PERGAMON

Applied Mathematics Letters 12 (1999) 13–19

Applied Mathematics Letters

Approximating Likelihoods Under Low but Variable Rates Across Sites

M. STEEL AND P. J. WADDELL
Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800, Christchurch, New Zealand

(Received and accepted October 1998)

Abstract—We present formulae for estimating the likelihood of sequence data in which sites evolve independently at low but variable rates under a tree-indexed Markov process. These methods can be much faster than those presently employed.

Keywords—Likelihood, Phylogenetic trees, DNA sequences, Markov processes.

1. INTRODUCTION

The use of Markov-style processes to analyse genetic sequences, and reconstruct evolutionary trees is a major enterprise in molecular and evolutionary biology [1]. They seem especially useful when matched with the maximum likelihood criterion. Under these Markov models, sites undergo random mutations along the edges of the evolutionary tree connecting the extant species under study (whose sequences label the leaves of the tree). It is usually assumed that these mutations occur independently between sites, and according to some Markov-style process, which varies between sites only by being scaled up or down according to an associated *rate* parameter λ. At each site, we have a corresponding *site pattern* f which maps each sequence to the corresponding nucleotide which occurs at that site in the sequence. Thus, once the details of the model are described (including the underlying tree T), it is possible to calculate the probability of generating any given site pattern f, if the site evolves at rate λ, and we will denote this probability as: $\mathbb{P}[f|\lambda]$. If we now regard λ as a random variable, then the expected probability of generating f, $\mathbb{P}[f]$ is just $\int \mathbb{P}[f|\lambda]\, dF(\lambda)$. We wish to quickly estimate these probabilities, particularly when the number of sequences n is large. When there is no variation of rates across sites, the calculation of likelihoods is straightforward, using the standard algorithm of [2]. With a continuous distribution of site rates, and for certain classes of models, exact likelihood methods have been independently developed by Yang [3–5], and (using the Hadamard representation) by Steel *et al.* [6], Waddell [7], and Waddell and Penny [8]. Presently, only approximate integrations are practical with large numbers of taxa, such as using a discrete approximating distribution, as suggested by Yang [4]. An alternative approach to dealing with site rate heterogeneity is that of Felsenstein and Churchhill [9] using hidden Markov models. Here, we explore a different type

The authors thank the New Zealand Marsden Fund (M1010) for supporting this research.

0893-9659/99/$ - see front matter. Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
PII: S0893-9659(99)00072-5

of approximation, in which the continuity of the rate distribution is irrelevant to the complexity of the calculations. It depends, instead, on the total expected number of mutations in the tree being small (e.g., sequences from within populations). The approximations we describe here are applicable to all reversible models, but were inspired by an approximation to the Hadamard conjugation approach [6,10]. They may also be useful in estimating ML branch lengths for a tree, as we indicate at the end of this paper.

2. STATIONARY, REVERSIBLE MODELS

Suppose we have a stationary reversible model on r states, with rate matrix R and associated equilibrium vector $\pi = [\pi_\mu]$. Let $p_e^{\mu\nu}(\lambda)$ denote the probability that on edge e, a site evolving at (relative) rate λ is in state ν at one end of the edge, given that it was in state μ at the other end (by reversibility we need not distinguish these two ends). By the spectral theory of reversible Markov processes [11], we have

$$p_e^{\mu\nu}(\lambda) = \sum_{i=1}^{r} c_i^{\mu\nu} \exp(-b_i \gamma_e \lambda), \tag{1}$$

where $\{b_i\}$ are the eigenvalues of R (with $b_1 = 0$) and where γ_e is a characteristic "length" associated with edge e. Consequently, using the expansion $\exp(-x) = \sum_{j=0}^{\infty}((-1)^j/j!)x^j$ and rearranging we get: $p_e^{\mu\nu}(\lambda) = \sum_{j=0}^{\infty} a_j^{\mu\nu}\gamma_e^j\lambda^j$, where $a_j^{\mu\nu} = ((-1)^j/j!)\sum_{i=1}^{r} c_i^{\mu\nu}b_i^j$. Note that $a_0^{\mu\nu} = 1$ if $\mu = \nu$, otherwise $a_0^{\mu\nu} = 0$. For brevity we will let a_j^{μ} denote $a_j^{\mu\mu}$. Thus,

