Syst. Biol. 50(5):713–722, 2001

Maximum Likelihood Estimation of Phylogenetic Trees Is Consistent When Substitution Rates Vary According to the Invariable Sites plus Gamma Distribution

JAMES S. ROGERS

Department of Biological Sciences, University of New Orleans, New Orleans, Louisiana 70148, USA;
E-mail: jsrogers@uno.edu

Abstract.—Maximum likelihood estimation of phylogenetic trees from nucleotide sequences is completely consistent when nucleotide substitution is governed by the general time reversible (GTR) model with rates that vary over sites according to the invariable sites plus gamma ($I + \Gamma$) distribution. [Consistency; general time reversible model; maximum likelihood; phylogeny estimation; rate heterogeneity.]

The maximum likelihood (ML) method of estimating phylogenetic trees from discrete characters such as nucleotide sequences has been widely advocated and used by many workers (e.g., Felsenstein, 1973, 1978, 1981; Huelsenbeck and Hillis, 1993; Swofford et al., 1996), in part because it is believed to have the statistical property of consistency. In the context of phylogeny reconstruction this means that, given the correct model of character change (except for the values of a finite number of parameters such as tree topology, branch lengths, relative substitution rates, and so forth, which are to be estimated by ML), the method will converge with certainty on the correct tree as the number of characters (nucleotide sites) increases without limit. Felsenstein (1973) based his argument for consistency of ML tree estimation on Wald's (1949) general proof of the consistency of ML estimates. Several workers (e.g., Nei, 1987:325; Saitou, 1988: Yang, 1994, 1996a; Yang et al., 1995; Russo et al., 1996; Siddall, 1998; Farris, 1999) have expressed misgivings about this argument, concerned that the discrete, unordered nature of a tree topology "variable" prevents it from being the sort of parameter required by Wald's (1949) proof. Their concern seems to arise from the mistaken assumption that Wald's criteria for the consistency of ML estimation include "that the likelihood function is everywhere continuous and continuously differentiable with respect to the parameter of interest" (Siddall, 1998). If this were true, a discrete variable such as tree topology obviously would not qualify. But, as I will show in the Appendix of this paper, that assumption is groundless (also see Swofford et al., in press). Further, Chang's (1996) proof of the consistency of ML tree reconstruction, which he calls "a customized variant of the fundamental consistency result of Wald," explicitly treats topology as one of the parameters. In the Appendix I present a somewhat different adaptation of Wald's (1949) assumptions and proof of the consistency of ML estimation that make the assumptions conform to the requirements of phylogenetic tree estimation.

A key assumption in Wald's (1949) proof is that all of the parameters of the likelihood model are *identifiable* from the true probability distribution of the data. For phylogenetic reconstruction from sequence data, this means that only one combination of tree topology, branch lengths, substitution model parameters, and so forth could generate any particular set of expected nucleotide site pattern frequencies. Chang (1996) demonstrated that the tree topology and the nucleotide substitution matrices for all branches of the tree are identifiable for very general substitution models that are not assumed to be stationary or time reversible. However, his proof is not completely general for continuous-time Markov models (see Chang, 1996:66). Rogers (1997) gave a proof of the identifiability of the tree topology for continuous-time, reversible Markov models. Both of these proofs assume that rates of substitution are uniform over all sites of a sequence, which is known not to be generally true (e.g., Yang, 1996b; Sullivan et al., 1999). Steel et al. (1994) showed that the true tree is identifiable from infinite sequence data with rate heterogeneity over sites if (1) the distribution of relative rate parameters is known, or (2) rate heterogeneity is restricted to a certain unknown proportion of invariable sites, with the remainder of sites evolving at the same rate, or (3) a molecular clock is in effect. Some authors (e.g., Farris, 1999) apparently have concluded that these conditions are *necessary* for identifiability. That is incorrect. Steel et al. (1994) simply demonstrated that they are *sufficient* for identifiability—which leaves open the possibility of identifiability under other conditions.

The Steel et al. (1994) demonstration of identifiability is restricted to the Cavender–Farris substitution model for two-state characters and the Kimura three-substitution-types model for four-state sequences. Both of these are equal-frequency models; that is, both assume that the stationary frequencies of all character states are equal. However, that is rarely ever true (e.g., Hasegawa et al., 1985).

In this paper I demonstrate the identifiability of all parameters, including topology, branch lengths, substitution rates, and rate heterogeneity parameters for the stationary, general time reversible (GTR) substitution model with rate heterogeneity determined by the invariable sites plus gamma distribution (I+Γ) model.

Nucleotide Site Patterns

As noted in the Appendix, the basic random variable for ML estimation of phylogenetic trees from sequence data is the nucleotide site pattern, that is, the ordered set of nucleotides observed at a given site of the sequences under consideration. For example, the four short sequences shown in Figure 1 have four distinct site patterns: AAAA, AATT, ATAT, and ATCG. Using the symbolism of the Appendix, we can designate these four patterns X_1, X_2, X_3, and X_4. In four very long sequences, each of these distinct patterns would occur with a certain frequency. If the four sequences were infinitely long, the observed frequencies of these four patterns would equal their expected frequencies, $f(X_1, \theta_0)$,$f(X_2, \theta_0)$,$f(X_3, \theta_0)$, and $f(X_4, \theta_0)$, where, again as explained in the Appendix, θ_0 is the vector of true parameter values, including the true tree topology. All other possible distinct site patterns for four sequences would have similar frequencies. Using this symbolism, the question of identifiability is whether the set of true frequencies $f(X_1, \theta_0)$,$f(X_2, \theta_0)$, $f(X_3, \theta_0)$,$\ldots$, $f(X_N, \theta_0)$ of all N possible distinct site patterns completely determines θ_0.

