{\mathscr{D}}(-l_{ij}/\gamma))$, which is a monotonic function of l_{ij} (for any $\mathscr{D}$), the d values will also be ultrametric and additive on the underlying tree.

We note also that our result does not imply that tree reconstruction is hopeless without prior or independent knowledge of the shape parameter, since Allman et al. (2008) have established that identifiability holds for this model (generically for all $r \geqslant 2$, and exactly when $r = 4$ which is the case that applies for DNA sequence data) and so methods such as maximum likelihood will be statistically consistent. Moreover, their result shows that just 3–way sequence comparisons are sufficient to identify the shape parameter. This suggests that it may be possible to develop statistically consistent but fast modifications of distance-based tree reconstruction methods (such as neighbor joining) that some allow triple-wise calculations.

Finally, it would also be interesting to check whether Theorem 3.1 remains true if one replaces the equal input model by the GTR model with any fixed (and given) rate matrix R. This seems quite likely, though the calculations appear to be more involved when the rate matrix has many different eigenvalues. The question of whether $\mathscr{T}'$ can have an arbitrary topology different to $\mathscr{T}$ in Theorem 3.1 (i.e. not just a nearest-neighbor interchange of $\mathscr{T}$) could also be of interest.

Acknowledgments

I thank Joe Felsenstein for several helpful comments, and whose talk at the Sante Fe Institute (April 2008) on a related problem motivated the present study. I also thank the two anonymous referees for their helpful suggestions. This work is supported by the *Allan Wilson Centre for Molecular Ecology and Evolution*.

Appendix A. Proof of Lemmas 3.2 and 3.3

Proof of Lemma 3.2. By a Maclaurin series expansion, we have, for $s = u - u' = v' - v$:

$$f(u') = f(u - s) = f(u) - sf'(u) + \frac{1}{2}s^2 f''(\theta),$$

where $\theta \in [u', u]$ and:

$$f(v') = f(v + s) = f(v) + sf'(v) + \frac{1}{2}s^2 f''(\theta'),$$

where $\theta' \in [v, v']$. Thus:

$$f(u') + f(v') = f(u) + f(v) + s(f'(v) - f'(u)) + \frac{1}{2}s^2(f''(\theta) + f''(\theta')).$$

Now $f'(v) - f'(u) > 0$ since f' is increasing (by the positivity of f''), and so, since $s > 0$:

$$f(u') + f(v') > f(u) + f(v) + \frac{1}{2}s^2(f''(\theta) + f''(\theta')) > f(u) + f(v)$$

where the last inequality follows from the positivity of f''. □

Proof of Lemma 3.3. Since the branch lengths of t are clock-like, it follows that l and hence τ is an ultrametric, i.e. for any three leaves $(i.j, k)$ of t we have

$$\tau_{ij} \leqslant \max\{\tau_{ik}, \tau_{jk}\}.$$

It follows that τ^ρ (where $\rho = k/k'$) satisfies precisely the same ultrametric conditions as τ and so we can assign (unique) positive branch lengths to t that realize τ^ρ and which are clock-like. From (10), these branch lengths satisfy (25) of Lemma 3.3. Moreover, since t has at least two leaves, we can select two leaves (say i, j) so that the path connecting i and j contains ρ. Then $l_{ij} = 2h(t, l)$, and $l'_{ij} = 2h(t, l')$. Now $\tau'_{ij} = (\tau_{ij})^\rho$ and so, by (8) and (9), we have

$$\left(1 + \frac{2h(t, l)}{k\gamma}\right)^\rho = 1 + \frac{2h(t, l')}{k'\gamma},$$

from which equality (26) of Lemma 3.3 now follows. □

References

Allman, E.S., Rhodes, J.A., 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J. Comput. Biol. 13 (5), 1101–1113.
Allman, E.S., Rhodes, J.A., 2008a. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. Math. Biosci. 211 (1), 18–33.
Allman, E.S., Rhodes, J.A., 2008b. The identifiability of covarion models in phylogenetics IEEE/ACM Trans. Comput. Biol. Bioinf., in press.
Allman, E.S., Ané, C., Rhodes, J.A., 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. Appl. Probab. 40 (1), 228–249.
Baake, E., 1998. What can and what cannot be inferred from pairwise sequence comparisons? Math. Biosci. 154, 1–21.
Bandelt, H.J., Fischer, M., 2008. Perfectly misleading distances from ternary characters. Syst. Biol. 57 (4), 540–543.
Chang, J.T., 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137, 51–73.
Evans, S.N., Warnow, T., 2004. Unidentifiable divergence times in rates-across-sites models. IEEE/ACM Trans. Comput. Biol. Bioinfo. 1, 130–134.
Felsenstein, J., 2003. Inferring Phylogenies. Sinauer Press.
Huson, D.H., Steel, M., 2004. Distances that perfectly mislead. Syst. Biol. 53 (2), 327–332.
Hakimi, S.L., Patrinos, A.N., 1972. The distance matrix of a graph and its tree realization. Quart. Appl. Math. 30, 255–269.
Matsen, F.A., Steel, M., 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56 (5), 767–775.
Rogers, J.S., 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Syst. Biol. 50 (5), 713–722.
Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.
Steel, M.A., 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. 7 (2), 19–24.
Steel, M., Penny, D., 2000. Parsimony, likelihood and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17 (6), 839–850.
Steel, M.A., Székely, L., Hendy, M.D., 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comput. Biol. 1 (2), 153–163.
Steel, M.A., Penny, D., Hendy, M.D., 1988. Loss of information in genetic distance. Nature 336 (6195), 118.
Waddell, P.J., Steel, M.A., 1997. General time reversible distances with unequal rates across sites. Mol. Phyl. Evol. 8 (3), 398–414.
Yang, Z., 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.