$$p_e^{\mu\mu}(\lambda) = 1 + a_1^{\mu}\gamma_e\lambda + a_2^{\mu}\gamma_e^2\lambda^2 + \cdots + a_k^{\mu}\gamma_e^k\lambda^k + O\left(\gamma_e^{k+1}\right), \tag{2}$$

$$p_e^{\mu\nu}(\lambda) = a_1^{\mu\nu}\gamma_e\lambda + a_2^{\mu\nu}\gamma_e^2\lambda^2 + \cdots + a_k^{\mu\nu}\gamma_e^k\lambda^k + O\left(\gamma_e^{k+1}\right), \qquad \mu \neq \nu. \tag{3}$$

Let m_p denote the p^{th} moment of the rate distribution. We may assume that this distribution has mean 1. Note that γ_e is proportional to the expected number of substitutions on edge e per site, and (with $m_1 = 1$) the constant of proportionality is $-\sum_\mu \pi_\mu R_{\mu\mu}$. Thus, $m_0 = m_1 = 1$, $m_2 = \sigma^2 + 1$ where σ^2 is the variance of the rate distribution, and $M(x) = \sum_{p\geq 0}(m_p/p!)x^p$ is the associated moment generating function. Now, suppose site pattern f assigns state μ to sequence 1. Then, the probability $P[f|\lambda]$ of site pattern f evolving on the tree under a corresponding site rate λ is given by $\mathbb{P}[f|\lambda] = \pi_\mu \sum_F \prod_{e=(u,v)} p_e^{F(u)F(v)}(\lambda)$, where the summation is over all extension F of f to all vertices of the tree. Thus, from equation (1), $\mathbb{P}[f|\lambda] = \sum_{i=0}^{\infty} u_i(f)\lambda^i$ for values $u_i(f)$ that depend on T, R, and $\{\gamma_e\}$. Thus, $\mathbb{P}[f] = \sum_{i=0}^{\infty} u_i(f)m_i$. Let $\mathbb{P}^{(k)}[f] := \sum_{i=0}^{k} u_i(f)m_i$, which we will call the *k^{th}-order approximation to* $\mathbb{P}[f]$, since from equations (2),(3), we have $\mathbb{P}[f] = \mathbb{P}^{(k)}[f] + O(\gamma^{k+1})$ where $\gamma := \max_e\{\gamma_e\}$. For each k, these sum to 1 as we show formally now.

LEMMA 1. *For each $k \geq 1$, $\sum_f \mathbb{P}^{(k)}[f] = 1$.*

PROOF. Let F_0 denote the set of site patterns that assign the same state to all leaves in the tree. Then, $\sum_{f\in F_0} u_0(f)m_0 = 1$, while for any $f \notin F_0$, we have $u_0(f) = 0$. Thus, $\sum_{f\notin F_0}(\sum_{i=1}^{\infty} u_i(f)m_i) = 0$. Rearranging the order of summation, we see that $\sum_{i=1}^{\infty}\beta_i m_i = 0$, where $\beta_i = \sum_{f\notin F_0} u_i(f)$. Thus, $0 = \beta_1 = \beta_2 = \beta_3 = \cdots$ since we can take $m_i = \alpha^{1-i}$ for variable $\alpha : 0 < \alpha < 1$ by selecting the rate distribution for which $1-\alpha$ proportion of sites evolve at 0 rate, and α proportion evolve at rate α^{-1} (thus, $m(x) = 1 - \alpha + \alpha\exp(-x/\alpha)$). Hence, $\sum_f \mathbb{P}^{(k)}[f] = \sum_f \sum_{i=0}^{k} u_i(f)m_i = \sum_{f\in F_0} u_0(f) + \sum_{i=0}^{k}\beta_i m_i = 1$. ∎