	Sites			
	1	2	3	4
Sequences				
1	A	A	A	A
2	A	A	T	T
3	A	T	A	C
4	A	T	T	G

FIGURE 1. Four hypothetical nucleotide sequences of four sites each.

The GTR Model of Nucleotide Substitution with Rate Heterogeneity

In the notation of Swofford et al. (1996), the GTR model can be expressed by a 4 × 4 stochastic matrix **Q** such that Q_{ij} is the instantaneous rate of replacement of nucleotide i by nucleotide j for $i \neq j$, and $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. **Q** has the additional property that, if $\Pi = \text{diag}(\pi_A, \pi_C, \pi_G, \pi_T)$ is the diagonal matrix of stationary frequencies of the four nucleotides, then $\Pi\mathbf{Q} = \mathbf{Q}^T\Pi$, that is, **Q**, is reversible (Lanave et al., 1984; Rodríguez et al., 1990). Reversibility allows us to work with unrooted trees.

The matrix of substitution probabilities over time τ for nucleotide sites with relative substitution rate r, $\mathbf{P}\ (r\tau)$, can be found from **Q** by the equation

$$\mathbf{P}(r\tau) = \exp(r\tau\mathbf{Q}), \qquad (1)$$

where $P_{ij}(r\tau)$ is the probability that nucleotide i will be replaced by nucleotide j after τ units of time at relative rate r. The matrix $\mathbf{F}(r\tau)$, where $F_{ij}(r\tau)$ is the expected frequency of finding the pair of nucleotides i and j at a site with relative substitution rate $r$in two sequences separated by time τ, is given by

$$\mathbf{F}(r\tau) = \Pi\mathbf{P}(r\tau), \qquad (2)$$

Conversely,

$$\mathbf{P}(r\tau) = \Pi^{-1}\mathbf{F}(r\tau). \qquad (3)$$

The matrix of substitution probabilities for all sites, $\bar{\mathbf{P}}$, can be found by taking the expectation with respect to r, that is,

$$\bar{\mathbf{P}}(\tau) = \mathbf{E}[\mathbf{P}(r\tau)] = \int_{\text{all}\, r} \phi(r)\mathbf{P}(r\tau)\, dr, \qquad (4)$$

where $\phi(r)$ is the probability or probability density of rate r. The function $\phi(r)$ may be discrete, continuous, or a mixture. Correspondingly, the "integral" of Eq. 4 may be a summation over discrete values of r, a definite integral over a continuous range of r values, or a mixture. Then, from Eqs. 1, 3, and 4,

$$\bar{\mathbf{P}}(\tau) = \mathbf{E}[\exp(r\tau\mathbf{Q})], \qquad (5)$$

and

$$\bar{\mathbf{P}}(\tau) = \Pi^{-1}\bar{\mathbf{F}}(\tau) = \Pi^{-1}\mathbf{E}[\mathbf{F}(r\tau)]. \qquad (6)$$

Because $\mathbf{Q}$ is a stochastic matrix, it is diagonalizable; that is, it can be represented as $\mathbf{Q} = \mathbf{U}\Lambda\mathbf{U}^{-1}$, where $\Lambda = \mathbf{diag}(0, -\lambda_2, -\lambda_3, -\lambda_4)$ is the diagonal matrix of eigenvalues of $\mathbf{Q}$, with $\lambda_i > 0$ for $i = 2, 3, 4$. $\mathbf{U}$ is the matrix of associated eigenvectors. Making this substitution, Eq. 5 becomes

$$\begin{aligned}\bar{\mathbf{P}}(\tau) &= \mathbf{E}[\mathbf{U}\exp(r\tau\,\mathbf{L})\,\mathbf{U}^{-1}]\\ &= \mathbf{U}\mathbf{diag}(1, \mathbf{E}[e^{-r\tau\lambda_2}], \mathbf{E}[e^{-r\tau\lambda_3}],\\ &\quad \mathbf{E}[e^{-r\tau\lambda_4}])\mathbf{U}^{-1}. \qquad (7)\end{aligned}$$

Thus, $\mathbf{U}$ is the matrix of eigenvectors of $\bar{\mathbf{P}}$, and the terms $\mathbf{E}[e^{-r\tau\lambda}]$ are eigenvalues of $\bar{\mathbf{P}}$. $\mathbf{E}[e^{-r\tau\lambda}]$ is also the moment-generating function of the distribution of r evaluated at $-\tau\lambda$.

Identifiability of the Stationary GTR Model with Rate Heterogeneity

The identifiability problem is to find the true values of all parameters of the model, given the set of true frequencies of site patterns $f(X_1, \theta_0), f(X_2, \theta_0), \ldots, f(X_N, \theta_0)$. In the following sections I demonstrate the identifiability of all of the types of parameters for the GTR+I+Γ model.

Stationary nucleotide frequencies.—The matrix of nucleotide frequencies, Π, can be easily determined. Letting s be the number of terminal taxa or observed sequences and n_{ij} the number of times that nucleotide i occurs in site pattern j, then the frequency of i is given by