We now describe a quadratic approximation to $\mathbb{P}[f]$. In principle, higher-order approximations are possible, but they become increasingly complicated to describe (however, Theorem 2 provides a closed form power series for $\mathbb{P}[f]$ for a certain class of models). Let $S = S^{(1)} := \sum_e \gamma_e$; $S^{(2)} := \sum_e \gamma_e^2$. Given site pattern f, let $L(f,T)$ denote the *parsimony score* of f on T—that is, the minimal number of edges that must be assigned different states at their endpoints in order to extend f to all vertices of T.

THEOREM 1. *We distinguish four cases, depending on $L = L(f,T)$.*

1. *$L = 0$. In this case, f is unvaried, and so assigns a single state—say μ—to all the species. Then,*

$$\mathbb{P}^{(2)}[f] = \pi_\mu \left(1 + a_1^\mu S + a_2^\mu m_2 S^{(2)} + \frac{1}{2} m_2 (a_1^\mu)^2 \left(S^2 - S^{(2)}\right)\right).$$

2. *$L = 1$. In this case, f can be generated by a change of state on just one edge e of T, say from state μ on one side (R = right) of e to a different state ν on the other side (L = left). Without loss of generality, we may suppose that there are two edges e_1, e_2 on the R-side of e each incident with e. If e is an internal edge, then there are also two edges e_3, e_4 on the L-side of e each incident with e. Then,*

$$\mathbb{P}^{(2)}[f] = \pi_\mu a_1^{\mu\nu}\gamma_e + \pi_\mu a_2^{\mu\nu} m_2 \gamma_e^2 + \pi_\mu a_1^{\mu\nu} m_2 \gamma_e \left(\sum_{e' \in R} a_1^\mu \gamma_{e'} + \sum_{e' \in L} a_1^\nu \gamma_{e'}\right) + a_1^{\mu\nu} a_1^{\nu\mu} m_2 (\pi_\mu \gamma_{e_1}\gamma_{e_2} + \pi_\nu \gamma_{e_3}\gamma_{e_4}).$$

3. *$L = 2$. In this case, f can be generated by changes of states on just two edges (however, there may be more than just one such pair of edges).*

$$\mathbb{P}^{(2)}[f] = \pi_\mu m_2 \sum a_1^{\mu\nu} a_1^{\nu\nu'} \gamma_{e_1}\gamma_{e_2},$$

where the summation is over the (at most three) pairs of edges e_1, e_2 on which f can be generated at the leaves of T by inducing a change from μ to ν on edge e_1 and a change ν to ν' on edge e_2.

4. *$L > 2$. In this case, $\mathbb{P}^{(2)}[f] = 0$.*

PROOF.

CASE 1. We have $\mathbb{P}[f|\lambda] = \pi_\mu \prod_e p_e^{\mu\mu}(\lambda) + O(\gamma^3) = \pi_\mu \prod_e (1 + a_1^\mu \gamma_e \lambda + a_2^\mu \gamma_e^2 \lambda^2 + \cdots) + O(\gamma^3)$. Expanding the product term, and integrating according to the rate distribution we get: $\mathbb{P}^{(2)}[f] = \pi_\mu(1 + a_1^\mu S + a_2^\mu m_2 S^{(2)} + (a_1^\mu)^2 m_2 \sum_{\{e,e'\}: e \neq e'} \gamma_e \gamma_{e'})$ and noting that the last summation term is just $1/2(S^2 - S^{(2)})$, we obtain the result.

CASE 2. We have

$$\begin{aligned}\mathbb{P}[f|\lambda] = {} & \pi_\mu p_e^{\mu\nu}(\lambda) \prod_{e' \in R} p_{e'}^{\mu\mu}(\lambda) \prod_{e' \in L} p_{e'}^{\nu\nu}(\lambda) \\ & + \pi_\mu p_{e_1}^{\mu\nu}(\lambda) p_{e_2}^{\nu\mu}(\lambda) p_e^{\nu\nu}(\lambda) \prod_{e' \in R_1} p_{e'}^{\mu\mu}(\lambda) \prod_{e' \in R_2} p_{e'}^{\mu\mu}(\lambda) \prod_{e' \in L} p_{e'}^{\nu\nu}(\lambda) \\ & + \pi_\nu p_{e_3}^{\nu\mu}(\lambda) p_{e_4}^{\mu\nu}(\lambda) p_e^{\mu\mu}(\lambda) \prod_{e' \in L_1} p_{e'}^{\nu\nu}(\lambda) \prod_{e' \in L_2} p_{e'}^{\nu\nu}(\lambda) \prod_{e' \in R} p_{e'}^{\mu\mu}(\lambda) + O\left(\gamma^3\right),\end{aligned}$$

where L_1, L_2 are the two subtrees of L incident with edges e_1, e_2 while R_1, R_2 are the two subtrees of R incident with e_3, e_4. The result now follows from equations (2),(3).

Cases (3) and (4) are straightforward. ∎

3. A POWER SERIES EXPANSION

We now describe a generic power series expansion which covers all trees, and handles a generalization of the Kimura 3ST model that allows nonstationary rate matrices, and nonstationary base composition. It also applies to submodels of this model, including the Kimura 3ST, 2ST, and Jukes-Cantor (= Neyman 4-state) models [1]. An analogous, and simpler treatment exists for the Neyman 2-state model. Under the ordinary Kimura 3ST model, four site patterns now

have equal probability (assuming a uniform distribution of states). Adding the probabilities of these site patterns together gives a vector s with 4^{n-1} components. From [6], we can write this vector (even for the nonstationary generalizations) as $s = H^{-1}(M(H\gamma))$ where H is a Hadamard matrix, and where γ is the "tree spectrum", most of whose entries are 0. We index each component of s and γ and each row and column of H by a *quadripartition* of $\{1,\ldots,n-1\}$—that is, a pair $\theta = (\sigma_1, \sigma_2)$ of subsets of $\{1,\ldots n-1\}$ as in [12]. For example, $s_{(\emptyset,\emptyset)}$ is the expected proportion of sites where all sequences take the same state (the "unvaried" sites). Let $s(T) := \{\rho^{(i)} : \rho \in T,\ i = 1,2,3\}$, where $\rho^{(1)} = (\rho, \emptyset)$, $\rho^{(2)} = (\emptyset, \rho)$, and $\rho^{(3)} = (\rho, \rho)$. Then, $\gamma_\theta = 0$ for $\theta \notin s(T) \cup \{\emptyset,\emptyset\}$, $\gamma_{(\emptyset,\emptyset)} = -\sum_{\theta \in s(T)} \gamma_\theta$.

THEOREM 2.

$$s_\theta = \delta_\theta + \sum_{p\geq 1}(-1)^p \frac{m_p}{p!} \sum_{(\theta_1,\ldots,\theta_p):\forall i,\theta_i \in s(T)} \nu_\theta(\theta_1,\ldots,\theta_p) \prod_{i=1}^{p} \gamma_{\theta_i},$$

where

$$\nu_\theta(\theta_1,\ldots,\theta_p) = \sum_{S\subseteq\{1,\ldots,p\}:\theta=\oplus_{i\in S}\theta_i} (-1)^{|S|}, \qquad \delta_\theta = \begin{cases} 1, & \text{if } \theta = (\emptyset,\emptyset), \\ 0, & \text{else.} \end{cases}$$