$$\pi_i = \frac{1}{s}\sum_j n_{ij} f(X_j, \theta_0). \qquad (8)$$

*Instantaneous rate matrix, **Q**.*—Let τ be the sum of path lengths between any two terminal taxa on the true unrooted tree. The matrix of dinucleotide frequencies, $\bar{\mathbf{F}}(\tau)$, for these two taxa also is determined by the frequencies of the site patterns. $\bar{F}_{ij}(\tau)$, the frequency of sites with nucleotide i in one sequence and nucleotide j in the other, can be found by summing the frequencies of site patterns with this pairing. Then the substitution probability matrix $\bar{\mathbf{P}}(\tau)$ can be found from Eq. 6. From Eq. 7 we see that $\bar{\mathbf{P}}(\tau)$ will have the same eigenvector matrix, $\mathbf{U}$, as $\mathbf{Q}$. Its vector of eigenvalues will be $(1, E[e^{-r\tau\lambda_2}], E[e^{-r\tau\lambda_3}], E[e^{-r\tau\lambda_4}])$. Let one of the nonunitary eigenvalues of $\bar{\mathbf{P}}(\tau)$ be $m_i = E[e^{-r\tau\lambda_i}]$. As noted above, this eigenvalue is also equivalent to the moment-generating function of the distribution of r evaluated at $-\tau\lambda_i$, which may be represented as $\mu(\tau\lambda_i)$. From the definition of a moment-generating function, then

$$\mu(\tau\lambda_i) = \int_{\text{all}\, r} \phi(r)e^{-r\tau\lambda_i}dr. \qquad (9)$$

μ $(\tau\lambda_i)$ therefore is a nonlinear function of $\tau\lambda_i$ that is equal to 1 when $\tau\lambda_i = 0$ and decreases monotonically and asymptotically to some constant $\gamma \geq 0$ as $\tau\lambda_i$ increases. Then, letting μ^{-1} be the functional inverse of μ,

$$\mu^{-1}(m_i) = \tau\lambda_i. \qquad (10)$$

Therefore, if the form and correct values of the parameters of μ are known, the values of $\tau\lambda_2$, $\tau\lambda_3$, and $\tau\lambda_4$ are determined. Because τ and λ_2, λ_3, and λ_4 always occur only in their products, we can arbitrarily set the path length between some pair of terminal taxa on the true unrooted tree to $\tau = 1$. Then Eq. 10 for this pair of taxa will determine the values of λ_2, λ_3, and λ_4. The rate matrix is then found from $\mathbf{Q} = \mathbf{U}\Lambda\mathbf{U}^{-1}$, where $\Lambda = \mathbf{diag}(0, -\lambda_2, -\lambda_3, -\lambda_4)$.

Tree topology and branch lengths.—Using the known value of any λ_i, Eq. 10 can be used to find the path distances on the true unrooted tree between all other pairs of taxa. The set of true distances between all pairs of taxa determines the topology and branch lengths of the true unrooted tree (Buneman, 1971; Chang, 1996).

The I + Γ Model of Rate Heterogeneity

The proof of the identifiability of the I + Γ model proceeds by showing that any I + Γ moment-generating function ν that differs from the true moment-generating function μ in the values of any of its parameters will not "fit" both $\bar{\mathbf{P}}(\tau)$ matrices for any two pairs of sequences that are separated by different path lengths τ on the true tree. This result is illustrated graphically later in Figure 3. In the case where $\nu = \mu$ the curve on the graph would be a straight line.

Proof.—Assume that ν is a possible moment-generating function of r that is taken as an estimator of the true, but unknown, function μ. Assume further that ν has variables t and l_i that estimate the parameters τ and λ_i, respectively, of μ. If $\nu = \mu$, then

$$\nu(tl_i) = m_i = \mu(\tau\lambda_i) \tag{11}$$

and

$$tl_i = \nu^{-1}(m_i) = \nu^{-1}(\mu(\tau\lambda_i)) = \tau\lambda_i \tag{12}$$

for all pairs of terminal taxa. Assume that τ_1 and τ_2 are the distances between two different pairs of terminal sequences on the true tree. Then for any of the three nonzero eigenvalues i,

$$\nu^{-1}(m_{1i})/\nu^{-1}(m_{2i}) = \tau_1\lambda_i/\tau_2\lambda_i = \tau_1/\tau_2, \tag{13}$$

where m_{1i} is the ith eigenvalue of $\bar{\mathbf{P}}(\tau_1)$ and m_{2i} is the corresponding eigenvalue of $\bar{\mathbf{P}}(\tau_2)$. But μ and ν are nonlinear functions. So, if $\nu \neq \mu$, $\nu^{-1}(\mu(\tau\lambda))$ is almost certainly a nonlinear function of $\tau\lambda$. If ν is a "reasonable" function with relatively few parameters, it will have only as many degrees of freedom as parameters and so will be unlikely to satisfy Eq. 13 for very many distinct values of τ and λ. However, improbability does not equal impossibility. In the following paragraphs I will demonstrate impossibility for the case in which ν and μ are both moment-generating functions for I + Γ rate heterogeneity models that differ only in the values of their parameters.

The I + Γ model of rate heterogeneity assumes that a proportion π of sites are invariable ($r = 0$) and that the remainder, with frequency $1 - \pi$, have varying r distributed according to the gamma probability function (Gu et al., 1995). Because r is a relative rate variable, we can assume that its mean, or expected value, among the variable sites is $E[r] = 1$. In this case, the gamma distribution has only a single parameter, α, the "shape" parameter. Then, if we let $\mu(x)$ be the moment-generating function of the true I + Γ model evaluated at $-x$,

$$\mu(x) = \pi + (1 - \pi)(1 + x/\alpha)^{-\alpha}. \tag{14}$$

Letting ν be an I + Γ distribution with known parameter values p and a that estimates μ with unknown true parameter values π and α, then

$$\nu(y) = p + (1 - p)(1 + y/a)^{-a}. \tag{15}$$

From Eqs. 12, 14, and 15,

$$\begin{aligned} y &= \nu^{-1}(\mu(x)) \\ &= a\left\{\left[\frac{\pi - p + (1 - \pi)(1 + x/\alpha)^{-\alpha}}{1 - p}\right]^{-\frac{1}{a}} - 1\right\}. \end{aligned} \tag{16}$$

Obviously, if $p = \pi$ and $a = \alpha$, then $y = x$.