PROOF. By definition, $s_\theta = (H^{-1}M(H\gamma))_\theta = 4^{1-n}\sum_{\theta'} h_{\theta,\theta'} r_{\theta'}$ where $r_{\theta'} = M((H\gamma)_{\theta'}) = \sum_{p\geq 0}(m_p/p!)(H\gamma)^p_{\theta'}$. Thus,

$$s_\theta = 4^{1-n}\left(\sum_{\theta'} h_{\theta,\theta'} + \sum_{p\geq 1}\frac{m_p}{p!}\sum_{\theta'} h_{\theta,\theta'}(H\gamma)^p_{\theta'}\right) = \delta_\theta + \sum_{p\geq 1}\frac{m_p}{p!}B_{\theta,p},$$

where $B_{\theta,p} = 4^{1-n}\sum_{\theta'} h_{\theta,\theta'}(H\gamma)^p_{\theta'}$. Now, $(H\gamma)_{\theta'} = -\sum_{\theta''\in s(T)}(1 - h_{\theta',\theta''})\gamma_{\theta''}$. Thus,

$$B_{\theta,p} = (-1)^p 4^{1-n}\sum_{\theta'} h_{\theta,\theta'}\left(\sum_{\theta''\in s(T)}(1-h_{\theta',\theta''})\,\gamma_{\theta''}\right)^p$$

$$= (-1)^p \sum_{(\theta_1,\ldots,\theta_p):\forall i,\theta_i\in s(T)} \prod_{i=1}^{p}\gamma_{\theta_i}\left(4^{1-n}\sum_{\theta'} h_{\theta,\theta'}\prod_{j=1}^{p}(1-h_{\theta',\theta_j})\right).$$

Now, we invoke the identity

$$4^{1-n}\sum_{\theta'} h_{\theta,\theta'}\prod_{i=1}^{p} h_{\theta',\theta_i} = \begin{cases} 1, & \text{if } \theta = \oplus_{i=1}^{p}\theta_i, \\ 0, & \text{else,} \end{cases}$$

to deduce that $4^{1-n}\sum_{\theta'} h_{\theta,\theta'}\prod_{j=1}^{p}(1-h_{\theta',\theta_j}) = \nu_\theta(\theta_1,\ldots,\theta_p)$ and the result now follows. ∎

Theorem 2 leads to the following first-order and second-order approximations for s_θ. Let $S := \sum_{\theta\in s(T)}\gamma_\theta, S^{(2)} := \sum_{\theta\in s(T)}\gamma_\theta^2$, $s(T)^2 = \{\{\theta_1,\theta_2\} : \theta_1,\theta_2 \in s(T)\}, s(T)^{(2)} = \{\theta_1 \oplus \theta_2 : \theta_1,\theta_2 \in s(T)\}$ and recall that $m_2 = 1 + \sigma^2$. Then we have the first-order approximation

$$s^{(1)}_\theta = \begin{cases} \gamma_\theta, & \text{if } \theta \in s(T), \\ 1 - \sum_{\theta\in s(T)}\gamma_\theta, & \text{if } \theta = (\emptyset,\emptyset), \\ 0, & \text{else,} \end{cases}$$

and the second-order approximation

$$s^{(2)}_\theta = \begin{cases} \gamma_\theta - m_2\gamma_\theta S + m_2 \sum_{\{\theta_1,\theta_2\}\in s(T)^2:\theta_1\oplus\theta_2=\theta}\gamma_{\theta_1}\gamma_{\theta_2}, & \text{if } \theta \in s(T), \\ 1 - S + \frac{1}{2}m_2\left(S^2 + S^{(2)}\right), & \text{if } \theta = (\emptyset,\emptyset), \\ m_2 \sum_{\{\theta_1,\theta_2\}\in s(T)^2:\theta_1\oplus\theta_2=\theta}\gamma_{\theta_1}\gamma_{\theta_2}, & \text{if } \theta \in s(T)^{(2)}, \\ 0, & \text{else.} \end{cases}$$

Moreover, in either approximation, the s_θ sum to 1 by virtue of the following.

THEOREM 3. *Let $s_\theta^{(r)}$ denote the r^{th}-order approximation for s_θ obtained by allowing p in Theorem 1 to range from 0 to r. Then*

$$\sum_\theta s_\theta^{(r)} = 1.$$

PROOF. Let $m(x) := \sum_{p=0}^{r}(m_p/p!)x^p$. Then, $s_\theta^{(r)} = (H^{-1}m(H\gamma))_\theta$ and since the elements in the vector $H^{-1}m(H\gamma)$ sum to $m(0) = 1$, so, too, do the elements $s_\theta^{(r)}$. ∎

3.1. Example: The Neyman 2-State Model

To illustrate these formulae, we consider the simple case of the Neyman 2-state model [13], in which there are just two states, and the rate matrix is symmetric. Such a model can be regarded as either a stationary reversible model (in which case Theorem 1 applies) or the simpler analogue of Theorem 2 (for the Neyman 2-state model) could alternatively be invoked. We have $p_e^{\mu\mu}(\lambda) = (1/2)(1 + \exp(-2\lambda\gamma_e))$, and so $a_1^\mu = -1$, $a_2^\mu = 1$. For $\mu \neq \nu$, we have $p_e^{\mu\nu}(\lambda) = (1/2)(1 - \exp(-2\lambda\gamma_e))$ and so, $a_1^{\mu\nu} = 1$, $a_2^{\mu\nu} = -1$. We specialise further now, and consider a fully resolved tree with just $n = 4$ leaves, labelled $1, \ldots, 4$ in which leaves 1,2 are separated from 3,4 by a central edge. Assign length γ_i to the edge e_i incident with leaf i, and assign length γ_5 on the central edge. In this example, we let p_i be the probability of generating the partition induced by a single mutation on edge e_i, and let p_{ij} be the probability of generating the partition induced by single mutations on edges e_i and e_j. Let s_0 denote the uniform pattern (all leaves in the same state). Thus, $s_{12} = s_5$ is the probability that leaves 1,2 are in one state, and leaves 3,4 are in the other state. Note that each s value is a sum of precisely two equal $\mathbb{P}[f]$ values. Then, we have the following second-order approximation to the s values. As before, let $S := \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5$ and $\gamma := \max_e\{\gamma_e\}$. Let

$$\begin{aligned}
s_1^{(2)} &= \gamma_1 - m_2(\gamma_1 S - \gamma_2\gamma_5),\\
s_2^{(2)} &= \gamma_2 - m_2(\gamma_2 S - \gamma_1\gamma_5),\\
s_3^{(2)} &= \gamma_3 - m_2(\gamma_3 S - \gamma_4\gamma_5),\\
s_4^{(2)} &= \gamma_4 - m_2(\gamma_4 S - \gamma_3\gamma_5),\\
s_5^{(2)} &= \gamma_5 - m_2(\gamma_5 S - \gamma_1\gamma_2 - \gamma_3\gamma_4),\\
s_{13}^{(2)} &= m_2(\gamma_1\gamma_3 + \gamma_2\gamma_4),\\
s_{14}^{(2)} &= m_2(\gamma_1\gamma_4 + \gamma_2\gamma_3),\\
s_0^{(2)} &= 1 - S + m_2\sum_\mu \gamma_\mu^2 + m_2 \sum_{\{\mu,\nu\}:\mu\neq\nu} \gamma_\mu\gamma_\nu.
\end{aligned}$$

Then, $s_\mu = s_\mu^{(2)} + O(\gamma^3)$ and $\sum_\mu s_\mu^{(2)} = 1$.