Now, assuming that $p \neq \pi$ or $a \neq \alpha$ and taking first and second derivatives of y with respect to x,

$$\begin{aligned} \frac{dy}{dx} &= \left(\frac{1 - \pi}{1 - p}\right)(1 + x/\alpha)^{-(\alpha+1)} \\ &\quad\times\left[\frac{\pi - p + (1 - \pi)(1 + x/\alpha)^{-\alpha}}{1 - p}\right]^{-\left(\frac{1}{a}+1\right)} > 0 \end{aligned} \tag{17}$$

and

$$\frac{d^2y}{dx^2} = \left(\frac{1-\pi}{1-p}\right)\left(\frac{\alpha+1}{\alpha}\right)(1+x/\alpha)^{-(\alpha+2)} \times\left[\frac{\pi-p+(1-\pi)(1+x/\alpha)^{-\alpha}}{1-p}\right]^{-(\frac{1}{a}+1)} \bullet\left\{\left(\frac{\alpha(a+1)}{a(\alpha+1)}\right)\left(\frac{1-\pi}{1-p}\right)(1+x/\alpha)^{-\alpha} \times\left[\frac{\pi-p+(1-\pi)(1+x/\alpha)^{-\alpha}}{1-p}\right]^{-1}-1\right\}. \tag{18}$$

Setting the second derivative equal to zero and solving for x gives

$$x = \alpha\left\{\left[\frac{(1-\pi)(\alpha-a)}{a(\alpha+1)(\pi-p)}\right]^{\frac{1}{\alpha}}-1\right\}. \tag{19}$$

Depending on the values of π, α, p, and a, Eq. 19 has either no real solutions or only one. In the latter case, the solution may be either negative, zero, or positive. Because we are interested only in cases where x is some product $\tau\lambda$ of a tree path length and an eigenvalue of $\mathbf{Q}$, only cases in which $x \geq 0$ are relevant. Equations 16–19 show that, in the range $x \geq 0$, y is a continuous, monotonically increasing function of x having at most one inflection point. From Eq. 16 we see that $y = 0$ when $x = 0$. If $\pi > p$, then

$$y \to a\left[\left(\frac{\pi-p}{1-p}\right)^{-\frac{1}{a}}-1\right] \quad \text{as} \quad x \to \infty. \tag{20}$$

If $\pi < p$, then

$$y \to \infty \quad \text{as} \quad x \to \alpha\left[\left(\frac{p-\pi}{1-\pi}\right)^{-\frac{1}{\alpha}}-1\right]. \tag{21}$$

Figure 2 shows the graph of y for (α, a, π, p) = (10, 0.4, 0.5, 0.2). For these values there is an inflection point at $x = 1.378$ and $y \to 4.24496$ as $x \to \infty$. This graph is concave upward for $0 < x < 1.378$ and concave downward for $x > 1.378$. For values of α, a, π, and p such that $\pi > p$ and there is no point of inflection for $x > 0$, the graph will be entirely concave downward. For cases in which $\pi < p$, the graph will be entirely concave upward if there is no inflection point, or will be concave downward for $0 < x <$ inflection point and concave upward for $x >$ inflection point. Figure 2 also shows two hypothetical "true" path length × eigenvalue products, $\tau_1\lambda_i$ and $\tau_2\lambda_i$, and their corresponding y values. For the two cases illustrated we can assume that $\lambda_i = 1$, $\tau_1 = 1$, and $\tau_2 = 10$. Holding the two latter values constant, varying λ_i over the range $0 \to \infty$, and plotting the resulting pairs of

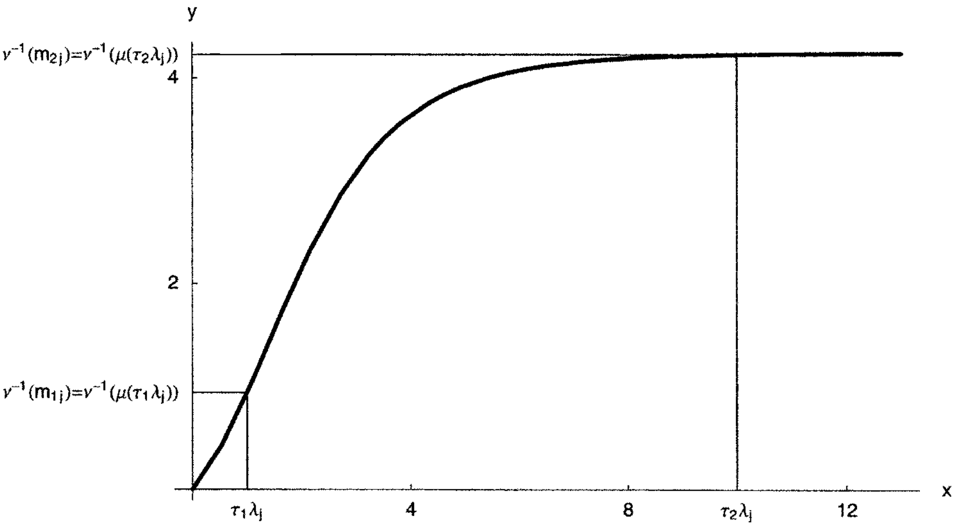


FIGURE 2. The graph of equation (16) for the case (α, a, π, p) = (10, 0.4, 0.5, 0.2).