3.2. Numerical Example

Let the 4-species tree described above have $[\gamma_1, \ldots, \gamma_5] = [0.01, 0.02, 0.03, 0.04, 0.05]$, while the site rates follow a gamma distribution with mean = 1 and shape parameter $k = 2$ (so $\sigma^2 = 1/2$). Then, as a test example, under the Neyman 2-state model the approximate calculation gives the site pattern vector

$$[s_0 : 0.87100, s_1 : 0.00925, s_2 : 0.01625, s_3 : 0.02625,$$
$$s_4 : 0.03325, s_5 : 0.04085, s_{13} : 0.00165, s_{14} : 0.00150],$$

while the exact probabilities are

$$[s_0 : 0.86858, s_1 : 0.00929, s_2 : 0.01681, s_3 : 0.02658,$$
$$s_4 : 0.03410, s_5 : 0.04204, s_{13} : 0.00135, s_{14} : 0.00125].$$

Here, even though the most divergent sequences are 11% apart (over twice the average divergence between human and chimpanzee genomes, e.g., [14]) the probabilities are still quite close. A useful measure of the difference of these sequence probability vectors is via the likelihood ratio statistic: $G^2 = \sum_i cs_i \ln(s_i^{(2)}/s_i)$, where c is the sequence length and s_i are the exact probabilities. This is also equal to the expected difference in log likelihood ($\ln L$). Assuming the sequences are 1000 sites long, then for the model given above, the difference in the likelihood of the whole data due to the approximations is only 0.0963 (out of a total $\ln L$ of -596.65). In a real situation, such a small difference would be insignificant. The Pearson X^2 statistic gives similar results. To evaluate how much the tree shape, the distribution of site rates, and the overall rate of evolution affect the accuracy of these calculations, we present the results in Figure 1. The tree is alternatively one obeying equal rates of evolution $\gamma_i = 0.1$ for all i, or a Felsenstein [15] tree with γ_i values [0.05, 0.175, 0.175, 0.05, 0.05] where rates of evolution in different lineages are highly unequal (note, the sum of edge lengths is equal in both cases). Site rates are assumed to be either identical rates (mean = 1, $\sigma^2 = 0$), or to be highly unequal in rate and follow a gamma distribution (mean 1 and $\sigma^2 = 1$). The x axis is equal to the total percentage evolution across the tree (e.g., at 10%, 10% of all sites are expected to have changed).

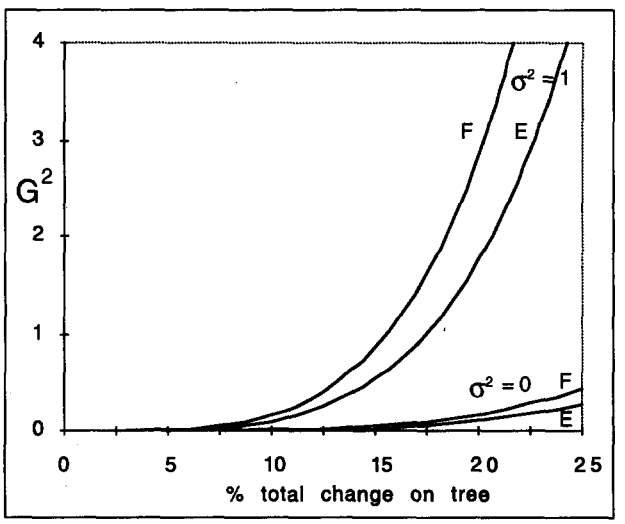

Figure 1. A plot of S (the total expected number of substitutions) on the model tree versus difference of log likelihood for sequences 1000 long (E: equal edge rates, F: Felsenstein tree).

Figure 1 shows the approximations are very close up to moderate rates of substitution, when site rates are identical. Note that while the shape of the tree makes some difference, the variance of site rates is overall more important. However, even under the worst combination, the calculated likelihoods are quite accurate at divergences found within species, between closely related species, and for some types of change (e.g., transversions) between species in different families (e.g., humans to monkeys). If we use other site rate distributions in place of the gamma (e.g., the inverse Gaussian [10]), then we get similar results.

3.3. Approximating ML Branch Lengths

In maximum likelihood analyses, one frequently seeks branch lengths (and/or parameters) that maximize the likelihood score on a given tree (given a collection of sequences, and a model of site substitution). It is, therefore, useful to have good estimates of these optimal branch lengths in order to speed up any search algorithm, and one such approach was described recently in [16]. Here, we point out that the approximations described above can also provide initial settings for maximum likelihood algorithms. Consider for example, the first-order approximation for the

Kimura 3ST model, described in Section 3. Then, the settings of the γ_θ for $\theta \in s(T)$ that maximize the likelihood score when each $s_\theta^{(1)}$ is substituted for s_θ (and sites of parsimony length > 1 are ignored) is provided by setting $\gamma_\theta = \max\{0, (x_\theta / x_{(\emptyset,\emptyset)}) - c\}$ where x_θ is the number of sites inducing the quadripartition θ, and where $c = (|s(T)| + 1)^{-1} \sum_{\theta \in s(T)} (x_\theta / x_{(\emptyset,\emptyset)})$.

REFERENCES

1. D.L. Swofford, G.J. Olsen, P.J. Waddell and D.M. Hillis, Phylogenetic inference, In *Molecular Systematics*, 2nd Edition, (Edited by D.M. Hillis, C. Moritz and B.K. Marble), pp. 407–514, Sinauer Associates, (1996).
2. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach, *J. Mol. Evol.* **17**, 368–376, (1981).
3. Z. Yang, Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Mol. Biol. Evol.* **10**, 1396–1401, (1993).
4. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods, *J. Mol. Evol.* **39**, 306–314, (1994).
5. Z. Yang, Among-site rate variation and its impact on phylogenetic analyses, *Trends in Ecology and Evolution* **11**, 367–372, (1996).
6. M.A. Steel, L.A. Székely, P.L. Erdős and P.J. Waddell, A complete set family of phylogenetic invariants for any number of taxa under Kimura's 3ST model, *N. Z. J. Bot.* **31**, 289–296, (1993).
7. P.J. Waddell, Statistical methods of phylogenetic analysis, Ph.D. Thesis, Massey University, Palmerston North, New Zealand, (1995).
8. P.J. Waddell and D. Penny, Evolutionary trees of apes and humans from DNA sequences, In *Handbook of Human Symbolic Evolution*, (Edited by A.J. Lock and C.R. Peters), Clarendon Press, Oxford, (1996).
9. J. Felsenstein and G.A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, *Mol. Biol. Evol.* **13**, 93–104, (1996).
10. P.J. Waddell, D. Penny and T. Moore, Hadamard conjugations and modeling sequence evolution with variable rates across sites, *Mol. Phyl. Evol.* **8**, 33–50, (1997).
11. J. Keilson, Markov chains models: Rarity and exponentiality, In *Applied Mathematical Sciences*, Volume 28, Springer-Verlag, New York, (1979).
12. M.A. Steel, M.D. Hendy, L.A. Székely and P.L. Erdős, Conjugate spectra and a closest tree method for genetic sequences, *Appl. Math. Lett.* **5** (6), 63–67, (1992).
13. J. Neyman, Molecular studies of evolution: A source of novel statistical problems, In *Statistical Decision Theory and Related Topics*, (Edited by S.S. Gupta and J. Yackel), pp. 1–27, Academic Press, New York, (1971).
14. W.J. Bailey, K. Hayasaka, C.G. Skinner, S. Kehoe, L.C. Sieu, J.L. Slightom and M. Goodman, Reexamination of the African Hominoid trichotomy with additional sequences from the primate b-globin gene cluster, *Mol. Phyl. Evol.* **1**, 97–135, (1992).
15. J. Felsenstein, Cases in which parsimony or compatibility will be positively misleading, *Syst. Zool.* **27**, 401–410, (1978).
16. J.S. Rogers and D.L. Swofford, A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences, *Syst. Biol.* **47** (1), 77–89, (1998).