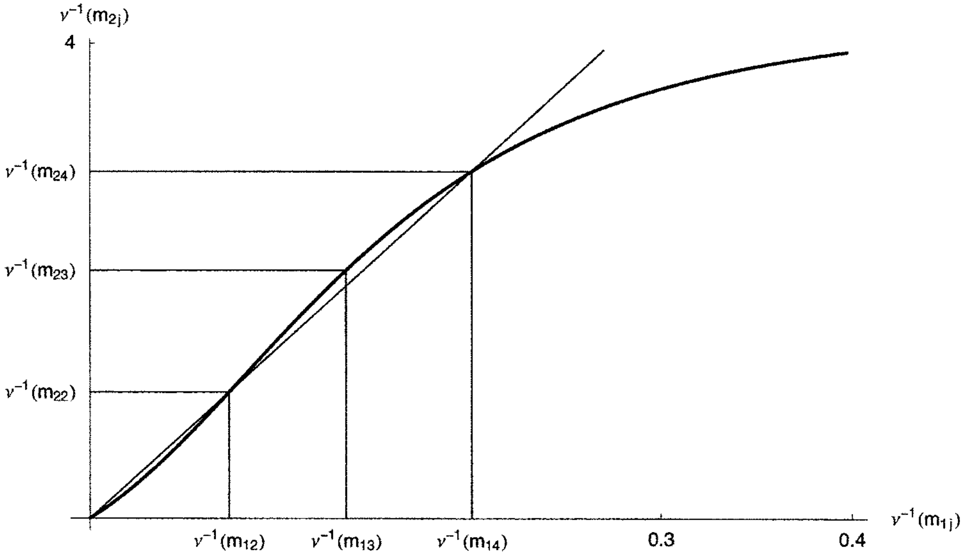


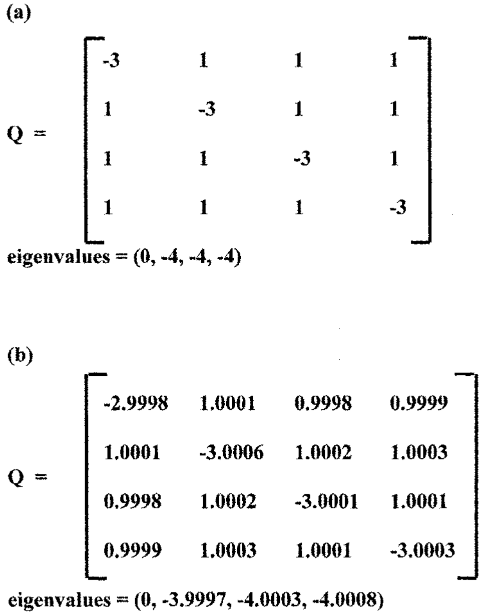
FIGURE 3. Illustration that when the two I + Γ distribution functions μ and ν are not equal the three ratios of equation 22 cannot all be equal. The straight line represents a case in which $\nu = \mu$. The curve is for the case illustrated in Figure 2 in which $\nu \neq \mu$.

values of $\nu^{-1}(\mu(\tau_1\lambda_i))$ and $\nu^{-1}(\mu(\tau_2\lambda_i))$, we obtain the graph shown in Figure 3. As in Figure 2, the graph has an inflection point, is concave upwards before the inflection point, and is concave downward after the inflection point. Similar graphs will be produced for any pair of path distances such that $\tau_2 > \tau_1$.

Assume that there are two pairs of terminal sequences with one pair connected by true path length τ_1 and the other by true path length τ_2, such that $\tau_2 > \tau_1$. This will always be true for any tree that is not a star tree (all internal branch lengths equal to zero) with all external branches of equal length (i.e., a phylogenetically uninformative set of sequences). As noted above, the transition probability matrices $\bar{\mathbf{P}}(\tau_1)$ and $\bar{\mathbf{P}}(\tau_2)$ and their sets of nonzero eigenvalues (m_{12}, m_{13}, m_{14}) and (m_{22}, m_{23}, m_{24}) for these two pairs of sequences are uniquely determined by the true frequencies of site patterns $f(X_1, \theta_0)$, $f(X_2, \theta_0)$, ..., $f(X_N, \theta_0)$. Because ν is a function with known parameter values, the values $\nu^{-1}(m_{12})$, $\nu^{-1}(m_{13})$, $\nu^{-1}(m_{14})$, $\nu^{-1}(m_{22})$, $\nu^{-1}(m_{23})$, and $\nu^{-1}(m_{24})$ are determined as well. Now, from Eqs. 11–13, if $\nu = \mu$, then it must be true that

$$\nu^{-1}(m_{12})/\nu^{-1}(m_{22}) = \nu^{-1}(m_{13})/\nu^{-1}(m_{23}) = \nu^{-1}(m_{14})/\nu^{-1}(m_{24}) = \tau_1/\tau_2 \,. \qquad (22)$$

But if $\nu \neq \mu$, then no more than two of the three ratios can be the same, as shown by Figure 3. Therefore, if the substitution rate matrix **Q** has three distinct nonzero eigenvalues, the parameters of the I + Γ rate heterogeneity will be uniquely determined. This, of course


(a)

$$\mathbf{Q} = \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}$$

eigenvalues = (0, -4, -4, -4)

(b)

$$\mathbf{Q} = \begin{bmatrix} -2.9998 & 1.0001 & 0.9998 & 0.9999 \\ 1.0001 & -3.0006 & 1.0002 & 1.0003 \\ 0.9998 & 1.0002 & -3.0001 & 1.0001 \\ 0.9999 & 1.0003 & 1.0001 & -3.0003 \end{bmatrix}$$

eigenvalues = (0, -3.9997, -4.0003, -4.0008)

FIGURE 4. Two hypothetical nucleotide substitution rate matrices and their eigenvalues. Matrix (a) is an exact Jukes-Cantor rate matrix. Matrix (b) is an approximate Jukes-Cantor matrix.

leaves open the possible cases in which **Q** has only one or two distinct nonzero eigenvalues, such as the Jukes–Cantor (one) or Kimura two-parameter (two) models. For real data sets, however, it is very unlikely that any two or all three of the eigenvalues will be exactly identical. The proof given above applies to any case in which two or more of the eigenvalues are as close in value to each other as we wish, short of absolute identity. So the GTR model can be made to approximate these two simpler models as closely as we wish by making two (Kimura two-parameter) or all three (Jukes–Cantor) of the eigenvalues of **Q** approximately equal to any degree short of absolute identity. Figure 4 contains a numerical example for the Jukes–Cantor model.

ACKNOWLEDGMENTS

I have benefited greatly from numerous discussions about ML with David Swofford, who also read an earlier version of this manuscript and made several suggestions that have improved it significantly. I thank Michael Sanderson, Junhyong Kim, and an anonymous reviewer for their thorough reviews and several helpful suggestions. My research on likelihood has been supported by grant DEB-9628835 from the National Science Foundation.

REFERENCES

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387–395 *in* Mathematics in the archaeological and historical sciences (F. R. Hodson and D. G. Kendall, eds.). Edinburgh Univ. Press, Edinburgh, Scotland.

CHANG, J. T. 1996. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Math. Biosci. 137:51–73.

FARRIS, J. S. 1999. Likelihood and inconsistency. Cladistics 15:199–204.

FELSENSTEIN, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

GOLDMAN, N. 1993. Statistical tests for models of DNA substitution. J. Mol. Evol. 36:182–198.

GU, X, Y. FU., AND W. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

LANAVE, C., G. PREPARATA, C. SACCONE, AND G. SERIO. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

RODRÍGUEZ, F., J. L. OLIVER, A. MARÍN, AND J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

ROGERS, J. S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. 46:354–357.

RUSSO, C. A. M., N. TAKEZAKI, AND M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13:525–536.

SAITOU, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. 27:261–273.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. Cladistics 14:209–220.

STEEL, M. A., L. A. SZÉKELY, AND M. D. HENDY. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comp. Biol. 1:153–163.

SULLIVAN, J., D. L. SWOFFORD, AND G. J. P. NAYLOR. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16:1347–1356.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, AND J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol.

WALD, A. 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20:595–601.

YANG, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.

YANG, Z. 1996a. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

YANG, Z. 1996b. Among-site rate variation and its impact on phylogenetic analyses. TREE 11:367–372.

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

Received 25 May 2000; accepted 11 November 2000
Associate Editor: M. Sanderson

Appendix: Adaptation of Wald's (1949) Proof of the Consistency of ML Estimation to Estimation of Phylogenetic Trees from Nucleotide Sequences

Initially I give Wald's definition of each term and assumption in italics, followed by comments regarding its applicability to phylogeny estimation in unitalicized text. Then I present modified versions of Wald's lemmas and theorems that demonstrate the consistency of ML estimation of phylogenetic trees.

Definitions

$X_1, X_2, \ldots, X_n$ = n *independently and identically distributed random variables.*

In the case of nucleotide sequence data, these are the nucleotide site patterns over all s sequences at the n nucleotide sites of the sequences.

$\theta^1, \theta^2, \ldots, \theta^k$ = *a finite number* k *of parameters, the values of which are to be estimated by the maximum likelihood analysis.*

In phylogenetic estimation these may be tree topologies, branch lengths, substitution rates, rate heterogeneity parameters, and so forth. As discussed by Chang (1996), the tree topology parameter may be given arbitrary numerical values, 0, 1, 2, ..., etc. Only the $K = (2s-5)(2s-7)\cdots 1$ completely bifurcating topologies need be included. Degenerate topologies, those with polytomies, correspond to certain bifurcating topologies with internal branch lengths equal to zero. Nucleotide frequencies and the proportion of invariable sites fall in the closed interval [0, 1]. The topology parameter falls in the closed interval [0, K]. All other parameters, branch lengths, and so on, fall in the open interval [0, ∞). These latter parameters of the GTR + I + Γ model can be rescaled by the transformation $\theta^i = 1 - e^{-z_i}$ so that they fall in [0, 1). And because all of the probabilities $f(x, \theta)$, defined below, are finite and fall in [0, 1] even when any untransformed parameter $z_i = \infty$, the transformed parameters can be assumed to fall in the closed interval [0, 1]. This transformation greatly facilitates the proofs given below

$\theta = (\theta^1, \theta^2, \ldots, \theta^k)$, *a parameter point in the* k*-dimensional Cartesian space.*

Ω = *the parameter space, the set of all possible parameter points; a subset of the* k*-dimensional Cartesian space.*

$F(x, \theta)$ = *prob* $(X_i < x)$

$f(x, \theta)$ = *density of* $F(x, \theta)$ *at* x *if* $F(x, \theta)$ *is absolutely continuous.*

= *prob*$(X_i = x)$, *if* $F(x, \theta)$ *is discrete.*

For sequence data, $F(x, \theta)$ is always discrete.

$f(x, \theta, \rho)$ = *least upper bound (lub) of* $f(x, \theta')$ *with respect to* θ' *when* $|\theta - \theta'| \leq \rho$, *for any positive* ρ.

$f^*(x, \theta, \rho) = f(x, \theta, \rho)$, *when* $f(x, \theta, \rho) > 1$, *and* $= 1$ *otherwise.*

For discrete data, such as sequences, $f(x, \theta)$ and $f(x, \theta, \rho)$ are always ≥ 1. So $f^*(x, \theta, \rho)$ always equals 1 and is, therefore, unnecessary.

$\varphi(x, r)$ = *lub of* $f(x, \theta)$ *with respect to* θ *when* $|\theta| > r$, *for any positive* r.

This function is also unnecessary for sequence data.

$\varphi^*(x, r) = \varphi(x, r)$, *when* $\varphi(x, r) > 1$, *and* $= 1$ *otherwise.*

As with $\varphi(x, r)$, this function is also unnecessary.

Comment on the Discrete Nature of the Topology Parameter

In the introduction to his paper Wald (1949:595) stated explicitly that his proof makes "no differentiability assumptions (thus, not even the existence of the likelihood equation is postulated)..." As pointed out by Swofford et al. (in press), if we define the likelihood function as

$$L(\theta) = \prod f(X_i, \theta),$$

then the likelihood equation referred to by Wald is

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0.$$

Thus, Wald's proof does not depend upon the differentiability of L with respect to the parameters of the model, contrary to the assertions of several workers (e.g., Yang, 1996a; Siddall, 1998; Farris, 1999). So the discrete nature of the tree topology parameter does not invalidate the application of the proof to tree estimation.

Assumptions

Assumption 1. $F(x, \theta)$ *is either discrete for all* θ *or is absolutely continuous for all* θ.

This is true for sequence data because nucleotide site patterns are discrete objects.

Assumption 2. *For sufficiently small* ρ *and sufficiently larger* r, *the expected values* $\int_{-\infty}^{\infty} \log f^*(x, \theta, \rho)\, dF(x, \theta_0)$ *and* $\int_{-\infty}^{\infty} \log \varphi^*(x, r)\, dF(x, \theta_0)$ *are finite where* θ_0 *denotes the true parameter point.*

This assumption is unnecessary for sequence data.

Assumption 3. *If* $\lim_{i \to \infty} \theta_i = \theta$, *then* $\lim_{i \to \infty} f(x, \theta_i) = f(x, \theta)$ *for all* x *except perhaps on a set that may depend on the limit point* θ *(but not on the sequence* θ_i*) and "whose" probability measure is 0 according to the probability distribution corresponding to the true parameter point* θ_0. [All assumptions are direct quotes from Wald (1949).]

Stated simply, this means that there is no permissible parameter point θ at which $f(x, \theta)$ is discontinuous, exceptperhaps for values of x that have a probability of 0 of occurring under the true probability distribution of x. This is true for the GTR + I + Γ nucleotide substitution model considered in this paper.

Assumption 4. *If* θ_1 *is a parameter point different from the true parameter point* θ_0, *then* $F(x, \theta_1) \neq F(x, \theta_0)$ *for at least one value of* x.

This is the identifiability assumption. This assumption is true if it can be shown that, given the true probability distribution $F(x, \theta_0)$, the true parameter point θ_0 is uniquely determined, or *identifiable*. Identifiability of the tree topology and all other parameters of the GTR + I + Γ model is demonstrated above in the body of the paper.

Assumption 5. *If* $\lim_{i\to\infty}|\theta_i| = \infty$, *then* $\lim_{i\to\infty} f(x, \theta_i) = 0$ *for any* x *except perhaps on a fixed set (independent of the sequence θ and "whose" probability is 0 according to the probability distribution corresponding to the true parameter point* θ_0.

This obviously is not true for substitution models on phylogenetic trees. In these models, when the branch lengths approach infinity, the expected pattern frequencies $f(x, \theta)$ approach some nonzero value for all site patterns x. But with all parameters rescaled to fall in closed intervals, it is not necessary.

Assumption 6. *For the true parameter point* θ_0 *we have* $\int_{-\infty}^{\infty} |\log f(x, \theta_0)|\, dF(x,\theta_0) < \infty$.

For discrete data this is equivalent to $-\sum_x f(x, \theta_0) \log f(x, \theta_0) < \infty$. Letting $0 \times \log(0) = 0$, this assumption is obviously always true because it is always true that $0 \le f(x, \theta) \le 1$ for all values of x and θ.

Assumption 7. *The parameter space* Ω *is a closed subset of the* k*-dimensional Cartesian space.*

This is true for substitution models on phylogenetic trees because any model parameter θ^i, including the tree topology, can be quantified or transformed so that $0 \le \theta^i \le c$ for a finite constant c. In this case Ω is bounded as well as closed, and so is *compact*.

Assumption 8. $f(x, \theta, \rho)$ *is a measurable function of* x *for any* θ *and* ρ.

As Wald notes, this assumption is unnecessary for discrete data.

Lemmas

Lemma 1. *For any* $\theta \neq \theta_0$ *we have* $E[\log f(X, \theta)] < E[\log f(X, \theta_0)]$, *where* X *is a chance variable with distribution* $F(x, \theta_0)$.

Wald's proof of Lemma 1 can be replaced by the following shorter proof in the discrete case.

Proof. For nucleotide data, the inequality above is equivalent to

$$\sum_x f(x, \theta_0) \log f(x, \theta) < \sum_x f(x, \theta_0) \log f(x, \theta_0). \quad (23)$$

From Assumption 4, if $\theta \neq \theta_0$, then $f(x, \theta) \neq f(x, \theta_0)$ for at least one value of x. Goldman (1993) has shown that, under this condition, Inequality 23 holds for nucleotide data. This completes the proof of Lemma 1.

Lemma 2. $\lim_{\rho\to 0} E\,[\log f(X, \theta, \rho)] = E\,[\log f(X, \theta)]$.

Wald's proof of Lemma 2 can be replaced by the following shorter proof in the discrete case.

Proof. From the $f(x, \theta, \rho)$,

$$\log f(x, \theta, \rho) \ge \log f(x, \theta), \quad (24)$$

Then

$$\begin{aligned} E[\log f(X, \theta, \rho)] &= \textstyle\sum_x f(x, \theta_0) \log f(x, \theta, \rho) \\ &\ge \textstyle\sum_x f(x, \theta_0) \log f(x, \theta) \\ &= E[\log f(X, \theta)]. \quad (25) \end{aligned}$$

And from Assumption 3

$$\lim_{\rho\to 0} \log f(x, \theta, \rho) = \log f(x, \theta). \quad (26)$$

Lemma 2 follows from Inequalities (25) and (26).

Lemma 3. *The equation* $\lim_{r\to\infty} E[\log \varphi(X, r)] = -\infty$ *holds*. This Lemma is unnecessary for nucleotide data.

Theorems

The following theorems and proofs are modified versions of Wald's originals.

Theorem 1. *Let* ω *be a subset of all points* θ *in* Ω such that $|\theta - \theta_0| \ge \varepsilon > 0$ *for some finite value of* ε. *Then for observations* $x_1, \ldots, x_n$,

$$Prob\left[\lim_{n\to\infty} \frac{\underset{|\theta-\theta_0|\ge\varepsilon}{lub}[f(x_1, \theta)\cdots f(x_n, \theta)]}{f(x_1, \theta_0)\cdots f(x_n, \theta_0)} = 0\right] = 1.$$

Proof. For each point θ in ω, let ρ_θ be a positive number such that

$$E\big[\log f(X, \theta, \rho_\theta)\big] - E\big[\log f(X, \theta_0)\big] < 0. \quad (27)$$

The existence of ρ_θ follows from Lemmas 1 and 2 and the definition of ω. Because ω is compact, there is a finite number of points $\theta_1, \theta_2, \ldots, \theta_h$ in ω such that the set $S(\theta_1, \rho_{\theta_1}) \cup \cdots \cup S(\theta_h, \rho_{\theta_h})$ has ω as a subset. $S(\theta, \rho)$ represents the "sphere" with center θ and radius ρ. Then

$$\begin{aligned} 0 &\le \underset{|\theta-\theta_0|\ge\varepsilon}{lub}[f(x_1, \theta)\cdots f(x_n, \theta)] \\ &\le \sum_{i=1}^{h} f(x_1, \theta_i, \rho_{\theta_i})\cdots f(x_n, \theta_i, \rho_{\theta_i}). \quad (28) \end{aligned}$$

Therefore, to prove Theorem 1, it is sufficient to show that

$$\text{Prob}\left[\lim_{n\to\infty} \frac{f(x_1, \theta_i, \rho_{\theta_i})\cdots f(x_n, \theta_i, \rho_{\theta_i})}{f(x_1, \theta_0)\cdots f(x_n, \theta_0)} = 0\right] = 1$$
$$(i = 1, \ldots, h) \quad (29)$$

or

$$\text{Prob}\left[\lim_{n\to\infty} \sum_{j=1}^{n} [\log f(x_j, \theta_i, \rho_{\theta_i}) - \log f(x_j, \theta_0)] = -\infty\right] = 1 \quad (30)$$

From the strong law of large numbers and the assumption that the $X_1, \ldots, X_n$ are independently and identically distributed, Eq. 30 is equivalent to

$$\text{Prob}\left[\lim_{n\to\infty} (n\, E\,[\log f(X, \theta_i, \rho_{\theta_i}) - \log f(X, \theta_0)]) = -\infty\right] = 1. \quad (31)$$

Equation 31 follows from Inequality 27. Therefore, Theorem 1 is proved.

Theorem 2. *Let* $\bar{\theta}_{\mathrm{n}}$ $(\mathrm{x}_1, \ldots, \mathrm{x}_{\mathrm{n}})$ *be a function of the observations* $\mathrm{x}_1, \ldots, \mathrm{x}_{\mathrm{n}}$ *that has the property*

$$\frac{f(\mathrm{x}_1, \bar{\theta}_n) \cdots f(\mathrm{x}_n, \bar{\theta}_n)}{f(\mathrm{x}_1, \theta_0) \cdots f(\mathrm{x}_n, \theta_0)} \geq c > . \qquad (32)$$

for all n *and for all* $\mathrm{x}_1, \ldots, \mathrm{x}_{\mathrm{n}}$. *Then*

$$\mathrm{Prob}\Big[\lim_{n\to\infty} \bar{\theta}_n = \theta_0\Big] = 1. \qquad (33)$$

Proof. Equation 33 is equivalent to

$$\mathrm{Prob}\Big[\lim_{n\to\infty} |\bar{\theta}_n - \theta_0| \leq \varepsilon\Big] = 1, \text{ for any } \varepsilon > 0. \qquad (34)$$

If the sequence $\{\bar{\theta}_n\}_{n\to\infty}$ has a limit point $\bar{\theta}$ such that $|\bar{\theta} - \theta_0| > \varepsilon$, then

$$\underset{|\theta-\theta_0|\geq\varepsilon}{lub} [f(x_1, \theta) \cdots f(x_n, \theta)] \geq f(x_1, \bar{\theta}_n) \cdots f(x_n, \bar{\theta}_n) \qquad (35)$$

for all n and all $x_1, \ldots, x_n$. But then

$$\lim_{n\to\infty} \frac{\underset{|\theta-\theta_0|\geq\varepsilon}{lub} [f(x_1, \theta) \cdots f(x_n, \theta)]}{f(x_1, \theta_0) \cdots f(x_n, \theta_0)} \geq c > 0. \qquad (36)$$

Because, from Theorem 1, Inequality 36 has probability zero, Eq. 34 and Theorem 2 are proved.

A ML estimate $\hat{\theta}_n(x_1, \ldots, x_n)$ satisfies Inequality 32 with $c = 1$. Therefore, Theorem 2 proves the consistency of $\hat{\theta}$ as an estimator of θ_0.